SPECIFICITY AND CONSTRAINTS IN PEPTIDE-PROTEIN BINDINGS IN THE MOUSE PROTEOME

Report for 3rd Year Research Project

February 18, 2016

Dhruv SHARMA

ÉCOLE
POLYTECHNIQUE
UNIVERSITÉ PARIS-SACLAY

Contents

Part I
Introduction

This report details the work done towards the fulfillment of the requirements of the department of Physics at Ecole Polytechnique. I undertook this research project under the guidance of Dr. Remi Monasson at the Laboratoire de Physique Theoriqu at the Ecole Normale Superieure in Paris.

The aim of the project was to study the interactions between short peptide chains and a specific section of signalling proteins called PDZ Domains. One of the aspects that we study here is the specificities of interactions between peptides and PDZ domains. It is well known that macromolecules such as proteins and enzymes interact in a specific manner with other macromolecules and biomolecules. What interested us over the course of the study are the constraints present in the peptide sequences due to the specificity of their interactions with PDZ domains. We will also have an occasion to understand similar constraints on the PDZ domain sequences.

This report is organized as follows. After a brief introduction to the biological importance of PDZ Domains, we explain the experiments performed by **Insert reference here**. Using these experiments, Stifler et al created a model which is capable of predicting whether a peptide will bind to a PDZ domain given the sequence of the peptide. We shall explain the data that Stifler et al have provided. The first two models that we propose utilise the data provided by Stifler et al.

Once the data presented and the biological context established, we present a first model which seeks to understand the constraints imposed on the peptide sequences under the effect of mutations. We present the results derived from this model and discuss the limitations. A second improved model is then proposed which considers error rates as probabilities. We present some interesting observations on the basis of this model. In particular, we show how certain positions are particularly constrained over all peptides and present a simple way of the calculating the level of constraint.

Finally, to render the study of peptide-PDZ domain specificity complete, we explain how we could integrate the PDZ Domain sequences into the modeling. This is done by a regression method called the *Lasso*. We shall have the chance to present the lasso method in more detail in the relevant section.

We conclude with a summary of our findings and possible directions of further improvements.

1 PDZ Domains

Let us begin by explaining PDZ Domains, their importance, their structure and how they bind to other macromolecules. PDZ domains are short sections of proteins composed of 80-90 amino acids. PDZ Domains are usually found in signaling proteins where they regulate processes such as the separation of cell membranes. PDZ itself is an acronym for the first three letters of three proteins: Post-synaptic density protein (PSD95), Drosophilia disc large tumor suppressor

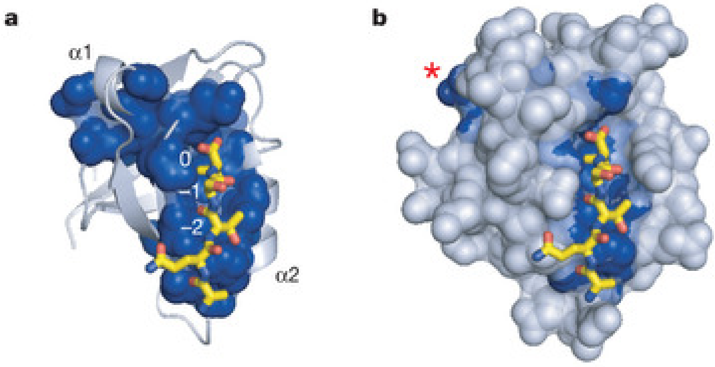

*Figure 1: **Left**: We show the PDZ domain (in blue) with the target peptide (yellow). The C-terminal of the target peptide is numbered 0 whilst the preceding amino acids are numbered in descending order. **Right**: The position of the PDZ-Domain within the parent protein is shown. The position marked with a little star shows that during the binding process there are other parts of the PDZ domain which are involved and not just the binding pocket*

(Dlg1) and zonula occludens-1 protein (zo-1)

PDZ Domains are usually composed of 80-90 amino acids and many PDZ domains can be present within a single protein. Multiple domains within the same protein can have similar functions or different functions, each of which will bing to a different part of the target protein or a different protein altogether.

PDZ Domains always bind to the C-terminal of certain specific sections of the target proteins. The binding occurs within a binding pocket formed by the domain. The binding pocket are usually formed by the first 20 amino acids. A diagram showing the positions of the binding pocket as well as the target peptide are shown below.

2 Explanation of the data of Stifler et al

Now that we have given a small introduction to PDZ domains we shall explain the work done by Stifler and co-authors which we sought to extend and improve upon.

The article that we have considered is titled "PDZ domain binding specificity is optimized across the mouse proteome". In this article, Michael Stifler and co-authors sought to characterise the selectivity between 157 PDZ domains and 217 target peptides found in the mouse proteome. To this end, they created a model which could predict whether a given PDZ domain would bind to a target peptide given the sequence of the target peptide.

In their model, the last 5 positions near the C-terminal of the target peptide are considered. The model can be summarized in the following simple equation:

$$\phi_i = \sum_{p,q} A_{p,q}\theta_{i,p,q} > \tau_i \tag{1}$$

where for a given PDZ Domain i, we can calculate a binding score ϕ_i. A is an indicator of the peptide sequence : $A_{p,q} = 1$ if the amino acid at position p of the peptide is q and $A_{p,q} = 0$

otherwise. The numbers $\theta_{i,p,q}$ are the parameters that need to be fit to the experimental data. τ_i is a scoring threshold which is different for each domain i. The values $\theta_{i,p,q}$ are to be interpreted in the following fashion: $\theta_{i,p,q} > 0$ if the PDZ Domain i prefers amino acid q at position p more than the other PDZ domains, negative if it prefers it less, and 0 if it has no bias relative to the other domains. τ_i is defined as the mth percentile of the ϕ_i's for all the peptides in the model that bound to the domain i. To make predictions using this model, it suffices to know the sequence of the target peptide and compute $\phi_i - \tau_i$ for the domain i. If the domain binds with the target protein then this number is positive and negative otherwise.

The data used to fit the model was obtained by protein micro-array assays. Each PDZ Domain-peptide pair under consideration was verified for binding during the experiment. The researchers were able to model 74 domains among the 157 domains experimented upon. The experimental results of these experiments are available to us in the form of an interaction matrix which tells us whether a PDZ domain i binds to a target peptide j or not.

For each of the 74 domains successfully modeled, we also possess the parameters $\theta_{i,p,q}$ allowing us to test the model on newer peptide sequences. We also possess the error rates for the model. These are given in terms of false positive and false negative rates. These error rates interpreted as empirical probabilities were used in one of the models presented later in the report.

3 Questions asked and answered

Part II
First Model

Part III
Improvements over first model: Bayesian Modeling

Part IV
Integrating PDZ Domain sequences

Part V
Conclusion

Acknowledgements

I would like to thank Dr. Remi Monasson for guiding me throughout the project. He showed great patience with me and made himself available for me during the project. I am especially grateful also for the numerous conversations surrounding the emerging domain of deep learning, a field which interests us both. I would also like to thank Simona Cocco for her pertinent remarks. Finally I want to thank David, Antoine and Andre for listening to me patiently (more or less) as I struggled to explain my project to them.