Charles University
Faculty of Science

Study programme: Bioinformatics
Branch of study: Bioinformatics

Kateřina Břicháčková

Use of residue-level annotations for structural prediction of protein-ligand binding sites

Využití anotací primární struktury pro strukturní predikci protein-ligand aktivních míst

Master thesis

Supervisor: RNDr. David Hoksza, Ph.D.

Prague, 2020

Prohlášení

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem uvedla všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze, XXX Kateřina Břicháčková

Poděkování

dedication

Acknowledgement

dedication

Abstract

abstract

Keywords: keyword1, keyword2

Abstrakt

abstract

Klíčová slova: keyword1, keyword2

Contents

1. Introduction

1.1 Section1

2. Method

To find properties mapped on the protein primary structure which are possibly important for prediction of protein-ligand binding sites, statistical analysis will have the crucial role. It is a great way to explore the big amounts of accessible data and it can potentially help to discover underlying patterns and draw inferences from the data.

This chapter describes the method that was used to analyse the *statistical significance* of the properties and to distinguish the ones that stand out in the known protein-ligand binding sites.

2.1 Hypothesis testing

Hypothesis testing is a method of statistical inference. Its goal is to infer properties of a *statistical population*, a set of similar items or events. To give an example of a population, in this work, two populations will be compared: we take values of a property for all the amino acids and compare the ones in binding sites and outside of binding sites.

A dataset usually contains a subset collected from a larger population, rather than the whole population. This subset is called a *statistical sample.* It should represent the population well and be unbiased.

A *hypothesis* makes a statement about an unknown population parameter. In a hypothesis testing problem, an experimenter states two complementary hypotheses: the *null hypothesis* and the *alternative hypothesis*, denoted by H_0 and H_1, respectively. The null hypothesis comprises a subset of possible parameters and the alternative hypothesis comprises the supplement, so that all the possible parameters are covered.

In a hypothesis testing problem, an experimenter should come to one of the conclusions: to either accept H_0, or to reject H_0 and accept H_1.

To decide which one of two complementary hypotheses is true, an experimenter employes a suitable *hypothesis test.* A hypothesis test is a rule that specifies for which sample values the H_0 is accepted as true and for which sample values it is rejected, and therefore H_1 is accepted as true. A hypothesis test is usually specified in terms of a test statistic (i.g. a function of the sample). [1]

As one may expect, the tests are not error-proof and a mistake can be made in the decision of whether to accept or reject the null hypothesis. There are two types of errors in hypothesis testing, commonly known as *Type I Error* and *Type II Error.* The test has made a Type I Error if it incorrectly rejects a true null hypothesis. If, on the other hand, a null hypothesis is accepted and it is not true, a Type II Error has been made. Both situations are depicted in the Table 2.1. The ideal test would have both error probabilities equal to zero. However, in most cases it is not possible to make both error probabilities arbitrarily small for a fixed sample size. [1]

To control statistical significance of the result, a study defines a threshold called *significance level*, a constant denoted by α. It is the probability of making a Type I Error, in other words, the probability that the study rejects the null hypothesis when it is true.

One way to report the result of the test would be simply to tell whether the null hypothesis was accepted or rejected at the given significance level. However, most researchers choose to report a certain kind of test statistic (function of a sample X), the so-called *p-value.* Smaller values of $p(X)$ give stronger evidence for rejecting the null hypothesis. The null hypothesis is rejected when $p(X) \leq \alpha$. Hence, we are able to determine the smallest significance level at which the hypothesis would be accept-

		Prediction	
		Accept H0	**Reject H0**
Truth	**H0**	Correct (true positive)	Type I Error (false positive)
	H1	Type II Error (false negative)	Correct (true negative)

Table 2.1: Type I and II Error in hypothesis testing.

ed/rejected. P-value gives us an idea of how strongly the data contradict the null hypothesis and furthermore, it allows other researchers to make a decision according to the significance level of their choice. [1, 2, 3]

It is suggested that the significance level for a study is set prior to any data collection. [4] The typical choices are $\alpha = 0.01$, 0.05 or 0.10. [1]

Be aware that by fixing the significance level of the test, the experimenter is controlling only the Type I Error probabilities. The probability of the Type II Error is subject to factors beyond the experimenter's control, such as the accuracy and completeness of the data. [2]

Let's suppose an experimenter has a research hypothesis that he or she hopes to prove, but does not want to risk accepting it without convincing data support. In this case, the test should be set up in such a way that the research hypothesis corresponds to the alternative hypothesis, not the null hypothesis. By specifying a small significance level α, the experimenter thus controls the probability of the Type I Error. In other words, the probability of accepting the research hypothesis when in is not true would be α at most. [1]

2.1.1 Welsch's test

2.1.2 Fisher's exact test

2.1.3 Chi-squared test

2.2 Data

plus significance level, null hypothesis, alternative hypothesis

2.3 Implementation in Python

Conclusion

Conclusion.

List of Abbreviations

AA	Amino acid
atd	a tak dale

Bibliography

[1] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[2] Pak C. Sham and Shaun M. Purcell. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*, 15(5):335–346, apr 2014.

[3] Joseph P. Romano Erich L. Lehmann. *Testing Statistical Hypotheses*. Springer New York, 2008.

[4] Jerzy Neyman and Egon S Pearson. *The testing of statistical hypotheses in relation to probabilities a priori*, volume 29. 1933.

A. Attachments

The attached CD contains two attachments:

A.1 attachment1

blabla

A.2 attachment2

blabla