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

Ondrovi :P

Acknowledgement

Ondrovi :P

Abstract

abstract

Keywords: keyword1, keyword2

Abstrakt

abstract

Klíčová slova: keyword1, keyword2

Contents

1 Introduction **2**

1.1 Section1 2

2 Method **3**

2.1 Hypothesis testing 3

2.1.1 Welch's test 4

2.1.2 Fisher's exact test and Chi-squared test of independence 5

2.2 Data 6

2.3 Implementation in Python 6

3 Method **7**

3.1 Hypothesis testing 7

3.1.1 Welch's test 8

3.1.2 Fisher's exact test and Chi-squared test of independence 9

3.2 Data 10

3.3 Implementation in Python 10

Conclusion **11**

List of Abbreviations **12**

Bibliography **13**

A Attachments **15**

A.1 attachment1 15

A.2 attachment2 15

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*, i.e. a set of similar items or events. In this work, two populations will be compared: we take values of a property for all the amino acids across all the proteins in the dataset and then compare the ones in binding sites and outside of binding sites.

A dataset usually contains a subset sampled 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.e. a function of the sample) [3].

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 3.1. The ideal test would have both error probabilities equal to zero. Nevertheless, in most cases it is not possible to make both error probabilities arbitrarily small for a fixed sample size [3].

To control statistical significance of the result, a study defines a threshold called *significance level*, a constant denoted by α. It represents 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 accepted/re-

		Prediction	
		Accept H_0	Reject H_0
Truth	H_0	Correct (true positive)	**Type I error** (false positive)
	H_1	**Type II error** (false negative)	Correct (true negative)

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

jected. P-value gives an idea of how strongly the data contradict the null hypothesis; furthermore, it allows other researchers to make a decision according to the significance level of their choice [3, 5, 11].

It is suggested that the significance level for a study is set prior to any data collection [9]. The typical choices in practice are $\alpha = 0.01$, 0.05 or 0.10 [3]. One should 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 such as the accuracy and completeness of the data and most importantly, the true effect size [11].

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 [3].

2.1.1 Welch's test

Welch's unequal variances t-test, or Welch's test in short, is a two-sample hypothesis test used to decide whether two populations have different central tendencies (means or medians). The decision is made based on the samples from the two populations. It is a more robust alteration of the widely used Student's t-test [12].

Both Student's and Welch's t-test assume that the two examined populations follow a normal distribution [12]. Nevertheless, when testing for the equality of means of "large enough samples", the normality assumption can be violated thanks to the large sample theory and the Central Limit Theorem [5]. It has been shown in previous studies that for large samples, the statistical significance level is protected not only for normally distributed data, but also for many non-normal distributions; moreover, in case of Welch's test, this is true even for unequal variances [7, 15, 16]. According to Lehmann and Romano [5], the Type II error is also relatively insensitive to non-normality. Many articles and textbooks mention that when the sample sizes are small, nonparametric tests (i.e. tests that do not assume a specific distribution) such as the Mann-Whitney test [**?**] should be considered as an alternative to t-tests. However, t-tests become superior when sample sizes increase [7, 13]. The simulations made by Lumley *et al.* [7] show that "sufficiently large sample size" means under 100 in most cases. Even for extremely non-normal data, the sufficient size is at most 500. This suggests that the choice of Welch's test is legitimate for this work.

The problem of the Student's t-test is that it performs badly when the variances of the two compared populations are unequal. Both Type I and Type II errors are negatively affected by violation of the equal variances assumption. The unequal variances can be less problematic if sample sizes are similar, but in practice, that is not always the case [10].

Unlike Student's t-test, Welch's test does not assume equal variances of the populations. It performs well when the samples have unequal variances; furthermore, it can be used even when the samples have unequal sizes [4].

Some researchers tend to pre-test for variance equality by a preliminary test of variances (such as Levene's [6], Bartlett's [**?**] or Brown-Forsythe test [**?**]) and then choose whether to use Student's or Welch's t-test. However, although this approach persists in some textbooks and software packages, it is not recommended by statisticians. As a preliminary test itself is subject to Type I and II errors, this two-stage procedure would not protect the significance level and could lead to incorrect decisions. One should be aware of the fact that even if the test suggested that the samples variances are nearly equal, it would not mean that the whole population variances could not differ to a larger extent [14]. Some researchers may try to make the significance level of a preliminary test more strict, so that they could be more confident about the choice of the subsequent test; however, as the significance level decreases, the performance of the compound test paradoxically gets worse. According to Zimmerman [14], "a higher Type I error rate of the preliminary test actually improves the performance of the compound test". This suggests that using the preliminary test is not correct in principle.

Welch's test should be used whenever the researcher is not sure that the variances are truly equal. Ruxton [10] even suggests the routine use of Welch's test. When the sample sizes and variances are equal, both tests perform similarly. When dealing with unequal variances and unequal sample sizes, Welch's test is more robust than Student's t-test and the Type I error rate does not deviate far from the nominal value [4]. Hence, Welch's test can be applied without any significant disadvantages to Student's t-test.

For all the reasons stated above, Welch's test seems to be the best choice for the purpose of this study. It has the best combination of performance and ease of use, the calculation is straightforward and it is available in commonly used statistics packages.

2.1.2 Fisher's exact test and Chi-squared test of independence

A different kind of tests will be needed for the analysis of categorical features. An example of a categorical feature is XXX. Moreover, quantitative features can be grouped into categories and analysed in the same way as categorical features. In this section, two widely-used tests of such kind will be presented and discussed.

Both Fisher's exact test and Chi-squared test of independence are well-known hypothesis tests used for the analysis of data in contigency tables. A *contigency table* is a table displayed in a form of a matrix where cells represent a frequency distribution of samples in the categories. An example of a contingency table can be seen in Table 3.2. The sums of frequencies in rows and columns are called *marginal totals*.

	PTM	Without PTM	Total
Binding sites	XX	XX	XX
Non-binding sites	XX	XX	XX
Total	XX	XX	XX

Table 2.2: A 2×2 contingency table. TODO real data

The null hypothesis assumes independence of the groups; in our case, the assumption is that there is no difference in the proportions of the analysed feature between binding sites and non-binding sites.

Fisher's exact test belongs to a class of so-called *exact tests*; it means that the p-value is calculated accurately, not approximately, as is the case of many tests including Welch's test and Chi-squared test. Fisher's test is mostly used for 2 × 2 contigency tables, although the principle of the computation can be extended to a general $m \times n$ table [8]. The principle of the test lies in computing the probability of obtaining a table that is more or equally extreme in the departure from the null hypothesis than the analysed table and has identical marginal totals [2].

Chi-squared (χ^2) test if independence is able to decide whether the difference between the observed frequencies and the "expected frequencies" is statistically significant. The expected frequencies are computed for every cell as $\frac{row\ total \times column\ total}{grand\ total}$. It can be imagined as the average frequencies we would get in the long run with the same marginal totals, assuming the null hypothesis is true (i.e. there is no association between groups). The result of the test tells how likely are we to observe given data under the assumption of the true null hypothesis [2]. TODO vysvetlit chi-sq rozdeleni a to jak se pocita ta testova statistika

The biggest difference between the two mentioned tests is that the chi-squared test is based on a aproximation approach; therefore, it needs a "large enough" sample. TODO FISHERUV TEST OBECNE NA MENSI VZORKY, ... W. G. Cochran (1954) [**?**] proposed a set of recommendations about the minimum expectations to be used in χ^2 tests and about the choice between Fisher's test and χ^2 test:

These recommendations are presented in several textbooks and articles as a rule of thumb [] and recommended to be used in practice.

2.2 Data

plus significance level, null hypothesis, alternative hypothesis

2.3 Implementation in Python

3. 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.

3.1 Hypothesis testing

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

A dataset usually contains a subset sampled 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.e. a function of the sample) [3].

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 3.1. The ideal test would have both error probabilities equal to zero. Nevertheless, in most cases it is not possible to make both error probabilities arbitrarily small for a fixed sample size [3].

To control statistical significance of the result, a study defines a threshold called *significance level*, a constant denoted by α. It represents 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 accepted/re-

		Prediction	
		Accept H_0	Reject H_0
Truth	H_0	Correct (true positive)	**Type I error** (false positive)
	H_1	**Type II error** (false negative)	Correct (true negative)

Table 3.1: Type I and II Error in hypothesis testing.

jected. P-value gives an idea of how strongly the data contradict the null hypothesis; furthermore, it allows other researchers to make a decision according to the significance level of their choice [3, 5, 11].

It is suggested that the significance level for a study is set prior to any data collection [9]. The typical choices in practice are $\alpha = 0.01$, 0.05 or 0.10 [3]. One should 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 such as the accuracy and completeness of the data and most importantly, the true effect size [11].

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 [3].

3.1.1 Welch's test

Welch's unequal variances t-test, or Welch's test in short, is a two-sample hypothesis test used to decide whether two populations have different central tendencies (means or medians). The decision is made based on the samples from the two populations. It is a more robust alteration of the widely used Student's t-test [12].

Both Student's and Welch's t-test assume that the two examined populations follow a normal distribution [12]. Nevertheless, when testing for the equality of means of "large enough samples", the normality assumption can be violated thanks to the large sample theory and the Central Limit Theorem [5]. It has been shown in previous studies that for large samples, the statistical significance level is protected not only for normally distributed data, but also for many non-normal distributions; moreover, in case of Welch's test, this is true even for unequal variances [7, 15, 16]. According to Lehmann and Romano [5], the Type II error is also relatively insensitive to non-normality. Many articles and textbooks mention that when the sample sizes are small, nonparametric tests (i.e. tests that do not assume a specific distribution) such as the Mann-Whitney test [**?**] should be considered as an alternative to t-tests. However, t-tests become superior when sample sizes increase [7, 13]. The simulations made by Lumley *et al.* [7] show that "sufficiently large sample size" means under 100 in most cases. Even for extremely non-normal data, the sufficient size is at most 500. This suggests that the choice of Welch's test is legitimate for this work.

The problem of the Student's t-test is that it performs badly when the variances of the two compared populations are unequal. Both Type I and Type II errors are negatively affected by violation of the equal variances assumption. The unequal variances can be less problematic if sample sizes are similar, but in practice, that is not always the case [10].

Unlike Student's t-test, Welch's test does not assume equal variances of the populations. It performs well when the samples have unequal variances; furthermore, it can be used even when the samples have unequal sizes [4].

Some researchers tend to pre-test for variance equality by a preliminary test of variances (such as Levene's [6], Bartlett's [**?**] or Brown-Forsythe test [**?**]) and then choose whether to use Student's or Welch's t-test. However, although this approach persists in some textbooks and software packages, it is not recommended by statisticians. As a preliminary test itself is subject to Type I and II errors, this two-stage procedure would not protect the significance level and could lead to incorrect decisions. One should be aware of the fact that even if the test suggested that the samples variances are nearly equal, it would not mean that the whole population variances could not differ to a larger extent [14]. Some researchers may try to make the significance level of a preliminary test more strict, so that they could be more confident about the choice of the subsequent test; however, as the significance level decreases, the performance of the compound test paradoxically gets worse. According to Zimmerman [14], "a higher Type I error rate of the preliminary test actually improves the performance of the compound test". This suggests that using the preliminary test is not correct in principle.

Welch's test should be used whenever the researcher is not sure that the variances are truly equal. Ruxton [10] even suggests the routine use of Welch's test. When the sample sizes and variances are equal, both tests perform similarly. When dealing with unequal variances and unequal sample sizes, Welch's test is more robust than Student's t-test and the Type I error rate does not deviate far from the nominal value [4]. Hence, Welch's test can be applied without any significant disadvantages to Student's t-test.

For all the reasons stated above, Welch's test seems to be the best choice for the purpose of this study. It has the best combination of performance and ease of use, the calculation is straightforward and it is available in commonly used statistics packages.

3.1.2 Fisher's exact test and Chi-squared test of independence

A different kind of tests will be needed for the analysis of categorical features. An example of a categorical feature is XXX. Moreover, quantitative features can be grouped into categories and analysed in the same way as categorical features. In this section, two widely-used tests of such kind will be presented and discussed.

Both Fisher's exact test and Chi-squared test of independence are well-known hypothesis tests used for the analysis of data in contigency tables. A *contigency table* is a table displayed in a form of a matrix where cells represent a frequency distribution of samples in the categories. An example of a contingency table can be seen in Table 3.2. The sums of frequencies in rows and columns are called *marginal totals.*

	PTM	Without PTM	Total
Binding sites	XX	XX	XX
Non-binding sites	XX	XX	XX
Total	XX	XX	XX

Table 3.2: A 2 × 2 contingency table. TODO real data

The null hypothesis assumes independence of the groups; in our case, the assumption is that there is no difference in the proportions of the analysed feature between binding sites and non-binding sites.

Fisher's exact test belongs to a class of so-called *exact tests*; it means that the p-value is calculated accurately, not approximately, as is the case of many tests including Welch's test and Chi-squared test. Fisher's test is mostly used for 2 × 2 contigency tables, although the principle of the computation can be extended to a general $m \times n$ table [8]. The principle of the test lies in computing the probability of obtaining a table that is more or equally extreme in the departure from the null hypothesis than the analysed table and has identical marginal totals [2].

Chi-squared (χ^2) test if independence is able to decide whether the difference between the observed frequencies and the "expected frequencies" is statistically significant. The expected frequencies are computed for every cell as $\frac{row\ total \times column\ total}{grand\ total}$. It can be imagined as the average frequencies we would get in the long run with the same marginal totals, assuming the null hypothesis is true (i.e. there is no association between groups). The result of the test tells how likely are we to observe given data under the assumption of the true null hypothesis [2]. TODO vysvetlit chi-sq rozdeleni a to jak se pocita ta testova statistika

The biggest difference between the two mentioned tests is that the chi-squared test is based on a aproximation approach; therefore, it needs a "large enough" sample. TODO FISHERUV TEST OBECNE NA MENSI VZORKY, ... W. G. Cochran (1954) [**?**] proposed a set of recommendations about the minimum expectations to be used in χ^2 tests and about the choice between Fisher's test and χ^2 test:

These recommendations are presented in several textbooks and articles as a rule of thumb [] and recommended to be used in practice.

3.2 Data

plus significance level, null hypothesis, alternative hypothesis

3.3 Implementation in Python

Conclusion

Conclusion.

List of Abbreviations

AA	Amino acid
atd	a tak dale

Bibliography

[1]

[2] Martin Bland. *An introduction to medical statistics.* Oxford University Press, Oxford New York, 1987. ISBN 0192615025.

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

[4] B. Derrick and P. White. Why welch's test is type i error robust. *The Quantitative Methods for Psychology*, 12(1):30–38, jan 2016. doi: 10.20982/tqmp.12.1.p030.

[5] Joseph P. Romano Erich L. Lehmann. *Testing Statistical Hypotheses.* Springer New York, 2008. ISBN 0387988645. URL `https://www.ebook.de/de/product/3186913/erich_l_lehmann_joseph_p_romano_testing_statistical_hypotheses.html`.

[6] Howard Levene. Robust tests for equality of variances. In Ingram Olkin, editor, *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pages 278–292. Stanford University Press.

[7] Thomas Lumley, Paula Diehr, Scott Emerson, and Lu Chen. The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, 23(1):151–169, may 2002. doi: 10.1146/annurev.publhealth.23.100901.140546.

[8] Cyrus R. Mehta and Nitin R. Patel. A network algorithm for performing fisher's exact test in r × c contingency tables. *Journal of the American Statistical Association*, 78(382):427, jun 1983. doi: 10.2307/2288652.

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

[10] Graeme D. Ruxton. The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. *Behavioral Ecology*, 17(4):688–690, may 2006. doi: 10.1093/beheco/ark016.

[11] 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. doi: 10.1038/nrg3706.

[12] B. L. WELCH. THE GENERALIZATION OF 'STUDENT's' PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED. *Biometrika*, 34(1-2):28–35, 1947. doi: 10.1093/biomet/34.1-2.28.

[13] Donald W. Zimmerman. Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *The Journal of Experimental Education*, 67(1):55–68, jan 1998. doi: 10.1080/00220979809598344.

[14] Donald W. Zimmerman. A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, 57(1):173–181, may 2004. doi: 10.1348/000711004849222.

[15] Donald W. Zimmerman and Bruno D. Zumbo. Rank transformations and the power of the student t test and welch t' test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 47(3):523–539, 1993. doi: 10.1037/h0078850.

[16] Bruno D. Zumbo and Daniel Coulombe. Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of reaction time. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 51(2):139–150, 1997. doi: 10.1037/1196-1961.51.2.139.

A. Attachments

The attached CD contains two attachments:

A.1 attachment1

blabla

A.2 attachment2

blabla