Classification of nucleic acid binding proteins by non-parametric statistical methods and machine learning

Gerhard Dürnberger

October 24, 2012

Contents

1 Introduction 1

2 Theory 5
- 2.1 Types of measurement scales 5
- 2.2 Non-parametric statistics 6
  - 2.2.1 U test 8
  - 2.2.2 Resampling methods 9
    - Permutation tests 10
    - Bootstrap 11
    - Cross-validation 12
- 2.3 Machine learning 13
  - 2.3.1 Support Vector Machines (SVMs) 14

3 Methods and Background 16
- 3.1 Nucleic acid structure prediction 16
- 3.2 Mass Spectrometry (MS) 17
  - 3.2.1 Ion sources 17
  - 3.2.2 Mass Analysers and Detectors 18
  - 3.2.3 Protein identification by mass spectrometry 21
  - 3.2.4 Normalisation of MS data 24
- 3.3 Statistical tests to identify protein specificities 26
  - 3.3.1 Resampling methods 26
- 3.4 Gene Ontology (GO) 28
  - 3.4.1 Enrichment of Gene Ontology Terms 29

4 Experimental methods 31
- 4.1 Nucleic acid affinity purification 31
- 4.2 Protein identification 32
  - 4.2.1 Mass spectrometry 32
  - 4.2.2 Peptide identification and protein grouping 32

5 Results **34**
5.1 Experimental approach . . . 34
5.1.1 Initial characterisation of the experimental data set . . 39
Filtering frequent hitters . . . 42
Identification of secondary interactors . . . 44
Searching for nucleic acid binding protein domains . . . 47
Searching for characterised nucleic acid binding protein domains . . . 47
Searching for uncharacterised nucleic acid binding protein domains . . . 48
5.1.2 Rsc normalisation . . . 52
5.2 Identifying the most powerful test to detect protein specificities 54
5.2.1 Generation of synthetic data sets . . . 58
5.2.2 Introduction of specificities . . . 59
5.2.3 Comparing statistical methods on synthetic data . . . 61
5.2.4 Comparing statistical methods on synthetic data for multiple cell types . . . 69
5.2.5 Comparing statistical methods on annotated data . . . 75
5.2.6 Choice of the statistical test . . . 78
5.3 Identification of protein specificities . . . 81
5.3.1 Validation based on external annotation . . . 82
5.3.2 Experimental validation . . . 82
5.4 Prediction of NABPs from Amino Acid Sequence . . . 88
5.4.1 Computation of charge profiles . . . 90
5.4.2 Model training and performance . . . 91
5.4.3 Model learning . . . 93
5.4.4 Parameter selection . . . 96
5.4.5 Application to proteomes of model organisms . . . 96

6 Conclusion and Discussion **100**

Bibliography **102**

Erklärung zur Verfassung der Arbeit

Gerhard Dürnberger
Weimarer Strasse 51/2/2, 1180 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschliesslich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

(Ort, Datum) (Unterschrift Verfasser)

Acknowledgements

This work at the interface between statistics, machine learning, informatics and biology was only possible due to the help of many people. First I want to thank my colleagues in the Colinge group at the Research Center for Molecular Medicine (CeMM) for great support and many helpful discussions. I am also very grateful to my colleagues in the Superti-Furga and Bennett lab for supporting me in the wet lab and all the patience to give me an insight in the every day work of a biologist. Especially I want to thank Jacques Colinge and Giulio Superti-Furga for their advice and being great supervisors. They also made it possible to get some experience in the lab as well as enabling me to perform part of the domain search with Tobias Doerks in the group of Peer Bork at the EMBL in Heidelberg, which was a great experience for me and to whom I am also very grateful.

Special thanks also goes to my academic supervisors Hans Lohninger and Gerhard F. Ecker for taking so much time for my issues and lots of good advice in numerous meetings.

Finally I also want to thank my friends and family who gave me lots of support which enabled me to carry out this project.

Thank you!

Vielen Dank!

Abstract

Interactions between proteins and nucleic acids play a fundamental role in many biological processes. Besides nuclear gene transcription, many processes including RNA homeostasis, protein translation and pathogen sensing for innate immunity involve protein-nucleic acid interactions. Transcriptional regulation is mainly facilitated by nucleic acid binding proteins (NABPs) that bind to specific nucleic acid sequence motifs. Therefore most work has focused on these sequence specific interactions, although sequence specific NABPs constitute only a fraction of all NABPs.

The aim of this study was to get an unbiased classification of both, sequence specific and non-sequence specific NABPs. Affinity purification in combination with high-resolution mass spectrometry allowed to cover a large portion of NABPs in the human proteome and their interactions with nucleic acid. Here, 25 systematically designed oligonucleotides were used to probe three human cell lines for NABPs. Overall, more than 10,000 interactions were detected between the nucleic acid baits and almost one thousand unique proteins.

Statistical methods to derive a classification based on this experimental proteomics data set are evaluated in this work. Non-parametric statistical methods allowed the classification of experimentally detected NABPs with high sensitivity. Application of these methods permitted to determine the specificity of 174 NABPs for different classes of nucleic acids. These findings were validated by additional experiments and available annotations. Among these we could show cytosine methylation specific binding of Y-box-binding protein 1 (YB-1). A novel finding with potential implications for cancer research.

Sequence analysis of the detected proteins revealed candidate nucleic acid binding protein domains. Furthermore, the extracted classification allowed us to implement a support vector machine that can predict NABPs with high specificity solely from the amino acid sequence.

Kurzfassung

In vielen grundlegenden biologischen Prozessen sind Interaktionen zwischen Proteinen und Nukleinsäuren beteiligt. Neben Gentranskription, sind diese Interaktionen unter anderem auch an Prozessen wie RNA-Homöostase, Proteintranslation und der Erkennung von Pathogenen in der Immunabwehr beteiligt. Transkription wird oft von Nukleinsäure-bindenden Proteinen (NABPs) reguliert, die an spezifische DNA-Sequenzmotife binden. Bisher hat sich die Forschung groteils auf diese Proteine fokussiert, obwohl sequenzspezifische NABPs nur einen Bruchteil aller NABPs ausmachen.

Ziel dieses Projekts ist es, beide Gruppen - sowohl sequenzspezifische, als auch nicht-sequenzspezifische NABPs - gemeinsam zu klassifizieren. Moderne Massenspektrometrische-Methoden erlauben es, einen breiten Überblick über diese Interaktionen zu erreichen. Dazu wurden 25 systematisch entworfene Oligonukleotide verwendet um NABPs aus drei verschiedenen menschlichen Zelllinien anzureichern. Insgesamt konnten mehr als zehntausend Interaktionen zwischen diesen synthetischen Nukleinsäuren und fast eintausend verschiedenen Proteinen gemessen werden.

In dieser Arbeit werden statistische Methoden evaluiert, um eine möglichst umfassende Klassifikation aus dem experimentellen Datensatz zu erzielen. Parameterfreie statistische Methoden haben sich als sehr sensitiv herausgestellt und erlauben es, eine grosse Anzahl an Proteinen zu klassifizieren. Durch diese Methoden konnten Speziftäten von 174 NABPs erkannt werden. Die gewonnenen Ergebnisse wurden mit verfügbarer Annotation und zusätzlichen Experimenten validiert. Unter anderem konnten wir zeigen, dass YB-1 an methylierte DNA bindet. Diese neue Erkenntnis könnte Auswirkungen in der Krebsforschung haben.

Mittels Sequenzanalyse der neu entdeckten NABPs konnten potentielle Nukleinsäure-bindende Proteindomänen identifiziert werden. Weiters konnte anhand der generierten Klassifikation ein auf Support Vector Machines basierender Klassifikator implementiert werden, der es erlaubt, NABPs mit hoher Spezifität lediglich anhand der Aminosäuresequenz vorherzusagen.

Chapter 1

Introduction

Proteins and nucleic acids are the key molecules of life. Many proteins have enzymatic functions and thereby conduct important biological tasks. Furthermore they also play a role as structural and signalling molecules. Also nucleic acids have multiple roles in the cell. DNA contains genes which encode the blueprint for the assembly of proteins, whereas RNAs, besides their involvement in transcription, also fulfil structural and enzymatic roles (i.e. rRNA, tRNA, ...).

Elucidation of enzymatic protein functions often reveals that many proteins do not accomplish their function in isolation but rather form complexes. These protein complexes, which consist of multiple different proteins, then fulfil a functional role [46, 45].

The importance of protein-protein interactions and protein complexes will now be illustrated by two examples. The ribosome is a huge complex composed of two subunits. Both of these subunits themselves consist of an RNA polymer - the ribosomal RNA (rRNA) - and multiple proteins that interact with the rRNA. Together these molecules form a stable structure which is able to synthesize proteins from a messenger RNA (mRNA) template. This process is also known as translation, as here the genetic nucleotide information is translated into an amino acid sequence.

DNA polymerase is another example. It also requires the interaction of multiple cofactors to fulfil its biological function. Many interactions are required to allow transcription, ranging from interaction with transcription factors in the recruitment phase to helicases that assist the transcription process by unwinding the double stranded DNA template.

Both examples are very interesting and have great relevance in biology. They are also interesting from another perspective as they do not only involve interactions between proteins, but also interactions between proteins and nucleic acid polymers are involved. More specifically these interactions occur between the ribosomal proteins and the rRNA which together form the ribosome, but also with mRNA as a substrate in case of the ribosome. In the case of polymerase interactions occur between the polymerase and the DNA template as well as a primer, which is required to initiate the transcription process. These examples also demonstrate that focusing on interactions between proteins neglects an important part of functionally relevant interactions which involve nucleic acids.

In general interactions between proteins and nucleic acids are involved in many biological processes. These processes include transcription and translation as described above, but range much further to regulation of the lifespan of nucleic acid molecules via stabilisation or controlled degradation [76]. Even sensing pathogenic factors by recognising specific aspects of viral nucleic acids and subsequent triggering of immune response involves interactions between proteins and nucleic acids. Therefore nucleic acid binding proteins (NABPs), which are also at the focus of this study, have always been of key interest for biologists.

Besides focused research on key biological processes, as for example in transcription and translation, a technique called Chromatin Immunoprecipitation (ChIP) [48, 68] has provided lots of information about interactions between proteins and nucleic acids. Here a protein of interest is cross-linked to interacting DNA fragments, which subsequently allows characterization of the DNA binding sites. ChIP is extensively used to determine distribution of epigenetic marks such as histone modifications and cytosine methylation as well as to elucidate the binding sites of transcription factors. Analysing ChIP samples by current high-throughput methods such as DNA microarrays (ChIP-chip [60, 75, 96, 98]) or next generation sequencing (ChIP-seq [6, 61, 108]) allows mapping of protein-DNA interactions on a genomic scale. An example of a transcription factor binding motif for the neuron-restrictive silencer factor (NRSF also known as REST) derived by ChIP-seq is shown in Figure 1.1.

More system wide approaches to decipher protein-nucleic acid interactions were performed using protein arrays and protein binding DNA-microarrays. Hu and colleagues extracted almost 4200 recombinant proteins which were spotted onto a protein array. These chips were probed against a selection of 460 known and predicted DNA motifs [59]. Analysis of this massive data set revealed sequence specific binding for many unexpected proteins. This by itself already indicates the value of large scale studies performed nowadays. In another effort the group of Martha Bulyk performs a contrary approach [12]. She initially showed the ability of custom DNA microarrays to reveal the sequence specificity of NABPs. Therefore diverse sequences are spotted onto an array and a protein is probed against these potential binding sites.

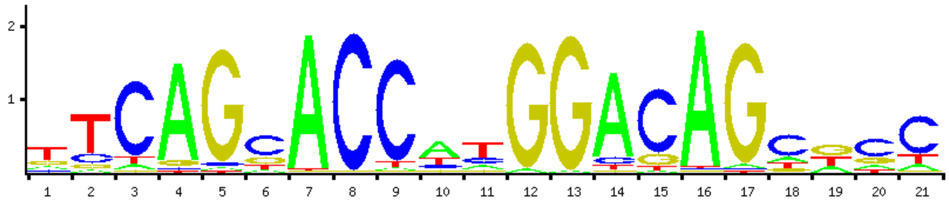


Figure 1.1 – Transcription factor binding motif for NRSF/REST [61], downloaded from the JASPAR database [94]. Individual nucleotides are scaled according to their conservation in the transcription factor binding site which is derived from the position specific enrichment in the ChIP-seq data.

Interactions can be measured by detecting the bound protein by fluorescence. Meanwhile, this approach also has been expanded to chips containing all possible octamers [115] and yields results complementing ChIP-experiments, without being biased to DNA motifs over-represented in the used model organism.

All of the mentioned techniques have a strong focus on measuring the sequence specificity of selected proteins. Querying Gene Ontology (GO) [4] annotation of the human proteome which in total contains 20422 different proteins [1] reveals that only 550 of the 2018 DNA binding proteins are known to bind specific DNA sequences (see Table 1.1). This clearly shows that focusing on sequence specific binders omits a large portion of DNA binding proteins.

Table 1.1 – Number of proteins classified to different sub-categories of the GO term "nucleic acid binding" within the human proteome. Sequence specific DNA binding proteins only account for 550 out of 2018 DNA binding proteins.

GO id	GO term	Proteins	percentage
GO:0003676	nucleic acid binding	3015	14.8
GO:0003677	DNA binding	2018	9.9
GO:0043565	sequence-specific DNA binding	550	2.7
GO:0003723	RNA binding	821	4.0

Although this annotation is incomplete, it clearly shows that sequence specific protein-nucleic acid interactions only account for a fraction of all protein-nucleic acid interactions, which in turn motivates broader studies which try to measure protein-nucleic acid interactions more comprehensively.

Going back to protein-protein interactions (ppi), these are most frequently measured by two techniques, yeast two-hybrid (Y2H) and affinity purification coupled to mass spectrometry (AP-MS). Meanwhile first attempts to measure protein-nucleic acid interactions by AP-MS have been undertaken. For example Lambert and colleagues have developed a method called modified chromatin immuno precipitation (mChIP) [71] to purify whole chromatin complexes that contain a genomic DNA fragment and interacting protein complexes. Therefore a DNA binding protein is tagged and used to precipitate interacting DNA fragments. Mild washing conditions allow to co-purify also protein complexes that interact with the enriched DNA fragments. Applying this approach to 102 chromatin related proteins in yeast could identify interactions to more than 700 prey proteins, revealing new partners for low affinity transcription factors [70] and proofs the value of such an approach for identification of novel chromatin associated protein complexes and further downstream analysis. Also the group of Matthias Mann could demonstrate the power of AP-MS approaches in two "proof of principle" papers. Quantifying interacting NABPs by Stable Isotope Labelling by Amino acids in Cell culture (SILAC [89]) - a quantitative proteomics approach - they could confirm a sequence specific transcription factor/DNA interaction. Furthermore they could also show that this approach is sensitive enough to reveal specific binding of a protein called Kaiso to methyl-cytosine in the MTA2 gene promoter [85]. In another study they could demonstrate that AP-MS approaches are also capable of selectively measuring interactions between proteins and RNA [15]. These results demonstrate that meanwhile AP-MS approaches are sensitive enough to measure protein/nucleic acid interactions and thereby allow to complement other established techniques.

Our lab has applied similar approaches to identify novel sensors of the innate immune system which sense for nucleic acids of pathogens and thereby trigger immune response [93, 13, 64]. For example we could show that a protein called AIM2 acts as a sensor for DNA in the cytoplasm and triggers downstream immune response [13]. Here we want to employ a similar approach on a larger scale with the aim to derive a novel classification of NABPs. In contrast to most of the studies described above the approach will be performed in human samples which increases complexity but also scientific relevance of the study. Of course the analysis of large data sets deserves statistical care-taking, therefore this study will evaluate adequate statistical methods to extract meaningful results from a medium size systems biology data set.

Parts of this thesis are taken from a paper [34] describing this project.

Chapter 2

Theory

This chapter tries to briefly introduce the theory behind the statistical methods applied throughout this thesis.

2.1 Types of measurement scales

Scientific progress in empirical sciences relies on measurements. These measurements can be acquired on different scales. The theory of scales as used nowadays was first described by Stanley Smith Stevens in 1946 [104]. Here four different levels of scales were distinguished:

1. The nominal scale basically allows assigning objects to different groups but does not capture relationships between the groups. The only comparison which can be done for objects on the nominal scale is comparing for equality, and thereby identifying if two objects belong to the same group.

2. On the ordinal scale relationships are defined and therefore it is possible to rank objects. This is enabled because operations for the comparison of objects like smaller and larger are defined on the ordinal scale which in turn permits ranking. Ordinal data can be further subdivided into singular- and grouped ordinal scaled data. This depends on the possibility of ties in the measurements, so if no ties can occur, the data is of singular ordinal scale.

3. On the interval scale further operations like subtraction and addition are defined. Still the zero point on this scale can be arbitrary which makes numerical operations like multiplication or division meaningless. The classical example for an interval scale is the Celsius temperature

scale. Here the difference between 10 and 20 degrees is as large as between 20 and 30 degrees which is not necessarily true on an ordinal scale. But 20 degrees does not mean that it is twice as warm as at 10 degrees.

4. The highest level of measurement is the ratio scale. Here multiplication and division are also valid. The Kelvin temperature scale is an example for a ratio scale, in contrast to the Celsius temperature scale, which is an interval scale.

These four levels are hierarchical, which means that measurements from a higher scale can be transformed onto a lower scale, but the opposite is not possible, e.g. it is possible to transform data from an interval scale onto ordinal scale but the reverse transformation cannot be performed. As a consequence of scale properties not all statistical measures make sense on each scale. For example it is not meaningful to compute the mean of a nominal scaled parameter. Here computing the mode is more appropriate.

2.2 Non-parametric statistics

For choosing a statistical test it is important to determine if the distribution, from which the samples were drawn, is known. A variety of tests, which allow testing of different hypotheses, have been developed for multiple background distributions. Probably the best known example is the t-test, which allows testing for a difference in means between two normally distributed samples. As the assumed distributions can be specified by their parameters and therefore the tests relies on the parameters of the distribution, these tests are known as parametric tests. If the data points follow a standard statistical distribution, for which a test that can evaluate the hypothesis of interest has been described, the corresponding parametric test should be applied. Otherwise, if it is not clear from which background distribution the samples were drawn or there is not sufficient similarity to a standard distribution, non-parametric tests are more appropriate. This decision is sometimes difficult as parametric tests are somewhat robust to a deviation from the assumed distribution. Also, if the data was measured on an ordinal scale, non-parametric methods have to be applied. This is a clear advantage of non-parametric methods, as for most parametric methods it is required that the data is measured at least at an interval scale due to the design of their test statistic.

As mentioned above, parametric tests require the sample data to be derived from a defined statistical distribution. Non-parametric tests are more flexible, they do not have this requirement and can be applied to both kinds of data, independent whether the data was derived from a known distribution or not. This is a big advantage, as they can be applied to a wider range of data sets without requiring to check this precondition. Of course this versatility comes at a price. As parametric tests have higher prior knowledge about the measurements, they are more powerful in detecting deviations from the null hypothesis in the data. This means, that an incorrect null hypothesis is correctly rejected. The power of a test is computed by the fraction of examples in which a test correctly identifies an incorrect null hypothesis at a given effect size. This value is usually computed from simulated data using Monte Carlo methods. The second disadvantage of applying non-parametric tests on data that would be suited to application of a parametric test is that the efficiency of the test is reduced [72]. The efficiency of a test allows power comparisons of statistical tests with identical alternative hypothesis. It is the ratio of the population sizes that are required to significantly reject the null hypothesis at a given effect size. Here the population size of the stronger (parametric) test is used as the numerator and hence the efficiency of a test that requires a larger data set is below one. As this measure depends on the given effect size Δ and the level of significance α more generic measures have been developed. It is possible to either increase the population size of the stronger test towards infinity and get an asymptotic relative efficiency (ARE) or to use Monte-Carlo-Methods to simulate the difference in efficiency. Here sample data that does not fulfil the null hypothesis is simulated and the number of correct rejections of the null hypothesis is counted. The ratio between these two numbers is then called the (local) relative efficiency. The local relative efficiency is dependent on the sample size. Parametric tests are usually more efficient on data, which fulfils the assumptions underlying the parametric test. If these prior assumptions are not fulfilled, non-parametric methods rapidly catch up in efficiency and outperform parametric methods, although parametric methods are to a certain degree robust against violations of the prior assumptions. This is especially pronounced for small sample sizes [56]. The superior performance of parametric methods on data that fulfils the requirements for a parametric test is also obvious to a certain degree, as most non-parametric methods are employed on an ordinal scale and therefore information is lost compared to parametric methods which rather use the interval scale directly. Also parametric methods take advantage of more prior knowledge, as the distribution of the data is pre assumed. This disadvantage of non-parametric methods is of course more prominent in idealised text book examples and usually less pronounced on not so well distributed "real life" data sets.

In the opposite case, applying a parametric test to data that does not fulfil the requirements of the parametric test i.e. is not distributed according to the tests precondition, the results will be incorrect.

On the other hand, applying a parametric test to data which does not fulfil the requirements of the parametric test will cause incorrect results. As mentioned above, there is some robustness to deviations from the preconditions in parametric tests, but to be on the safe side applying non-parametric methods is advisable. In practice the requirements for parametric tests are often not fulfilled and applying non-parametric methods is more suitable. Therefore non-parametric analogues to most parametric statistical methods have been developed.

Non-parametric statistical methods will now be illustrated by an example of a non-parametric test which will also be used later in this thesis: the U test.

2.2.1 U test

The U test is a frequently used non-parametric test which can test for difference in population medians. In contrast to a standard t-test, it allows to compare samples that are not derived from a normally distributed population. Also when the data is measured on an ordinal scale or outliers are expected in the data, the U test is the appropriate method to compare samples. Of course it can also be applied to normally distributed samples, but then it suffers from reduced power compared to the t-test, which is better suited in this case. The test was originally developed by Wilcoxon [109, 77] as well as Mann & Whitney [81] and is therefore also known as the Wilcoxon test or the Mann-Whitney test. The null hypothesis which is tested by the U test is, if there is a significant difference between the average rank of the two samples (see Equation 2.1) after they were transferred onto a common rank scale [109].

$$H_0 : E(\overline{R_1}) = (\overline{R_2}) \tag{2.1}$$

The underlying test statistic U is computed by summing the number of measurements in the second sample that are larger than the measurement in the first population for all measurements in the first population. The same statistic is computed after switching the groups (U'). These sums can be efficiently computed by Equations 2.2 and 2.3 respectively.

$$U = N_1 N_2 + \frac{N_1(N_1+1)}{2} - T_1 \tag{2.2}$$

$$U' = N_1 N_2 + \frac{N_2(N_2+1)}{2} - T_2 \tag{2.3}$$

Here N_1 and N_2 are the sample sizes of the two groups and T_1 and T_2 are the sums of the ranks of the respective group. The minimum of these two values U and U' is used as the test statistic. The distribution of the test statistic can be computed from combinatorial considerations and hence the smaller of both values is compared with the test statistic at a given significance level α. If both sample sizes are larger than 20 the distribution of U can be approximated by a normal distribution [52]. The mean and variance for this normal distribution are computed according to Equations 2.4 and 2.5.

$$E(U) = \frac{N_1 N_2}{2} \tag{2.4}$$

$$\sigma_U = \sqrt{N_1 N_2 \frac{N_1 + N_2 + 1}{12}} \tag{2.5}$$

A test which is mathematically equivalent was developed by Wilcoxon [110] as already mentioned earlier. This test statistic is based on the sum of the ranks of the smaller sample. This test is also known as the Wilcoxon rank sum test. If ties occur in the sample data, the variance of the test statistic is reduced and a correction needs to be applied [25].

2.2.2 Resampling methods

Resampling methods are a group of statistical methods that allow analysis by sampling from the sample data. As the samples of the resampling method are drawn from the original sample data, these methods usually do not rely on assumptions regarding the distribution of the data. To obtain accurate results, resampling methods require taking a high number of samples of the original data and performing computations on them individually. This repeated procedure obviously requires a higher number of computations compared to standard statistical methods and was a limiting factor for the application of resampling methods before the availability of computers. With the propagation and progress of computational power, resampling methods have become more attractive and widely used. This is also reflected in an increase of scientific publications in the field since the eighties of the last century.

Resampling methods allow to estimate the precision of a statistical measure (confidence intervals), hypothesis testing and also to assess the performance of predictive methods. There are different methods that belong to the family of resampling methods some of which will now be introduced in further detail.

Permutation tests

Permutation tests - or exact tests - were first described by Fisher and Pitman. The beauty of permutation tests is, that they provide exact significance levels with only few mathematical assumptions, which makes them widely applicable. On the other hand their calculation is computation intensive, which has limited their use initially.

The basic idea behind permutation tests is to permute the data and thereby extract a distribution of the test statistic. This distribution is then used to determine the significance level of the value of the test statistic on the original data which was not permuted. The derived significance level can be used to perform hypothesis tests. Permutation tests do not require the sample data to be derived from a certain distribution and therefore belong to the group of non-parametric tests. Instead of inferring the distribution of the test statistic that is observed under certain preconditions, the original data is sampled to obtain the background distribution from the original sample data which preserves and also adapts to the distribution of the sample data. The permutation test is most frequently used in the two-sample problem. Here, each data point is connected to a group assignment. Now the permutation test shuffles these group assignments to guarantee that the null hypothesis is fulfilled. Subsequently the test statistic is computed for each of these permuted data sets. These values form the distribution of the test statistic under the null hypothesis. Comparing the value of the test statistic for the original - unpermuted - data to this distribution allows to infer the significance level. This is achieved by calculating the fraction of cases, where the test statistic of the permuted data is more extreme than the original value of the test statistic on the unpermuted data set. This measurement is called the achieved significance level (ASL). As described above, it is the fraction of events where the value of the test statistic of the permuted data sets $\hat{\theta}^*$ is more extreme than the statistic computed on the original data $\hat{\theta}$.

$$ASL = Prob_{H_0}\{\hat{\theta}^* \geq \hat{\theta}\} \tag{2.6}$$

The achieved significance level is equivalent to a P value and is used to decide if the null hypothesis is rejected at a certain significance threshold α. As the permutation test tests all possible combinations of the original data it provides an accurate P value. An accurate P value means that the test does not falsely reject a null hypothesis by delivering a too small P value. Although the P value is always accurate, the choice of the test statistic critically influences the power of the test [41, p. 211]. So by using a proper test statistic one can increase the power of the test that can better discriminate between data sets where the null hypothesis is true and others where it is not.

The number of possible permutations dramatically increases with the size of the data set. If the two groups in the two sample problem have size n and m, the number of possible permutations of the group assignments is

$$\binom{m+n}{m} = \frac{(m+n)!}{n!m!} \tag{2.7}$$

This number grows rapidly and therefore it is not always feasible to compute the test statistic of all possible permutations within reasonable time. Hence Monte Carlo methods that sample a random subset of all possible permutations have been developed [35]. Here the test statistic is only computed on a randomly chosen subset of all permutations and the distribution of these values is used as an approximation of the distribution of the test statistic for all permutations. Consequently, the result of the permutation test is not completely accurate any more, but if a sufficient number of permutations is performed the derived result is still very accurate.

Bootstrap

Bootstrap methods were first described by Bradley Efron in 1979 [40]. The bootstrap is mainly used to estimate confidence intervals for statistical measures. This is achieved by sampling data points from the original sample data and computing the measure on this resampled data set. Iteratively repeating this procedure generates a distribution of the measure of interest. This distribution can be used to infer confidence intervals of the parameter in the distribution from which the sample data was obtained. So for example the bootstrap allows computing a confidence interval for the mean of a distribution from which sample values were drawn, only by the use of these sample values. A huge advantage of the bootstrap is that these confidence intervals can also be computed for test statistics for which no analytical solution is known, e.g. the median or trimmed mean. As the bootstrap does not assume any distribution from which the samples were drawn it belongs to the family of non-parametric statistical methods.

Although computing confidence intervals of statistical measures is quite useful by itself, the bootstrap can also be used to perform hypothesis tests. This is achieved by computing the value of the test statistic on a sufficient number of bootstrap samples. To compute the significance level of the test, the expected value of the test statistic under the null hypothesis is compared to the values of the test statistic on the bootstrap samples. The quantile of the expected value of the test statistic in the bootstrap values reflects the significance level of the hypothesis test. For example if we want to test for equality of means by using the test statistic $\theta = \overline{x} - \overline{y}$ then the expected θ under the null hypothesis is $\theta = 0$. Now computing bootstrap replicates of the difference of means from the sample data provides a distribution. The quantile of the expected value of the test statistic $\theta = 0$ within the bootstrap values of the test statistic allows to compute the P value under which the null hypothesis should be rejected.

In the case of hypothesis testing, the bootstrap and permutation tests are closely related. The key difference is that each data point in the bootstrap data sets is sampled from the complete original data set, whereas in permutation tests the data is shuffled and essentially only the order is affected. Consequently a data point from the original data set can occur multiple times or not at all in a bootstrap sample, on the other hand each data point occurs exactly once in the permuted data set of a permutation test. So from a combinatorial point of view the difference between both methods is, that the sample data points are drawn with replacement by the bootstrap method, whereas they are drawn without replacement in the permutation test. This difference leads to an even faster growing number of possible resampling data sets compared to permutation tests (see Equation 2.7), because a single data point might occur multiple times. The number of possible data sets grows rapidly with the size of the data set (see Equation 2.8).

$$(n+m)^{(n+m)} \tag{2.8}$$

Therefore bootstrap is usually performed on a certain number of randomly chosen samples from all possible data sets and not all possible bootstrap data sets are evaluated, similar to Monte-Carlo permutation tests.

Cross-validation

Cross validation also belongs to the family of non-parametric methods, although the aim behind this method is quite different from methods introduced above. Whereas bootstrap and permutation tests are used to infer the population distribution of test statistics and thereby allow to derive confidence intervals and perform hypothesis tests, cross validation is used to measure the performance of classification algorithms.

A classifier with a high number of parameters can model the aspects of a given training data set in more detail. As the training data set besides the actual signal also contains noise, a complex classifier will start to model the noise in the signal. The aim in classification is to model a general trend in the data set - generalisation - without adapting to noise. A powerful classifier with too many parameters is likely to adapt to noise in the training set and therefore looses the ability to generalise. When a new entity is presented to this classifier, the performance will be dramatically below the performance estimated on the training set.

Cross validation is a technique to check against this overfitting. Here the data set is split into partitions to circumvent this problem. One partition is used as a training set, whereas the other partitions are used to evaluate the model performance on unseen entities. This partition, which was left out during the training, is called the test set. After training, model performance is measured on the test set. Here performance measures lead to more realistic results, as the data points in the test set were not seen by the classifier in the training phase. This procedure is iteratively repeated until each partition was used as the test set once.

If the data set is small it can be split into partitions of size one which is referred to as "Leave-one-out cross-validation" or "full cross-validation". This is computationally expensive for larger data sets, as the number of model trainings is equal to the number of data points. Therefore for larger data sets N-fold cross-validation was developed. Here the data set is split into N partitions that contain multiple data points of the training set.

2.3 Machine learning

Machine learning is a discipline of artificial intelligence. Here methods try to infer properties of a data universe from a training set which is a subset of the data universe. Extracting knowledge from this training set is achieved by finding a simplified model that can describe the training set. This process of identifying a simplified model is called generalization. Generalization allows to apply the model to novel entities in the data universe and predict their properties correctly. According to the principle of parsimony, the learned models should be as simple as possible to allow correct prediction on novel entities. If too complex models are chosen, they tend to overfit the training data and lose predictive power on novel entities. This phenomenon can be antagonised by splitting the training set in partitions to generate a training and a test data set, like for example by cross validation as described above.

Machine learning can be applied to both, labelled and unlabelled training data sets. Working on labelled data sets, where a desired outcome is provided in the training set is referred to as supervised learning, whereas when no labels are available, methods are known as unsupervised learning.

2.3.1 Support Vector Machines (SVMs)

Support vector machines are a supervised machine learning method which was invented by Cortes and Vapnik in 1995 [27]. They can be used for both, classification and regression problems [107]. Here only the use of SVMs for classification will be discussed.

To predict the desired output parameter, SVMs try to find an optimal hyperplane that can separate the entities of the training data into the groups specified in the output parameter of the training data set (see Figure 2.1a). This hyperplane is also referred to as decision surface, as it separates the data set into two groups and allows to decide on which side of the hyperplane a given data point lies. Usually the method is used to separate a data set into two groups (binary classification), but also extensions for multiple groups do exist [67].

If there are multiple hyperplanes which allow to separate the data set into the desired groups, the best hyperplane is selected by surrounding the hyperplane by a margin, in which no data points are allowed (see Figure 2.1b). The hyperplane with the largest margin in which no data points do occur, is chosen as the best hyperplane. As a consequence of this maximisation, SVMs belong to the group of maximum margin classifiers. Maximising the margin with the constraint that the decision surface is linear can be formulated as an optimisation problem. Such constrained optimisation problems can be efficiently solved using Lagrange optimisation. The derived solution is solely dependent on the few data points that lie exactly on the margin, whereas the remaining data points do not influence the solution. As these data points are sufficient to compute the optimal hyperplane, they are also referred to as support vectors.

Often the training data set contains noisy data points which are more likely to be support vectors. Therefore an extended approach known as soft margin classifiers [27] has been developed, which allows data points to be present within the margin (see Figure 2.1c). Depending on how far data points penetrate into the margin, they are penalised by so called slack variables. This penalty is also taken into account in the optimisation of the hyperplane. The magnitude of this penalty can be defined by the user in a parameter called cost coefficient. Of course a small penalty for data points in the margin creates a larger set of solutions and leads to a solution with a larger margin. Allowing data points to penetrate the margin usually makes the solution more robust and leads to improved classification performance, although the presence of data points within the margin can also lead to incorrect predictions if a data point is beyond the decision surface.

As the data set is separated by a hyperplane in feature space, it needs to

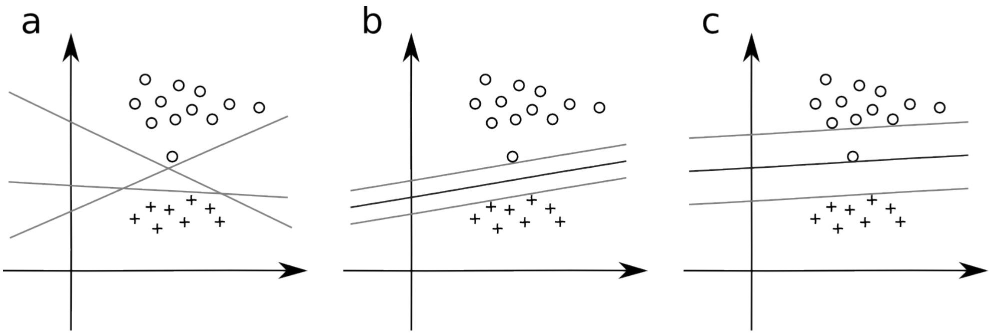


Figure 2.1 – Decision surfaces for a two dimensional data set. **(a)** Potential linear decision surfaces. **(b)** Maximum margin decision surface. The margin is shown in light grey, whereas the central decision surface is dark grey. **(c)** A soft margin decision surface. Here a potential outlier is located in the margin

be linearly separable for the algorithm to find a solution. Linearly separable means that the data points can be separated by a hyperplane into the two desired groups. This restriction can be circumvented by applying the so called "kernel trick" [2, 10], which transforms the input data into a higher dimensional feature space. Lagrange optimisation which is used to solve the optimisation problem and derive the optimum decision surface also has the advantage that it can be nicely combined with the kernel trick, which turns the linear into a very efficient non-linear classifier. Transforming a data set which is not linearly separable in the original feature space into a higher dimensional feature space using the kernel trick often allows to make the data set linearly separable in this higher dimensional feature space. Therefore different kernels such as linear-, polynomial- and Gaussian-kernels are available and frequently used. Unfortunately it is not possible to predict which kernel will lead to an optimal solution, therefore different kernels have to be evaluated. As using kernels increases the complexity of the model this can also lead to overfitting, which can be circumvented by evaluating model performance with cross validation.

Chapter 3

Methods and Background

3.1 Nucleic acid structure prediction

Nucleic acids are polymers, composed of a linear chain of nucleotides. DNA and RNA contain four different nucleotides, both species share adenine(A), cytosine(C) and guanine(G). DNA also contains thymine(T), whereas RNA contains uracil(U) as the fourth nucleotide base instead. The ability to form hydrogen bonds leads to increased affinity between Cytosine and Guanine, as well as adenine and thymine or Uracil respectively in RNA. For this reason the antisense strand in double stranded DNA contains the complementary base to maximize binding forces between both strands. In single stranded nucleic acids this leads to the formation of folding patterns. Upon a certain length, single stranded nucleic acid chains fold into structures depending on their nucleotide sequence, trying to pair as many nucleotides with affinity (C&G; A&T/U) besides other structural constraints like bending angles. These structures have implications on the function and recognition of nucleic acid molecules and therefore predicting this fold from the nucleotide sequence is an important task in computational biology. Many methods to predict RNA structures have been developed [116, 100, 28, 84, 29, 83].

As DNA occurs predominantly double stranded, research was focusing on the prediction of RNA structures, nevertheless single stranded DNA also forms structures. ViennaRNA [57] is a software package which allows the prediction of both, DNA and RNA secondary structures. It facilitates a dynamic programming approach to compute the secondary structure with minimum free energy which is also the most probable fold of the nucleic acid chain.

3.2 Mass Spectrometry (MS)

Mass spectrometry (MS) is a technology that allows to measure the mass of charged molecules with high accuracy. This is achieved by exposing the charged sample molecules to a defined force which in turn leads to acceleration of the ions. If the same force is applied to different ions, their acceleration is solely dependent on their mass [87]. Conversely, this enables the separation of molecules by their mass and thereby allows measuring the mass of charged molecules. As the force acting on the ion is also dependent on its charge, multiply charged molecules are exposed to a multiple of the force. Consequently the MS instrument does not measure the mass itself but the mass to charge ratio ($\frac{m}{z}$). Many different types of MS instruments have been developed which use the described principle.

Biological samples are often pre-separated by liquid chromatography (LC) to reduce their complexity at a given time-point of the analysis. Measuring exact masses allows identification of proteins. This can be achieved in two ways. Either intact proteins are measured "top-down-proteomics", or the protein can be digested with a protease and the resulting peptides are measured by MS "bottom-up". As different peptides can have identical masses, peptides are usually further fragmented and the mass of these fragments is used to identify the peptide ("Tandem-MS" or "MS/MS"). The coupling of liquid chromatography to Tandem Mass spectrometry - LC-MS/MS - is probably the most frequently used setup for protein identification nowadays. As the signal intensity for a certain molecule correlates with its abundance, also quantitative approaches are being developed.

In general an MS instrument consists of three major parts:

- an ion source, to charge the analyte molecules
- an analyser, to separate the ions based on their mass-to-charge-ratio
- and a detector to detect the ions

For each of these components multiple variants have been developed. Some of these with relevance in protein identification are further described here.

3.2.1 Ion sources

Biological samples are mostly ionised by electrospray ionization (ESI) [44] and matrix-assisted laser desorption/ionization (MALDI) [62]. MALDI and ESI provide soft ionization techniques that do not destroy fragile large molecules which is important in the analysis of biological polymers such as proteins and peptides.

Matrix-assisted Laser desorption/ionization (MALDI)

MALDI was developed by Karas in 1987 who observed, that mixing peptides with solvent and tryptophan allows ionization of alanine side chains in the peptides by pulsed laser stimulation [63]. Meanwhile multiple new matrix substances have been identified instead of tryptophan. These matrix substances increase laser absorption and allow to ionise more molecules in the sample. Although efficiency of the MALDI approach has been improved, additional sample preparation is required, which is the major disadvantage of MALDI.

Electro Spray Ionization (ESI)

Electro Spray Ionization was invented by John Fenn in 1989 [44], for which he also was rewarded the Nobel Prize in Chemistry in 2002. Here the sample is ejected through a small needle tip which is connected to a high voltage. This generates a spray containing charged droplets. Besides charged analyte molecules, these droplets also contain solvent. To get rid off the solvent in the droplets the spray is exposed to a strong counter flow of warm gas, which supports evaporation of the solvent which in turn decreases droplet size. Reduction of droplet size leads to increased charge density within the droplet and this furthermore leads to instability of the droplet when the forces within the droplet due to charge exceed the surface tension. Finally this causes a so called “Coulomb explosion” and subsequent desorption of the sample ions. The ESI approach has undergone substantial miniaturisation which also led to a reduction in sample flow rates. Along these lines technologies like “microspray” [42] and “nanospray” [111] have been developed. Of course this miniaturisation is beneficial for biological experiments with limited sample amount, e.g. when tiny model organisms or patient biopsies are analysed.

After sample ionisation the charged sample molecules are accelerated and focused to form an analyte beam within the mass spectrometer.

3.2.2 Mass Analysers and Detectors

Sector instruments

One approach to allow mass measurements is to send the ion beam through a magnetic or electro static field which applies a force to the ions and thereby influences their motion. The magnitude of this deflection is dependent on the mass (m) and the charge number (z) of the ionised molecule as well as the strength of the field. This leads to a spatial spreading of the beam. Altering the strength of the field or the initial momentum of the ions allows to control this deflection. This can be used to “scan” a range of masses over time by a single detector.

Quadrupole The quadrupole constitutes of four parallel, rod shaped electrodes which are arranged in a square. An alternating voltage (AC) is applied to two opposite electrodes and a second alternating voltage is applied to the two remaining electrodes. This generates a rotating electrostatic field which makes the path of ions passing through the quadrupole instable and only ions of a selected $\frac{m}{z}$ can pass through the quadrupole. Obviously this can be used to filter ions. After filtering, selected ions that passed the quadrupole can be measured by a detector. Similar to sector instruments this is used to scan the spectrum over time. Alternatively a quadrupole can also be used to filter ions ahead of fragmentation in Tandem MS.

Time of flight (TOF)

Another possibility to measure the mass-to-charge ratio of charged molecules in an MS instrument is to accelerate the ions by an electric field and measure the time it takes the particles to travel a known distance. Instruments which measure the mass-to-charge ratio by measuring the travelling time of the ions are referred to as “time of flight” (TOF) instruments. As the ions are exposed to the same electric field the acceleration is dependent on the ion mass. Heavier ions are less accelerated and will require a longer time to travel a certain distance. Of course also the charge of the ion affects the acceleration and doubly charged ions will be accelerated with double force. The acceleration can be applied into the original moving direction of the ions in the beam - therefore the beam needs to be gated, but also fields which accelerate the beam ions sideways (Orthogonal acceleration TOF) or even invert their moving direction (Reflectron TOF) are used. Using a reflectron allows to extend the path of the ions. Because travelling time can be measured at a certain precision, this leads to increased resolution of the instrument as a longer ion path leads to a stronger separation of sample ions within the time scale.

Orbitrap

Another possibility to determine the mass of the ions is to trap the ions in an orbit around a strong charged electrode. Mass analysers applying this principle are called Orbitraps. Alexander Makarov invented Orbitrap instruments in the year 2000 [80]. The idea of orbitrap combines principles already used in other instrument types (the Kingdon trap and Fourier Transform Ion

Cyclotron Resonance) in a sophisticated way. Here sample ions are captured in a collection quadrupole. After a short collection period of about 100 ms, the collected ions are injected into the orbitrap analyser by a very rapid pulse which is shorter than a microsecond [58]. This is important in order to deliver the sample ions within a package of limited size into the orbitrap. The orbitrap analyzer constitutes of a central spindle electrode and an outer barrel formed electrode. These two electrodes generate a strong electric field that can stabilise the injected ions on orbital trajectories around the central electrode. Figure 3.1 shows a schematic overview of the construction of an orbitrap instrument (schema taken from [58]). In contrast to a quadrupole where all the four electrodes are outside the sample the orbitrap applies a static field, to stabilise the ions on circular trajectories. The rotational speed of the sample ions here is dependent on mass-to-charge ratio $\frac{m}{z}$. The rotating ions create an alternating current which can be detected by sensor electrodes. As the rotational speed is dependent on the mass-to-charge ratio, the frequency of the detected current ("image current") allows to calculate the mass-to-charge ratio of the sample ions. This is achieved by calculating the frequencies of the recorded image current using Fast Fourier Transformation [26]. The resulting frequency spectrum can then be converted into a mass spectrum. As the sample ions can be circulated for a high number of orbital revolutions the frequency can be determined very precisely which in turn allows very precise mass measurements. This represents a big advantage of Orbitrap instruments.

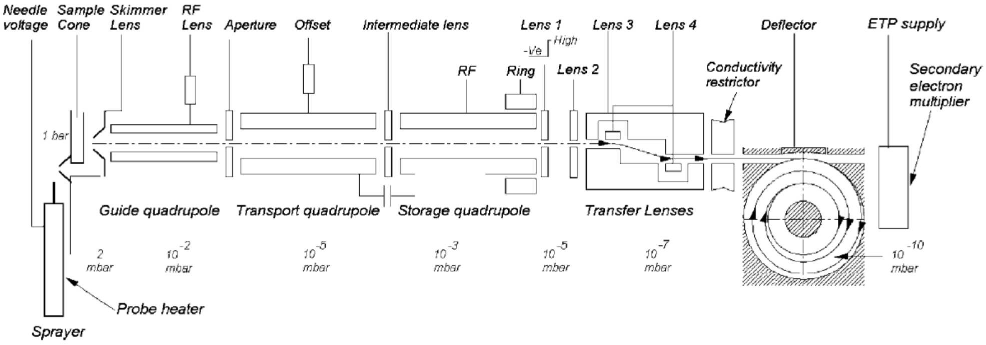


Figure 3.1 – Schematic overview of an orbitrap instrument, also indicating the pressure at individual components. (taken from [58])

A frequently used setup is to couple an ESI ion source to an orbitrap mass analyser [53]. Biological samples in this work were also analysed according to this setup.

3.2.3 Protein identification by mass spectrometry

Besides the identification of chemical compounds, the identification of proteins in biological samples is one of the main applications of MS. Here also the identification of protein modifications - like phosphorylation or ubiquitination - becomes increasingly important, as these are key regulatory mechanisms in biology.

Peptide mass fingerprinting

Identification of proteins can be achieved by an approach called "peptide mass fingerprinting". This approach was developed in 1993 by multiple groups [82, 54, 90]. Here the proteins in a biological sample are digested by a protease into peptides. To generate well defined peptides a protease with specific cleavage properties is used. Trypsin has been established as a frequent cleavage enzyme for protein digestion. It cleaves proteins at the carboxy side of lysine or arginine, except when these are followed by proline. Also chymotrypsin, which cleaves proteins at amino acids that contain an aromatic ring (tyrosine, tryptophan and phenylalanine), is used for protein digestion. The defined cleavage site of these enzymes allows to calculate the mass of expected peptide fragments from protein sequences. The masses measured in an MS experiment can then be compared to theoretical fragment masses of proteins in a protein sequence database.

Tandem MS and MSn

The information obtained from a peptide mass finger printing experiment is often too little to allow to uniquely identify a protein from large protein sequence databases. Therefore fragmentation approaches have been developed that allow to obtain more information about the peptides present in the sample and thereby enable more precise protein identification. Here, the peptide mixture injected into the instrument is analysed by an MS scan (MS1). Afterwards a subset of the observed peptides is selected for fragmentation. This choice is based on the observed spectra and can be adapted by targeted-aspects to focus the analysis on peptides/proteins of specific interest. The fragmentation process breaks chemical bonds within the peptide, which enables the measurement of fragment masses in a second MS run (MS2). This workflow is usually repeated in cycles, where one MS1 scan is followed by multiple MS2 scans. In total up to several thousands of spectra are acquired during the analysis of a single sample. Also approaches that perform multiple fragmentation cycles (MSn) are being used to further fragment molecules of interest.

Fragmentation techniques

Tandem MS and MSn approaches require further fragmentation of sample molecules. Therefore a variety of different approaches have been developed. Collision-induced dissociation (CID) and electron-transfer dissociation (ETD) are two frequently used techniques for peptide fragmentation. In CID the selected sample peptide is fragmented by collision with a neutral gas. The collision induces breakage of chemical bonds. These bond breaks can occur at different positions in the sample peptide. To improve identification of the sample peptide, bond breaks at peptide bonds are desired. Also bond breaks between the peptide backbone and the amino acid side chain do occur, which are less informative for peptide identification. A nomenclature for peptide fragments has been established which terms the ions dependent on breakage position and if the amino- or carboxy-terminus of the initial peptide is contained in the fragment ion.

In ETD [105] the sample peptide undergoes a chemical reaction with anions. This leads to fragmentation of the peptide and the transfer of an electron from the anion to the sample peptide. As an electron is transferred to the peptide fragments, ETD only works for multiply charged peptides. Singly charged ions would be neutralised by the electron transfer and therefore resulting fragments are not detectable in a mass spectrometer. The advantage of ETD is, that peptides are predominantly fragmented at the peptide bond and c and z ions are generated. These are more informative for peptide identification as well as for the identification of post translational protein modifications [23].

Additional technical aspects

To increase the analytical resolution of MS approaches, complex protein samples are usually pre-separated by liquid chromatography (LC) which is directly coupled to the MS instrument (online LC-MS). For further increase of analytical resolution it is also common to perform offline pre-separation of the complex protein mixture in the sample by gel electrophoresis or another LC run that uses a different gradient than the second - online - gradient. These offline separation techniques are performed before the actual MS analysis and the fractions generated are usually analysed as independent injections. Although all these improvements meanwhile have led to the identification of the complete proteome - all proteins - of a yeast cell [86], abundance of different protein species spans a wide range over several orders of magnitude. This high dynamic range of protein abundance in biological samples is usually still an issue in MS analysis and leads to the loss of identification of low abundant proteins.

Besides the identification of proteins themselves, the identification of protein modifications has become increasingly important. These post translational modifications (PTMs) play an important role in cellular regulation. As the modification leads to a predictable alteration of the molecular mass, mass spectrometry allows to identify these modifications. In practice many modifications are present only in a small fraction of the sample which makes initial enrichment of modified proteins a requirement.

Also quantitation of proteins by MS is becoming increasingly important. As regulation of biological systems is often mediated by regulating the abundance of protein species, determining and comparing quantitative proteomics data is instrumental to study regulatory processes in biological systems. Multiple approaches such as SILAC [89] or iTRAQ [101] have been developed. These techniques require chemical labelling of the samples and allow to measure abundance by comparing ion intensities in the MS spectra. The chemical labelling approaches are significantly more cost- and/or labour-intensive, resulting in the development of computational methods for “label-free” quantitation of proteins from MS data without prior labelling.

Database search

The measured spectra are subsequently used to identify sample proteins. Tandem MS data allows to infer the amino acid sequence of a peptide from the peptide fragment spectra. As no additional information is required, this approach is known as “de novo sequencing”. As the mass of leucine and isoleucine is identical, it is not possible to resolve the difference between these amino acids by de novo methods. Also a huge number of amino acid modifications and the combinatorial growth of the search space makes de novo methods cumbersome.

Alternatively it is possible to compare measured spectra with hypothetical spectra from a protein database. In the age of many completed genomes and gene loci prediction, complete protein databases get available and database searching is mainly used for protein identification. Therefore peptide sequences of all sequences in a protein database are calculated. This “in silico digestion” is possible due to the predictable cleavage properties of digestion enzymes used in the sample preparation. For the generated peptides, fragment masses are calculated and compared to measured MS2 spectra. The match between the experimental and the theoretical spectrum is judged by a scoring function. Of course the set of possible modifications increases the size of the search space, but the knowledge of known peptide sequences derived from the protein database leads to substantial reduction of search space, with obvious advantages and disadvantages - as for example the inability to identify peptides with point mutations, as these are not represented in the database. Multiple database search engines like SEQUEST [43], Mascot [92] or Phenyx [24] that work according to the above described principle have been developed.

To determine score thresholds for reliable peptide identification, the experimental spectra can be queried against a database with reverse protein sequences. Inverse peptide sequences should not be found in the sample and are therefore used as a negative control set for identification. This allows to adjust the threshold in order to achieve a desired False Discovery Rate (FDR). Also approaches that generate a negative control protein database by shuffling amino acids and therefore also leaving amino acid abundance ratios constant have been proposed.

After the reliable identification of peptides, these are mapped back to underlying protein sequences in the database to determine the proteins present in the sample. The number of spectra, that can be matched onto a certain protein, is referred to as "spectral count". This measurement can be used to quantify protein abundance [88], but due to reduced accuracy in comparison methods that quantify protein abundance by chemical labelling these label-free measures are referred to as "semi-quantitative" measures of protein abundance.

Peptide sequences exist which are present in multiple proteins, mostly due to homology. These "shared peptides" cannot uniquely be mapped on a single protein, but are present in multiple proteins in the database. This is of course a problem for protein identification. In such a case, grouping of potentially identified proteins - which share the non-unique peptide sequence - based on detected peptides is necessary. Often an additional peptide, which is specific for a protein within the group is also identified in the MS data and can be used to partially resolve this problem and allow protein inference. The presence of such a "specific" peptide allows to infer that the protein containing this peptide is indeed present, but does not allow to conclude that the other proteins which contain the shared peptide were not present in the sample. The problem of shared peptides is also relevant in protein quantitation by chemical labelling, as it can lead to different abundance ratios between shared and unique peptides if multiple proteins that contain the shared peptide were present in the sample.

3.2.4 Normalisation of MS data

Data normalisation is an important aspect in the analysis of quantitative (labelled), but also for semi-quantitative (label-free) MS data sets. As the features of quantitative and semi-quantitative (spectral-count) data sets are very distinct, different methods have been developed for these approaches. As the experimentally obtained data in this work is of semi-quantitative nature, this section will focus on the normalisation of semi-quantitative data sets.

The number of spectra (spectral count), which led to the identification of a protein, correlates with the proteins abundance in the sample. Therefore normalisation methods for semi-quantitative are usually based on the spectral count. Rsc [88] and NSAF [117] are two commonly used methods for normalising protein quantitation of label-free shotgun data. Also methods that take into account ion intensities in the MS1 spectrum either by summing up intensities or by computing the area under the elution curve in the ion chromatogram of the individual peptides have been developed (e.g. [49]). These methods provide promising results, but have been neglected in this work as ion intensities were not accessible in the analysis pipeline that was used here. Instead Rsc normalisation was applied, which allows normalisation of semi-quantitative MS data based on spectral counts.

All these semi-quantitative methods do not allow absolute quantitation of protein amount, but allow to capture relative changes in protein abundance between multiple experiments and therefore allow to normalise semi-quantitative MS data.

Rsc normalisation

Rsc normalisation was originally developed for serial analysis of gene expression (SAGE) data by Beissbarth et al [7]. This approach was adapted to normalise mass spectrometric spectral count data by Old et al [88]. Both techniques have in common that they produce discrete values as measurements and this was also acknowledged by Beissbarth by implementing a correction factor in the formula (see Equation 3.1).

$$R_{sc} = log_2 \frac{n_2 + f}{n_1 + f} + log_2 \frac{t_1 - n_1 + f}{t_2 - n_2 + f} \tag{3.1}$$

Here, R_{sc} is the logarithmic abundance ratio for a given protein between two samples, whereas n_1 and n_2 represent the spectral count of that protein in the two samples and t_1 and t_2 are the total spectral counts in the two experiments. f is a correction factor, here typically $f = 0.5$ is used.

3.3 Statistical tests to identify protein specificities

In order to identify specificities of proteins to different classes of baits, statistical methods can be applied to identify significantly increased abundance in a group of experiments which is a hallmark of protein specificity. Therefore the experimental measurements of one protein are divided into two groups depending on the property, for which specificity should be tested (see Table 5.3 for a list of bait properties). As the experiments were performed in multiple cell types, the statistical methods should also allow to incorporate this additional information. The statistical methods applied in the Results chapter (page 34) will now be introduced here.

3.3.1 Resampling methods

Besides the t- and U test, which are used as a benchmark, also resampling methods are applied on the experimental data set. Here bootstrap, which samples the experimental data with replacement to compute the distribution of the test statistic under the null hypothesis, and permutation tests, which perform in principle the same sampling but without replacement are used. The applied test statistic critically influences the power of the analysis, and therefore a variety of test statistics will be analysed.

Test statistics

The test statistics need to allow the resampling test to identify if there was a significant difference in the abundance of the two classes of experiments. Here basically two test statistics will be used:

- the difference in mean abundance (see Equation 3.2)

  $$\frac{\sum_e m_{pe} I_e}{N_1} - \frac{\sum_e m_{pe}(1 - I_e)}{N_2} \tag{3.2}$$

  Here m_{pe} is the measurement of protein p in experiment e and I_e is an indicator variable which indicates if experiment e was performed with a bait for which we want to test specificity. N_1 and N_2 are the group sizes of the two groups of experiments and are connected to I_e by $N_1 = \sum_e I_e$ and $N_2 = \sum_e (1 - I_e)$. Methods using this test statistics are recognisable by the suffix `meanDiff` in the method name (see power- and ROC curves in Figures 5.13-5.24 in the Results chapter).

- and the test statistic of the U test (see Equations 2.2 and 2.3 for computation of the test statistic of the U test). The test statistic of the U test uses the sum of the ranks of one group after the measurements were transformed onto a rank scale and also allows to identify significant differences between samples.

Plugging both test statistics into a permutation test or the bootstrap allows to identify significant differences between two groups of experiments. All four versions of the test (permutation test/bootstrap & mean difference/U test statistic) will be evaluated in Section 5.2.

As the measurements in the experimental data set were obtained in multiple cell types, it is interesting to evaluate if more sophisticated test statistics, which take into account the structure of the experimental data benefit from this information. Therefore test statistics which use the cell type information are designed.

Test statistics for multiple cell types

One possibility to incorporate cell type information is to compute the test statistic for the measurements of each cell type separately. This produces one value of the test statistic for each cell type. As protein specificity is assumed to be identical between cell types these three values of the test statistic need to be combined into one combined test statistic again to allow the computation of one single P value for protein specificity. A straightforward way to combine the three values of the test statistic is to compute their sum. Test statistics, which take into account cell lines are indicated with `CL` in the method name (see Figures 5.21-5.24 in the Results chapter). Another alternative is to compute the weighted average of the test statistics, to increase the weight of cell types where the protein was detected with higher abundance, the weights can be chosen according to the sum of the proteins measurements in that respective cell type to increase the influence of cell types with higher protein abundance. Test statistics, which use a weight average to combine the test statistics of the individual cell lines are indicated with the suffix `weighted` (see Figures 5.21-5.24 in the Results chapter). This has the advantage that the influence of cell types for which much data is present is increased, whereas the contribution of cell types with little or no measurements is decreased. Both options will be used in the specificity analysis to detect protein specificities from multiple cell type data.

3.4 Gene Ontology (GO)

Gene Ontology (GO) [4, 47] provides a dynamic, controlled vocabulary to annotate knowledge about biological entities like genes and proteins. GO is divided into three independent sub ontologies handling different classes of biological information. These three sub categories allow to classify

- biological processes in which the protein participates
- molecular functions which the protein fulfils
- cellular components where the protein is located

These three entities (biological process, molecular function and cellular component) are also the names of the root nodes of the three respective sub ontologies. Each of these ontologies contains nodes that specify roles within the given focus of that ontology. These nodes are connected by directed edges, which enables to distinguish between the more generic (closer to the root) and the more specific node (more distant from the root). Nodes in the ontology are also referred to as GO terms. The hierarchical organisation of nodes allows to annotate proteins at different levels of knowledge, from rather general roles close to the root of an ontology to very specific roles further down in the directed ontology graph. This is an important property to support the annotation process in ongoing research and allows to organise information for well studied proteins with very specific functions to less well studied proteins with more generic GO terms. The definition of GO prohibits cycles occurring in the ontology graph and therefore GO is also referred to as a "directed acyclic graph" which in turn simplifies data analysis.

One reason for the initiationof GO was, because functions between homologous proteins are often conserved between species and a common nomenclature was required. The generated ontology should unify the annotation of biological findings and should also be amenable to computational analysis. Meanwhile almost 10.000 citations underline the presence of this need. The concept of transferring knowledge between homologous proteins was very useful in the early days of large scale biological research. Meanwhile a high majority of annotations is inferred from electronic annotation [32]. Now also disadvantages of this practice become apparent by a subset of propagated annotations which are not evolutionary conserved and therefore lead to incorrect annotations that in turn might be misleading for research. This insight has led to the fact that biologists treat annotations with care and rather use them as an indicator instead of strictly relying on them.

One branch of Gene Ontology, which is very important for this work is located in the molecular function sub ontology. "Nucleic acid binding" is

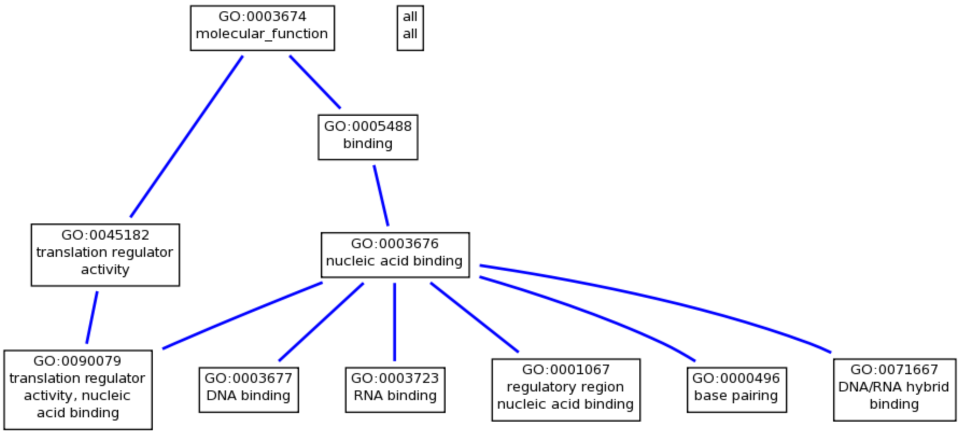


Figure 3.2 – Fragment of the “molecular function“ ontology in Gene Ontology showing the position of ”nucleic acid binding“ and its first level child nodes. Image created by Gene Ontology browser AMIGO [16].

represented here as a node with id GO:0003676 in the second level of the molecular function ontology. Figure 3.2 shows the location of “nucleic acid binding” and its direct child nodes within a fragment of the top layers of the molecular function ontology. In total “nucleic acid binding” is connected to 299 child nodes which give more specific definitions to different aspects of “nucleic acid binding”. 156 of these child nodes are children of “RNA binding” and 136 are children of “DNA binding”, whereas the other child nodes of “nucleic acid binding” contribute less sub annotation nodes. GO annotation distributes upward in the directed ontology graph, this means that genes or proteins which are annotated as “DNA binding” are hereby also defined to be “nucleic acid binding” which is the ancestor node of “DNA binding”. This is also valid across multiple layers of annotation, so proteins annotated as “DNA binding” are implicitly also defined as “binding”, which lies two layers upstream (see Figure 3.2). This leads to a number of currently 48222 gene products which are annotated as “nucleic acid binding” in multiple species.

3.4.1 Enrichment of Gene Ontology Terms

Since the propagation of high-throughput methods in biological research it has become more and more required to support the interpretation of biological data by computational methods. The development of Gene Ontology supports the interpretation of biological measurements by computational methods. Biological experiments often deliver a (sorted) list of proteins or genes. A straightforward approach to interpret such lists is to annotate the entities in the list using Gene Ontology and then compute enrichment probabilities for individual GO terms. Here mostly a hypergeometric test is used to compute the statistical significance of the enrichment of a certain GO term (see Equation 3.3).

$$Prob(X \leq b) = \sum_{i=b}^{min(n,B)} \frac{\binom{n}{i}\binom{N-n}{B-i}}{\binom{N}{B}} \tag{3.3}$$

Here N is the total number of proteins, B of which are annotated with the GO term of interest. The given list of proteins has a size of n proteins b of which are annotated with this same GO term. This computation is performed either for all GO terms or for a limited set. Usually a one-sided test for enrichment is performed, as depletions are likely to be less informative. A number of tools for different applications have been developed that follow this approach, for example [114, 37, 18, 97]. Also methods that allow to analyse enrichments in the context of biological networks/graphs are under development [79].

Chapter 4

Experimental methods

4.1 Nucleic acid affinity purification

Oligonucleotides were synthesized by Microsynth. The sense strand was biotinylated at the 5' end, whereas the antisense strand was not modified. Double-stranded baits were annealed by heating to 80 °C for 10 minutes, followed by slow cooling to 25 °C. For generating the affinity resin, Ultralink immobilized Streptavidin Plus Gel (provided by Pierce) was washed three times with PBS. 4 nmol of nucleic acid (single-stranded or double-stranded) were then added to the streptavidin resin equilibrated in PBS, followed by incubation at 4 °C for one hour on a rotary wheel to allow binding of the biotinylated oligonucleotides. Next, the resin was washed twice with PBS and twice with TAP lysis buffer (50 mM Tris, pH 7.5, 100 mM NaCl, 5% (vol/vol) glycerol, 0.2% (vol/vol) Nonidet-P40, 1.5 mM $MgCl_2$, 25 mM NaF, 1 mM Na_3VO_4 and protease inhibitor 'cocktail' (Complete; provided by Roche)) for the removal of unbound oligos. Cells were lysed in TAP lysis buffer. For every 4 nmol immobilized nucleic acid, 6 mg cell extract was used for nucleic acid affinity purification. Additionally 10 μg/mL poly(I:C) (for DNA baits) or 10 μg/mL calf-thymus DNA (for RNA baits) were added as soluble competitor. Cell extracts were combined with the immobilized nucleic acids, followed by incubation for two hours at 4 °C on a rotary wheel. Unbound proteins were removed by three consecutive washes in TAP lysis buffer. Bound proteins were eluted with 300 μL 1M NaCl.

For the validation of XRCC6, HNRNPR and NCL were detected by immunoblotting using available antibodies (AB1358, 05-620, 05-565; provided by Millipore). Myc-tagged C20orf72, AIM2, UHRF1 and YB-1 were overexpressed in HEK293 cells and visualized by immunoblotting using anti-Myc-IRDye800 (provided by Rockland). Bound proteins were eluted in SDS sample buffer for validation experiments.

4.2 Protein identification

4.2.1 Mass spectrometry

Samples were analyzed on a hybrid Linear Trap-Quadrupole (LTQ) Orbitrap XL mass spectrometer (ThermoFisher Scientific) coupled to a 1200 series high-performance liquid chromatography (HPLC) system (Agilent Technologies) with an analytical column packed with C18 material.

4.2.2 Peptide identification and protein grouping

Database search and integration of protein identifications were performed as already described earlier in [14]. Therefore data generated by tandem MS were searched against the human UniProtKB/Swiss-Prot database version 57.12 [113] using the Mascot [92] and Phenyx [24] search algorithms. Both search engines were operated using 4 ppm as parent and 0.3 Da as fragment mass tolerance for tryptic peptides with maximum one missed cleavage. All cysteins in the peptide sequence were modified to carbamidomethyl cystein and the oxidation was allowed as variable modification. Theoretical spectra of peptides shorter than six amino acids were excluded from the in silico digested database. Results of both search engines were parsed separately and a minimum of two distinct peptides above a score threshold was required. Also proteins identified by a single peptide were accepted if the identification score was above a more stringent threshold and the peptide accounts for at least 2.5% of the amino acids of the protein (sequence coverage). The score threshold was chosen by searching a protein database with reverse sequences which guarantees identical peptide size distribution. In this reverse database no proteins should be identified and so the score which leads to a FDR of 1% was determined. The resulting score thresholds can be seen in Table 4.1.

Table 4.1 – Score thresholds to achieve 1% FDR at a peptide level obtained by searching a reverse database, for single and multiple peptide hits separately

hit type (peptides/protein)	Mascot ion score	Phenyx z-score
multiple peptides	18	4.5
single peptide	50	6

The union of the peptide identifications of both search engines was used to combine their results. Subsequently proteins were grouped based on shared peptides.

Chapter 5

Results

5.1 Experimental approach

The aim of this work is to study the interaction between proteins and nucleic acids in an unbiased fashion. Furthermore it is desirable to achieve a maximum coverage of the studied proteome - the set of all proteins which are expressed in an organism. To identify these interactions a affinity purification approach was employed. In combination with state of the art mass spectrometry this should enable to detect a significant fraction of nucleic acid binding proteins (NABPs).

The one step affinity purification protocol to enrich for NABPs was already established in the lab [13]. Here, synthetic oligonucleotides containing a biotin moiety at the 5'end of the nucleic acid chain are coupled to streptavidin beads, which exhibit high affinity to biotin. This matrix is used to purify NABPs from a complex mixture of many different protein species. Subsequently interacting proteins are eluted and subjected to mass spectrometric analysis to identify NABPs.

A substantial fraction of NABPs do not bind all different nucleic acid molecules equally well but instead exhibit varying binding affinity for different nucleic acid molecules. Therefore probing with diverse nucleic acid baits leads to the identification of additional NABPs and for this reason an array of multiple nucleic acid baits has to be probed against a biological sample to achieve satisfactory coverage of the proteome. To maintain the association between baits and preys, biological samples cannot be pooled and hence each additional nucleic acid bait requires an additional mass spec analysis. Consequently the set of nucleic acid baits is a compromise between achieving optimal coverage also for NABPs with very specific binding preferences and analytical workload and also cost.

Using synthetic oligonucleotides allows complete control of bait properties. Hence it is desirable to design nucleic acid baits that cover nucleic acid sequence space with maximum diversity. This aim was achieved by designing nucleic acid baits that contain only two different nucleotides and synthesizing all combinatorial dinucleotide combinations according to this bait design. This leads to six different nucleic acid baits per single stranded nucleic acid species $\binom{4}{2} = 6$ and four species for double stranded species, as due to strand complementarity there are two hypothetical baits that differ by being biotinylated on the opposite strand but share nucleotide composition of individual strands otherwise (i.e.. AG:TC is equivalent to TC:AG in terms of nucleotide composition).

In Transfac [112] - a database of sequence specific protein-DNA interactions - NABPs cover between six and eight nucleotides of their nucleic acid interaction partner. Therefore, baits were designed with a length of 30 nucleotides to allow comfortable binding of NABPs also including some spacer from the streptavidin beads. As already described in Section 3.1, single stranded nucleic acid molecules form secondary structures. This structure also has critical influence on protein-nucleic acid interactions. The space of different nucleic acid structures - which is determined by the nucleotide sequence - is huge. Already a large number of different nucleic acid structures have been described [50]. The experimental approach, where each additional bait requires independent mass spectrometric analysis, does not allow to study both parameters - nucleotide composition and for each composition independent structures - at the same time. Therefore, baits were designed to form only a minimum of secondary structure. According to these requirements, a bait sequence was designed that should fulfil the following requirements:

- the sequence of one strand should be composed of two different nucleotides only (dinucleotide baits)
- the length of the bait sequences is 30 nucleotides
- both nucleotides should appear in the bait in a one-to-one ratio (15 occurrences of each nucleotide)
- the bait sequence should form as little secondary structure as possible to avoid structure specific interactions

The number of possible permutations to construct a sequence of 30 characters based on two characters, where each character occurs 15 times, is extremely high $\binom{30}{15} \approx 2 * 10^{20}$ and is not amenable to enumerate each individual sequence and predict its secondary structure. Hence a set of one million randomly chosen bait sequences was designed that fulfils the above mentioned requirements. For each dinucleotide combination of these one million candidate patterns the secondary structure and minimum free energy was predicted using ViennaRNA [57]. The formation of secondary structures is also reflected in the free energy (G) of the minimum free energy structure of the nucleic acid sequence. Here a high free energy in the minimum free energy structure reflects a structure which does not form a strong secondary structure. So to achieve an unstructured bait it is desirable to have a high free energy. To accomplish this for all dinucleotides, the free energy of the respective structures was summed up (see Equation 5.1). Hence a scoring function was designed that sums up the free energy (G) of all possible dinucleotide combinations for a particular candidate bait sequence and the sequence that achieved the highest total G ($\sum G$) was chosen for synthesis.

$$\sum G = G_{AU} + G_{AC} + G_{AG} + G_{UC} + G_{UG} + G_{CG} \tag{5.1}$$

The sequence with the highest free energy within the pool of one million random sequences was `XXXIIXIIXIIXXXXIIXIIXIIXIXIIXX`, where `X` and `I` can be replaced with any dinucleotide combination and were selected to improve legibility. To demonstrate the effect of this optimisation, for example the free energy of an UA-RNA sequence following this pattern is -0.38 kcal/mol. Here `X` was replaced by uracil and `I` by adenosine resulting in the RNA sequence `UUUAAUAAUAAUUUUAAUAAUAAUAUAAUU`. In contrast, the free energy of a sequence with alternating adenine and uracil nucleotides (`AUAUAUAUAUAUAUAUAUAUAUAUAUAUAU`) is -9.42 kcal/mol. The high symmetry in the second sequence promotes the formation of secondary structure and consequently leads to the formation of a hairpin-like structure (the predicted minimum free energy structures of both sequences is shown in Figure 5.1).

This determined pattern was synthesized for all six dinucleotide combinations in its single-stranded DNA (ssDNA) and single-stranded RNA (ssRNA) form. The four distinct double-stranded DNA forms were synthesized as well. To allow the detection of proteins reading epigenetic marks, also cytosine-methylated CG-DNA oligos were included in the list of baits. Furthermore also mononucleotide DNA baits and polyA-RNA - due to the presence in polyA tails of mRNA probably the most abundant mononucleotide form of RNA - were included in the list of baits. These mononucleotide baits are constructed of 30 nucleotide long repeats of the same nucleotide. Not all forms of DNA mononucleotides were amenable to synthesis by the manufacturer as for example poly G DNA cannot be synthesized and also for example CG rich baits are difficult to synthesize, probably due to the formation of secondary structures in the synthesis process. This was also observed in reduced yield

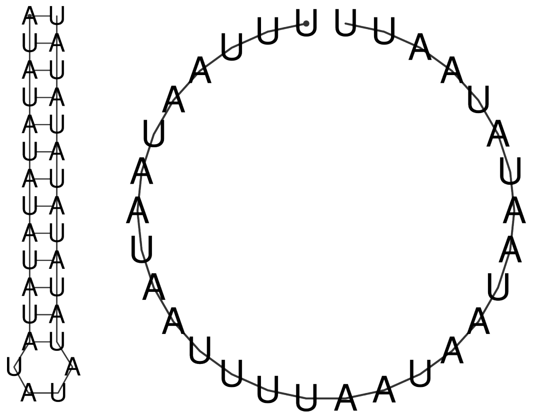
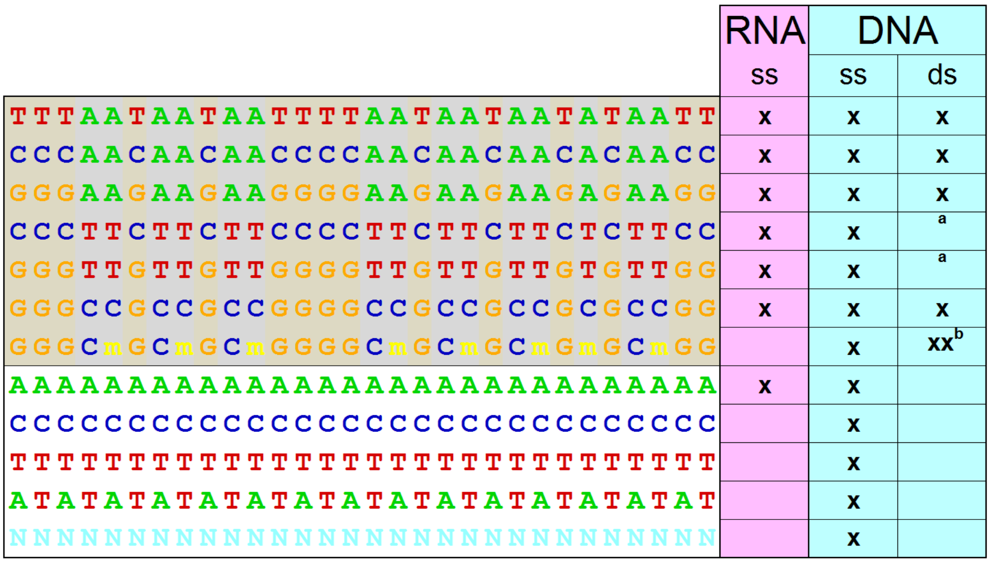


Figure 5.1 – Minimum free energy structures of an alternating AU-RNA and the designed bait sequence. (Figure produced using the ViennaRNA Websuite [51])

for baits that contain nucleotides with high affinity (CG). The full list of all 25 distinct nucleic acid baits can be seen in Figure 5.2. This consequent design for maximum sequence diversity within a limited number of nucleic acid baits should allow to capture a maximum fraction of NABPs in the sample. Furthermore this systematic experimental design also guarantees differential binding of the interacting proteins solely due to nucleotide composition differences, as the baits are identical in length and unstructuredness otherwise.

Another important aspect to obtain satisfactory coverage of NABPs is to ensure the availability of a huge variety of different protein species in the biological sample which is probed against the nucleic acid column. Individual cell types only express a subset of the complete proteome of an organism. As a consequence diverse human cell types are required to obtain a representative fraction of the human proteome. During the development of an organism an early embryo separates in germ layers, gene and protein expression diverges in the subsequent development of an organism. Consequently choosing cell types of different germ layers should guarantee a huge variety of different protein species in the samples. So to increase the coverage of the human proteome we performed our affinity purification experiments against whole cell lysates of three cell lines picked from the three different germ layers. We were using U937 (a human lymphoma cell line), HepG2 (a human liver carcinoma cell line) and HaCat (a human keratinocyte cell line), three well established and widely used cell lines. Whole cell lysates of these cell lines were generated and affinity purifications performed against the de-



	RNA	DNA	
	ss	ss	ds
TTTAATAATAATTTTAATAATAATATAATT	x	x	x
CCCAACAACAACCCCAACAACAACACAACC	x	x	x
GGGAAGAAGAAGGGGAAGAAGAAGAGAAGG	x	x	x
CCCTTCTTCTTCCCCTTCTTCTTCTCTTCC	x	x	[a]
GGGTTGTTGTTGGGGTTGTTGTTGTGTTGG	x	x	[a]
GGGCCGCCGCCGGGGCCGCCGCCGCGCCGG	x	x	x
GGGCmGCmGCmGGGGCmGCmGCmGmGCmGG		x	xx[b]
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	x	x	
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCC		x	
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTT		x	
ATATATATATATATATATATATATATATAT		x	
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNN		x	

Figure 5.2 – The complete list of the 25 nucleic acid baits used in the study. Baits in the top seven rows follow the same dinucleotide pattern which was designed for minimal secondary structure by maximising the free energy of the minimum free energy structure.

[a]) The top six rows represent all possible dinucleotide combinations for single stranded baits, whereas in dsDNA CT:GA and GA:CT as well as GT:CA and CA:GT are equivalent.

[b]) The bait in row seven includes eight cytosine methylated sites and is otherwise equivalent to the pattern in row six. This bait was synthesised as a ssDNA as well as hemi- and dimethylated dsDNA.

scribed list of 25 oligonucleotide baits. To distinguish proteins binding to our streptavidin matrix we also performed affinity purifications using the matrix without coupled nucleic acid bait in each of the cell lines. These 78 biological samples were then analysed by gel-free shotgun Mass Spectrometry (MS).

5.1.1 Initial characterisation of the experimental data set

The raw data obtained in these 78 MS analysis was searched against the human version of SwissProt. This database contains all 20422 currently known human proteins as well as their isoforms. Mascot [92] and Phenyx [24] were used as search engines. Database search led to the identification of 10810 proteins, which means that on average ≈ 140 proteins could be identified per sample. Combining these results we could detect 952 unique proteins, which in contrast means that each protein was on average detected in eleven different experiments. This already indicates that it is inevitable to use a diverse set of nucleic acid baits to probe for NABPs.

Figure 5.3 supports this in a histogram which shows that most proteins only interact with a low number of baits. Furthermore also using multiple cell lines is supported by this Figure, as many proteins are only detected in a single cell line.

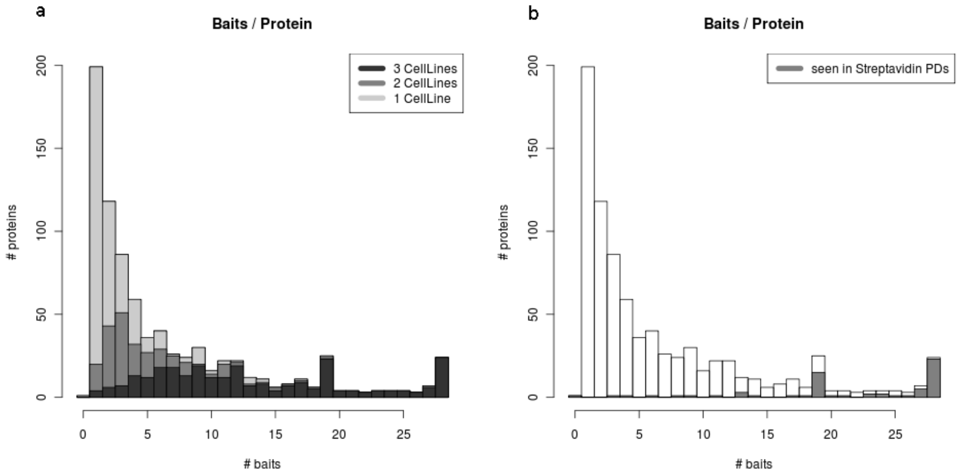


Figure 5.3 – Frequency of protein detection. (**a**) Proteins which interact with a low number of different baits are often only detected in a single cell line. (**b**) Proteins which were also detected in Streptavidin pulldowns, are detected in a high number of experiments.

To estimate the enrichment in NABPs achieved by the experimental approach the identified proteins were annotated using Gene Ontology and proteins classified as nucleic acid binding were marked. Figure 5.4 and Figure 5.5 show that the majority of proteins identified in the individual experiments were actually known NABPs. Figure 5.4 shows the absolute numbers of proteins identified in the individual experiments, whereas Figure 5.5 reports the relative fraction of NABPs. Comparing this fraction to MS results of complete cell lysates of these cell lines shows the enrichment of NABPs in the experimental data set. These complete cell lysates were analysed without subsequent affinity purification - the input material of the affinity purification experiments. This data set of whole cell lysates was available from another study we performed earlier [14]. Also the overall fraction of NABPs in the human proteome is indicated (dashed line). While in the unpurified whole cell lysate experiments around 20% of the detected proteins are known to be nucleic acid binding, which is comparable to the fraction of known NABPs in the human proteome, affinity purification increases this fraction to ≈ 75% (see Figure 5.5). Altogether these figures confirm, that the affinity purification experiments were successful in enriching for NABPs.

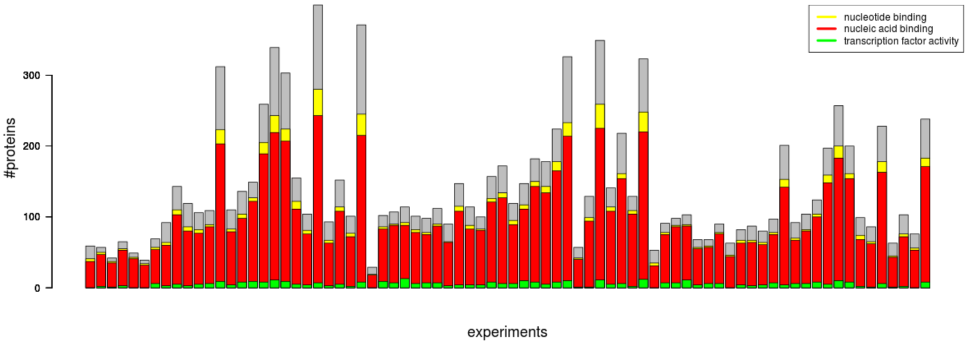


Figure 5.4 – Bar Graphs illustrating the fraction of known NABPs (according to GO) per experiment. Proteins annotated as transcription factors - sequence specific DNA binding - are shown in green and nucleotide binding proteins in yellow.

As expected, transcription factors which mostly bind to specific nucleic acid sequence motifs were not enriched in our experiments (see Figure 5.5, right panel) as these motifs were not present in the limited set of the nucleic acid baits.

After showing that the experimental procedure is capable to enrich for NABPs the next question is to find out why proteins that are not known

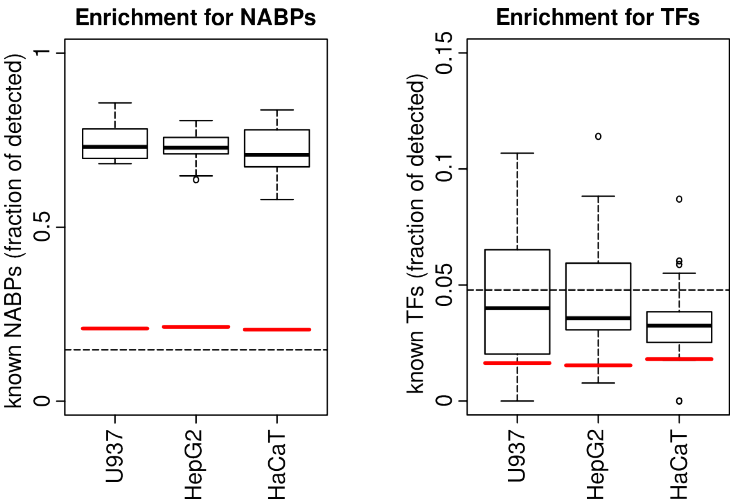


Figure 5.5 – Boxplots illustrating the relative fraction of NABPs in the individual experiments, compared to experiments of unpurified samples of the same cell lines (red) and the human proteome (SwissProt database, shown as black dashed line). The right panel shows the same information for transcription factors.

to be nucleic acid binding do occur in the data set. These proteins might originate from multiple sources:

- Proteins might be nucleic acid binding but this function is still unknown or not yet represented in the ontology used to infer protein functions. Also misannotation is a potential source for proteins falling in this category.

- Proteins might interact with a protein that is binding to the nucleic acid bait. These “secondary interactors” which do not bind to the nucleic acid bait directly are also enriched in samples generated by affinity purification.

- Identification of proteins by MS requires the protein to be present in the sample at a certain - minimal - level of abundance to exceed the detection limit. As proteins in living cells are not equally abundant but their abundance spans several orders of magnitude, and the purification procedure does not perfectly deplete very abundant proteins, residual amounts these proteins are frequently detected in qualitative and semi-quantitative MS analysis. These proteins are also referred to as “frequent hitters”.

Filtering frequent hitters

Proteomics experiments analysed by Mass Spectrometry are known to also identify proteins which are present at very high abundance in cells. The relative abundance of these proteins is dramatically reduced by the affinity purification procedure, but their sheer abundance in the starting material - often several orders of magnitude higher than other proteins - leads to a fraction which remains in affinity purification samples and these proteins are therefore also detected in the MS analysis. These proteins are therefore called frequent hitters, as they are observed in many MS data sets, independent of the used matrix. A common approach to get rid of these proteins in a data set is to perform an affinity purification experiment, which uses the column without attached bait. This also allows to get rid of proteins, which bind to the streptavidin matrix itself and not to the coupled bait.

The three streptavidin pulldowns, without coupled nucleic acid bait led to the identification of 72 proteins. Some of the abundant proteins identified in the streptavidin pulldowns are well known nucleic acid binding proteins. Annotating these experiments with GO classifies 41 of the 72 (57%) identified proteins as nucleic acid binding and therefore the proteins detected in the streptavidin experiments should not be removed entirely from the experimental data set. Hence, the spectral count in the streptavidin experiments was compared with the spectral count in the nucleic acid experiments. For both groups of experiments the highest spectral count was used to measure maximum abundance of the protein per group (see Equation 5.2).

$$r_{Strep} = \frac{max(sc_{NA})}{max(sc_{Strep})} \tag{5.2}$$

Here sc_{NA} represents the spectral counts of a protein in the nucleic acid experiments whereas sc_{Strep} represents the number of spectra measured in the negative control streptavidin experiments. The fraction of these two values r_{strep} gives an indication for the enrichment of the streptavidin binding protein in the nucleic acid experiments. Table 5.1 shows this ratio for the 72 proteins, which were detected in the streptavidin experiments. As expected proteins which are known to be nucleic acid binding are more abundant in the nucleic acid purifications than in the streptavidin experiments. To remove frequent hitters, proteins with a spectral count that is less than five fold higher in a nucleic acid experiment than in the streptavidin experiment were removed from the experimental data set. This filtering step allows to remove 31 proteins from the experimental data set and reduces its overall size to 921 proteins.

Table 5.1 – Ratio between the maximum spectral count in the nucleic acid experiments compared to Streptavidin only pulldowns for protein detected in Streptavidin experiments.

gene name	r_{Strep}	NA binding		gene name	r_{Strep}	NA binding
ACTN4	0.0	FALSE		RPS25	5.4	TRUE
ACTBL2	0.7	FALSE		RPS17	5.5	FALSE
MYO1G	0.8	FALSE		RPL23	5.7	FALSE
CORO1C	0.8	FALSE		RPS10	5.7	FALSE
NSUN5	0.8	FALSE		RPL36A	6.0	FALSE
EIF6	1.3	TRUE		RPL27A	6.8	TRUE
SELH	1.5	FALSE		RPS7	6.8	TRUE
TUBA1B	1.6	FALSE		RPS3	7.5	TRUE
RPL13	2.0	TRUE		RPS4X	7.6	TRUE
NACA	2.0	TRUE		MPO	7.7	FALSE
RPL22L1	2.0	FALSE		RPL35	8.0	TRUE
RPL30	2.3	TRUE		RPS23	8.0	FALSE
RPL37A	2.8	FALSE		RPS18	8.0	TRUE
RPL9	2.8	TRUE		RPS29	8.0	FALSE
ZNF593	2.8	TRUE		RPL38	8.0	TRUE
CANX	3.0	FALSE		RPL24	8.0	TRUE
RPS5	3.3	TRUE		RPS13	8.4	TRUE
RPL11	3.4	TRUE		RPS8	8.5	FALSE
RPL29	3.5	TRUE		RPL31	9.0	TRUE
RPL14	4.0	TRUE		RPL17	12.0	FALSE
TFB1M	4.0	TRUE		HSPA9	12.0	FALSE
NSUN5	4.0	FALSE		LYZ	12.0	FALSE
DCAF13	4.0	FALSE		SRP14	14.5	TRUE
DIMT1L	4.0	TRUE		RPS6	15.0	FALSE
FAU	4.1	TRUE		HIST1H1B	15.3	TRUE
UBA52	4.5	FALSE		HSPA8	18.0	FALSE
RPS19	4.7	TRUE		THOC4	18.0	TRUE
RPL36AL	4.7	FALSE		H1FX	18.0	TRUE
RPS20	4.8	TRUE		RPS14	20.7	TRUE
RPS16	4.8	TRUE		RPS3A	22.7	TRUE
HSPA5	4.8	FALSE		SRP9	23.0	TRUE
RPS26	5.0	TRUE		RPL22	25.0	TRUE
RPS28	5.0	FALSE		RPL23A	26.5	TRUE
RPS24	5.2	FALSE		RPL8	28.0	TRUE
RPS11	5.3	TRUE		RPS15A	36.0	TRUE
FGFBP1	5.3	FALSE		DDX5	44.5	TRUE

Identification of secondary interactors

After removing frequent hitters from the data set, the remaining proteins can be classified in proteins which are known to interact with nucleic acid, their interactors (secondary interactors of nucleic acid) and additional proteins for which the mode of interaction with nucleic acid is unknown.

For this purpose all 921 proteins of the experimental data set were queried against Gene Ontology and it is checked if they are annotated as nucleic acid binding (GO:0003676). This procedure identified 495 proteins of the experimental data set as known NABPs. This annotation now allows to identify secondary interactors by querying public interaction databases for interactions between the direct known NABPs in the data set and the remaining proteins. Therefore a compendium of multiple public interaction databases including BioGRID [11], MINT [19], InnateDB [78], HPRD [66] and IntAct [65] was compiled. These databases contain known protein-protein interactions (ppis) that originate from different experimental sources. The majority of interactions here is contributed by large scale studies that facilitate technologies like affinity purification mass spectrometry (AP-MS) or yeast two hybrid (Y2H). To increase knowledge for a specific organism, interactions are also transferred from model organisms by sequence homology. In total interactions for 175 of the remaining 426 proteins with NABPs could be found.

Overall, this annotation process allows to classify the 921 experimentally detected proteins into

- 495 known NABPs,
- 175 secondary interactors and
- 251 proteins not known to interact with nucleic acid before.

To validate this classification, isoelectric points (pI) for all proteins were computed. As expected proteins known to be nucleic acid binding tend to have a higher pI. Figure 5.6 shows that the pI of NABPs in the human proteome (green dashed line) is higher than for all proteins (black dashed line). The same shift is observed for NABPs in the experimental data set (green), whereas proteins classified as secondary have a lower pI (blue). Proteins classified as novel NABPs (red) experience an even more pronounced shift as known NABPs which provides additional evidence that these indeed interact with nucleic acid.

Figure 5.7 illustrates protein detections and their classification as known, secondary or novel interactors in an interaction graph.

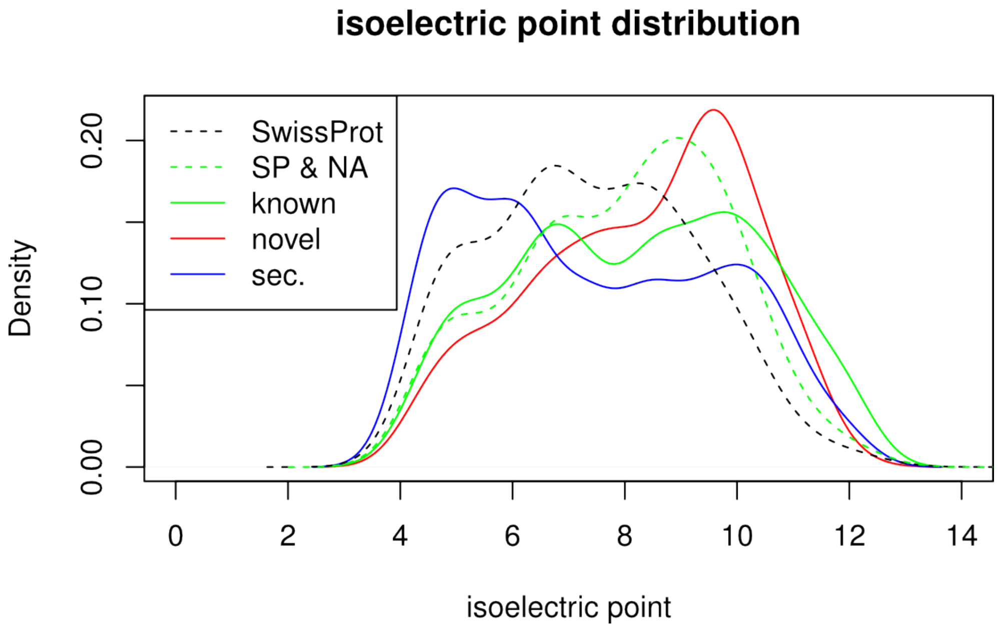


Figure 5.6 – Distribution of isoelectric points for the different classes of proteins. NABPs within the human proteome (dashed green line) have a higher pI than the set of all human proteins (dashed black line). Experimentally detected proteins known to be nucleic acid binding (green) are also shifted, whereas secondary interactors (blue) have a smaller pI. Also proteins classified as novel (red) have an increased pI.

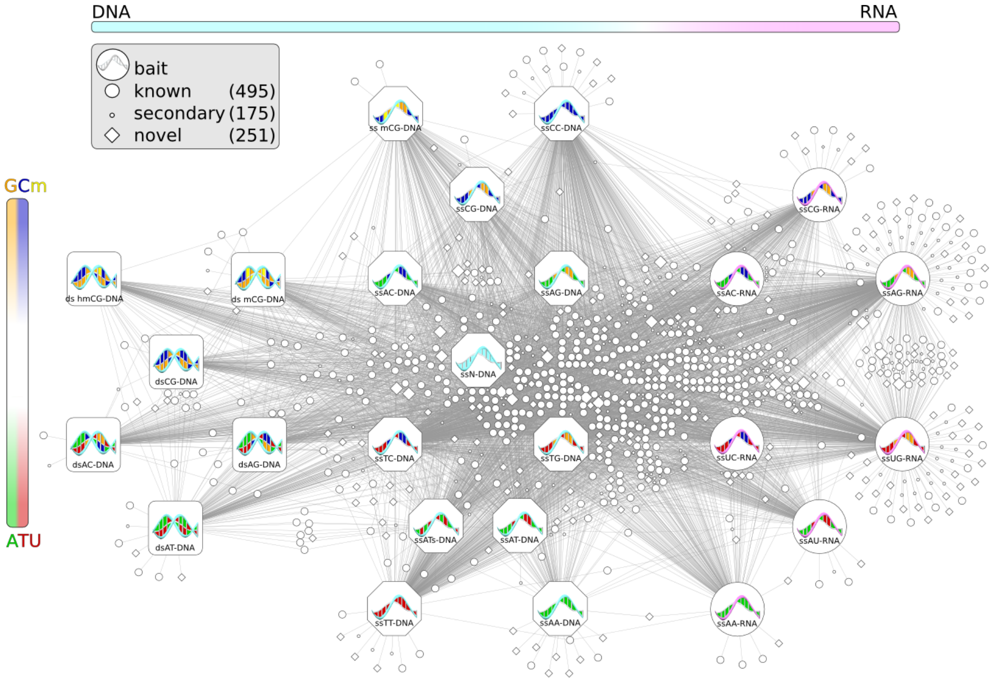


Figure 5.7 – Graph representation of protein identifications linked to the experimentally used baits. Here protein node shape indicates known, secondary and novel NABPs.

Searching for nucleic acid binding protein domains

Proteins often contain sequence fragments which occur in multiple proteins in a very similar form - in terms of amino acid sequence. These fragments are called protein domains. Protein domains are associated with performing distinct functions. These functional building blocks of proteins allow a protein to fulfil its function. For example a protein could be composed of two domains, one that binds DNA and a second one with helicase activity, which allows the protein to bind DNA by the DNA binding domain and to unwind it afterwards in the helix domain. As all members of a protein domain are considered to fulfil the same function, this also allows to annotate multiple proteins that share a certain domain for which the function is known. This is also true for NABPs which mostly contain a defined nucleic acid binding domain.

Classifying the experimentally detected proteins in known, secondary and novel NABPs allows to ask the question, if the novel proteins contain a protein domain, that was not known be nucleic acid binding before. This domain could be already characterised on a sequence level but its function still undeciphered or the domain might also be completely uncharacterised. Subsequently two approaches will be described that allow to search for nucleic acid binding protein domains in the subset of novel NABPs of the experimental data set.

Searching for characterised nucleic acid binding protein domains

To identify nucleic acid binding protein domains that have already been characterised on the sequence level - known domains - it is sufficient to perform an enrichment analysis of protein domains in the set of novel NABPs identified earlier. Therefore the binomial distribution is applied to compute the probability, that an experimentally detected domain is over-represented in the set of novel NABPs compared to a control data set. The set of novel NABPs is derived from the classification determined earlier on Page 44. The complete cell lysate control data set already used in Figure 5.5 was used to estimate the abundance of the individual domains in proteins identified by mass spectrometry. This should provide a more realistic representation of the observable proteome compared to using the domain abundance in the complete human proteome and consequently avoid a bias in the analysis. Figure 5.8 shows an illustration of the enrichment analysis.

To avoid probabilities of zero for domains which do not occur in the control data set, the number of occurrences in the control data set of each domain was increased by one. Overall all 330 domains occurring in the set

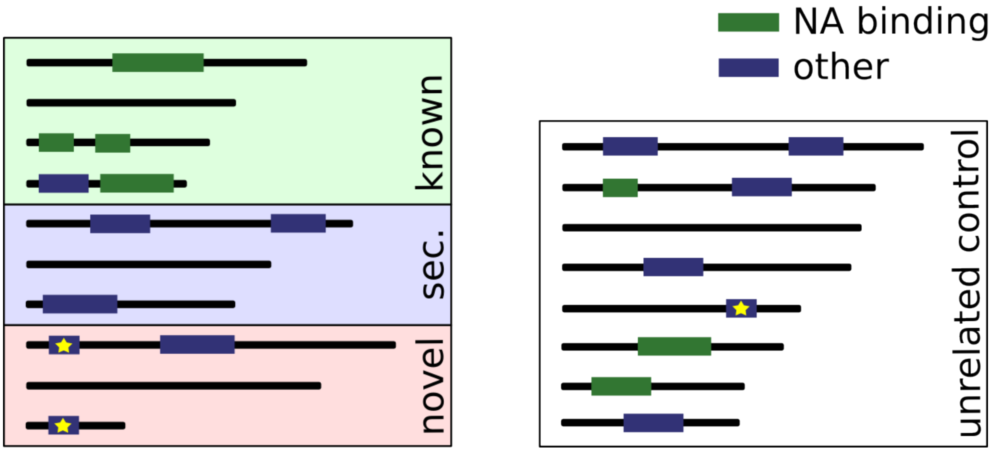


Figure 5.8 – Schematic representation of the domain enrichment analysis to identify characterised nucleic acid binding domains. The abundance of each domain in the group of novel NABPs is compared to its abundance in control experiments [14] to derive an enrichment score. The starred domain is enriched in the novel NABPs compared to its abundance the control experiments (right).

of novel NABPs were checked for over-representation. After Bonferroni correction [33] for multiple testing, nine domains were identified to be enriched at a significance level of 0.05 (see Table 5.2).

Often proteins within a protein family share the same domain composition and therefore sometimes groups of domains are found enriched that occur in the same set of proteins (e.g. the FERM adjacent and the FERM domain are shared by four proteins - EPB41L1, EPB41L2, EPB41L5, FARP1). The enrichment analysis cannot detect which of these domains within a group is likely to be nucleic acid binding, but at least one of them is likely to bind nucleic acid.

Searching for uncharacterised nucleic acid binding protein domains

Although protein domains have been extensively studied and a huge number of protein domains are described in public databases like Pfam [95] or SMART [73] our experimental data set has the potential to reveal novel nucleic acid binding protein domains. To test this the protein classification into known, secondary and novel interactors established before was used. As the group of known NABPs have mostly been annotated due to the presence of a known nucleic acid binding domain and the group of proteins classified as secondary interactors are most likely not in direct interaction with the

Pfam ID	**group**^{a}	**domain name**	**P value**	**proteins**
PF08736	1	FERM adjacent	0.00094	EPB41L1, EPB41L2, EPB41L5, FARP1
PF09380		FERM domain	0.02521	EPB41L1, EPB41L2, EPB41L5, FARP1
PF10239	2	Domain of unknown function	0.00473	FAM98A, FAM98B, FAM98C
PF00106	3	Short-chain dehydrogenase	0.00507	DCXR, DECR1, DECR2, DHRS2, HSD17B4, HSD17B8, HSDL2, PECR
PF03914	4	CBF/Mak21 family	0.01800	NOC3L, NOC4L
PF04900	5	Fcf1	0.01800	FCF1, UTP23
PF09532	6	DFDF motif	0.01800	LSM14A, LSM14B
PF09542		FFD and TFG box motifs	0.01800	LSM14A, LSM14B
PF12701		Lsm N-terminal domain of mRNPs	0.01800	LSM14A, LSM14B

Table 5.2 – Domains that were enriched in the set of newly identified NABPs.
^{a})Domains cooccuring in the same set of experimentally detected proteins could not be separated.

nucleic acid baits, these were excluded from the domain screen.

The remaining 251 proteins which were classified as novel NABPs before were subjected to a domain search pipeline. The design of this pipeline is following the design of pipelines which already proofed utility in identifying protein domains [30, 31]. Each protein was annotated with its known protein domains. None of these domains was annotated as nucleic acid binding. Of course this is not surprising, because proteins which contain a known nucleic acid binding protein domain would be annotated as nucleic acid binding, which would have led to a classification in the class of known NABPs earlier. Next, known domains were excluded from the domain search pipeline, as this analysis should identify novel domains. If the known domains contained in the class of novel NABPs are nucleic acid binding but not yet annotated they should have been detected in the enrichment analysis in the last section. Subsequently amino acid sequences outside known domains were then subjected to the domain search pipeline.

In this pipeline first short fragments - below ten amino acids - which are too short to be an independent domain were removed from the query sequences. Subsequently homologous protein sequences to the query sequences were searched using Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST [3]). PSI-BLAST is a homology search algorithm which can identify homologous sequences in a sequence database. The advantage of using PSI-BLAST is, that an iterative approach is performed. Homologous sequences identified in an initial BLAST search are compiled into a sequence profile, which is used as the query in the next iteration. This leads to increased sensitivity [91] as the sequence profile allows the search algorithm to take account for highly conserved residues in the query. Repeating the search with the generated profile leads to a result with increased sensitivity, from this result a new profile can be generated and the search iteratively repeated. Two factors indicate the identification of a novel domain:

- a domain must be present in a sufficient number of different species, which in turn shows evolutionary conservation of the sequence fragment and indicates that the fragment fulfils a relevant function which led to evolutionary conservation. It is easy to check the presence of the domain in multiple species by just counting the number of different species of proteins identified in the PSI-BLAST search.

- a new domain also needs to co-occur with different other domains, which shows its independence. In contrast, if a domain candidate only co-occurs with the same domain, this would indicate that the candidate is functionally related to the other domain. To quantify domain co-occurrence the number of distinct domain compositions was computed.

Figure 5.9 shows a schematic overview of the domain search pipeline.

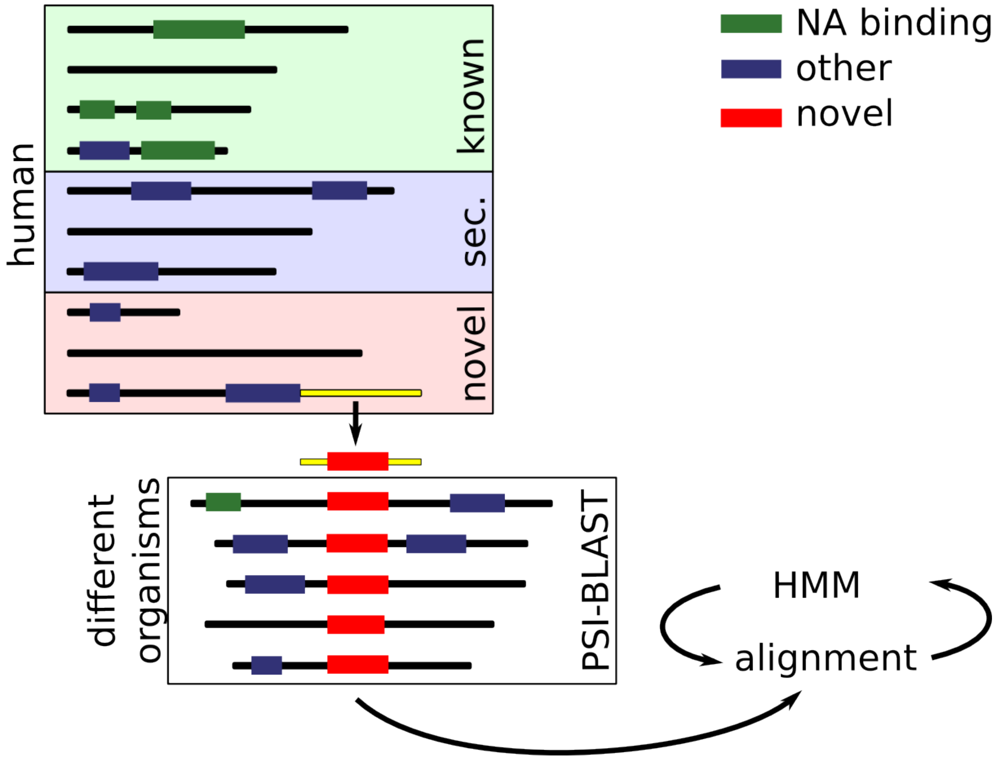


Figure 5.9 – Schematic representation of the pipeline used to identify novel nucleic acid binding protein domains. Sequences outside known domains (yellow) are queried against a multi-species database to identify evolutionary conserved regions. Resulting candidates are manually refined by iterative alignment and HMM searches.

An interesting candidate extracted by the pipeline is the uncharacterised protein C20orf72. As this protein was not annotated to contain known domains, the full-length (amino acids 1-344) of the protein was queried as a domain candidate sequence. This led to 2007 homologous hits, in 1451 different species and 39 different domain architectures after three PSI-BLAST iterations. Here the standard BLAST e-Value threshold of 10 was applied. The e-Value is a measure to assign statistical significance to hits in sequence analysis. It estimates the number of hits one would expect by chance in a database of the given size. After five iterations of PSI-BLAST 5463 hits in 2545 different species with 90 different architectures could be found. According to these numbers the candidate sequence fulfils the above mentioned criteria and was selected for subsequent manual refinement, as it furthermore represents a yet unstudied protein.

After removing very closely related sequences from the PSI-BLAST hits using cd-hit [74], an automatic multiple sequence alignment was generated by MUSCLE [38, 39]. This alignment was manually refined and a Hidden Markov Model (HMM) of the alignment was created. HMM models of a protein domain improve the sensitivity of the subsequent search. The resulting HMM model was queried against the NCBI non-redundant (nr) protein database using HMMer [36]. The nr database is the most comprehensive protein repository, which is essential here to generate a versatile domain model. Iteratively repeating these steps three times led to a satisfactory model of the candidate domain.

An interesting question that remains, is to deduce the function of the novel domain. Interestingly the novel domain co-occurs with a helicase domain called "UvrD-helicase" in a high fraction of the HMMer hits of the novel domain. To further investigate this the domain HMM was also queried against a protein sequence database of resolved 3D protein structures (pdb). This generated significant hits in two structures (pdb ids 1W36 and 3K70, both at an e Value of 3e-10), both structures of the same protein complex (see Figure 5.10).

This complex contains three yeast proteins (RecB, RecC and RecD), which were found to process DNA double strand breaks [103]. This function requires a helicase domain and a nuclease domain in the complex. The HMM hit of the novel domain maps to the nuclease domain of the structure of the protein complex. In combination with the finding that the novel domain often co-occurs with a helicase domain, this could point to a nuclease role of the novel domain. Here the role of the yeast complex could be performed by a single protein in other species, which contains both, a UvrD-helicase and the novel nuclease domain, which subsequently led to the co-occurrence.

Meanwhile the described domain was associated to a nuclease superfamily ("PD-(D/E)XK nuclease superfamily") in the recent version of Pfam [95] (Pfam 25.0), which supports the described observations.

5.1.2 Rsc normalisation

After these qualitative analysis normalisation is required to enable quantitative comparisons. To improve comparability between samples the raw spectral count data was normalised using Rsc normalisation [88]. Rsc normalisation allows to compare samples with each other. To enable the comparison of multiple samples a reference sample was created, which contained all experimentally detected proteins. The semi-quantitative abundance information for these proteins was generated by summing the spectral counts

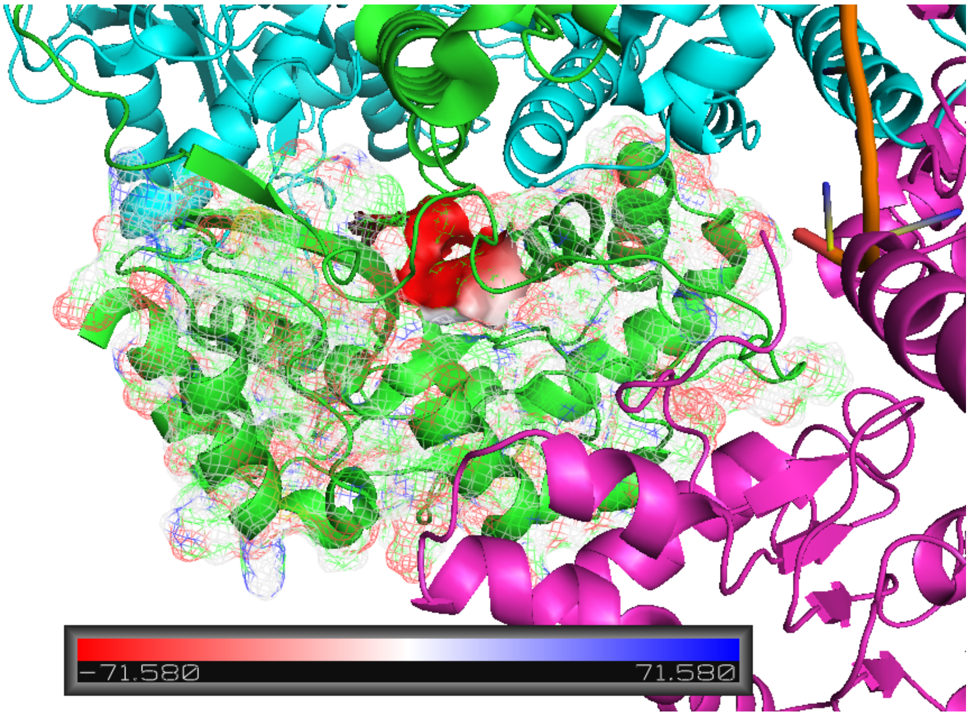


Figure 5.10 – The novel domain mapped onto the closest experimental 3D structure (PDB:3K70, represented as cartoon). Here a zoom into the mapped domain (shown as a grid) is shown, which lies in the helicase domain of the complex. The electrostatic potential of the active site which is highly conserved in the novel domain is shown as a solid surface.

of the individual experiments. As the Rsc normalisation computes the logarithm of the abundance ratio between the two samples, comparing individual samples to this reference sample results in a value which conveys if the protein was more or less abundant in the single sample than on average. After Rsc normalisation these values lie in a range between -10 and +10 and are real numbered in contrast to spectral counts that obviously are natural numbers.

5.2 Identifying the most powerful test to detect protein specificities

The systematic dinucleotide design of the synthetic nucleic acid bait sequences allows grouping of the baits based on their properties. Obviously baits can be separated into DNA- and RNA-baits. They can also be grouped according to their nucleotide content, because it is known for each nucleotide if it occurs in the bait sequence. Here also the dinucleotide design is of advantage, because the individual nucleotides do not occur in each bait, as it would most likely be the case in random sequences of the same length that allow the occurrence of all four nucleotides. Here, the use of all possible dinucleotide combinations leads to equally large groups of baits that do or do not contain a certain nucleotide which is of advantage for further analysis. Furthermore it is also possible to group the baits based on their methylcytosine content. Alternatively, methylcytosine could also be seen as an independent fifth nucleotide. The complete annotation of the baits and their corresponding properties can be seen in Table 5.3.

Given the unbiased experimental data set, we were interested if proteins in the data set were detected with higher abundance in a group of experiments, whose baits share a common property. This would indicate a higher affinity of this protein to nucleic acid with this specific property. For example, if a protein is detected with higher abundance in DNA purifications, this allows to infer that the protein is DNA specific. This also applies for single nucleotides, so if a protein is detected with higher abundance in baits that contain adenine(A) it is likely that the protein preferentially binds to nucleic acid sequences that contain adenine. Of course it is known for many proteins that they specifically bind to DNA or RNA. Also proteins binding to CpG islands upstream of transcribed genes or to the polyA tail are known - a long stretch of adenines at the 3'end of mRNA. But usually these specificities are detected in individual per protein studies. The data set generated here enables the unbiased detection of the specificities on a larger scale using data which also allows the comparison of different proteins from within a common

Table 5.3 – Properties of the 26 nucleic acid baits

Bait	**DNA**	**RNA**	**A**	**T**	**U**	**C**	**G**	**mCG**
strep								
dsDNA_dAdC:dGdT	X		X	X		X	X	
dsDNA_dAdG:dCdT	X		X	X		X	X	
dsDNA_dAdT:dAdT	X		X	X				
dsDNA_dCdG:dCdG	X					X	X	
dsDNA_dCdG:dCmdG	X					X	X	X
dsDNA_dCmdG:dCmdG	X					X	X	X
ssDNA_dA	X		X					
ssDNA_dAdC	X		X			X		
ssDNA_dAdG	X		X				X	
ssDNA_dAdT	X		X	X				
ssDNA_dAdTs	X		X	X				
ssDNA_dC	X					X		
ssDNA_dCdG	X					X	X	
ssDNA_dCdGm	X					X	X	X
ssDNA_dN	X		X	X		X	X	
ssDNA_dT	X			X				
ssDNA_dTdC	X			X		X		
ssDNA_dTdG	X			X			X	
ssRNA_A		X	X					
ssRNA_AC		X	X			X		
ssRNA_AG		X	X				X	
ssRNA_AU		X	X		X			
ssRNA_CG		X				X	X	
ssRNA_UC		X			X	X		
ssRNA_UG		X			X		X	

data set. Often also specificities reported in the databases were measured in a model organism and this knowledge was transferred to the human protein by sequence homology. This is a useful approach to increase knowledge of the individual proteins but usually happens without verification of the reported specificity in this species.

To detect specificities to bait classes with high sensitivity it is necessary to identify the best statistical method for this purpose. This method should take into account different aspects of the experimental data set. To identify the best suited statistical test it is essential to check if the experimental data is distributed according to a described standard statistical distribution, which would permit the application of a parametric test. Histograms are useful to visually inspect the distribution of the data and check if the measurements follow a statistical standard distribution. Figure 5.11 shows the histograms of four proteins that were detected in all the experiments. Obviously the data is not normally distributed, so it is not possible to perform a t-test to identify if there is a difference in the abundance means between the samples of two different bait classes. Also the measure - spectral counts - rather behaves as an ordinal measurement rather than interval scale as the measured count is not linearly related to the protein concentration in the sample. For parametric statistical methods data from at least an interval scale is required [9, p. 79]. Both observations, the data not being distributed according to a standard statistical distribution and measurements of ordinal scale, point to the fact that parametric statistical methods should not be applied on this data set.

The next question is now to find out which non-parametric statistical method is best suited to identify proteins that are specific to a certain subclass of nucleic acid baits within the experimental data set. This will be established in two steps.

First synthetic data sets are created to measure the performance and get an impression of the behaviour of different candidate statistical tests. These synthetic data sets are based on the actual experimental data to ensure realistic distribution of the data. The synthetic data sets are generated in a way that they fulfil the null hypothesis. Afterwards one group of measurements is modified at different orders of magnitude to violate the null hypothesis. This should simulate specificity in the synthetic data set. Then the different statistical tests are applied on this synthetic data sets. The results of the statistical test allow measuring the statistical power. Applying different candidate statistical tests on this synthetic data allows identifying the most powerful test to identify protein specificities.

In a second step the candidate methods are used to predict prior knowledge from the experimental data set. Here the different methods are applied

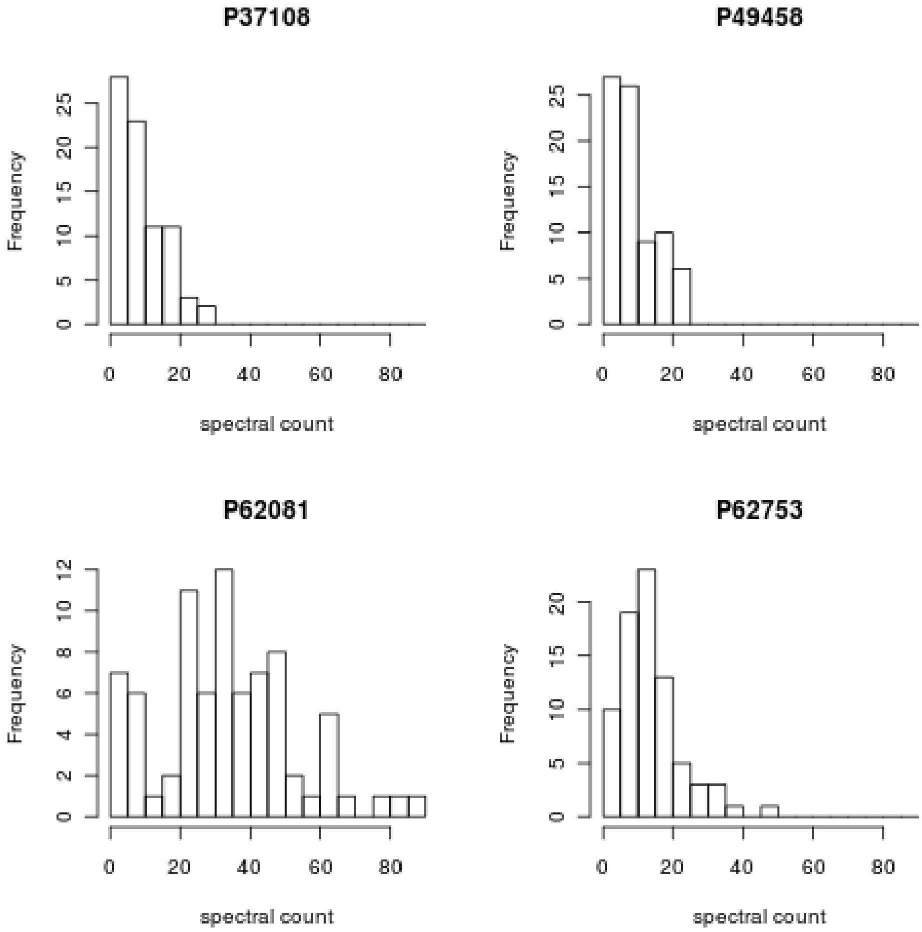


Figure 5.11 – Spectral count distribution of four frequently detected proteins.

to perform specificity analysis and identify proteins that specifically bind to DNA. The methods are evaluated on DNA binding proteins, because the annotation of DNA binding proteins is the most accurate among the bait classes that will be studied later. For most of the bait properties which are studied here no annotation is available in databases. Measuring the performance of the different methods should allow identifying the method which is best suited to identify proteins specificities. This method will then be applied on the complete data set and used to predict specificities for all bait classes (DNA, RNA, A, T, C, G, U and mCG).

5.2.1 Generation of synthetic data sets

Generating synthetic data sets can give valuable insight in the behaviour of a statistical method. To capture the specific aspects of the data set, it is beneficial to sample the data points of the synthetic data set from the original experimental data set. To obtain a realistic distribution of the measurements in the synthetic data sets, the actual experimental data is used to create a synthetic data set. Therefore a protein is randomly chosen from the experimental data set. Data of this protein measured in one randomly selected and shuffled. This ensures that the null hypothesis - no systematic difference in the proteins abundance between two groups of baits - is fulfilled. This procedure is repeated for 1000 randomly selected proteins to generate a synthetic data set of sufficient size The process of generating synthetic data sets is also illustrated in Figure 5.12.

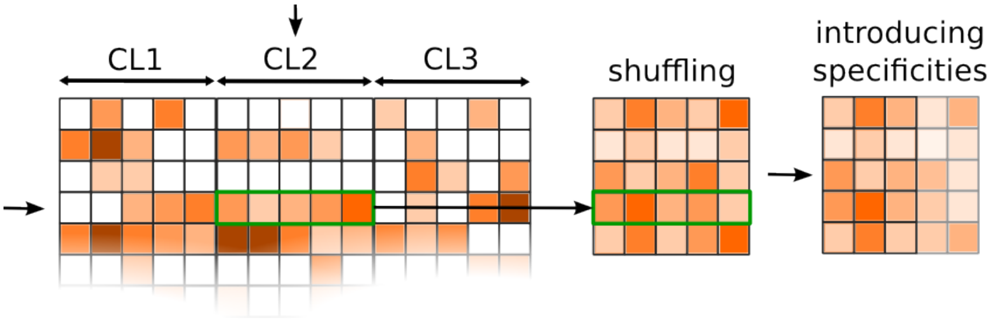


Figure 5.12 – Generating synthetic data sets. A random protein and cell line combination is selected and the data is shuffled. This step is repeated multiple times. Afterwards specificities between bait groups are introduced such that the synthetic data set does not fulfill the null hypothesis any more.

The number of simulated proteins is a compromise between accuracy of the simulation result and computation time. Obviously a larger synthetic data set allows estimating the statistical power of the different statistical methods more accurately. The synthetic data set generated here is a matrix with 26 columns and 1000 rows. The columns of this data set represent the measurements of the individual experiments of one cell type. As the original measurements are shuffled the number of columns between the synthetic and the real data set are identical, which should be advantageous to transfer observations in the simulation more easily to the real analysis. Furthemore the overall size of the synthetic data set - i.e. number of proteins - is comparable to the original experimental data set which is beneficial if we want to transfer observations made in the simulation onto the analysis of the real data set.

Under these conditions the null hypothesis is fulfilled and there is no difference - specificity - between two groups of measurements. Consequently the test should not reject the null hypothesis in more than α - the chosen significance level - cases in this data set.

5.2.2 Introduction of specificities

The synthetic data set generated above does not systematically violate the null hypothesis, and a statistical test should only identify a significant difference between sample groups in a number of cases that is dependent on the chosen significance level α. So if the synthetic data set is split into two groups of experiments applying a test at a significant level of $\alpha = 0.01$ should reject the null hypothesis in 10 out of 1000 proteins ($1000\alpha = 10$). This was ensured by the random sampling of the experimental data in the generation of the synthetic data set. To compare the statistical power of multiple tests, alterations in protein abundance need to be introduced in the data set in turn to violate the null hypothesis. Now the test which rejects the null hypothesis more often has higher statistical power than a test which cannot detect the violations of the null hypothesis as frequently.

We want to detect if there is a significant difference in protein abundance between two groups of experiments, the experiments which do have a certain property - e.g. DNA - and the other baits which do not have this property. As another layer of complexity we have to account for the fact that bait properties are not equally abundant in the list of baits (see Table 5.3). Therefore one parameter of the simulation is the size of the experimental group. The total number of experiments in one cell type is constant across the whole data set and predefined by the experimental design. This predefines also the size of the second experimental group, which contains the remaining experiments, as they do not have the bait property.

After splitting the experiments of the synthetic data set into two groups for different group sizes, the measurements of one group need to modified in a way that the currently fulfilled null hypothesis is violated. As spectral counts are an ordinal measurement with a minimum value of zero, there are different arithmetic possibilities to modify the measurements. Introducing specificities after generating a synthetic data set is also illustrated in Figure 5.12.

Introduction of specificities by linear modification

One option to introduce systematic difference between the measurements in the two groups is to multiply the measurements of one group with a factor smaller than one and thereby simulate a decreased abundance of the measurements in this group (see Equation 5.3).

$$m_{pe} = f u_{pe} \tag{5.3}$$

Here m_{pe} is the simulated - modified - abundance of the protein p in experiment e of the synthetic data set after modification and u_{pe} is the unmodified measurement.

Introduction of specificities by exponential function

As the relation between protein abundance and measured spectral count is not necessarily linear other possibilities to modify abundance are tested as well. Another option is to modify the spectral counts by an exponential function (see Equation 5.4). If the exponent x is chosen below one, this also decreases signals. In contrast to the modification by a linear function, this relation does not model a linear relationship between spectral counts and the proteins abundance.

$$m_{pe} = {u_{pe}}^{x} \tag{5.4}$$

As multiplication by a floating point factor or transformation with an exponential function can lead to real numbered values in the result which are not present in the original data, the values are rounded after modification, to maintain the discrete nature of the original measurements and preserve the property that the data set is composed of natural numbers $\mathbb{N}_0$.

Introduction of specificities by addition

Another option to introduce significant differences between the experiment groups of the synthetic data set is to add or subtract a constant value from the measurements of one group (see Equation 5.5)

$$m_{pe} = u_{pe} + c \tag{5.5}$$

Here it is important to not allow values below zero, as the spectral counts in the original data are also natural numbers and we want to maintain this property. Rounding is not necessary if the values of c are chosen as integer numbers and thereby no real numbers are introduced. Values below zero are replaced by zero after modification by Equation 5.5, to get rid of negative numbers and maintain the properties of the original experimental data set.

As the non-linear relationship between protein abundance and measured spectral counts does not suggest a clear method to simulate differential protein abundance all three of the above mentioned methods will be facilitated to modify the synthetic data set and introduce difference between experimental groups. Furthermore, testing different abundance/measurement relationships also will show robustness of the statistical methods.

Introduction of specificities in Rsc normalised data

Rsc normalisation [88] introduces major alterations to the characteristics of the measured data. This makes it necessary to consider if the possibilities to modify the synthetic data set described above are also applicable to modify Rsc normalised data. The positive integer valued count data is transformed to signed real valued measurements due to the logarithm based Rsc normalisation. Consequently, altering the data by multiplication (Eq. 5.3) or an exponential function (Eq. 5.4) is not meaningful any more and the only reasonable alteration that remains is adding or subtracting constant values (Eq. 5.5) to introduce specificities. As Rsc normalised data also contains negative values it is not necessary to replace negative values by zero to maintain the data properties, as it was the case for spectral count data.

5.2.3 Comparing statistical methods on synthetic data

After all these considerations about how to construct synthetic data sets, the statistical methods introduced earlier will be benchmarked on several synthetic data sets constructed as described above.

All these simulations are performed on synthetic data sets containing 1000 proteins. Hypothesis tests were performed one-sided at a significance level of $\alpha = 0.01$ to identify increased protein abundance in the unmodified group of experiments. Resampling methods were performed by performing 1000 random samples to compute the achieved significance level and infer the P value. Using 1000 bootstrap samples or performing 1000 random permutations for permutation tests respectively does not provide very accurate P values and was a compromise to achieve reasonable computation time. But as the analysis always considers the complete set of 1000 proteins and does not investigate individual proteins within the synthetic data set, the mistake which is made on individual proteins should even out over the complete set of simulated proteins.

Simulation on spectral count data

In a first example a synthetic data set containing 1000 proteins was generated from all the proteins in the experimental data set as described earlier. The synthetic data set was split into two groups of experiments. The first group contains twelve experiments, the size of the second group is predefined with a size of 14 experiments, as all the proteins simulated in the synthetic data sets sample the complete set of 26 measurements of a protein in a single cell type (26-12=14). The group of twelve synthetic experiments was kept unmodified and the synthetic measurements of the other 14 experiments were modified by multiplying spectral counts with a factor ranging from 0.5 to 1.5 in steps of 0.1 according to Equation 5.3. The result of this first simulation can be seen in Figure 5.13.

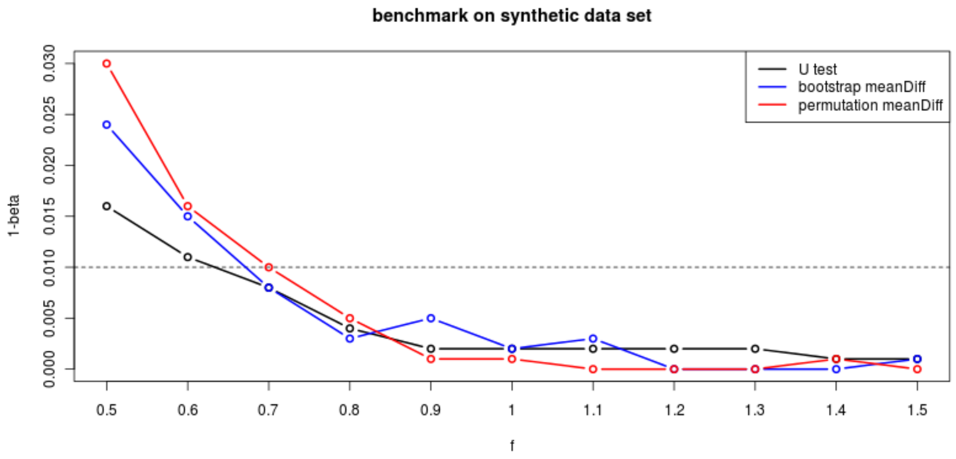


Figure 5.13 – Power curves for three tests, on the first synthetic data set

The plot shows the factor f which was used for modification on the x axis. A factor of 1 means, that there was no modification made to the original synthetic data set, which does not violate the null hypothesis due to column-wise shuffling of the original experimental measurements. The y axis shows the fraction of cases where the null hypothesis was rejected in the synthetic data set of 1000 proteins. So in cases where the original synthetic data set was modified to violate the null hypothesis this is one minus the type two error $1 - \beta$. The methods benchmarked on this first synthetic data set was the standard U test, a permutation test using the difference in group means as test statistic and a bootstrap hypothesis test also using the difference in means as test statistic. As described above, the resampling methods were performed using 1000 resampling iterations. This does not give very accurate results, but the graphs show a clear trend for all three methods with only modest sampling noise, which would be the case if the curves were not smooth. Surprisingly none of the methods rejects the null hypothesis for unmodified data ($f = 1$) at the correct level of α which was chosen as $\alpha = 0.01$ - also indicated by a dashed line in the plot.

Closer inspection of the results led to the source of this inaccuracy. If the experimental data, from which the synthetic data set is generated, is pre-filtered for proteins that were detected in at least ten experiments the tests correctly rejects the null hypothesis in 1% of the cases. This in reverse means, that the statistical test cannot identify a significant difference between synthetic experiment groups, when too little data is available and many of the data points are zero, which is the case for spectral counts when a protein was not detected in an MS experiment.

To facilitate a comparison of both cases a second synthetic data set was generated. Here, the experimental data was pre-filtered, and proteins that were detected in less than ten of the 78 experiments were excluded. This reduces the size of the experimental data set to 315 proteins. Based on these 315 experimentally detected proteins another synthetic data set was generated by randomly picking a cell type from a randomly chosen protein and permuting all the 26 data points measured within this cell type. This procedure was iterated 1000 times as described earlier, resulting in another synthetic data set that again contains 1000 proteins but now these were sampled only from proteins that were frequently detected (in at least ten experiments). The reduction of the sample space for generating the synthetic data set is not critical, although the number of starting proteins has now reduced to far below 1000, but there is still data for three cell types and this data is then permuted again, which still leads to a extremely high number of potential synthetic proteins.

Benchmarking the statistical methods on this second synthetic data set shows that now the tests correctly reject the null hypothesis at a fraction of $1 - \beta = 0.01$ in the unmodified ($f = 1$) synthetic data set (see Figure 5.14).

Here, the results of the simulation on the first synthetic data set where proteins were sampled from all experimentally detected proteins are shown as dashed lines, and the results of the second data set, where proteins were only sampled from frequently detected proteins are shown as solid lines.

The performance of the different statistical methods - different colors -

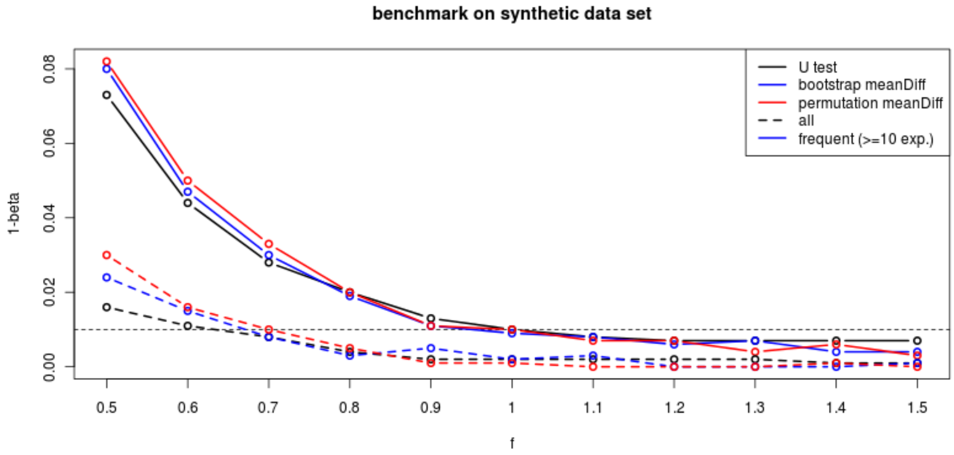


Figure 5.14 – Comparison of the power curves of three tests, on the first synthetic data set (dashed lines) and a synthetic data set that was generated from a pre-filtered experimental data set which only contains frequently detected proteins (solid lines)

is very similar on the simulated data set. As this second synthetic data set now allows to correctly evaluate the performance of different statistical tests further methods are benchmarked.

As a next step the U test statistic is introduced and plugged into the bootstrap and permutation test methods. Figure 5.15 shows that switching the test statistic from difference in means to the value of the U test statistic also does not have a big effect on the performance of the resampling methods.

To see the influence of property abundance, which is used to divide the data set into two groups of experiments, the same synthetic data set as used before was split into groups of size five and 21 $(26 - 5 = 21)$. As bait properties are not equally abundant, this should provide information about the performance of the test for differently abundant properties.

The results of this simulation in Figure 5.16 show that resampling methods using the difference in means (`meanDiff`) as a test statistic are still as powerful in rejecting an incorrect null hypothesis (with a value of $1-\beta \approx 0.08$ at $f = 0.5$) as before when the property group sizes were chosen 12:14. On the other hand the standard U test and resampling methods using the test statistic of the U test have decreased power compared to the last simulation.

So far differences between experiment groups were introduced by multiplying the data points of one experimental group by a factor different from one (as described in Equation 5.3). In the next step further types of mod-

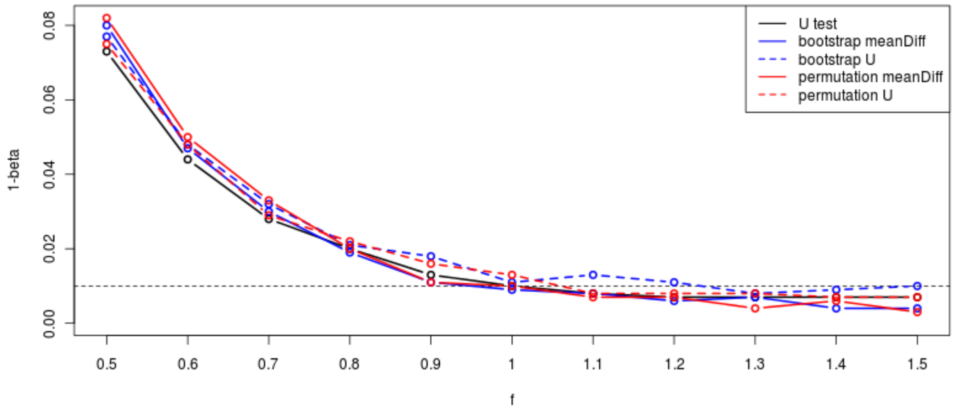


Figure 5.15 – Power curves for tests used before in comparison to using the test statistic of the U test as the test statistic of the bootstrap.

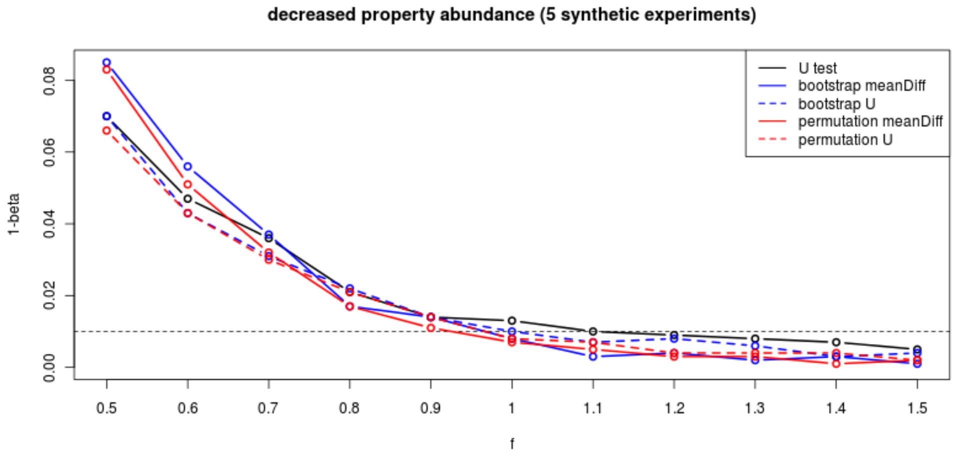


Figure 5.16 – Power curves for synthetic data sets, where bait property group sizes were changed. Here the abundance of the bait property is 5:21 in comparison to 12:14 used before.

ifications discussed above are tested to evaluate the behaviour of the test statistics under these circumstances.

First the synthetic data set used before is again split into groups of twelve and fourteen experiments and the measurements in the fourteen experiments are modified by an exponential function as described in Equation 5.4. The exponent x of this function is chosen between 0.5 and 1.5 in steps of 0.1. As $x^1 = x$, modifying by an exponent of one leaves the data points unmodified. The results of this simulation are summarised in Figure 5.17 in a similar way as before.

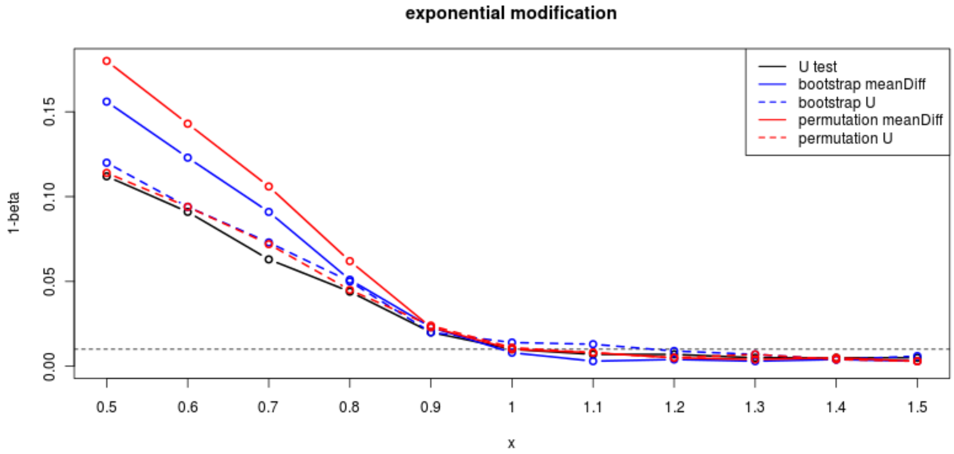


Figure 5.17 – Power curves of synthetic data sets, when abundance is modified by an exponential function (see Equation 5.4).

Now the exponent x is plotted on the x axis, instead of the factor f before. Similar to the last simulation with decreased property abundance, resampling methods using the difference in group means again show increased performance in this simulation.

The third way to introduce differential abundance between experiment groups discussed here, is to reduce the synthetic spectral count by adding a negative value to all the measurements of one group (see Equation 5.5). The same synthetic data set as described before was used and split into two groups of twelve and fourteen experiments. The measurements of the experiments in the second group were modified by adding a value that was chosen between -20 and +10 in increasing steps of one. Obviously, adding zero to the synthetic data set leaves the data set unmodified and should lead to $1 - \beta = 0.01$.

The results of this simulation in Figure 5.18 indicate - in contrast to

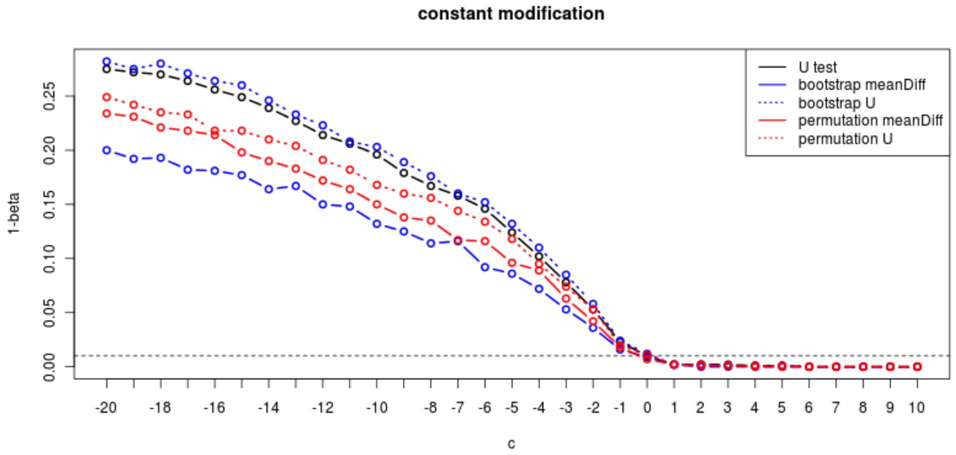


Figure 5.18 – Power curves of synthetic data sets, when abundance is modified by adding a constant value (c on the x-axis; see also Equation 5.5).

previous simulations - that the test statistic of the U test outperforms using the difference in group means as test statistic. Here the constant which was added to one group of measurements in the synthetic data set c (as described in Equation 5.5) is plotted on the x axis. Also the standard U test performs surprisingly well on this synthetic data set.

Simulation on Rsc normalized data

So far only the raw spectral count data was facilitated to simulate the performance of different statistical methods on synthetic data sets. As the data was normalised using the Rsc method [88] also this normalised data matrix can be used to perform benchmarks and simulate the performance of the statistical methods. As Rsc normalisation produces positive and negative real numbered data points, only modification by adding a constant value (Equation 5.5) allows to modify positive and negative measurements equally. To benchmark the performance of the different tests, the Rsc normalised experimental data was used and a synthetic data set was created similar as described before. A cell type was randomly chosen for a random protein that was detected in at least ten experiments, then the Rsc normalised measurements of this protein in that cell type were permuted. A synthetic data set containing 1000 proteins that were detected in at least ten experiments was created this way. Afterwards this data set was modified by adding a constant value ranging from -1 to +1 in steps of 0.2 to one group of the experiments (see Equation 5.5).

Figure 5.19 shows the simulation results when the synthetic data set is grouped into two groups of experiments of size 12:14 and Figure 5.20 shows the result for a less abundant property where the experiments are grouped in the ratio 5:21.

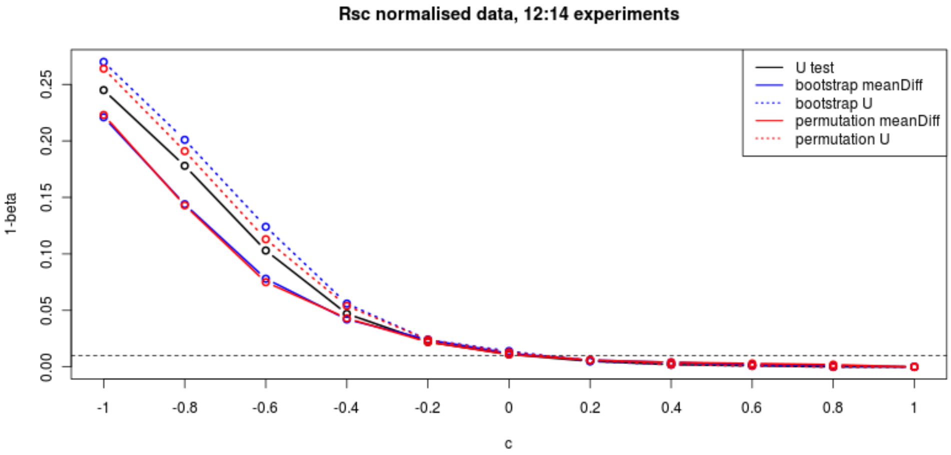


Figure 5.19 – Power curves for a synthetic data set generated from Rsc normalised data. Here only introducing abundance differences by subtracting a constant (Equation 5.5) can be applied.

Essentially the methods perform very similar, which is likely to be a result of the normalisation. So probably if the data is well normalised the choice of the statistical test does not have a big influence on the power of the analysis.

Summary of simulation results

To summarise the results of this round of simulations:

- The performance of the different methods in synthetic data modified by multiplication is similar (Equation 5.3 and Figure 5.15).
- The performance of resampling methods that use the difference in group means as test statistic is slightly reduced, when the property abundance is reduced from twelve to five (Figure 5.16).
- The performance of resampling methods that use the difference in group means as test statistic is increased, when the synthetic data is modified by an exponential function (Equation 5.4 and Figure 5.17).

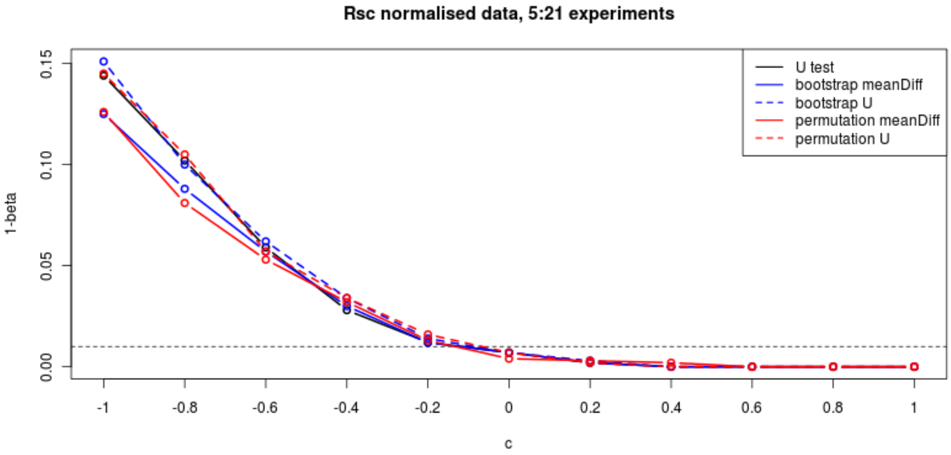


Figure 5.20 – Power curves for a synthetic data set generated from Rsc normalised data. Here bait property abundance was again altered from 12:14 to 5:21.

- The performance of resampling methods that use the test statistic of the U test as test statistic is increased, when the synthetic data is modified by adding a constant value (Equation 5.5 and Figure 5.18).

- The standard U test performs well when differences are introduced by adding a constant (Equation 5.5 and Figure 5.18), but shows rather average power otherwise.

- When the data is normalized using Rsc, all statistical methods perform very similar.

- Apart from the first simulation, were synthetic proteins were sampled from all - even rarely - experimentally detected proteins all the simulations correctly reject the null hypothesis in $1 - \beta \approx 0.01$ cases in unmodified data ($f = 1$, $x = 1$ or $c = 0$).

5.2.4 Comparing statistical methods on synthetic data for multiple cell types

So far the statistical methods were compared on synthetic data sets containing 1000 proteins and 26 experiments. These 26 experiments were derived from data that was measured within a single cell type. The experimental screen was performed in three cell types to increase protein coverage. As protein specificity is identical across cell types, it is advantageous if the statistical method to detect protein specificities can incorporate data from multiple cell types and thereby improve the confidence or increase the sensitivity of the analysis. Therefore in the next step synthetic data sets, which simulate multiple cell types are generated and used to benchmark different statistical methods to detect protein specificities.

These data sets are generated in a similar way as the synthetic data sets were generated before. A protein is randomly chosen from the pre-filtered experimental data set, where proteins that were detected in less than ten out of the 78 experiments were removed. The experimental data from a randomly selected cell type of this protein is permuted to generate the data of the first synthetic cell type. Afterwards the exactly same data - from the same protein in the same cell type - is permuted again to generate synthetic measurements for a second cell type in the synthetic data set. All the measurements of the second synthetic cell type are multiplied by an arbitrary selected factor of $f = 0.2$ according to Equation 5.3 to simulate different expression levels of that protein in the second cell type. After modification the synthetic measurements are rounded to maintain the discrete natural number characteristics of the original experimental data set in the synthetic data set. This procedure is repeated 1000 times to generate a synthetic data set of multiple proteins.

As the synthetic data was generated by permutation of the experimental data, the synthetic data set fulfils the null hypothesis, which is that there is no systematic difference between the mean of different groups of experiments.

Now significant differences of increasing magnitude are introduced between two groups of experiments to violate the null hypothesis and perform power analysis. As the list of baits, that were used to perform the experiments, was identical across cell types, the group size of the two groups of experiments also needs to be identical in different cell types. Therefore splitting the 26 synthetic experiments needs to be done in both synthetic cell types equivalently to maintain this characteristics of the experimental data set. Hence, both synthetic cell types are split into two groups at the same ratio, for example 12:14 as mostly used before.

Several statistical methods, that take into account that the data was measured in multiple cell types, were proposed in the Methods chapter (Chapter 3). These methods were not benchmarked in the power analysis of the synthetic data sets earlier, as these data sets simulated only data of a single cell type. Here, with synthetic data sets simulating multiple cell types, these methods will be included in the benchmark. Also the statistical methods benchmarked before are included for comparison, although they do not take into account cell types explicitly. Figure 5.21 shows the results, when bootstrap methods are benchmarked on the synthetic data set described above, where synthetic data for two cell types is generated and the expression of the second synthetic cell type is reduced by a factor of $f = 0.2$. This Figure is

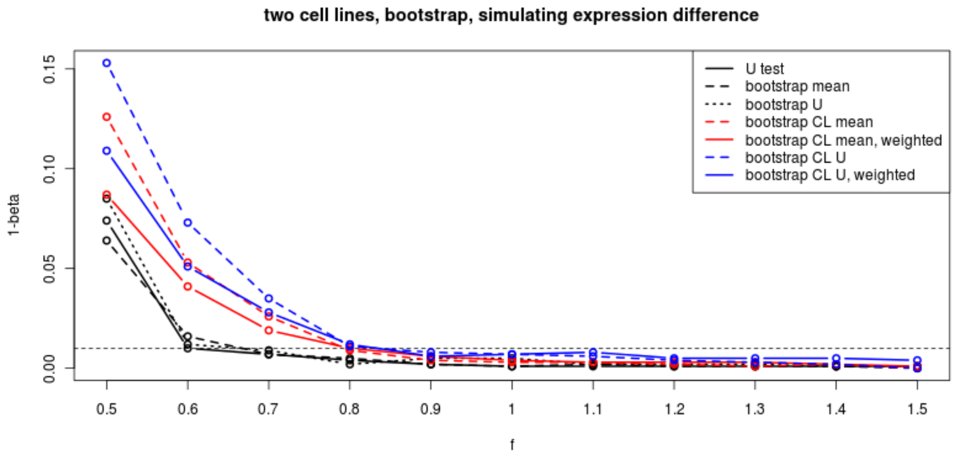


Figure 5.21 – Power curves for bootstrap methods applied to synthetic data sets simulating multiple cell types. Here a second cell type is simulated with 80% reduced signal intensity ($f = 0.2$ in Equation 5.3). Test statistics which account for multiple cell lines are indicated by `CL`. Basically here the test statistic is computed for each cell line individually and combined afterwards by either computing the sum or the weighted mean (`weighted`). Details see Page 27.

comparable to Figure 5.15 with the difference, that here two cell types are simulated. The methods not taking into account multiple cell types in their design (shown in black) do not reject the null hypothesis correctly in one percent of the cases (α was set to 0.01 again) in unmodified data ($f = 1$). Most importantly, the methods that use the test statistic of the U test per cell type, are more powerful compared to methods where the difference in means is used as the test statistic by the bootstrap.

Figure 5.22 shows the power curves, when performing the same simulation on permutation test methods. Here all the methods are too conservative and reject the null hypothesis too often when the synthetic data was unmodified ($f = 1$). Methods that perform weighting of the test statistic per cell type based on the amount of expression/detection have increased power compared to the other methods. Again, methods that were designed to analyse data of multiple cell types are more powerful than methods not taking into account multiple cell types. Comparing the simulation results of synthetic data for two cell types (Figure 5.21 and 5.22) with previous simulations, where only one cell type was simulated (see Figure 5.15 for direct comparability), shows

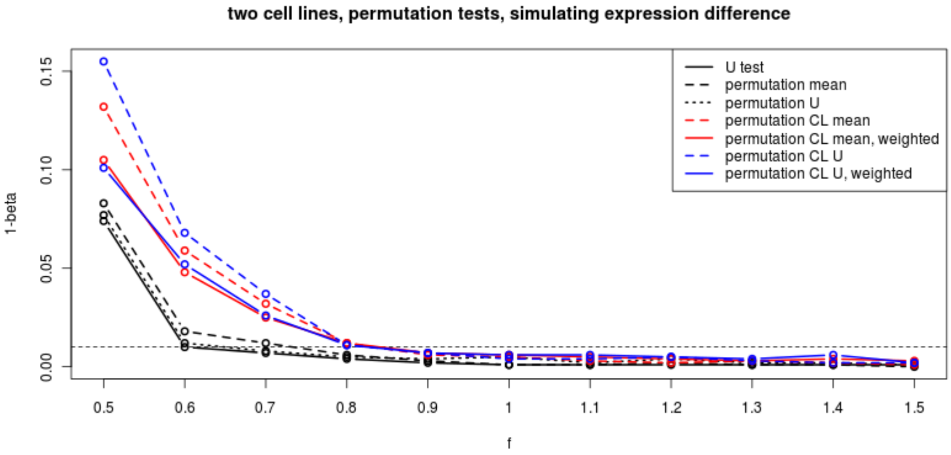


Figure 5.22 – Permutation tests applied to synthetic data sets simulating multiple cell types as done in Figure 5.21 for Bootstrap methods.

that methods that were designed to incorporate data from multiple cell types take advantage from the information contained in the second cell type. This leads to increased power compared to methods which do not take into account the experimental setup (i.e. the separation in multiple cell types). Interestingly, the power of methods that do not take into account cell type information remain at a similar level of performance ($1-\beta \approx 8\%$ at $f = 0.5$) as for single cell line data, whereas test statistics that take advantage of the cell type information gain power ($1-\beta$ up to 15% at $f = 0.5$).

Estimating the influence of noise

Simulating data of multiple cell types also provides an easy opportunity to estimate the influence of noise on the performance of the different test statistics. Therefore the synthetic data of the second simulated cell type, in which the abundance was reduced by 80% by multiplying with a factor of $f = 0.2$, is not modified when significant differences are introduced. So here the synthetic data set is generated by twice permuting data of a random protein to generate data for two synthetic cell types. In the second cell type the abundance is decreased by multiplying by a factor of $f = 0.2$. Afterwards, when the data is modified to violate the null hypothesis, only the data of the first synthetic cell type is altered. The data of the second cell type remains unchanged and therefore does not violate the null hypothesis. In contrast to the data of the first synthetic cell type which violates the null hypothesis after alteration, the second cell type can be considered as noise, which does not contribute to rejecting the null hypothesis in the statistical analysis. As the abundance in the second cell type is reduced, statistical methods should still be able to identify violations to the null hypothesis, but probably in a reduced number of cases. This should lead to reduced power in the analysis compared to the last simulation.

Benchmarking the bootstrap methods on this synthetic data set, in which the second cell type only contributes noise and does not violate the null hypothesis, shows that here methods that use the test statistic of the U test correctly reject the null hypothesis in 1% of the proteins when the synthetic data was unmodified ($f = 1$), whereas the other methods do not reject the null hypothesis often enough (see Figure 5.23). Again methods that take into account that the measurements were derived from multiple cell types perform better than simple test statistics which do not take cell types into account. As expected the power of all methods is reduced upon the introduction of noise, as now the most powerful method can reject the null hypothesis in $\approx$ 7% of the cases as compared to $\approx$ 15% for the best method before in Figure 5.15.

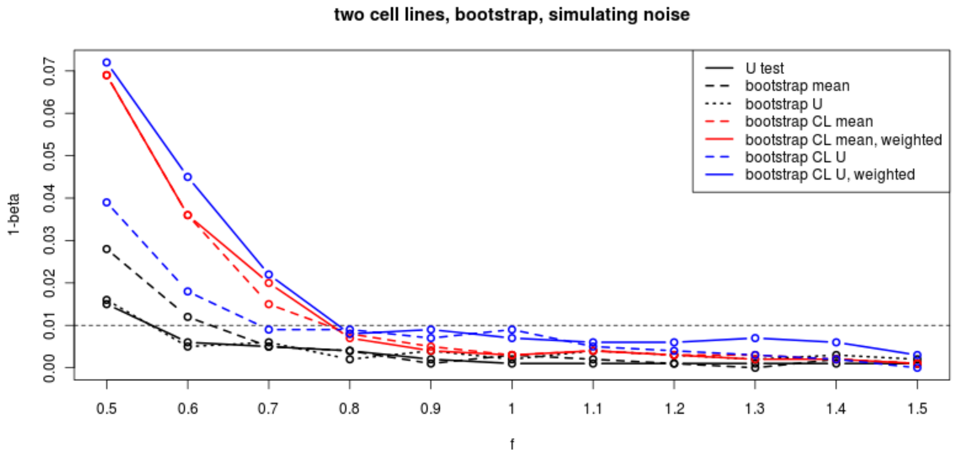


Figure 5.23 – Power curves for bootstrap methods applied to synthetic data sets simulating multiple cell types. Here a second cell type is simulated with 80% reduced signal intensity. Data of this second - less abundant - synthetic cell type is not modified to violate the null hypothesis, which introduces noise.

Comparing the power of permutation tests (see Figure 5.24) reveals that all permutation methods tend to reject the null hypothesis in slightly too little cases. Here permutation tests which use the difference in means as test statistic are more powerful than the tests statistic of the U test. Again, the

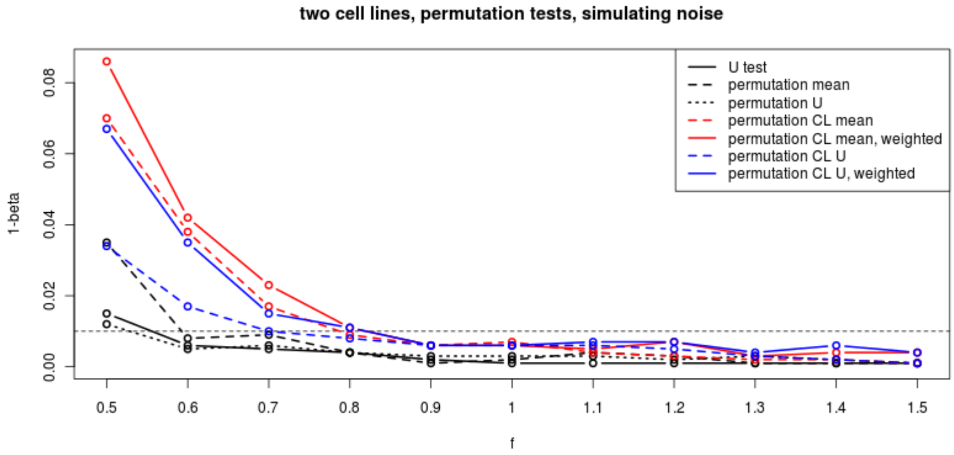


Figure 5.24 – Permutation tests applied to synthetic data sets simulating multiple cell types as done in Figure 5.23 for Bootstrap methods.

power of all methods is reduced upon the introduction of noise via the second cell type.

Summary of simulation in multiple cell types results

The key results of the simulation of multiple cell types using synthetic data sets can be summarized as follows:

- Bootstrap methods that use the test statistic of the U test per cell type are more powerful compared to methods which use the difference in mean as the test statistic (see Figure 5.21).
- Permutation test methods that do not weight the test statistic have higher power than methods which weight the test statistic per cell type (see Figure 5.22).
- Many methods are too conservative and do not reject the null hypothesis at the correct level of α when data of multiple cell types that does not violate the null hypothesis is analysed ($f = 0$). This is more prominent for methods that do not take into account cell types.
- As expected the introduction of noise reduces the power of all statistical methods.
- Statistical methods that were designed to take cell types into account perform better on data that contains multiple cell types.

Simulations on synthetic data sets allow to dissect the influence of different factors on the statistical analysis. Here the influence of:

- the relation of protein abundance and measured spectral counts
- Rsc normalisation [88]
- the abundance of bait properties and thereby the effect of different group sizes
- multiple cell types including expression differences
- experimental noise

for different proposed statistical methods have been studied. Test statistics which were designed to incorporate data from multiple cell types clearly outperform other methods ignoring cell type information. Although these simulations are very informative and provide some insight on the behaviour of the methods to different aspects of experimental data, no method turned out to be superior in all the studied aspects. Another option to benchmark different methods is to evaluate their performance to predict prior knowledge. This aspect will be handled in the next section.

5.2.5 Comparing statistical methods on annotated data

Besides the rather artificial considerations to benchmark different statistical tests on synthetic data sets to determine the most appropriate candidate for data analysis, it is also possible to use available knowledge about proteins within the experimental data set to compare the performance of different statistical methods and determine the best suited method.

Gene Ontology (GO) [4] provides a platform which allows to annotate the function of genes and proteins. “DNA binding” (GO:0003677) and “RNA-binding” (GO:0003723) are represented in Gene Ontology. Both are child nodes of “nucleic acid binding” (GO:0003676) inside the “Molecular Function” sub-branch of Gene Ontology (see Figure 3.2). Genes and proteins can be annotated by functions represented in GO. The SwissProt database [1, 8], which was used to identify proteins in the experimental data set, facilitates GO annotation. Therefore a protein entry incorporates a link to a node identifier in Gene Ontology to annotate the protein with a certain function. So a DNA binding protein contains a link to the node “DNA binding” in its Gene Ontology annotation. As “DNA binding” is connected to multiple child nodes in GO, also proteins which are connected to one of these child nodes by their annotation are declared as “DNA binding”. For example the protein PARP1 is annotated as “DNA binding”, whereas the protein XRCC5 is annotated as “double-stranded DNA binding” and “telomeric DNA binding”. These functions themselves are child nodes of “DNA binding” in the GO graph and therefore XRCC5 is also specified as “DNA binding”. This hierarchical structure allows to annotate proteins at different levels of precision, which can be useful if limited information is available as well as for defining broad functional roles. Of course, the same is also true for RNA-binding proteins where proteins annotated with a child node of RNA-binding are RNA-binding. As DNA- and RNA-binding both are child nodes of “nucleic acid binding”, all the proteins annotated to one of these child nodes are annotated as “nucleic acid binding” automatically.

The experimental design of this study allows to group experiments based on their bait composition into groups of DNA-, RNA-, adenine-, ... methylcytosine-containing baits (see Table 5.3). The statistical analysis of DNA vs. RNA containing baits should reveal DNA specific proteins. As Gene Ontology provides information about DNA- and RNA-binding proteins, this allows to create a set of proteins that should be identified by the analysis. DNA- and RNA-binding are the only two nodes that can be used to create a set of true positives this way, as the other specificities of the statistical analysis are either only annotated in a very small number of proteins of the SwissProt database or are not represented in Gene Ontology at all. The annotation process also allows to annotate a protein as both, DNA-binding and RNA-binding. This annotation does not allow to infer that the protein is DNA-specific, and therefore proteins that are annotated as DNA- and RNA-binding are excluded from the true positive set. Unfortunately Gene Ontology annotation is not complete, so there are DNA-binding proteins which are not annotated as such. The information represented in the annotation is also not perfectly accurate, so there might be false positives in the annotation. Furthermore the annotation is also biased for well-studied proteins, for which more annotation is present. Nevertheless it provides a valuable starting point to benchmark the statistical methods on external knowledge.

2018 of the 20422 proteins in the SwissProt database [113], which we used to perform database search of the MS data and identify the proteins in the experimental sample, are annotated as DNA binding (see also Table 1.1). 190 of these proteins were actually identified in the experimental data set. A substantial fraction - 59 - are also annotated as RNA-binding. Now this knowledge provides a benchmark data set of 131 proteins that should be identified as DNA specific by the statistical analysis. As proteins rarely detected in the experimental data set cannot be significantly detected by the statistical analysis, the experimental data set was again pre-filtered for proteins that were detected in at least five experiments, which reduced the size of the experimental data set to 469 proteins, 73 of which are annotated as DNA-binding and not annotated as RNA-binding. Experiments were grouped into classes of DNA- and non-DNA-baits and the candidate statistical methods proposed to identify protein specificities were applied to this experimental data set to identify significant differences between the two experimental groups. The 73 proteins exclusively annotated as DNA-binding or one of its child nodes were used as true positives. Figure 5.25 shows the complete ROC curve of the simulation results in the first panel. As high false positive rates (FPR) are not desired the second panel provides a zoom in on the range of a FPR between 0 and 0.2.

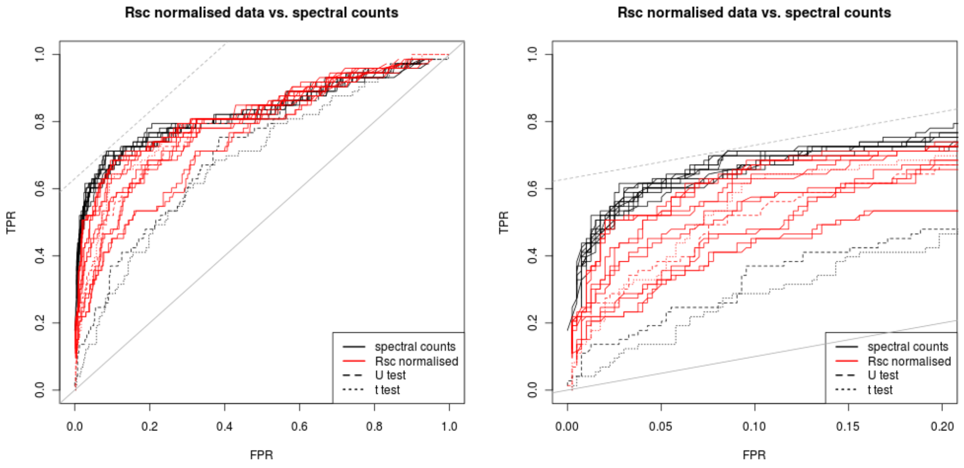


Figure 5.25 – ROC curves of multiple candidate tests, trying to infer DNA specificity when using GO annotation as gold standard.

As higher FPR rates are not relevant for the analysis, further ROC curves will also display this FPR range for comparability. Figure 5.25 allows to compare the performance of the statistical methods on raw spectral count measurements (in black) and Rsc normalised data (in red). Interestingly most statistical methods perform better on spectral count measurements as in Rsc normalised data. Test statistics that are designed to handle data from multiple cell types show a clear improvement compared to employing t- or U-test on the complete data set for spectral count data, whereas the improvement on Rsc normalised data is not very strong. As the quality of prediction results is clearly better when spectral counts are used as input data, these measurements will be facilitated to perform the specificity analysis. The next step is to find the best method to identify specificities from spectral count data.

Comparing permutation test methods with the bootstrap using the same set of test statistics, shows very similar performance for both methods (see Figure 5.26).

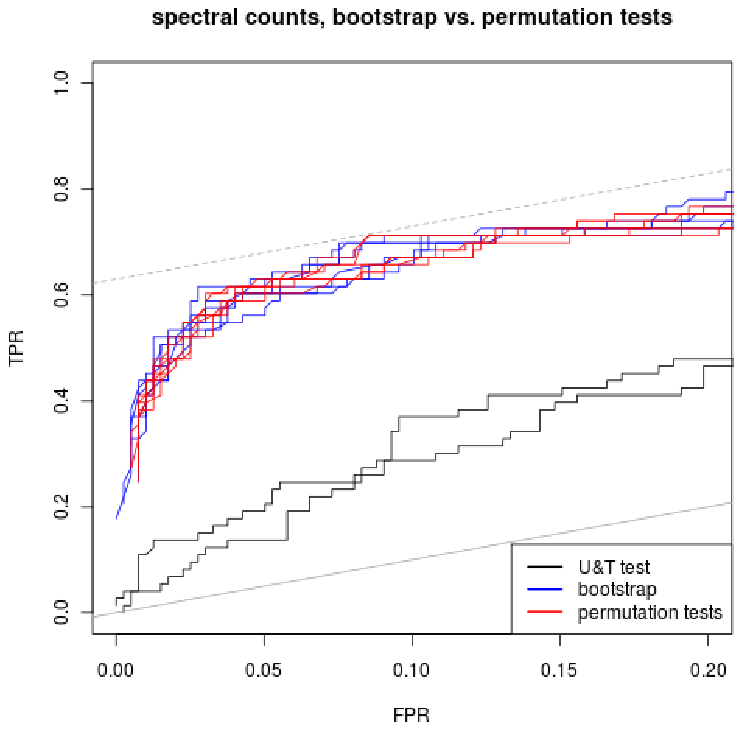


Figure 5.26 – ROC curves to compare the performance of bootstrap to permutation tests. Again GO annotation is used as a gold standard.

Figure 5.27 shows the comparison of methods which use the test statistic of the U test compared to methods which use the difference of experiment group means as test statistic. Here methods (permutation tests and bootstrap) which use the test statistic of the U test perform better than methods, which use the difference in means as the test statistic.

5.2.6 Choice of the statistical test

Benchmarking different statistical methods that employ multiple test statistics showed that:

- best performance is achieved on spectral count data
- test statistics which are designed to take multiple cell types into account (described in Section 3.3.1) perform better than standard t or U test

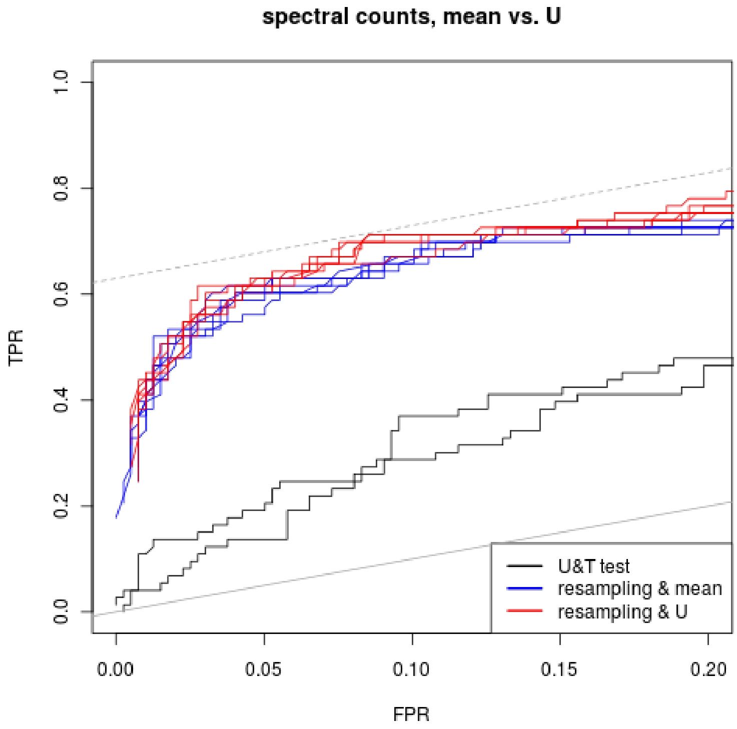


Figure 5.27 – ROC curves to compare the difference in mean to the test statistic of the U test plugged into either a permutation test or the bootstrap.

- the test statistic of the U test has slight advantage compared to difference in group means

Therefore the specificity analysis will be performed on spectral count data, using the test statistic of the U test on cell types individually. As the U test statistic does not weight cell types according to the signal magnitude, the test statistics of the individual cell types are weighted according to the signal intensity, which also increased performance on the synthetic data. The performance of this test statistic coupled to a permutation test is shown in Figure 5.28. This Figure shows the performance of all the tested methods to predict DNA binding proteins - according to Gene Ontology - from spectral count measurements in the experimental data set. Here the permutation test which uses the test statistic of the U test per cell type (`permClUW` shown in red) is compared to the performance of all the other methods evaluated.

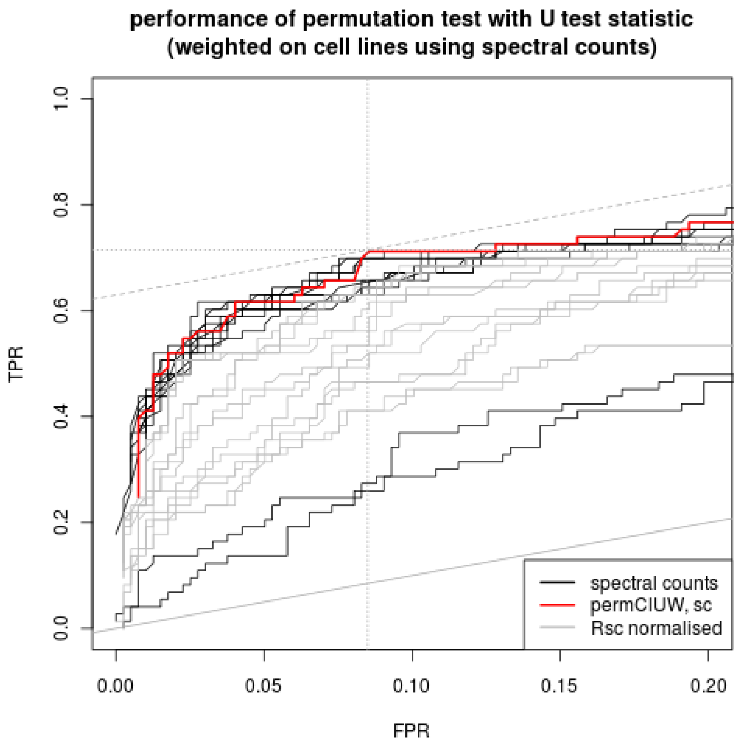


Figure 5.28 – Performance of the selected test - a permutation test which computes the test statistic of the U test on individual cell lines and combines these by weighted mean - in comparison to other methods tested, using GO annotation as a gold standard.

The selection of this test is also assured in the ROC curve in Figure 5.28. Selecting a classifier and the corresponding threshold is always a tradeoff between trying to maximise sensitivity (TPR) and specificity (1-FPR) at the same time. These two optimisation criterions are always counteracting, which means that increasing sensitivity reduces specifictiy. A classifier which assigns classification randomly would achieve identical TPR and FPR indicated by the grey line. The classifier which achieves maximum distance from this 45 degree line is best suited to classify the data. The selected permutation test (`permClUW`) achieves a TPR of 71.5% at a FPR of 8.5% (indicated by grey dashed lines in the ROC curve).

This method will be applied to analyse the spectral count data of the experimental data set. Now also the remaining bait properties according to Table 5.3 will be analysed - for most of which no annotation is available.

5.3 Identification of protein specificities

Now, after the best suited test to identify protein specificities was determined, this test is applied to identify protein specificities for all classes of baits defined in Table 5.3 on the experimental data set. Therefore a permutation test is applied. This test uses the non-parametric test statistic of the U test as a test statistic. To compute the test statistic the measurements are divided into two groups dependent on if the bait property of interest - for which specificity should be determined - was present in the nucleic acid bait that was used in the individual experiment. Due to different abundance levels across cell lines, this test statistic is computed on the three experimental cell lines individually. These three values are combined into a single test statistic by computing the weighted mean. These weights correspond to the sum of the signal in the corresponding cell line (as described in Section 3.3.1). This test was determined as the best suited statistical test to identify protein specificities in the previous Section.

The studied bait properties were DNA, RNA, adenine (A), thymine (T), cytosine (C), guanine (G), uracil (U) and methylated cytosine (mCG) (see Table 5.3 for exact assignment of bait properties to individual baits). Unfortunately the specificity for single- and double-stranded baits could not be determined, because the experiments were performed in two batches, where double-stranded pulldowns were performed in a separate batch and so identified specificities could also be due to experimental differences. Overall this analysis led to the identification of 254 protein specificities at a P value below $P < 0.01$. These specificities were identified in 174 of the 921 proteins. The number of proteins with significant specificities is smaller than the number of significant specificities because some proteins had multiple significant specificities. Table 5.4 shows the number of significant proteins

P values calculated in the specificity analysis were not corrected for multiple hypothesis testing.

Table 5.4 – Number of specific proteins for individual bait property groups ($P < 0.01$)

DNA	RNA	A	T	C	G	U	mCG
38	75	5	35	10	43	27	21

Figure 5.29 visualizes the identified protein specificities on the network of protein identifications. Here the layout is identical to Figure 5.7. The location of protein specificities in the force directed network layout co-locates with bait properties which were used to position the baits in groups in the layout.

5.3.1 Validation based on external annotation

To get a first impression if the statistical test can detect proteins which specifically bind to different types of nucleic acids, proteins which bind specifically to DNA were annotated using Gene Ontology (GO) [4]. Table 5.5 shows the result of this annotation process. Here 34 of 38 proteins are annotated as DNA binding in GO. Of the remaining four proteins two are obviously misannotations. These are TFAM, as it is a transcription factor which by definition bind DNA, and POLR2H, a DNA directed polymerase which also binds DNA. One of the remaining two proteins is a uncharacterised protein which has not been focus of intense studies and therefore almost no annotation is available. This again demonstrates the value of this unbiased study that has the ability to reveal knowledge about unstudied proteins. The last remaining protein in the list is Lysozyme C (LYZ), which was not implicated to be DNA binding and represents a potential false positive of the analysis. Overall this first annotation based validation demonstrates the power of the experimental approach in combination with focused statistical analysis. This validation step was focusing on DNA specific proteins. In a next step identified specificities for all bait classes are validated by additional experiments.

5.3.2 Experimental validation

In the next step, predicted specificities were tested by additional experiments. Four candidates were chosen which should be specific for DNA, RNA, A/T

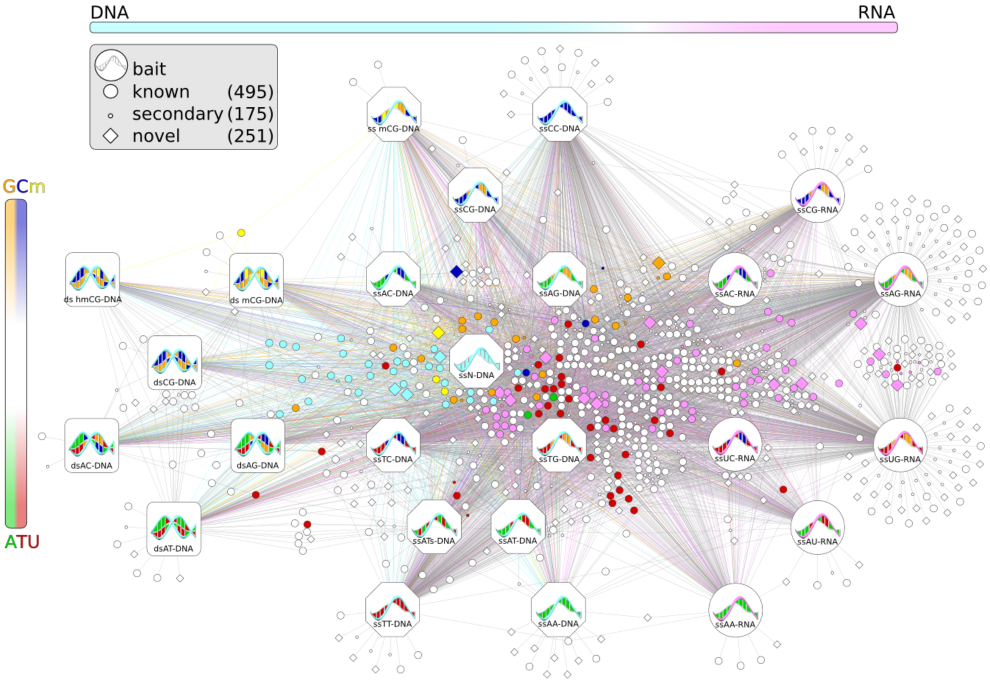


Figure 5.29 – Overview of the data set. Specificities detected in the statistical analysis are mapped onto the interaction network, already shown in Figure 5.7. Baits are indicated by large nodes. Nucleotide composition of the baits and specificity of proteins are color coded as before. In case of multiple specificities for a single protein the most significant one is reported. Interacting proteins are split into three groups (known, likely secondary, and novel) based on public annotation and interaction data.

Table 5.5 – Table of significantly DNA specific proteins ($p_{DNA} < 0.01$) and their GO annotation for "DNA binding" (GO:0003676).

protein	name	p_{DNA}	GO
PARP1	Poly [ADP-ribose] polymerase 1	0.00000	TRUE
XRCC6	X-ray repair cross-complementing protein 6	0.00000	TRUE
XRCC5	X-ray repair cross-complementing protein 5	0.00000	TRUE
XRCC1	DNA repair protein XRCC1	0.00000	TRUE
MSH3	DNA mismatch repair protein Msh3	0.00000	TRUE
HMGB2	High mobility group protein B2	0.00000	TRUE
MPG	DNA-3-methyladenine glycosylase	0.00000	TRUE
RFC4	Replication factor C subunit 4	0.00000	TRUE
RFC1	Replication factor C subunit 1	0.00000	TRUE
MSH2	DNA mismatch repair protein Msh2	0.00000	TRUE
RECQL	ATP-dependent DNA helicase Q1	0.00000	TRUE
LIG3	DNA ligase 3	0.00000	TRUE
SUB1	Activated RNA polymerase II transcriptional coactivator p15	0.00000	TRUE
PRKDC	DNA-dependent protein kinase catalytic subunit	0.00000	TRUE
TFAM	Transcription factor A, mitochondrial	0.00000	FALSE
C20orf72	Uncharacterized protein C20orf72	0.00000	FALSE
RPA3	Replication protein A 14 kDa subunit	0.00005	TRUE
SSBP1	Single-stranded DNA-binding protein, mitochondrial	0.00005	TRUE
RFC2	Replication factor C subunit 2	0.00015	TRUE
RPA2	Replication protein A 32 kDa subunit	0.00025	TRUE
POLG	DNA polymerase subunit gamma-1	0.00045	TRUE
RFC3	Replication factor C subunit 3	0.00050	TRUE
LYZ	Lysozyme C	0.00055	FALSE
DDB2	DNA damage-binding protein 2	0.00055	TRUE
RPA1	Replication protein A 70 kDa DNA-binding subunit	0.00060	TRUE
HIST1H1C	Histone H1.2	0.00075	TRUE
DDB1	DNA damage-binding protein 1	0.00090	TRUE
POLR3A	DNA-directed RNA polymerase III subunit RPC1	0.00125	TRUE
POLR3C	DNA-directed RNA polymerase III subunit RPC3	0.00175	TRUE
HMGB1	High mobility group protein B1	0.00190	TRUE
POLR1C	DNA-directed RNA polymerases I and III subunit RPAC1	0.00310	TRUE
POLR3B	DNA-directed RNA polymerase III subunit RPC2	0.00480	TRUE
H1F0	Histone H1.0	0.00490	TRUE
POLR2H	DNA-directed RNA polymerases I, II, and III subunit RPABC3	0.00545	FALSE
HMGB3	High mobility group protein B3	0.00551	TRUE
POLR3F	DNA-directed RNA polymerase III subunit RPC6	0.00584	TRUE
BANF1	Barrier-to-autointegration factor	0.00743	TRUE
SMARCAL1	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1	0.00944	TRUE

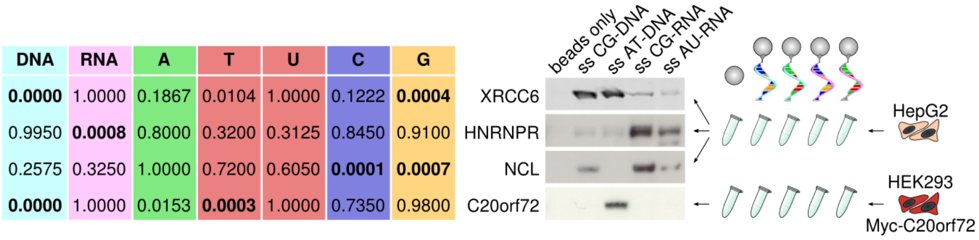
and C/G according to the experimental analysis. This allows testing the specificities by a set of four baits. Furthermore antibodies for the candidates should be available which reduces experimental effort. The number of A/T specific proteins in the results is quite limited and therefore a protein for which no antibody is available had to be selected. The list of selected candidates was:

- XRCC6 is specific for DNA and baits that contain Guanine, according to the statistical analysis
- HNRNPR is specific for RNA baits
- NCL is specific for baits that contain Cytosine or Guanine
- C20orf72 is specific for baits that contain DNA, Thymine and also the P value for Adenine containing baits is close to the significance level

Antibodies for XRCC6, HNRNPR and NCL were available. In contrast to C20orf72, for which no antibody was available. Therefore this protein was cloned and a tag was attached during the cloning procedure. The tag allows to quantify the protein by western blotting by using an antibody against the tag. The tagged form of C20orf72 was cloned into HEK293 cells, as these cells are easier accessible to genetic manipulation.

Selective baits to probe for statistically identified specificities were chosen. These baits are the single stranded DNA and RNA baits that are composed of either CG or AT(AU in the case of RNA). Using these baits, two experiments were performed. In the first experiment affinity purifications from HepG2 cell lysates against the four chosen baits and a streptavidin control were performed. Here the HepG2 cell line was chosen, because this cell line manifested the highest abundance of the three candidate proteins in the original screen and this should support detectability by antibodies. The samples generated here were split into three aliquots and probed against the antibodies of XRCC6, HNRNPR and NCL. The second experiment was performed on HEK293 cells that were transfected with tagged C20orf72 and the same set of nucleic acid baits. These samples were afterwards blotted with an antibody against the tag. This quantitative analysis by western blotting (see Figure 5.30) confirms the predicted specificities.

The specificity for methylated cytosine (mCG) is slightly more complicated to analyse as the baits containing methylated cytosine only contain cytosine and guanine and therefore form a subset of CG containing baits. This leads to a correlation between the P value of methylated cytosine (p_{mCG}) and the P values of cytosine (p_C) and guanine (p_G) specificity. As the specificity



DNA	RNA	A	T	U	C	G	
0.0000	1.0000	0.1867	0.0104	1.0000	0.1222	**0.0004**	XRCC6
0.9950	**0.0008**	0.8000	0.3200	0.3125	0.8450	0.9100	HNRNPR
0.2575	0.3250	1.0000	0.7200	0.6050	**0.0001**	**0.0007**	NCL
0.0000	1.0000	0.0153	**0.0003**	1.0000	0.7350	0.9800	C20orf72

Figure 5.30 – Four examples of specific binding affinities of NABPs represented with P-values in the statistical analysis and the Western blots in the experimental validation. (C20orf72 was purified with an Myc tag in HEK293 cells instead of a specific antibody in HepG2 cells).

for C and G is computed on all baits that contain C or G - also for example AG baits belong to the group of guanine containing baits and therefore contribute to p_G - and the correlation is manifested in both these P values (p_C and p_G). Therefore an additional statistical analysis was performed that measures the specificity of proteins for CG baits in which cytosine is not methylated (p_{CG}). This additional measure should allow to decompose the correlation between p_C, p_G and p_{mCG} and identify proteins with specificity for methylated cytosine. Table 5.6 lists the top ten proteins with the lowest P value for methylated cytosine (p_{mCG}).

The correlation between p_C, p_G and p_{mCG} is clearly visible in a subset of the proteins as proteins that have a significant P value for mCG also are more likely to be significant for C and/or G than expected by chance. These proteins can be considered to preferentially bind to any C/G containing bait, independent of cytosine methylation. Computing an additional P value that tests the specificity especially for unmethylated CG baits (p_{CG}), which is a complementary subset to the group of methylated (CG) baits within the group of all C/G containing baits, allows to separate proteins with general specificity for baits containing cytosine and/or guanine from the more specific proteins that bind to the CG baits that contained methylated cytosine. Consequently proteins that are specific for cytosine methylated DNA should have an insignificant P value for CG containing baits (p_{CG}) besides a significant P value for methylated baits (p_{mCG}) of course. The ten proteins with most significant p_{mCG} in Table 5.6 show exactly these two groups. Proteins of the RPA family, which have a small P value for methylated and unmethylated CG baits, and can therefore be considered as binding to all CG containing baits. Whereas UHRF1 only has a significant P value for methylated baits (p_{mCG}). UHRF1 was reported to specifically bind to methylated DNA [106]

Table 5.6 – Top ten most significant methylcytosine specific proteins (mCG). As methylcytosine baits form a subset of C/G containing baits, therefore a correlation with p_C and p_G is observed. To resolve this also a P value for specificity for unmethylated CG oligos was computed (p_{CG}). Proteins with low p_{mCG} and high p_{CG} are more likely to be methylcytosine specific.

rank	name	p_{mCG}	p_{CG}	p_C	p_G
1	RPA2	0.0000	0.0002	0.1375	0.1020
2	RPA1	0.0000	0.0006	0.1256	0.0900
3	**UHRF1**	0.0000	**0.5300**	0.0110	0.0113
4	CGGBP1	0.0000	0.0337	0.0043	0.0030
5	RPA3	0.0001	0.0019	0.2420	0.1090
6	PRKDC	0.0006	0.0743	0.0043	0.0021
7	CNBP	0.0007	0.1040	0.0056	0.0000
8	RECQL	0.0015	0.0142	0.0005	0.0011
9	**YB-1**	0.0017	**0.1800**	0.0867	0.0010
10	TFAM	0.0022	0.0220	0.0144	0.0005

earlier. YB-1 shows similar behaviour.

To test the specificity of YB-1 experimentally UHRF1, YB-1 and AIM2 were cloned into HEK293 cells. As the specificity of UHRF1 is already known it serves as a positive control. Furthermore AIM2, a nucleic acid binding protein with no reported specificity, was included in the experiment as an unspecific control. All proteins were cloned into HEK293 cells in a tagged form. To probe for methylation specificity AT, CG and methylated CG DNA were used as baits. The outcome of the experiments in Figure 5.31 shows that UHRF1 is specific for methylated DNA as described in the literature and confirmed in the statistical analysis. The experiment confirms, that YB-1 is specific to methylated DNA too, which was not known previously. This finding is especially interesting as YB-1 - which usually resides in the cytoplasm [22] - was shown to translocalise to the nucleus in various tumors [21, 5]. This might indicate a role of YB-1 in altering transcriptional regulation during cancer, especially as methylation is known to be altered in tumor cells [69].

By combining the P-values of specific proteins for the various binding specificities in vectors for each protein, we could cluster the NABPs and observed several protein families sharing specificities (see Figure 5.32). Here P values computed in the specificity analysis were log transformed. Due to

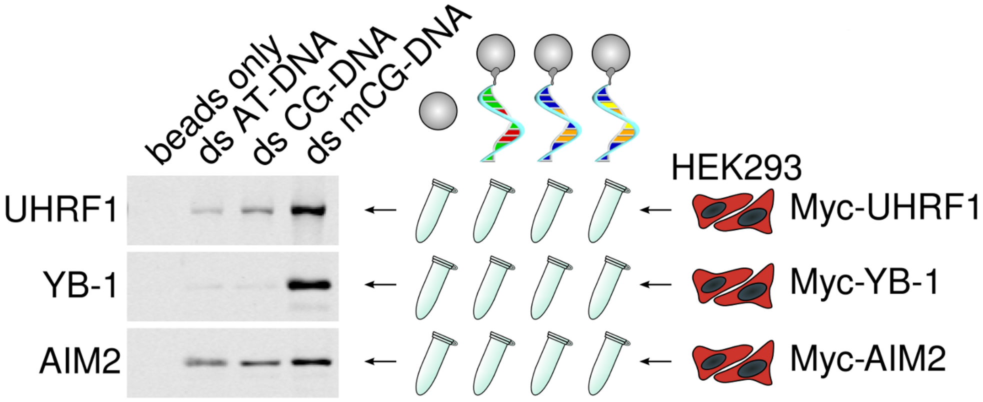


Figure 5.31 – Validation of methylcytosine specificities. UHRF1 (known) and YB-1 (unknown) are methylcytosine specific, whereas AIM2 serves as an unspecific control.

the limited number of permutations in the permutation test, also P values of zero occur in the result of the specificity analysis. Therefore 0.0005 was added to each P value before log transformation. Next, Euclidean distances were computed between these vectors of log transformed values and a dendrogram was generated by neighbour-joining [102] implemented in `phylip` [99].

5.4 Prediction of NABPs from Amino Acid Sequence

As already mentioned, biological annotation is not perfectly reliable and incomplete as it is subject of ongoing research. An additional value of our data set is, that it provides a body of experimentally detected NABPs. Therefore it is interesting to see how well this data set is suited to serve as a starting point for prediction of NABPs by computational methods. Of course numerous attempts to predict protein function have been undertaken. Analysis of protein domains, as already mentioned earlier, has established as a standard for inferring protein function(s). Here a protein sequence is compared against all known protein domain models (usually described in the form of HMMs). Identification of a protein domain with known function on the protein sequence allows to get an idea of the potential function of the query protein.

Here a completely different approach will be tested. As nucleic acid polymers confer a negative charge on the backbone, the charge distribution of

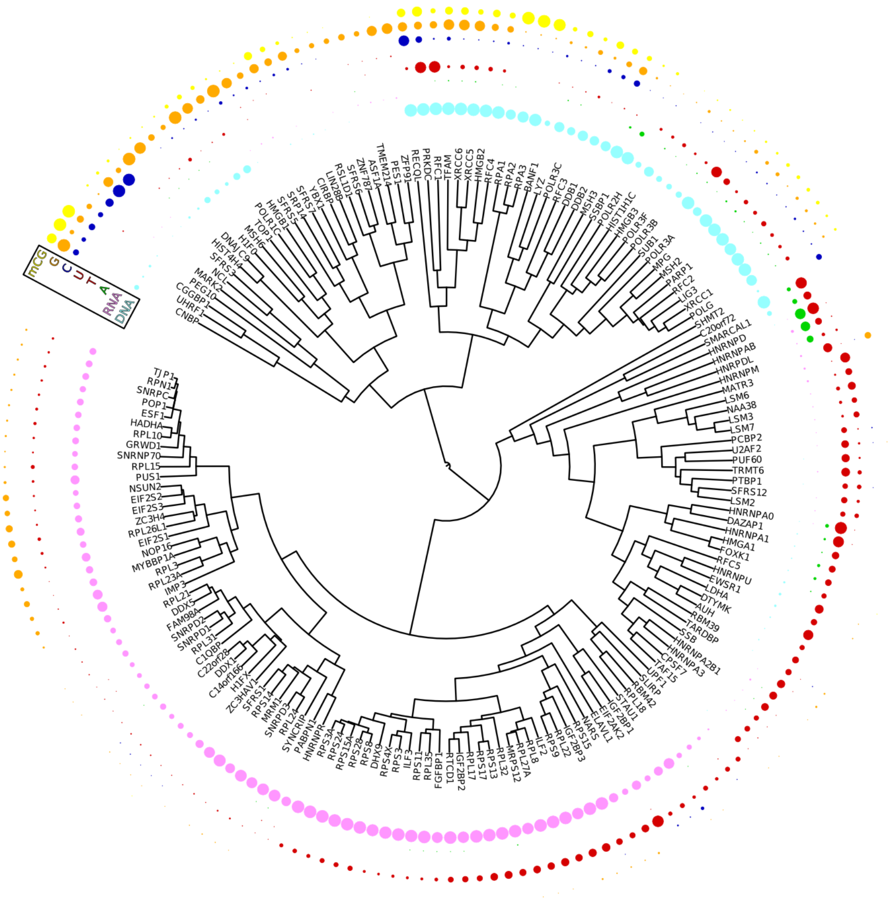


Figure 5.32 – The 174 proteins that were assigned a binding specificity for at least one nucleic acid class have been clustered to reflect similarities in specificities. Most protein families show similar specificities. H1FX was found to be RNA specific in contrast to the family members H1F0 and HIST1H1C that are DNA-specific.

a query protein is analysed to infer nucleic acid binding. Protein-substrate binding sites usually do not span the whole protein sequence, but rather a small proportion of the protein is directly involved in a specific interaction. The same is true for interactions between proteins and nucleic acids. Therefore local charge patterns of a protein are of great interest to infer nucleic acid binding. Due to three dimensional folding of a protein chain, the tertiary structure of a protein plays a crucial role to determine the charge of local binding sites. On the other hand protein structures are only available for a small fraction of proteins. Consequently to guarantee general applicability to all proteins, the prediction needs to be based on amino acid sequence alone. This is especially advantageous for less well studied proteins, where the likelihood of an available structure is further decreased. Due to this design decision a method based on amino acid sequence will be applicable to all proteins.

5.4.1 Computation of charge profiles

Based on the aforementioned considerations the following approach was implemented. Several protein properties are computed based on amino acid side chain properties and then used as an input for a machine learning method to predict nucleic acid binding. The amino acid sequence of the protein is converted into a numeric vector. Here amino acids with a positively charged side chain (i.e. Lysine, Histidine and Arginine) are represented by +1. Conversely negatively charged amino acids (Aspartic Acid and Glutamic Acid) are represented by -1. The remaining - uncharged - amino acids are represented as zeros within the vector. This vector is used as a first charge representation along the amino acid sequences (see the black bar graph in Figure 5.33 for an example). Based on this vector the fraction of positively and negatively charged amino acids is computed. The sum of these two numbers is equivalent to the mean of the initial signed vector elements (-1/0/+1) and is used to represent the overall charge of the query protein. As mentioned above, interaction surfaces of proteins are relatively small and therefore local charged regions play an important role in protein-nucleic acid interactions. To account for these local charged regions a moving average is computed along the vector. This moving average can identify regions in which a high number of identically charged amino acids do occur which indicates a highly charged local region. Subsequently the fraction of positions where this moving average lies above an arbitrary threshold of +0.3 and +0.5 as well as below -0.3 and -0.5 is calculated and used as an input parameter for the machine learning method. The rational behind using the fraction of positions that fulfil a certain condition (charge/moving average above a threshold) is, that this automatically provides a simple normalisation to account for protein sequence length, as larger proteins are more likely to have a higher number of charged amino acids. Consequently all the computed parameters lie in a range between zero and one, except the average charge, which can lie between plus and minus one. Finally also the minimum and maximum of the moving average within the given query proteins is calculated and used as an input parameter. The moving average is computed three times using different window sizes of 5, 10 and 20, which lies in the range of typical protein-nucleic acid interaction interfaces. All the described values are subsequently used as input parameters for a SVM classifier which tries to predict nucleic acid binding based on the calculated charge parameters.

Figure 5.33 shows the charge profile and the derived charge parameters for two exemplaric proteins. Beta-actin (ACTB) is a well known and very abundant protein which does not bind to DNA. In contrast to Histone H1 which is well known to bind to DNA in the nucleus and is also very abundant. Obviously H1F0 contains many positively charged amino acids, whereas ACTB is relatively neutral.

5.4.2 Model training and performance

The 21 aforementioned parameters were used as input features for a subsequent classification by a SVM. The training set was composed of nucleic acid binding proteins, detected in the previous analysis and proteins that do not bind to nucleic acid. For the set of NABPs in the training set eleven proteins that were classified as secondary interactors were removed from the 113 DNA and RNA-specific proteins detected in the previous specificity analysis. The resulting 102 proteins are incorporated into the machine learning training set as examples for proteins which bind to nucleic acid. To also incorporate proteins that do not bind nucleic acid as counter examples in the training set, proteins that were detected in the nucleic acid experiments were removed from proteins detected in abundance proteomics experiments [14] - abundant proteome of the same cell lines measured by MS without prior enrichment by a nucleic acid column. These experiments were performed in the same cell lines and already used earlier as a negative control (Figure 5.4). To obtain a clean negative control data set also proteins that were annotated as nucleic acid binding in Gene Ontology were removed from this set. Proteins that are annotated as nucleotide binding (GO:0000166) were removed as well, because these are highly similar to proteins annotated as nucleic acid binding. To eliminate some of the remaining misannotated proteins also proteins that contain “ribosome”, “zinc finger”, “nuclease”, “DNA” or “RNA” in their name were removed. Of course this could also remove proteins that are for

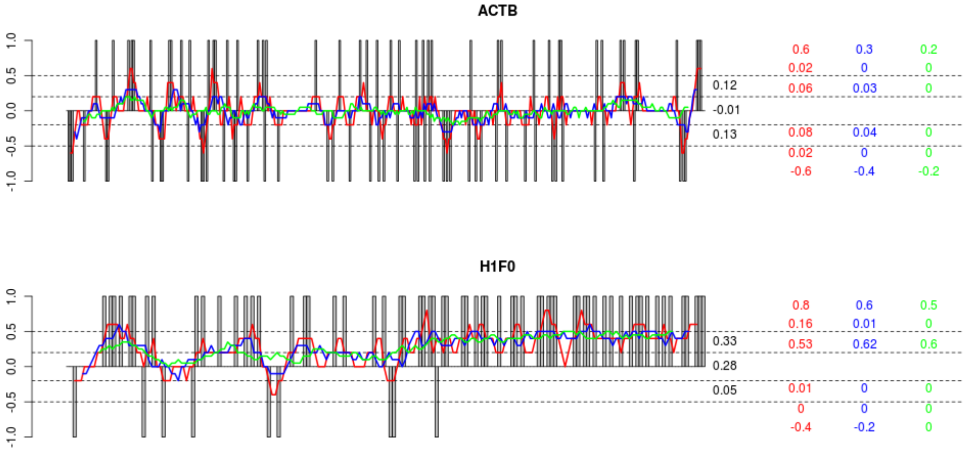


Figure 5.33 – Graphical representation of the computed charge parameters. Charged amino acids are indicated a grey bars (positive +1, negative -1). The fraction of charged amino acids is given on the right in black (positively at the top, negatively at the bottom and the overall average in the center. Moving averages for different window sizes (w) are also indicated in the graph ($w = 5$ in red, $w = 10$ in blue and $w = 20$ in green). The thresholds for the moving average are indicated as dashed lines at +0.5, +0.3, -0.3 and -0.5. The fraction of positions outside this threshold is also given on the right in the corresponding color outside the respective threshold as well as the minimum and maximum of the moving average in the very top (max) and in the very bottom (min). Histone H1 (H1F0) is DNA binding, whereas Beta-actin (ACTB) is not.

example called “non DNA binding protein 1”, but as negative nomenclature is not very common in biology this filtering will generate a cleaner training set. Furthermore loosing some non-NABPs in the set is not as dramatic as if the negative set contains proteins that bind to nucleic acid (false negatives). The filtering procedure for non-NABPs in the training set can also be seen in Table 5.7

Table 5.7 – Workflow and absolute protein numbers for the compilation of a non-NABP training set for machine learning.

description	# proteins remaining
abundance proteomics experiments [14]	3054
- proteins detected in nucleic acid experiments	2609
- proteins annotated as nucleic acid binding	2255
- proteins annotated as nucleotide binding	1840
- proteins that contain ribosome, zinc finger, nuclease, DNA or RNA in their description	1771

5.4.3 Model learning

The 21 charge parameters for these 1873 proteins were computed. This training data set was used to build a SVM model in R. Therefore the `svm` [20] function of the R package `e1071` was used. This function provides an implementation of a ”C-classification” SVM model, which is a soft margin classifier with user definable cost coefficient.

Table 5.8 – Performance of the first machine learning classification on the training set using all 21 charge features.

		prediction	
		NABP	non NABP
classification	NABP	19	83
	non NABP	0	1771

Applying the SVM on the training set with a cost coefficient of one ($C = 1$) shows that 19 of the 102 NABPs in the training set can be classified correctly, whereas 83 proteins were incorrectly classified as non-NABPs (false negatives). This results in a sensitivity of 18.6% ($\frac{19}{102}$) at a specificity of 100% ($= \frac{1771}{1771}$) (see confusion matrix in Table 5.8). Here no false positives were predicted, which is important as these misclassifications should be avoided for correct novel predictions. A classification with high specificity of predictions is important and preferred, also if this on the other hand decreases sensitivity. Overall the result is equivalent to a prediction error of 4.4% ($\frac{83}{102+1771}$). Of course evaluating the classification performance directly on the training set is misleading as here overfitting is not detectable. Therefore a ten fold cross validation which is also provided by the `svm`-function was applied to detect overfitting. Cross validation resulted in an accuracy of 95.4% which is very close to the calculated error of 4.4% and indicates that in this simple model overfitting is not yet a problem.

Next, model performance is evaluated on models with different cost coefficient parameters to determine if varying the cost coefficient and therefore altering the magnitude of the soft margin penalty (slack variables) can improve the performance of the classifier (see Table 5.9).

Table 5.9 – SVM performance for different cost coefficients. False Negatives (FN) and False Positives (FP) were measured on the training set without cross validation, whereas the accuracy was determined using ten fold cross validation implemented in the `svm`-function.

cost coefficient	time [sec]	FN (training)	FP (training)	accuracy [%] (cross validation)
0.01	0.4	86	0	95.3
0.1	0.6	83	0	95.4
1	1.8	83	0	95.4
10	11.1	82	0	95.4
40	146.3	83	0	95.4

Interestingly the performance is not strongly affected when altering the cost coefficient. A very low value of the cost coefficient led to a significant increase in false negatives. Increasing the cost coefficient above 40 was not possible, as algorithm run time dramatically increased and possibly would not have converged. A drawback of the internal implementation of the cross validation supplied by the `svm`-function is, that it only provides the accuracy of the classification, whereas the classifier sensitivity, specificity or confusion matrix is not provided. Therefore cross validation was implemented manually to determine the classification specificity, which is important in our setting.

The cumulative confusion matrix of the ten test set partitions can be seen in Table 5.10. As in Table 5.8 the cost coefficient was set to one. Annoyingly one false positive protein was observed, which reduces specificity to 99.9% ($\frac{1770}{1771}$). This does not seem very dramatic, but looking at the false discovery rate (FDR) this means an increase to 5% ($\frac{1}{1+18}$) compared to an FDR of 0 earlier when evaluating on the training set. This is especially annoying as this means that actually roughly 5% of the predicted NABPs might be incorrect, although the number of positively predicted proteins is quite low in the data set, which obviously leads to a high confidence interval of the calculated FDR. The accuracy computed on this confusion matrix is similar to the accuracy provided by the cross validation implemented in the `svm`-function (95.5%). Next, again the influence of the cost coefficient was determined (see Table 5.11). Interestingly, here reducing the cost coefficient led to a reduction of false positives, whereas contrary a higher cost coefficient increased the number of true positives.

Table 5.10 – Performance evaluated by ten fold cross validation. (Linear kernel and cost coefficient=1)

		prediction	
		NABP	non NABP
classification	NABP	18	84
	non NABP	1	1770

Table 5.11 – SVM performance for different cost coefficients and using a linear kernel. Here False Negatives (FN), False Positives (FP) and True Positives were measured by ten fold cross validation.

cost coefficient	FN (cross validation)	FP (cross validation)	TP (cross validation)
0.01	86	1	16
0.1	85	1	17
1	84	1	18
10	83	2	19
40	82	2	20

Next, a polynomial kernel of degree 2 is evaluated, to determine if a more complicated model can help to better model the data set (see results in Table 5.12).

Unfortunately this does not improve classification at all. Although the number of true positives is generally increased, but this comes at a cost of a drastic increase in the number of false positives. As specificity is the main goal of the classification, polynomial kernels are not appropriate to classify the data set at hand. Again a correlation between cost coefficient, number of true positives and number of false positives could be observed, as had been observed earlier in the cross validation of the linear kernel. Probably the poor performance of polynomial kernels is due to overfitting, which indicates that polynomial kernels are too powerful and can not achieve generalisation.

Table 5.12 – SVM performance for different cost coefficients when using a polynomial kernel of degree two. False Negatives (FN), False Positives (FP) and True Positives were measured by ten fold cross validation.

cost coefficient	FN (cross validation)	FP (cross validation)	TP (cross validation)
0.01	85	3	17
0.1	79	4	23
1	78	5	24
10	82	10	20
40	76	16	26

5.4.4 Parameter selection

Another method to improve classification performance is to reduce the set of features in the training set. Here “backward elimination” - a technique from regression [55] - was applied to evaluate if this can improve the classification. Therefore each of the 21 features in the training set was eliminated individually and the classifiers performance was determined again using the ten fold cross validation with a linear kernel and a cost coefficient of one.

All models led to exactly the same performance (TP=18, FP=2, FN=84) which is equivalent to an accuracy of 95.4% at a sensitivity of 17.6% and a specificity of 99.89% (FDR=10%). Because feature elimination of single features in the training set could not improve the classifiers performance, the approach was not pursued further.

5.4.5 Application to proteomes of model organisms

The SVM classification offers the possibility to predict if a protein binds to nucleic acid, solely based on the amino acid sequence of any given protein. This allows to apply the classifier to complete proteomes, i.e. the amino acid sequences of all the proteins known in a given organism. In this section this capability is examined on multiple model organisms.

Applying the classifier to model organisms produces a substantial number of additional predictions for NABPs, not only in the human proteome but also in mouse, rat, fly, E. coli, and yeast proteomes (see Table 5.13). Many predicted proteins that are not annotated as nucleic acid binding in GO turned out to be yet uncharacterised proteins and, as judged on the proportion of known NABPs, the performance is not limited by evolutionary distance.

Surprisingly, also ribosomal proteins were predicted which so far lack the

Table 5.13 – Prediction of NABPs in several model organisms by the SVM classifier. Proteins=size of the proteome in SwissProt, known=proteins annotated as nucleic acid binding (GO:0003676), pred=number of proteins predicted to be nucleic acid binding by the SVM, new=proteins not yet annotated as nucleic acid binding [a], ribo=ribosomal proteins [b], ZNF=zinc finger proteins [a,b], unch.=uncharacterised proteins [a,b]
[a]=subset of the new predicted NABPs (column **new**)
[b]=inferred by description

Species	Proteins	known	**pred.**	**new**	ribo.	ZNF	unch.
Homo sap.	20256	3209	257	96	23	1	28
Mus musc.	16401	2346	160	80	44	1	4
Rattus nor.	7660	839	108	59	40	0	1
Drosophila m.	3131	653	53	40	29	0	1
S. cerevisiae	6620	1100	122	104	37	0	54
E. coli	4430	733	31	20	9	0	7

corresponding annotation in Gene Ontology. As ribosomal proteins usually are well studied we wondered whether the annotation or the prediction is correct. Considering the latest ribosome structures of E. coli (for which in contrast to other model organisms many experimental structures are available), we found that all the predicted and not annotated proteins that were present in these structures (rpsU, rpsN, rpmI, rpmJ, rpmB and rpmD) were within a distance of 3 Angstrom to ribosomal RNA, comparable to the other ribosomal proteins that are annotated as nucleic acid binding (see Tables 5.14 & 5.15 for distances between individual proteins and nucleic aids in the structure). This indicates that the identified proteins also are in close contact to ribosomal RNA and thereby supports the predictions of the classifier.

Table 5.14 – Distance between the proteins (rows) of the small ribosomal subunit in E. coli and nucleic acids (columns) in PDB structure 3R8N. Proteins in bold font are proteins predicted to be nucleic acid binding by the SVM and not annotated as such.

	name (chain)		
name (chain)	16S rRNA (A)	tRNA (V)	mRNA (X)
rpsU (U)	2,35	27,38	2,83
rpsN (N)	2,38	27,31	33,50
rpsJ (J)	2,56	18,55	25,14
rpsK (K)	2,33	18,60	3,36
rpsL (L)	2,32	24,73	10,60
rpsM (M)	2,34	4,69	29,27
rpsO (O)	2,39	55,98	34,62
rpsP (P)	2,33	85,11	71,61
rpsQ (Q)	2,48	61,80	50,89
rpsR (R)	2,29	48,66	8,51
rpsS (S)	2,38	17,67	37,68
rpsB (B)	2,56	46,60	21,46
rpsT (T)	2,26	95,04	82,13
rpsC (C)	2,49	22,54	18,53
rpsD (D)	2,29	46,71	38,19
rpsE (E)	2,32	16,01	14,90
rpsF (F)	2,52	56,97	27,08
rpsG (G)	2,30	14,16	2,99
rpsH (H)	2,42	62,19	44,03
rpsI (I)	2,36	2,44	13,68

Table 5.15 – Distance between the proteins (rows) of the large ribosomal subunit in E. coli and nucleic acids (columns) in PDB structure 3R8S. Proteins in bold font are proteins predicted to be nucleic acid binding by the SVM and not annotated as such.

	name (chain)	
	23S rRNA (A)	5S rRNA (B)
name (chain)		
rpmI (3)	2,40	32,32
rpmJ (4)	2,54	19,75
rpmB (X)	2,27	59,94
rpmD (Z)	2,51	3,33
rplM (J)	2,38	24,61
rplN (K)	2,41	50,45
rplO (L)	2,17	34,69
rplP (M)	2,48	3,06
rplQ (N)	2,41	68,86
rplR (O)	2,29	2,30
rplS (P)	2,34	81,65
rplT (Q)	2,19	23,46
rplU (R)	2,68	25,79
rplV (S)	2,47	60,95
rplW (T)	2,51	81,35
rplX (U)	2,41	94,49
rplY (V)	2,73	2,42
rpmA (W)	2,52	2,98
rpmC (Y)	2,70	113,04
rpmF (0)	2,49	43,64
rpmG (1)	2,45	15,89
rpmH (2)	2,32	73,15
rplB (C)	2,16	54,98
rplC (D)	2,30	31,88
rplD (E)	2,50	49,25
rplE (F)	2,24	2,42
rplF (G)	2,49	27,05
rplI (H)	3,01	86,89
rplK (I)	2,29	48,34

Chapter 6

Conclusion and Discussion

This work has demonstrated the viability of state-of-the-art MS technology to screen protein-nucleic acid interactions. Although only a limited number of nucleic acid baits were screened here, a substantial fraction of known human NABPs were recovered in this study. In addition, a significant number of proteins not previously described as nucleic acid binding were detected. These findings are consistent with recent studies which indicate that only a subset of all NABPs are currently known [17], even in well studied organisms.

Among the set of novel proteins a nuclease domain in the uncharacterised protein C20orf72 was identified. Furthermore, enrichment analysis allowed for nucleic acid binding function to be proposed for several known protein domains.

Based on the experimental data set, a classification of NABPs in groups specific to nucleic acid sub-categories was established. To generate this classification, parametric- and non-parametric statistical methods were evaluated. Although parametric methods are known to provide some robustness to violation of the underlying assumptions, non-parametric methods turned out to achieve better performance. This allowed the specificities for 174 proteins in the experimental data set to be proposed. Some of these specificities were validated experimentally, or by comparison to publicly available annotation. This approach can annotate previously uncharacterised proteins, such as C20orf72 as (A)T-DNA specific, in addition to providing novel insight into well studied proteins like YB-1, for which methylcytosine specific binding was revealed. This demonstrates the value of system wide studies to generate unbiased hypothesis as a starting point for subsequent focused investigation.

Finally, a machine learning predictor was implemented using the determined classifications. Here, a novel approach was evaluated in which predictions are based on charge properties of the query protein and requires only the amino acid sequence as an input. This makes the method widely applicable. Incorporating extra levels of data, for example protein localisation, should improve sensitivity of this method, but on the other hand would also limit general applicability, as this information is only available for a subset of all proteins. We note, that analysis of proteins based on local charges is not applicable to all questions arising in sequence analysis, however, if does provide a powerful simplification when investigating NABPs. Furthermore, charge profiles can also be computed very efficiently. Application of this classifier to the human proteome identified 257 NABPs, 96 of which were not previously annotated as nucleic acid binding. The number of false positives within these novel predictions should be moderate, as a false discovery rate of 5% was estimated by cross validation on a training data set.

Interestingly, these novel predictions contain a high number of uncharacterised proteins. Application of the predictor in other species did not impair performance which indicates the general applicability of the approach beyond human proteins. As a next step, this method could be applied to protein fragments, potentially allowing the localisation of nucleic acid binding sites within NABPs.

The experimental approach could be easily extended in multiple directions. Besides further increasing the number of nucleic acid baits in system wide studies, also the continuously increasing sensitivity of MS technology will provide even larger data sets. Involving quantitative MS methods such as SILAC [89] or iTRAQ [101] should also improve the sensitivity of the method, as these methods allow the detection of less prominent alterations in protein abundance compared to semi-quantitative methods used in this study.

Besides system wide studies, this technology could also be used to address focused questions. For example, investigation of sequence alterations in highly homologous nucleic acid sequences upon NABP binding - e.g. in the context of sequence specific transcription factor binding. Also, studying protein complexes interacting with nucleic acid can give interesting insights as demonstrated for example in the studies of Jean-Philippe Lambert [71, 70].

Overall, the approach presented here has the potential to be a complementary method for studying the interactions between proteins and nucleic acids.

Bibliography

[1] Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic acids research, 40(Database issue):D71–5, January 2012.

[2] A. Aizerman, E. M. Braverman, and L. I. Rozoner. Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control, 25:821–837, 1964.

[3] S F Altschul, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17):3389–402, September 1997.

[4] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, Midori A Harris, David P Hill, L Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C Matese, Joel E Richardson, M Ringwald, Gerald M Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics, 25(1):25–9, May 2000.

[5] R C Bargou, K Jürchott, C Wagener, S Bergmann, S Metzner, K Bommert, M Y Mapara, K J Winzer, M Dietel, B Dörken, and H D Royer. Nuclear localization and increased levels of transcription factor YB-1 in primary human breast cancers are associated with intrinsic MDR1 gene expression. Nature medicine, 3(4):447–50, April 1997.

[6] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. Cell, 129(4):823–37, May 2007.

[7] Tim Beissbarth, Lavinia Hyde, Gordon K Smyth, Chris Job, Wee-Ming Boon, Seong-Seng Tan, Hamish S Scott, and Terence P Speed. Statis-

tical modeling of sequencing errors in SAGE libraries. Bioinformatics (Oxford, England), 20 Suppl 1:i31–9, August 2004.

[8] Brigitte Boeckmann, Marie-Claude Blatter, Livia Famiglietti, Ursula Hinz, Lydie Lane, Bernd Roechert, and Amos Bairoch. Protein variety and functional diversity: Swiss-Prot annotation in its biological context. Comptes rendus biologies, 328(10-11):882–99, 2005.

[9] J. Bortz, G.A. Lienert, and K. Boehnke. Verteilungsfreie Methoden in der Biostatistik. Springer-Lehrbuch. Springer, 2010.

[10] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, pages 144–152. ACM Press, 1992.

[11] Bobby-Joe Breitkreutz, Chris Stark, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, Michael Livstone, Rose Oughtred, Daniel H Lackner, Jürg Bähler, Valerie Wood, Kara Dolinski, and Mike Tyers. The BioGRID Interaction Database: 2008 update. Nucleic acids research, 36(Database issue):D637–40, January 2008.

[12] M L Bulyk, X Huang, Y Choo, and G M Church. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. Proceedings of the National Academy of Sciences of the United States of America, 98(13):7158–63, June 2001.

[13] Tilmann Bürckstümmer, Christoph Baumann, Stephan Blüml, Evelyn Dixit, Gerhard Dürnberger, Hannah Jahn, Melanie Planyavsky, Martin Bilban, Jacques Colinge, Keiryn L Bennett, and Giulio Superti-Furga. An orthogonal proteomic-genomic screen identifies AIM2 as a cytoplasmic DNA sensor for the inflammasome. Nature immunology, 10(3):266–72, March 2009.

[14] Thomas R Burkard, Melanie Planyavsky, Ines Kaupe, Florian P Breitwieser, Tilmann Bürckstümmer, Keiryn L Bennett, Giulio Superti-Furga, and Jacques Colinge. Initial characterization of the human central proteome. BMC systems biology, 5:17, January 2011.

[15] Falk Butter, Marion Scheibe, Mario Mörl, and Matthias Mann. Unbiased RNA-protein interaction screen by quantitative proteomics. Proceedings of the National Academy of Sciences of the United States of America, 106(26):10626–31, June 2009.

[16] Seth Carbon, Amelia Ireland, Christopher J Mungall, ShengQiang Shu, Brad Marshall, and Suzanna Lewis. AmiGO: online access to ontology and annotation data. Bioinformatics (Oxford, England), 25(2):288–9, January 2009.

[17] Alfredo Castello, Bernd Fischer, Katrin Eichelbaum, Rastislav Horos, Benedikt M Beckmann, Claudia Strein, Norman E Davey, David T Humphreys, Thomas Preiss, Lars M Steinmetz, Jeroen Krijgsveld, and Matthias W Hentze. Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. Cell, 149(6):1393–406, June 2012.

[18] C. I. Castillo-Davis and D. L. Hartl. GeneMerge–post-genomic analysis, data mining, and hypothesis testing. Bioinformatics, 19(7):891–892, May 2003.

[19] Gianni Cesareni, Andrew Chatr-aryamontri, Luana Licata, and Arnaud Ceol. Searching the MINT database for protein interaction information. Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.], Chapter 8:Unit 8.5, June 2008.

[20] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[21] Manik Chatterjee, Christoph Rancso, Thorsten Stühmer, Niels Eckstein, Mindaugas Andrulis, Christian Gerecke, Heike Lorentz, Hans-Dieter Royer, and Ralf C Bargou. The Y-box binding protein YB-1 is associated with progressive disease and mediates survival and drug resistance in multiple myeloma. Blood, 111(7):3714–22, April 2008.

[22] C Y Chen, Roberto Gherzi, Jens S Andersen, Guido Gaietta, K Jürchott, H D Royer, Matthias Mann, and Michael Karin. Nucleolin and YB-1 are required for JNK-mediated interleukin-2 mRNA stabilization during T-cell activation. Genes & development, 14(10):1236–48, May 2000.

[23] An Chi, Curtis Huttenhower, Lewis Y Geer, Joshua J Coon, John E P Syka, Dina L Bai, Jeffrey Shabanowitz, Daniel J Burke, Olga G Troyanskaya, and Donald F Hunt. Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. Proceedings of the National Academy

of Sciences of the United States of America, 104(7):2193–8, February 2007.

[24] Jacques Colinge, Alexandre Masselot, Marc Giron, Thierry Dessingy, and Jérôme Magnin. OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics, 3(8):1454–63, August 2003.

[25] W. J. Conover. Rank Tests for One Sample, Two Samples, and \$k\$ samples Without the Assumption of a Continuous Distribution Function. The Annals of Statistics, 1(6):1105–1125, November 1973.

[26] James W. Cooley and John W. Tukey. An Algorithm for the Machine Calculation of Complex Fourier Series. Mathematics of Computation, 19(90):297, April 1965.

[27] Corinna Cortes and Vladimir Vapnik. Support-vector networks. In Machine Learning, pages 273–297, 1995.

[28] Ye Ding and Charles E Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. Nucleic acids research, 31(24):7280–301, December 2003.

[29] Chuong B Do, Daniel A Woods, and Serafim Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics (Oxford, England), 22(14):e90–8, July 2006.

[30] T Doerks, P Bork, E Kamberov, O Makarova, S Muecke, and B Margolis. L27, a novel heterodimerization domain in receptor targeting proteins Lin-2 and Lin-7. Trends in biochemical sciences, 25(7):317–8, July 2000.

[31] Tobias Doerks, Saskia Huber, Erich Buchner, and Peer Bork. BSD : a novel domain in transcription factors and synapse-associated proteins. Trends in biochemical sciences, 27(4):168–170, 2002.

[32] Louis du Plessis, Nives Skunca, and Christophe Dessimoz. The what, where, how and why of gene ontology–a primer for bioinformaticians. Briefings in bioinformatics, 12(6):723–35, November 2011.

[33] Olive Jean Dunn. Multiple comparisons among means. Journal of the American Statistical Association, 56(293):52–64, 1961.

[34] Gerhard Dürnberger, Tilmann Bürckstümmer, Kilian Huber, Roberto Giambruno, Tobias Doerks, Evren Karayel, Thomas R Burkard, Ines Kaupe, Andre Müller, Gerhard F Ecker, Hans Lohninger, Peer Bork, Keiryn L Bennett, Giulio Superti-Furga, and Jacques Colinge. A global experimental survey for human nucleic acid interacting proteins. in preparation, 2012.

[35] Meyer Dwass. Modified Randomization Tests for Nonparametric Hypotheses. The Annals of Mathematical Statistics, 28(1):181–187, March 1957.

[36] Sean R Eddy. A new generation of homology search tools based on probabilistic inference. Genome informatics. International Conference on Genome Informatics, 23(1):205–11, October 2009.

[37] Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC bioinformatics, 10:48, January 2009.

[38] Robert C Edgar. MUSCLE : a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics, 19:1–19, 2004.

[39] Robert C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research, 32(5):1792–7, January 2004.

[40] B. Efron. Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics, 7(1):1–26, January 1979.

[41] B. Efron and R.J. Tibshirani. An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability. Taylor & Francis, 1994.

[42] Mark R. Emmett and Richard M. Caprioli. Micro-electrospray mass spectrometry: Ultra-high-sensitivity analysis of peptides and proteins. Journal of the American Society for Mass Spectrometry, 5(7):605–613, July 1994.

[43] Jimmy K. Eng, Ashley L. McCormack, and John R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry, 5(11):976–989, November 1994.

[44] J B Fenn, M Mann, C K Meng, S F Wong, and C M Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. Science (New York, N.Y.), 246(4926):64–71, October 1989.

[45] Anne-Claude Gavin, Patrick Aloy, Paola Grandi, Roland Krause, Markus Boesche, Martina Marzioch, Christina Rau, Lars Juhl Jensen, Sonja Bastuck, Birgit Dümpelfeld, Angela Edelmann, Marie-Anne Heurtier, Verena Hoffman, Christian Hoefert, Karin Klein, Manuela Hudak, Anne-Marie Michon, Malgorzata Schelder, Markus Schirle, Marita Remor, Tatjana Rudi, Sean Hooper, Andreas Bauer, Tewis Bouwmeester, Georg Casari, Gerard Drewes, Gitte Neubauer, Jens M Rick, Bernhard Kuster, Peer Bork, Robert B Russell, and Giulio Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. Nature, 440(7084):631–6, March 2006.

[46] Anne-Claude Gavin, Markus Bösche, Roland Krause, Paola Grandi, Martina Marzioch, Andreas Bauer, Jörg Schultz, Jens M Rick, Anne-Marie Michon, Cristina-Maria Cruciat, Marita Remor, Christian Höfert, Malgorzata Schelder, Miro Brajenovic, Heinz Ruffner, Alejandro Merino, Karin Klein, Manuela Hudak, David Dickson, Tatjana Rudi, Volker Gnau, Angela Bauch, Sonja Bastuck, Bettina Huhse, Christina Leutwein, Marie-Anne Heurtier, Richard R Copley, Angela Edelmann, Erich Querfurth, Vladimir Rybin, Gerard Drewes, Manfred Raida, Tewis Bouwmeester, Peer Bork, Bertrand Seraphin, Bernhard Kuster, Gitte Neubauer, and Giulio Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature, 415(6868):141–7, January 2002.

[47] Gene Ontology Consortium. The Gene Ontology project in 2008. Nucleic acids research, 36(Database issue):D440–4, January 2008.

[48] D. S. Gilmour and J. T. Lis. Detecting protein-dna interactions in vivo: distribution of rna polymerase on specific bacterial genes. Proc Natl Acad Sci U S A, 81(14):4275–4279, Jul 1984.

[49] Noelle M Griffin, Jingyi Yu, Fred Long, Phil Oh, Sabrina Shore, Yan Li, Jim a Koziol, and Jan E Schnitzer. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. Nature biotechnology, 28(1):83–9, January 2010.

[50] Sam Griffiths-Jones, Alex Bateman, Mhairi Marshall, Ajay Khanna, and Sean R Eddy. Rfam: an RNA family database. Nucleic acids research, 31(1):439–41, January 2003.

[51] Andreas R Gruber, Ronny Lorenz, Stephan H Bernhart, Richard Neuböck, and Ivo L Hofacker. The Vienna RNA websuite. Nucleic acids research, 36(Web Server issue):W70–4, July 2008.

[52] John Haigh and W. J. Conover. Practical Nonparametric Statistics. Journal of the Royal Statistical Society. Series A (General), 144(3):370, 1981.

[53] Mark Hardman and Alexander A Makarov. Interfacing the orbitrap mass analyzer to an electrospray ion source. Analytical chemistry, 75(7):1699–705, April 2003.

[54] W J Henzel, T M Billeci, J T Stults, S C Wong, C Grimley, and C Watanabe. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. Proceedings of the National Academy of Sciences of the United States of America, 90(11):5011–5, June 1993.

[55] R. R. Hocking. The Analysis and Selection of Variables in Linear Regression. 32(1):1–49+, 1976.

[56] J. L. Hodges and E. L. Lehmann. The Efficiency of Some Nonparametric Competitors of the t-Test. The Annals of Mathematical Statistics, 27(2):324–335, June 1956.

[57] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. Monatshefte fr Chemie Chemical Monthly, 125(2):167–188, February 1994.

[58] Qizhi Hu, Robert J Noll, Hongyan Li, Alexander Makarov, Mark Hardman, and R Graham Cooks. The Orbitrap: a new mass spectrometer. Journal of mass spectrometry : JMS, 40(4):430–43, April 2005.

[59] Shaohui Hu, Zhi Xie, Akishi Onishi, Xueping Yu, Lizhi Jiang, Jimmy Lin, Hee-sool Rho, Crystal Woodard, Hong Wang, Jun-seop Jeong, Shunyou Long, Xiaofei He, Herschel Wade, Seth Blackshaw, Jiang Qian, and Heng Zhu. Profiling the human protein-DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling. Cell, 139(3):610–22, October 2009.

[60] V R Iyer, C E Horak, C S Scafe, D Botstein, M Snyder, and P O Brown. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature, 409(6819):533–8, January 2001.

[61] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-DNA interactions. Science (New York, N.Y.), 316(5830):1497–502, June 2007.

[62] M. Karas, D. Bachmann, U. Bahr, and F. Hillenkamp. Matrix-assisted ultraviolet laser desorption of non-volatile compounds. International Journal of Mass Spectrometry and Ion Processes, 78:53–68, September 1987.

[63] Michael. Karas, Doris. Bachmann, and Franz. Hillenkamp. Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. Analytical Chemistry, 57(14):2935–2939, December 1985.

[64] Evren Karayel, Tilmann Bürckstümmer, Martin Bilban, Gerhard Dürnberger, Stefan Weitzer, Javier Martinez, and Giulio Superti-Furga. The TLR-independent DNA recognition pathway in murine macrophages: Ligand features and molecular signature. European journal of immunology, 39(7):1929–36, July 2009.

[65] Samuel Kerrien, Bruno Aranda, Lionel Breuza, Alan Bridge, Fiona Broackes-Carter, Carol Chen, Margaret Duesbury, Marine Dumousseau, Marc Feuermann, Ursula Hinz, Christine Jandrasits, Rafael C Jimenez, Jyoti Khadake, Usha Mahadevan, Patrick Masson, Ivo Pedruzzi, Eric Pfeiffenberger, Pablo Porras, Arathi Raghunath, Bernd Roechert, Sandra Orchard, and Henning Hermjakob. The IntAct molecular interaction database in 2012. Nucleic acids research, 40(Database issue):D841–6, January 2012.

[66] T S Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, Lavanya Balakrishnan, Arivusudar Marimuthu, Sutopa Banerjee, Devi S Somanathan, Aimy Sebastian, Sandhya Rani, Somak Ray, C J Harrys Kishore, Sashi Kanth, Mukhtar Ahmed, Manoj K Kashyap, Riaz Mohmood, Y L Ramachandra, V Krishna, B Abdul Rahiman, Sujatha Mohan, Prathibha Ranganathan, Subhashri Ramabadran, Raghothama Chaerkady, and Akhilesh Pandey. Human Protein Reference Database–2009 update. Nucleic acids research, 37(Database issue):D767–72, January 2009.

[67] Ulrich HG Kressel. Advances in kernel methods. chapter Pairwise classification and support vector machines, pages 255–268. MIT Press, Cambridge, MA, USA, 1999.

[68] M. H. Kuo and C. D. Allis. In vivo cross-linking and immunoprecipitation for studying dynamic protein:dna associations in a chromatin environment. Methods, 19(3):425–433, Nov 1999.

[69] P W Laird and R Jaenisch. DNA methylation and cancer. Human molecular genetics, 3 Spec No:1487–95, January 1994.

[70] Jean-Philippe Lambert, Jeffrey Fillingham, Mojgan Siahbazi, Jack Greenblatt, Kristin Baetz, and Daniel Figeys. Defining the budding yeast chromatin-associated interactome. Molecular Systems Biology, 6(448):1–16, December 2010.

[71] Jean-Philippe Lambert, Leslie Mitchell, Adam Rudner, Kristin Baetz, and Daniel Figeys. A novel proteomics approach for the discovery of chromatin-associated protein networks. Molecular & cellular proteomics : MCP, 8(4):870–82, April 2009.

[72] E.L. Lehmann. Elements of Large-Sample Theory. Springer Texts in Statistics. Springer, 1998.

[73] Ivica Letunic, Tobias Doerks, and Peer Bork. SMART 7: recent updates to the protein domain annotation resource. Nucleic acids research, 40(Database issue):D302–5, January 2012.

[74] Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics (Oxford, England), 22(13):1658–9, July 2006.

[75] J D Lieb, X Liu, D Botstein, and P O Brown. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nature genetics, 28(4):327–34, August 2001.

[76] S Linn and W Arber. Host specificity of DNA produced by Escherichia coli, X. In vitro restriction of phage fd replicative form. Proceedings of the National Academy of Sciences of the United States of America, 59(4):1300–6, April 1968.

[77] J T LITCHFIELD and F WILCOXON. A simplified method of evaluating dose-effect experiments. The Journal of pharmacology and experimental therapeutics, 96(2):99–113, June 1949.

[78] David J Lynn, Geoffrey L Winsor, Calvin Chan, Nicolas Richard, Matthew R Laird, Aaron Barsky, Jennifer L Gardy, Fiona M Roche, Timothy H W Chan, Naisha Shah, Raymond Lo, Misbah Naseer, Jaimmie Que, Melissa Yau, Michael Acab, Dan Tulpan, Matthew D Whiteside, Avinash Chikatamarla, Bernadette Mah, Tamara Munzner, Karsten Hokamp, Robert E W Hancock, and Fiona S L Brinkman. InnateDB: facilitating systems-level analyses of the mammalian innate immune response. Molecular systems biology, 4:218, January 2008.

[79] Steven Maere, Karel Heymans, and Martin Kuiper. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics (Oxford, England), 21(16):3448–9, August 2005.

[80] Alexander Makarov. Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. Analytical Chemistry, 72(6):1156–1162, March 2000.

[81] H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics, 18(1):50–60, March 1947.

[82] M Mann, P Hø jrup, and P Roepstorff. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biological mass spectrometry, 22(6):338–45, June 1993.

[83] Nicholas R Markham and Michael Zuker. UNAFold: software for nucleic acid folding and hybridization. Methods in molecular biology (Clifton, N.J.), 453:3–31, January 2008.

[84] David H Mathews, Matthew D Disney, Jessica L Childs, Susan J Schroeder, Michael Zuker, and Douglas H Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proceedings of the National Academy of Sciences of the United States of America, 101(19):7287–92, May 2004.

[85] Gerhard Mittler, Falk Butter, and Matthias Mann. A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. Genome research, 19(2):284–93, February 2009.

[86] Nagarjuna Nagaraj, Nils Alexander Kulak, Juergen Cox, Nadin Neuhauser, Korbinian Mayr, Ole Hoerning, Ole Vorm, and Matthias

Mann. System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. Molecular & cellular proteomics : MCP, 11(3):M111.013722, March 2012.

[87] I. Newton and P. Frost. Newton's Principia. Macmillan and Co., 1863.

[88] William M Old, Karen Meyer-Arendt, Lauren Aveline-Wolf, Kevin G Pierce, Alex Mendoza, Joel R Sevinsky, Katheryn a Resing, and Natalie G Ahn. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. Molecular & cellular proteomics : MCP, 4(10):1487–502, October 2005.

[89] Shao-En Ong, Blagoy Blagoev, Irina Kratchmarova, Dan Bach Kristensen, Hanno Steen, Akhilesh Pandey, and Matthias Mann. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Molecular & cellular proteomics : MCP, 1(5):376–86, May 2002.

[90] D J Pappin, P Hojrup, and A J Bleasby. Rapid identification of proteins by peptide-mass fingerprinting. Current biology : CB, 3(6):327–32, June 1993.

[91] J Park, K Karplus, C Barrett, R Hughey, D Haussler, T Hubbard, and C Chothia. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. Journal of molecular biology, 284(4):1201–10, December 1998.

[92] D N Perkins, D J Pappin, D M Creasy, and J S Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 20(18):3551–67, December 1999.

[93] Andreas Pichlmair, Caroline Lassnig, Carol-Ann Eberle, Maria W Górna, Christoph L Baumann, Thomas R Burkard, Tilmann Bürckstümmer, Adrijana Stefanovic, Sigurd Krieger, Keiryn L Bennett, Thomas Rülicke, Friedemann Weber, Jacques Colinge, Mathias Müller, and Giulio Superti-Furga. IFIT1 is an antiviral protein that recognizes 5'-triphosphate RNA. Nature immunology, 12(7):624–30, January 2011.

[94] Elodie Portales-Casamar, Supat Thongjuea, Andrew T Kwon, David Arenillas, Xiaobei Zhao, Eivind Valen, Dimas Yusuf, Boris Lenhard, Wyeth W Wasserman, and Albin Sandelin. JASPAR 2010: the greatly

expanded open-access database of transcription factor binding profiles. Nucleic acids research, 38(Database issue):D105–10, January 2010.

[95] Marco Punta, Penny C Coggill, Ruth Y Eberhardt, Jaina Mistry, John Tate, Chris Boursnell, Ningze Pang, Kristoffer Forslund, Goran Ceric, Jody Clements, Andreas Heger, Liisa Holm, Erik L L Sonnhammer, Sean R Eddy, Alex Bateman, and Robert D Finn. The Pfam protein families database. Nucleic acids research, 40(Database issue):D290–301, January 2012.

[96] J L Reid, V R Iyer, P O Brown, and K Struhl. Coordinate regulation of yeast ribosomal protein genes is associated with targeted recruitment of Esa1 histone acetylase. Molecular cell, 6(6):1297–307, December 2000.

[97] Jüri Reimand, Meelis Kull, Hedi Peterson, Jaanus Hansen, and Jaak Vilo. g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic acids research, 35(Web Server issue):W193–200, July 2007.

[98] B Ren, F Robert, J J Wyrick, O Aparicio, E G Jennings, I Simon, J Zeitlinger, J Schreiber, N Hannett, E Kanin, T L Volkert, C J Wilson, S P Bell, and R a Young. Genome-wide location and function of DNA binding proteins. Science (New York, N.Y.), 290(5500):2306–9, December 2000.

[99] J D Retief. Phylogenetic analysis using PHYLIP. Methods in molecular biology (Clifton, N.J.), 132:243–58, January 2000.

[100] E Rivas and S R Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. Journal of molecular biology, 285(5):2053–68, February 1999.

[101] Philip L Ross, Yulin N Huang, Jason N Marchese, Brian Williamson, Kenneth Parker, Stephen Hattan, Nikita Khainovski, Sasi Pillai, Subhakar Dey, Scott Daniels, Subhasish Purkayastha, Peter Juhasz, Stephen Martin, Michael Bartlet-Jones, Feng He, Allan Jacobson, and Darryl J Pappin. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Molecular & cellular proteomics : MCP, 3(12):1154–69, December 2004.

[102] N Saitou and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution, 4(4):406–25, July 1987.

[103] Martin R Singleton, Mark S Dillingham, Martin Gaudier, Stephen C Kowalczykowski, and Dale B Wigley. Crystal structure of RecBCD enzyme reveals a machine for processing DNA breaks. Nature, 432(7014):187–93, November 2004.

[104] S S Stevens. On the Theory of Scales of Measurement. Science (New York, N.Y.), 103(2684):677–80, June 1946.

[105] John E P Syka, Joshua J Coon, Melanie J Schroeder, Jeffrey Shabanowitz, and Donald F Hunt. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proceedings of the National Academy of Sciences of the United States of America, 101(26):9528–33, June 2004.

[106] Motoko Unoki, Toshihiko Nishidate, and Yusuke Nakamura. ICBP90, an E2F-1 target, recruits HDAC1 and binds to methyl-CpG through its SRA domain. Oncogene, 23(46):7601–10, October 2004.

[107] V.N. Vapnik. The Nature of Statistical Learning Theory. Statistics for Engineering and Information Science. Springer, 2000.

[108] Chia-Lin Wei, Qiang Wu, Vinsensius B Vega, Kuo Ping Chiu, Patrick Ng, Tao Zhang, Atif Shahab, How Choong Yong, YuTao Fu, Zhiping Weng, JianJun Liu, Xiao Dong Zhao, Joon-Lin Chew, Yen Ling Lee, Vladimir a Kuznetsov, Wing-Kin Sung, Lance D Miller, Bing Lim, Edison T Liu, Qiang Yu, Huck-Hui Ng, and Yijun Ruan. A global map of p53 transcription-factor binding sites in the human genome. Cell, 124(1):207–19, January 2006.

[109] Frank Wilcoxon. Individual Comparisons by Ranking Methods. Biometrics Bulletin, 1(6):80–83, December 1945.

[110] Frank Wilcoxon. Probability tables for individual comparisons by ranking methods. Biometrics, 3(3):119–22, September 1947.

[111] M Wilm and M Mann. Analytical properties of the nanoelectrospray ion source. Analytical chemistry, 68(1):1–8, January 1996.

[112] E Wingender, P Dietze, H Karas, and R Knüppel. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic acids research, 24(1):238–41, January 1996.

[113] Cathy H Wu, Rolf Apweiler, Amos Bairoch, Darren a Natale, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger,

Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J Martin, Raja Mazumder, Claire O'Donovan, Nicole Redaschi, and Baris Suzek. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic acids research, 34(Database issue):D187–91, January 2006.

[114] Barry R Zeeberg, Weimin Feng, Geoffrey Wang, May D Wang, Anthony T Fojo, Margot Sunshine, Sudarshan Narasimhan, David W Kane, William C Reinhold, Samir Lababidi, Kimberly J Bussey, Joseph Riss, J Carl Barrett, and John N Weinstein. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome biology, 4(4):R28, January 2003.

[115] Cong Zhu, Kelsey J R P Byers, Rachel Patton McCord, Zhenwei Shi, Michael F Berger, Daniel E Newburger, Katrina Saulrieta, Zachary Smith, Mita V Shah, Mathangi Radhakrishnan, Anthony a Philippakis, Yanhui Hu, Federico De Masi, Marcin Pacek, Andreas Rolfs, Tal Murthy, Joshua Labaer, and Martha L Bulyk. High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome research, 19(4):556–66, April 2009.

[116] M Zuker and P Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic acids research, 9(1):133–48, January 1981.

[117] Boris Zybailov, Amber L Mosley, Mihaela E Sardiu, Michael K Coleman, Laurence Florens, and Michael P Washburn. Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. Journal of proteome research, 5(9):2339–47, September 2006.