Eberhard Karls Universität Tübingen
Mathematisch-Naturwissenschaftliche Fakultät
Wilhelm-Schickard-Institut für Informatik

Bachelor Thesis Bioinformatics

Semi-supervised learning for nucleic acid cross-linking mass spectrometry

Emil Paulitz

14.08.2020

Reviewer

Prof. Oliver Kohlbacher
Department of Computer Science
University of Tübingen

Supervisor

Timo Sachsenberg
Address
University of Tübingen

Paulitz, Emil:
Semi-supervised learning for nucleic acid cross-linking mass spectrometry
Bachelor Thesis Bioinformatics
Eberhard Karls Universität Tübingen
Period: 14.04.2020-14.08.2020

Abstract

Write here your abstract.

Acknowledgements

Write here your acknowledgements.

Contents

List of Figures

List of Tables

List of Abbreviations

MS	Mass Spectrometry
LC	Liquid Chromatography
MS/MS	Tandem Mass Spectrometry
PSM	Peptide Spectrum Match
FDR	False Discovery Rate
ROC-curve	Receiver Operating Characteristic Curve

Chapter 1

Introduction

- Motivation

1.1 Background

Proteomics is an interdisciplinary research field analyzing the composition, interaction and impacts of the proteome (the entirety of proteins) of single cells or up to a whole organism [6, 10]. In this thesis, research was done in a related field, focusing on peptides cross-linked with RNA. The chemical bond between cross-linked molecules has been artificially induced, for example using UV light [10]. Applying this to peptides and RNA could possibly give insight into their *in vivo* interactions, and may also allow conclusions about protein-DNA interaction.
For quantitatively characterizing the proteome of a sample, large scale measuring techniques are needed. Mostly, mass spectrometry (MS) is used, or more specifically, as for the data in this thesis, tandem mass spectrometry (MS/MS) combined with liquid chromatography (LC). In order to analyze the protein sample with MS, its complexity has to be reduced as much as possible, for example using LC [10]. As Han et al. [6] explains, the mass spectrometer then produces mass spectra, which have to be analyzed further. It does so by first ionizing the substrate, because it can only detect charged particles. Then, the sample is separated in the mass analyzer by the ratio $\frac{m}{z}$, mass of the particles to their charge. The detector then quantifies the amount of a particle in the sample. The result is a mass spectrum, as shown in figure **??**.

Because mass alone does not give enough information about a peptide to determine its sequence, tandem mass spectrometry is often used to gather more detailed evidence. In this procedure, particles of similar $\frac{m}{z}$ ratio are selected for fragmentation in a collision cell after the first round of mass measurement [10]. In there, the substance collides with a gas to be broken down into smaller molecules. For proteins, fragmentation happens predominantly

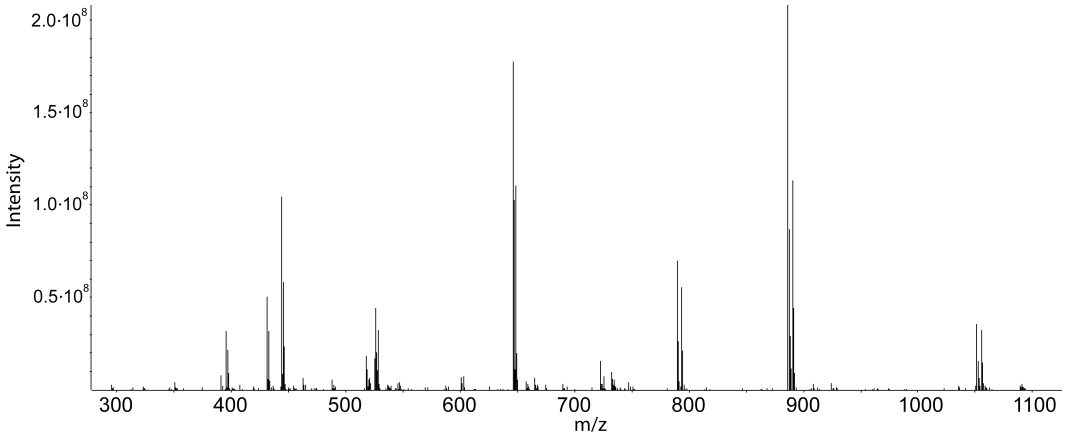

Figure 1.1: Example for a mass spectrum as recorded by a mass spectrometer. The ion intensity correlates with the amount of a molecule in the sample, $\frac{m}{z}$ is the mass-to-charge-ratio. From: Sachsenberg [10]

in their backbone, producing all possible sub-sequences of the peptide. This produces a spectrum that is almost unique for its protein, which allows for peptide identification using bioinformatics tools [2].
Algorithms like Sequest [4] or X! Tandem [3] compare the resulting spectra with theoretical spectra calculated from a list of possible peptides and compute a score based on their similarity. The peptides are generated by obtaining a list of proteins expected in the sample and calculating the peptides resulting from the, for example enzyme-based, degradation of the proteins. The best scoring peptide is then considered a peptide-spectrum-match (PSM). The scores produced by those algorithms often do not distinguish well enough between correct and incorrect matches [7], but they enable FDR estimation using decoy databases and serve as a basis for score re-calibration with the Percolator algorithm [7, 5].
Decoy databases are created from the target database, contain usually as many peptides [9, 8] in a reversed or shuffled order with respect to the amino acid sequence [1]. They are presented to the scoring algorithm either separately [5] or mixed with the target database [9]. It is assumed, that decoy and target peptides have similar features [8] and are not easily distinguishable by a scoring algorithm. When the actually fitting peptide for a given spectrum is not in the target database, and thus a wrong one will be chosen, the best scoring peptide will be a decoy approximately half of the time. This allows for an estimation of wrongly assigned targets, since the score distribution is assumed to be the same for decoys and false targets [1].
In practice, one estimates the probability of a PSM being a false target by counting the number of decoy-PSMs with the same or a higher score. It is

then assumed, there are as many false targets and thus a false discovery rate (FDR) can be estimated. This leads to the following formula[1] [5]:

$$FDR = \frac{\#\text{ false target PSMs}}{\#\text{ all target PSMs}} \approx \frac{\#\text{ decoy PSMs}}{\#\text{ all target PSMs}} \tag{1.1}$$

The q-value as a measure for a single PSM rather than a metric for a set of PSMs is then derived from this as the minimum FDR of all PSMs with a lower or equal score [5, 1]. It will be used for estimating the credibility for any one PSM.

As Käll et al. [7] say, separating correct from incorrect target PSMs with already mentioned algorithms works fine, but there is still room for improvement. This is because often not all information is used and considered jointly. Percolator [7, 5] tries to utilize as much information as possible by using scores from different algorithms, features of the peptide like its length, of the spectrum or the PSM itself. It joins them using a linear SVM and a semi-supervised approach with cross-validation to retain as many PSMs as possible. In every iteration, the top ranking, non-decoy PSMs up to a certain threshold of q-value are chosen as positive training examples, and the decoy PSMs are used as negative training set. The PSMs are then re-ranked using the SVM score, with the intend of getting a better separation of true and false PSMs. If that holds true, the positive training set of the next iteration better is of higher quality and the SVM can be trained even better. The algorithm usually converges within the first 10 iterations [7]. To avoid having to split the data into training and testing set and consequently losing possibly correct PSMs but also avoid overfitting, a nested cross-validation approach is being used [5]. - ROC Curve noch erklären?

[1] In this thesis, the following approximation is used:
$FDR \approx \frac{\#\text{ decoy PSMs}}{\#\text{ all PSMs}} = \frac{\#\text{ decoy PSMs}}{\#\text{ decoy PSMs}+\#\text{ target PSMs}}$
It is faster to calculate and yields results differing by the FDR, so in the relevant range of FDRs of 0 to 5% up to 5%:
$\frac{\frac{\#\text{ decoys}}{\#\text{ targets}}}{\frac{\#\text{ decoys}}{\#\text{ decoys}+\#\text{ targets}}} = \frac{\#\text{ decoys}+\#\text{ targets}}{\#\text{ targets}} = 1 + \frac{\#\text{ decoys}}{\#\text{ targets}} \approx 1 + FDR$

Chapter 2

Material and Methods

- Material: Was ich für ein Datensatz zum Testen benutzt habe und wo der herkommt

2.1 Implementation of the percolator algorithm

- Wie genau stelle ich das vor, siehe Mail? Wichtige Punkte wären:
- Verwendete Scipy-Methoden
- Abbruch wenn es nicht besser wird und dass ich die AUC als Metrik nutze
- feature normalization
- Wichtige Hilfsfunktionen (pseudoROC zB)

2.2 Improvements of the percolator algorithm for cross-link identification

To be able to monitor the difference any experiment makes, especially with respect to the cross-linked or non-cross-linked PSMs, following features were implemented:
First, in addition to the q-value, which is calculated as described in 1.1, the calculation of a class-specific q-value was implemented. This is done by splitting the dataset according to the class affiliation and calculating the q-value separately for both splits.
Secondly, a ROC curve using the *pseudoROC* function is calculated after every iteration of Percolator, for the whole dataset, only for cross-linked and only for non-cross-linked PSMs. Accordingly, the respective class-specific q-value is used. Thus, three plots containing the corresponding class(es) and every iteration are shown. This allows for fast visual detection of the impact a specific

change to the algorithm has on certain classes, iterations or general sensitivity.

2.2.1 How to deal with different Ranks

- OptimalRanking Option (Erst paar Iterationen Ränge verändern lassen und dann die schlechten entfernen)

2.2.2 Characteristics of cross-linking PSM datasets

- Verhältnis Targets:Decoys und XL:non-XL in inneren und äußeren splits gleich lassen und MinMaxMedian Auswertungen mithilfe von google colab cloud computing
- Imputation (kam zwar nichts raus ist aber trotzdem interessant)
- Trennung von Datensatz nach XL/nXL oder sogar cross-linking target falls Datensatz groß genug

2.2.3 Small datasets

- Ratio Testing (nicht-random aus ganzem Datensatz und random aus Top 10%. Liefert Erkenntnisse über die mögliche Größe des Datensatzes und eventuell die Sinnhaftigkeit, wann man die Datensätze einfach trennen kann → Für den Leser relevant)
- Einbau von Identifikationen bei 1% FDR als Metrik (Sinnhaftigkeit kann man ja diskutieren)

(- Performance auf anderem Datensatz
- Vergleich mit Entrapment FDR)

Chapter 3

Results

3.1 Implementation of the percolator algorithm

- Reimplementierung funktioniert wie Original
- feature normalization war wichtiger boost
- ROC nach jeder Iteration zeigen

3.2 Improvements of the percolator algorithm for cross-link identification

3.2.1 How to deal with different Ranks

- Ergebnisse von OptimalRanking

3.2.2 Characteristics of cross-linking PSM datasets

- Verhältnis Targets:Decoys und XL:non-XL verringt die Streuung: MinMax-Median Auswertungen
- Bei Imputation kam nichts heraus
- Großer Unterschied wenn man den (großen) Datensatz nach XL/nXL oder sogar cross-linking target aufteilt

3.2.3 Small datasets

- Sinnvolle Plots zu Ratio Testing
- Neue Metrik erlaubt es der Implementierung, auch auf kleineren Datensätzen

zu funktionieren

Chapter 4

Discussion and Outlook

- Methoden hinterfragen oder begründen, Ergebnisse interpretieren, Anwendbarkeit diskutieren, z.B.:
- Warum habe ich mich mit Rängen beschäftigt? (Bei XL-Datensätzen oft sinnvoll um erst percolator entscheiden zu lassen was auf Rang 1 steht (?)) -¿ Timo: "XL sind schwer. Richtiger XL oft auf Rang 2+, d.h. man muss alle Daten nutzen (= Percolator benutzen) um die maximale Anzahl an XLs zu identifizieren
- Falsche Formel für q-value
- C Parameter für jeden split neu optimieren führt zu overfitting? → Original-Algorithmus macht es auch so
- Wie sinnvoll ist die neue Metrik (idents bei 1%)?
- ScanNr Versuche: Gleiche Spektren (identifiziert anhand der ScanNr.) auf verschiedene splits verteilen verändert nichts, d.h. vermutlich sind die niedrigeren Ränge dann so schlecht, dass es nichts bringt die schonmal gesehen zu haben.
- Peptide Versuche: Schlechtere Ergebnisse, aber vllt ehrlicher?
- Mögliche weiterführende Experimente: mächtigere Klassifikatoren + monotonic constraints (wie von Timo ausprobiert), Ada-Boosting, feature selection

Bibliography

[1] Suruchi Aggarwal and Amit Kumar Yadav. False discovery rate estimation in proteomics. In *Methods in Molecular Biology*, pages 119–128. Springer New York, 2016. doi: 10.1007/978-1-4939-3106-4_7. URL https://doi.org/10.1007/978-1-4939-3106-4_7.

[2] Thomas E. Angel, Uma K. Aryal, Shawna M. Hengel, Erin S. Baker, Ryan T. Kelly, Errol W. Robinson, and Richard D. Smith. Mass spectrometry-based proteomics: existing capabilities and future directions. *Chemical Society Reviews*, 41(10):3912, 2012. doi: 10.1039/c2cs15331a. URL https://doi.org/10.1039/c2cs15331a.

[3] R. Craig and R. C. Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, February 2004. doi: 10.1093/bioinformatics/bth092. URL https://doi.org/10.1093/bioinformatics/bth092.

[4] 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*, 1994.

[5] Viktor Granholm, William Noble, and Lukas Käll. A cross-validation scheme for machine learning algorithms in shotgun proteomics. *BMC Bioinformatics*, 13(Suppl 16):S3, 2012. doi: 10.1186/1471-2105-13-s16-s3. URL https://doi.org/10.1186/1471-2105-13-s16-s3.

[6] Xuemei Han, Aaron Aslanian, and John R Yates. Mass spectrometry for proteomics. *Current Opinion in Chemical Biology*, 12(5):483–490, October 2008. doi: 10.1016/j.cbpa.2008.07.024. URL https://doi.org/10.1016/j.cbpa.2008.07.024.

[7] Lukas Käll, Jesse D Canterbury, Jason Weston, William Stafford Noble, and Michael J MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4(11):923–925, October 2007. doi: 10.1038/nmeth1113. URL https://doi.org/10.1038/nmeth1113.

[8] Roger E. Moore, Mary K. Young, and Terry D. Lee. Qscore: An algorithm for evaluating SEQUEST database search results. *Journal of the American Society for Mass Spectrometry*, 13(4):378–386, April 2002. doi: 10.1016/s1044-0305(02)00352-5. URL `https://doi.org/10.1016/s1044-0305(02)00352-5`.

[9] Junmin Peng, Joshua E. Elias, Carson C. Thoreen, Larry J. Licklider, and Steven P. Gygi. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *Journal of Proteome Research*, 2(1):43–50, February 2003. doi: 10.1021/pr025556v. URL `https://doi.org/10.1021/pr025556v`.

[10] Timo Sachsenberg. Computational methods for mass spectrometry-based study of protein-rna or protein-dna complexes and quantitative metaproteomics. 2017. URL `http://hdl.handle.net/10900/83311`.

Selbständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben von Quellen als Entlehnung kenntlich gemacht worden sind. Diese Bachelorarbeit wurde in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt.

Ort, Datum Unterschrift