Algorithm for Bioinformatics

Final Project

Student: Gabriele Pozzati **Course: Fall 2019/2020**

Introduction

Proteins are the essential machines of the cell and it is very common that biological mechanisms are mediated by interactions between them (PPIs). Subsequently, in order to successfully understand biological functions, powerful methods to study these interactions are required.

Between the residues that form the interface of the two proteins, some seems to be more important than others in determining the PPI. This importance is commonly determined by mutating a certain residue to an Alanine and measuring the PPI delta-G before and after the mutation. The subtraction of these two values give us the mutation delta-delta-G. It is commonly accepted that delta-delta-G values greater than 2 are adopted to indicate the mutated residue as a hotspot.

In the state of the art, many methods have been proposed to predict PPI hotspots based on these kind of data. Some degree of success has been achieved, in particular with the use of supervised machine learning methods. These methods are based on the mathematical modelling of a set of known samples referred as training set. When the amount of available data is large, these methods are highly effective in producing realistic models. Unfortunately, available alanine scanning that allows to identify hotspots are quite scarce in number (in the order of magnitude of hundreds of mutated residues). Subsequently, machine learning methods applied to hotspot prediction are very prone to be overfitted to the data used to train them. In this work, I propose a method which uses a graph representation of PPIs topology in order to predict potential hotspots. This method is only based on assumptions over the nature of PPIs and the resulting statistic is subsequently more robust, not involving any bias possibly present in the available data.

Material and methods

dataset - alanine mutation scan data has been obtained joining information from Alanine Scanning Energetics Database (ASEdb), SKEMPI database (Structural Kinetic and Energetic database of Mutant Protein Interactions) and dbMPIKT (the kinetic and thermodynamic database of mutant protein interactions). All the mutations are referred to a relative crystallized structure which has been downloaded from the PDB database. All the PDBs containing the selected mutations have been scanned for the relative PPIs. In this stage, mutated residues in contact with non-protein chains have been excluded from our dataset. The remaining 689 mutations have been mapped to 149 PPIs. In order to establish a gold standard, I adopted the commonly accepted classification system that consider hotspot all the mutated residues with delta-delta-G greater than 2Kcal/mol and non-hotspot the ones with delta-delta-G smaller than 0.4Kcal/mol.

network - to obtain a network from each one of the selected PPIs, I considered every contact between two residues on opposite chains as a node of the network. Then, if any couple of contacts shared a residue, an edge has been placed between the relative nodes (Fig.1). Network building and clique search algorithm used in this study are implemented in the python library networkx (https://networkx.github.io/, version 2.4).

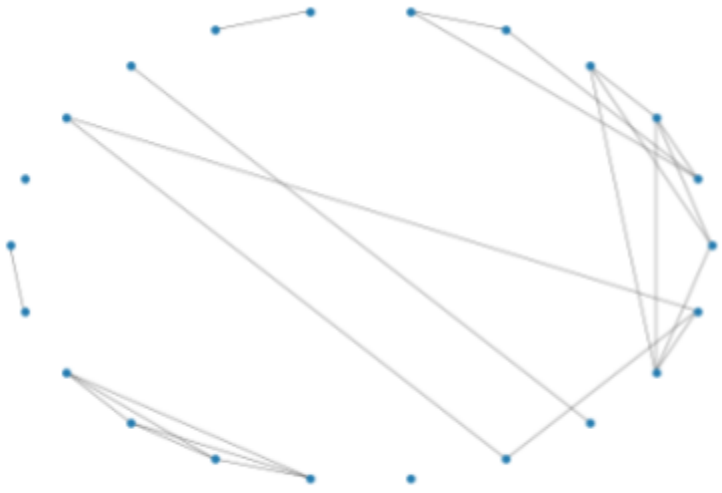

Fig1: PPI network example from the interface 1ais_A-B

hotspot detection - the final goal of hotspot prediction has been achieved by running a series of steps over each PPI network. First, a clique analysis have been performed in order to obtain a list of cliques (group of nodes fully connected to each other). Each clique composed only by two nodes has been discarded. Then, every residue from every node/contact in the resulting list of cliques has been counted in the number of total occurrences. Next, a fixed value has been added to each residue occurrence count, based on the physico-chemical properties of the residue: +3 for charged and large hydrophobic residues (Arg, His, Lys, Asp, Glu, Met, Phe, Tyr, Trp), +2 for small hydrophobic residues (Val, Ile, Leu), -2 for polar residues (Ser, Thr, Asn, Gln) and +0 for all the others (Cys, Gly, Pro, Ala). Finally, each residue with value greater than 4 has been considered a positive prediction.

statistical methods - in order to evaluate the goodness of the proposed method, the following metrics have been adopted:

- True Positives (TP) - number of correctly predicted hotspots
- True Negatives (TN) - number of correctly predicted non-hotspots
- False Positives (FP) - number of non-hotspots predicted as hotspots
- False Negatives (FN) - number of hotspots predicted as non-hotspots
- True Positive Rate (TPR) = TP / (TP+FN)
- True Negative Rate (TNR) = TN / (TN+FP)
- Accuracy (Acc) = (TP+TN) / (TP+TN+FP+FN)
- Precision (PPV) = TP / (TP+FP)
- Matthews Correlation Coefficient (MCC) = ((TP*TN)-(FP*FN)) / ((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))

codes - python implementation of all the methods described above can be retrieved in the following GitHub repository: https://github.com/gabrielepozzati/ABI.git

Results

The network design allows, as highlighted by the results in Table 1, a simple and efficient analysis of PPI in terms of hotspot prediction.

Table 1: results of the prediction using the clique protocol

/	TP	FP	TN	FN
/	97	33	344	113
TPR	TNR	ACC	PPV	MCC
0.46	**0.91**	0.75	**0.75**	0.43

I decided to repeat the analysis with a slight variation in order to verify the utility of the clique-filtering step. In particular, instead of grouping the contacts in cliques and filter according to that, I directly counted the occurrence of every residue in the complete network. Surprisingly, as shown in Table 2, this modification seems to improve the overall performance of the prediction.

Table 2: results of the prediction using the clique-less protocol

/	TP	FP	TN	FN
/	123	50	327	87
TPR	TNR	ACC	PPV	MCC
0.59	0.87	**0.77**	0.71	**0.48**

The increase in Accuracy and Matthew Correlation Coefficient indicate that considering all contacts in the adopted range (4A) provides a slightly better overall definition of the interface hotspots, consistently increasing the number of positives. However, this advantage is paid with a loss of precision due to a higher number of False Positives. This may be a hint toward the fact that cliques are effectively a good indicator for hotspots in protein interfaces. Indeed, in this network representation, cliques indicate group of residues from one chain, equally close between themselves and to the residue of the other chain with which they are in

contact. Subsequently, it is possible to hypothesize that this kind of topology may be useful in hotspot determination, although it is probably not very common and not as important as the complete physico-chemical surrounding.

Overall, the network representation seems to be very powerful when applied to the study of PPI topology and further investigations should be performed in this direction.