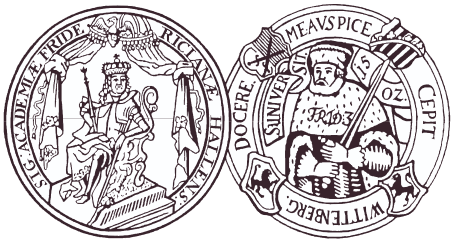
Bioinformatics tools for mass spectrometry, phylogenetic footprinting, and the integration of biological data



Dissertation

zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften (Dr.rer.nat.)

der Naturwissenschaftliche Fakultät III
Agrar- und Ernährungswissenschaften, Geowissenschaften und Informatik
der Martin-Luther-Universität Halle-Wittenberg

vorgelegt von

Hendrik Treutler

Geb. am 10.03.1985 in Karl–Marx–Stadt

Gutachter:
1. Prof. Dr. Ivo Grosse
2. Prof. Dr. Burkhard Morgenstern
3. Dr. Steffen Neumann

Tag der Verteidigung: 16. November 2017

Acknowledgements

My first thanks belong to my beloved wife Anna-Maria Treutler. Thanks to you I can submit this thesis before our fifth wedding anniversary. You encourage me all the time and you gave me three wonderful children. Many thanks also belong to my parents who formed my roots in this world and made it possible for me to fulfill my wishes.

I thank my instructors Ivo Grosse, Steffen Neumann, Stefan Posch, and Falk Schreiber for guiding me to this thesis. You formed my roots in the scientific world and your advice on technical, scientific, and political issues lead to own experiences which will be indispensable for me in the future.

Martin Nettling earns special thanks. The exchange of thoughts with you concerning technical, scientific, and private issues changed and widened my mind countless times. With him on your side things will basically work out.

I thank Gerd Balcke for accompanying me in the world of mass spectrometry. Your constructive ideas and your eagerness are an inspiration for me and lead to a great collaboration. I thank Karin Gorzolka for her passion to share thoughts in many positive discussions. I thank Susann Mönchgesang, Christoph Ruttkies, Daniel Schober, and Sarah Scharfenberg for their great support in technical and scientific issues and for their friendship. I thank Jesus Cerquides, Tobias Czauderna, Jan Grau, Eva Grafahrend-Belau, Anja Hartmann, Astrid Junker, Jens Keilwagen, Matthias Klapperstück, Matthias Lange, Christian Klukas, Hendrik Rohn, and Uwe Scholz for fruitful discussions and professional teamwork.

Last but not least I thank Kristian Peters for technical support with *MetFamily*. You do a really great job and I enjoy your presence for many reasons.

Contents

1 Summary

1.1 English version

Complex biological processes such as RNA splicing, protein degradation, and metabolic switches are a prerequisite for organisms to adapt to environmental stimuli, respond to pathogens, and absorb nutrients. These biological processes are studied in molecular biology and there are powerful techniques for the study of proteins, DNA sequences, and small molecules such as nuclear magnetic resonance (NMR) spectroscopy, high–throughput sequencing technologies, and mass spectrometry. Depending on the technique the amount of biological raw data can be in the order of hundreds of gigabytes and specialized tools are needed to process this data and extract interpretable information. A deep understanding of both biology and computer science is a prerequisite for the development of such tools and bioinformatics is an interdisciplinary field which emerged from this demand. Bioinformatics research involves the development of tools for the study of biological processes using methods from computer science.

Systems biology is a bioinformatics field which aims at an understanding of biological systems rather than investigating each biological process individually. Systems biology approaches are able to translate various kinds of biological data to parameters of mathematical models enabling, amongst others, the understanding of the complex interplay of different biological entities, the reproduction of emergent properties of metabolic pathways, and the prediction of the behaviour of cells under hypothetical conditions. Research in systems biology necessitates insights into the regulation of gene expression, the metabolism, and the integration of biological data from different sources. The publications in this cumulative thesis are rooted in these three fields, namely (i) "Phylogenetic footprinting" for the study of the regulation of transcriptional initiation, (ii) "Data Processing and Interpretation of Mass Spectrometry Data" for the study of metabolic fingerprints, and (iii) the "Integration of biological data" for the combined analysis of different data. Two publications reside in each of these fields and I will summarize these six publications subsequently.

Decoding the regulation of gene expression is essential for the understanding of life and

the transcriptional initiation is a vital sub–process. Transcriptional initiation is governed by the concerted binding of transcription factors (TFs) to transcription factor binding sites (TFBSs) and *de novo* motif discovery with "Phylogenetic footprinting" on basis of *ChIP–seq* data gives deep insights into this sub–process.

Unfortunately, *ChIP–seq* data is subject to the ubiquitous binding–affinity bias which leads to the prediction of distorted sequence motifs and biased sets of TFBSs. My colleagues and I developed a phylogenetic footprinting model which is capable of estimating and correcting the binding–affinity bias in *ChIP–seq* data. We found that the proposed phylogenetic footprinting model improves *de novo* motif discovery and that the corrected sequence motifs are softer than the uncorrected sequence motifs.

The increasing availability of sequence data and algorithms for *de novo* motif discovery results in the demand to compare different sequence motifs. Sequence motifs which have been extracted from, e.g., different species, tissues, and samples are often similar which makes the comparison challenging. My colleagues and I developed the *R*–based tool *DiffLogo* for the comparative visualization of sequence motifs in DNA sequences, RNA sequences, and protein domains. *DiffLogo* allows the detection of even small motif differences between two sequence motifs and also supports the pair–wise comparison of more than two sequence motifs. We demonstrated the utility of *DiffLogo* using sequence motifs of one transcription factor from different cell lines, sequence motifs from different transcription factors, and sequence motifs in protein domains from different phylogenetic kingdoms. *DiffLogo* was also used to show the effect of the binding–affinity bias in *ChIP–seq* data.

The study of small molecules in the cell gives insights into the chemical fingerprints of different biological processes. Metabolite profiles provide a snapshot of these chemical fingerprints in the samples of interest and Mass Spectrometry (MS) is an essential technology to prepare these metabolite profiles. The "Data Processing and Interpretation of Mass Spectrometry Data" is mandatory in order to gain biological insights from MS data and the identification and precise quantification of metabolites remains a major challenge.

Isotope clusters are sets of related signals in MS data and it has been shown that isotope clusters can be utilized to improve the identification and quantification of metabolites. My colleagues and I developed an approach for the prediction, detection, and validation of isotope clusters in MS data and we integrated this approach into the popular *R*–based tools *xcms* and *CAMERA*. We found that using the proposed approach it is possible to extract 37% more isotope signals from *Arabidopsis thaliana* measurements and in a mix of standard compounds the correct molecular formula could be predicted in 92% of the cases in the top three ranks.

The structural elucidation of metabolites in metabolite profiles is done one–by–one and known as one of the major bottlenecks for the research of the metabolism. My colleagues

and I developed the web application *MetFamily* for the discovery of regulated metabolite families in metabolite profiles providing a birds eye view on comparative studies. *MetFamily* supports the clustering of metabolites with respect to śtructural similarity and the discovery of group–discriminating metabolites which enables the discovery of metabolite families with biochemical relevance. In a study we compared metabolite profiles of tomato trichomes and trichome–free tomato leaves and we classified a multitude of unknown metabolites to metabolite families which are specific for tomato trichomes.

The life sciences face an continuously increasing amount of available data from diverse sources such as publications, biological databases, and own experiments. Challenging problems such as cancer development, the increase of crop yield, and diabetes necessitate the analysis of data of different types from various sources and the "Integration of biological data" is a prerequisite to tackle such problems.

The integration of biological data into biological networks enables the interpretation of wet lab data in a biological context. My colleagues and I extended the *VANTED* system for the network–assisted visualization and analysis of wet lab data. We added and refined several features such as file import capabilities, standardized graphical representations of biological networks, and the simulation of mathematical models representing biological networks.

Sharing datasets is crucial in most life science projects and the compliance to standards is a prerequisite for the integration of biological data. My colleagues and I developed the *DBE2 information system* for the sharing of biological data in a unified way for the *VANTED* system. The *DBE2 information system* supports a seamless integration into the *VANTED* system, the central sharing of biological data with a user right management, and the ontology–assisted standardization of biological terms.

In summary, my colleagues and I contributed to the development of tools for the research of complex biological processes in molecular biology. *De novo* motif discovery with "Phylogenetic footprinting" on basis of *ChIP–seq* data gives insights into the regulation of gene expression, the "Data Processing and Interpretation of Mass Spectrometry Data" reveals metabolic fingerprints in metabolic profiles, and the "Integration of biological data" enables a combined data analysis for a more holistic understanding of processes in the cell. The development of tools for the research of these three fields is a prerequisite to investigate biological questions using systems biology approaches. Advancements in the research of systems biology promise a more comprehensive understanding of biological processes and hence an understanding of the mode of existence of organisms.

1.2 German version

Komplexe biologische Prozesse wie das Spleißen von mRNAs, der Abbau von Proteinen und metabolische Schalter sind eine Voraussetzung allen Lebens, um auf Nährstoffe, Krankheitserreger und abiotische Reize zu reagieren. Die Molekularbiologie untersucht derartige Prozesse mittels leistungsfähiger Techniken wie zum Beispiel Kernspinresonanzspektroskopie, DNA-Sequenzierung im Hochdurchsatz, und Massenspektrometrie. Je nach Technik können hierbei Datenmengen von mehreren hundert Gigabyte anfallen, was spezialisierte Softwarewerkzeuge für die Extraktion interpretierbarer Informationen notwendig macht. Die Bioinformatik ist ein interdisziplinäres Feld welches Konzepte der Biologie und Informatik zusammenbringt und die Grundlage für die Entwicklung derartiger Werkzeuge bildet.

Die Systembiologie ist ein Teilgebiet der Bioinformatik und erforscht biologische Systeme, anstatt einzelne biologische Prozesse isoliert zu betrachten. Ansätze der Systembiologie versuchen unterschiedliche biologische Daten als Parameter von Modellen zu nutzen um, zum Beispiel, das komplexe Zusammenspiel verschiedener biologischer Faktoren zu verstehen, intrinsische Eigenschaften von Stoffwechselwegen zu simulieren oder das Verhalten von Zellen unter hypothetischen Bedingungen vorherzusagen. Fortschritte in der Systembiologie fußen auf Erkenntnissen der Genregulation, des Metabolismus und der Integration biologischer Daten von verschiedenen Quellen. Die Publikationen in dieser kumulativen Dissertationsschrift sind in diesen drei Bereichen angesiedelt, und zwar (i) “Phylogenetic footprinting” für die Untersuchung der Initiation der Transkription, (ii) “Data Processing and Interpretation of Mass Spectrometry Data” für die Untersuchung biochemischer Signaturen und (iii) “Integration of biological data” für die integrative Analyse von biologischen Daten. Diesen drei Gebieten steuere ich je zwei Publikationen bei und ich werde diese sechs Publikationen im Folgenden umreißen.

Die Entschlüsselung der Genregulation ist eine Voraussetzung für das Verständnis des Lebens und die Initiation der Transkription ist ein entscheidender Teilprozess der Genregulation. Die Initiation der Transkription wird von der gezielten Bindung von Transkriptionsfaktorbindestellen durch Transkriptionsfaktoren reguliert und die Vorhersage von Sequenzmotiven mittels “Phylogenetic footprinting” in ChIP-seq Daten gewährt umfassende Erkenntnisse über diesen Teilprozess.

Allerdings sind ChIP-seq Daten vom so genannten *binding–affinity bias* (BA bias) verzerrt, was zu ebenso verzerrten Sequenzmotiven und vorhergesagten Transkriptionsfaktorbindestellen führt. Meine Kollegen und ich entwickelten ein Phylogenetic Footprinting Modell, welches den BA bias in ChIP-seq Daten schätzen und korrigieren kann. Wir zeigten, dass das vorgeschlagene Modell zur Vorhersage von weicheren Sequenzmotiven führt und die Güte der Motivvorhersage verbessert.

Die stetig wachsende Menge von Sequenzdaten und Algorithmen für die Motivvorhersage führt zur Notwendigkeit, verschiedene Sequenzmotive miteinander zu vergleichen. Sequenzmotive, welche von verschiedenen Spezies, Geweben oder biologischen Proben gewonnen wurden, sind sich oft sehr ähnlich was den Vergleich schwierig gestaltet. Meine Kollegen und ich entwickelten das *R*–basierte Softwarewerkzeug *DiffLogo* für die vergleichende Visualisierung von Sequenzmotiven in DNA Sequenzen, RNA Sequenzen und Proteindomänen. *DiffLogo* ermöglicht die Erkennung von feinen Unterschieden zwischen zwei oder mehr Sequenzmotiven durch paarweise Vergleiche. Am Beispiel von Sequenzmotiven des gleichen Transkriptionsfaktors in unterschiedlicher Zelllinien, Sequenzmotiven unterschiedlicher Transkriptionsfaktoren und Sequenzmotiven unterschiedlicher Reiche demonstrierten wir die Möglichkeiten von *DiffLogo*. *DiffLogo* wurde benutzt um den Effekt des BA bias in ChIP-seq Daten aufzuzeigen.

Die Untersuchung niedermolekularer Verbindungen in der Zelle gewährt tiefe Einblicke in die biochemischen Signaturen von biologischen Prozessen. Metabolitenprofile stellen eine Momentaufnahme dieser biochemischen Signaturen in biologischen Proben dar und die Massenspektrometrie ist eine entscheidende Technologie für die Erstellung von Metabolitenprofilen. Das Themengebiet "Data Processing and Interpretation of Mass Spectrometry Data" bildet die Grundlage, um biologische Erkenntnisse aus Massenspektrometriedaten zu gewinnen. Besonders die Quantifizierung und Identifizierung von Metaboliten stellt hier eine große Herausforderung dar.

Isotopenmuster setzen sich aus einer Reihe verwandter Signale in Massenspektrometriedaten zusammen und es wurde gezeigt, dass Diese für eine genauere Quantifizierung und Identifizierung von Metaboliten genutzt werden können. Meine Kollegen und ich entwickelten einen Ansatz für die Vorhersage, Erkennung und Validierung von Isotopenmustern in Massenspektrometriedaten und integrierten diesen Ansatz in die verbreiteten Softwareerkzeuge *xcms* und *CAMERA*. Wir zeigten, dass unser Ansatz in Messungen von *Arabidopsis thaliana* 37% mehr Isotopensignale extrahiert und in einer Mischung von chemischen Standardsubstanzen sagten wir in 92% der Fälle die richtige Summenformel in den obersten drei Rängen vorher.

Die Strukturaufklärung von tausenden von Metaboliten in Metabolitenprofilen findet einzeln statt und stellt daher einen wesentlichen limitierenden Faktor in der Metabolomik dar. Meine Kollegen und ich entwickelten die Webanwendung *MetFamily* für die Erkennung von regulierten Metabolitenfamilien, um eine Vogelperspektive auf vergleichende Studien zu ermöglichen. *MetFamily* kann Metabolite anhand struktureller Ähnlichkeit gruppieren und gruppenunterscheidende Metabolite finden, was die Entdeckung von Metabolitenfamilien mit biochemischer Relevanz ermöglicht. Wir verglichen Metabolitenprofile von glandulären Trichomen der Tomatenpflanze mit trichomfreien Tomatenblättern und ordneten eine Vielzahl unbekannter Metabolite Metabolitenfamilien zu, welche spezifisch für glan-

dulärer Trichome der Tomate sind.

Wissenschaftlern der Lebenswissenschaften steht eine stetig wachsende Menge von Daten aus Publikationen, biologischen Datenbanken und hauseigenen Experimenten zur Verfügung. Große Probleme wie die Krebsbekämpfung, die Steigerung des Ertrags von Kulturpflanzen und Diabetes erfordern die Analyse verschiedener biologischer Daten aus heterogenen Quellen und das Themengebiet "Integration of biological data" beschäftigt sich mit dieser Anforderung.

Die Integration biologischer Daten mit biologischen Netzwerken ermöglicht die Interpretation von Labordaten im biochemischen Kontext. Meine Kollegen und ich erweiterten das Softwarewerkzeug *VANTED* für die netzwerkgestützte Visualisierung und Analyse von Labordaten. Wir erweiterten bestehende Funktionalitäten und fügten Neue hinzu bezüglich der Unterstützung verschiedener Datenformate, der standardisierten Darstellung biologischer Netzwerke und der Simulation mathematischer Modelle von biologischen Netzwerken.

Der Austausch von biologischen Daten ist ein wichtiger Bestandteil von Projekten in den Lebenswissenschaften und die Einhaltung von Standards ist eine Voraussetzung für die Integration dieser Daten. Meine Kollegen und ich entwickelten das *DBE2 information system* für den vereinheitlichten Austausch biologischer Daten in *VANTED*. Das *DBE2 information system* ist nahtlos in *VANTED* integriert, realisiert den zentralen Austausch von biologischen Daten inklusive Nutzerrechteverwaltung und standardisiert biologische Begrifflichkeiten basierend auf Ontologien.

Zusammenfassend trugen meine Kollegen und ich zur Entwicklung von Softwarewerkzeugen für die Erforschung von komplexen biologischen Prozessen in der Molekularbiologie bei. Entwicklungen im Themengebiet "Phylogenetic footprinting" ermöglichen Einblicke in die Genregulation anhand von ChIP-seq Daten, Entwicklungen im Themengebiet "Data Processing and Interpretation of Mass Spectrometry Data" ermöglichen die Erkennung von biochemischen Signaturen in Metabolitenprofilen und Entwicklungen im Themengebiet "Integration of biological data" ermöglichen die Untersuchung umfassenderer Fragestellungen zu den biologischen Prozessen in der Zelle. Softwarewerkzeuge für die Forschung in diesen drei Themengebieten sind eine Voraussetzung für die Ergründung biologischer Fragestellungen mittels Ansätze der Systembiologie. Fortschritte in der Systembiologie versprechen ein ganzheitliches Verständnis von biologischen Prozessen und damit der Lebensweise von Lebewesen.

2 Introduction

The pursuit to understand and influence the nature of organisms is as old as mankind. Based on pure empiricism, humans were able to change the nature of animals and plants more than 10,000 years ago by domestication and breeding of dogs, cattle, and fowl as well as grasses, herbage, and fruits. In turn, these achievements also fundamentally changed the human way of life as these finally moved us from the caves into apartment buildings (Sturgis, 2015). However, the human activities in this field were mostly not scientific and the basic principles of life in the cell remained concealed for a long time.

In the mid of the 19th century, Mendel suggested the existance of inheritable units in organisms and Darwin introduced in his work "On the Origin of Species" the revolutionary idea that species are subjected to evolution (Darwin, 1859). In the mid of the 20th century, theoretical and technical advances facilitated the research of the basic building blocks of life and Warren Weaver coined the term *Molecular biology* as the use of methods from physics such as X–rays, ultracentrifuges, and mathematics for the research of living things (Weaver, 1970). Francis Crick and James Watson published their revolutionary paper about the structure of the DNA double helix and Francis Crick constituted the *central dogma of molecular biology* which established the basis for today's understanding of life, i.e. DNA makes RNA and RNA makes protein (F. H. Crick, 1958; F. Crick, 1970). The deeper exploration of molecular biology continued quickly and more recent breakthroughs include the production of human insulin in the bacterium *Escherichia coli*, "DNA finger printing" in the law court, and the sequencing of the human genome (Kamionka, 2011; Chambers et al., 2014; Sawicki et al., 1993). In the future, cultural, political, and ethical questions become more and more important as nowadays possibilities involve genetically modified organisms (GMOs) in the environment, biowarefare, and the possibility to concertedly modify the human germline.

Computer–assisted data analysis is nowadays a prerequisite for the research in the life sciences. The theoretical basis of computer science was laid by Gottfried Wilhelm Leibniz with the formal definition of binary logic in the 18th century and George Boole with his boolean algebra in the 19th century. In the 20th century, Kurt Gödel founded the complexity theory, Alan Turing introduced the turing machine, and Claude Elwood Shannon

invented the modern digital circuit design. Starting with comparatively simple mechanical computers for arithmetics and the enciphering of messages, the breakthrough of computer science started with the invention of the first point-contact transistor invented by John Bardeen, Walter Brattain, and William Shockley (Shockley, 1952). The "von Neumann architecture" by John von Neumann paved the way for the first personal computer in the mid of the 20th century (Godfrey et al., 1993). In the 1960s the ARPANET managed the switching of packets of data in a network of local area networks. This installation fundamentally changed the way of exchanging data world–wide as the ARPANET is considered the predecessor of the Internet. Computer science changed economy, sciences, and last but not least our everyday life. Current research in computer science includes voice recognition for better interfaces, deep learning for the recognition of patterns in diverse kinds of data, and distributed systems for the management of *big data*. In the future, quantum computing, artificial intelligence, and advanced computer–human interfaces could change everything again.

Computer science changed the research in the life sciences in the 1950s as more and more biological sequences became available due to *Sanger sequencing* (Sanger et al., 1975). First theoretic and algorithmic developments for the comparison and alignment of DNA sequences, RNA sequences, and amino acid sequences emerged to the field of *bioinformatics* as proposed by Paulien Hogeweg and Ben Hesper in 1970. Originally, the term bioinformatics coined "the study of informatic processes in biotic systems" (Hogeweg, 2011), but with the rapid growth of this field in the 1990s the term bioinformatics was thenceforth rather used for "computational methods for comparative analysis of genome data" (Hogeweg, 2011). After the turn of the millennium, the field of bioinformatics published an enormous amount of papers describing the development and implementation of computer programs for the management of various types of data, statistical measures, and new algorithms to analyze large data sets. Typical problems in bioinformatics include the prediction of genes in genomic sequences, 3D modeling of proteins, and the clustering of protein sequences to protein families. Current research in bioinformatics includes phylogeny for the understanding of evolutionary mechanisms, the concerted reprogramming of genetic material, and the creation of holistic cell models. In the future, new insights into the structure and regulation of hundreds of genomes, precise predictions of drug targets, and continuing contributions to the theoretical biology in general could greatly extend our possibilities to understand and influence the nature of organisms. In this sense, I am honored to contribute to the field of bioinformatics in different aspects such as the network–assisted integration of biological data, the analysis of metabolite profiles, and the precise extraction of DNA motifs from wet lab data.

The manuscripts in this thesis touch different subjects in biology, computer science, and bioinformatics. In this chapter I give a brief introduction into the basics of these subjects.

In section 2.1, I will introduce the biological background of my works, in section 2.2 I will introduce the computer science background, and in section 2.3 I will introduce the bioinformatics background. Finally, I state the research objectives of this thesis in section 2.4.

2.1 Biological background

The manuscripts in this thesis touch different subjects in the field of biology and in this section I give a brief introduction into the basics of these subjects. Specifically, in subsection 2.1.1 I will introduce the central dogma of molecular biology and the major players thereof, in subsection 2.1.2 I will introduce the regulation of gene expression, and in subsection 2.1.3 I will introduce the metabolism and small molecules.

2.1.1 The central dogma of molecular biology

The central dogma of molecular biology holds for all living organisms and states that sequential genetic information can be translated into a chain of amino acids, but a chain of amino acids can not be translated into sequential genetic information (F. Crick, 1970). Possible translations of sequential information originally included three essential processes in the cell, namely DNA replication, transcription, and translation (see Figure 2.1). These three processes involve the three biopolymers DNA, RNA, and proteins.

The Deoxyribonucleic acid (DNA) encodes the potential of organisms to reproduce and exist in a dynamic environment (Watson et al., 1953). DNA is a biopolymer of four different nucleobases, namely adenine (A), cytosine (C), guanine (G), and thymine (T), and the order of these nucleobases encodes genetic information. Two reverse complement (+/-) strands of DNA form the well–known double–helix in higher organisms which means that the entire genetic information is encoded on each strand. All genetic information of an organism is called genome which includes genes, regulatory sequences, and even patches of genetic information from foreign species and *genomics* refers to the study of the genome. A prerequisite for the reproduction of organisms is DNA replication (cf. Figure 2.1) which denotes the process to produce two copies of DNA molecules from the original DNA molecule (Alberts et al., 2015). The implementation of genetic information is called gene expression and involves the production of Ribonucleic acids (RNAs) by transcription and the production of proteins by transcription followed by translation.

Transciption is the process to produce RNA from a segment of DNA (Solomon et al., 2007). RNA is a biopolymer of four different nucleobases, namely adenine (A), cytosine (C), guanine (G), and uracil (U), and the order of these nucleobases encodes genetic information.

There are various types of RNA such as messenger RNAs (mRNAs), microRNAs, and ribosomal RNAs and the volume of RNA in the cell represents the currently available genetic information. All mRNAs in the cell constitute the transcriptome and the study of the transcriptome is denoted *transcriptomics*.

Translation is the process to produce polypeptides from mRNAs (Berg et al., 2002). Polypeptides are a biopolymer of 22 different amino acids and the order of these amino acids in vital for the properties of the folded protein. Short polypeptides are called peptides and serve as hormones, signaling molecules, and antibiotics. Proteins can form big complexes and constitute, amongst others, the cell structure, membrane transporters, and enzymes. All proteins in the cell constitute the proteome and the study of the proteome is denoted *proteomics*.

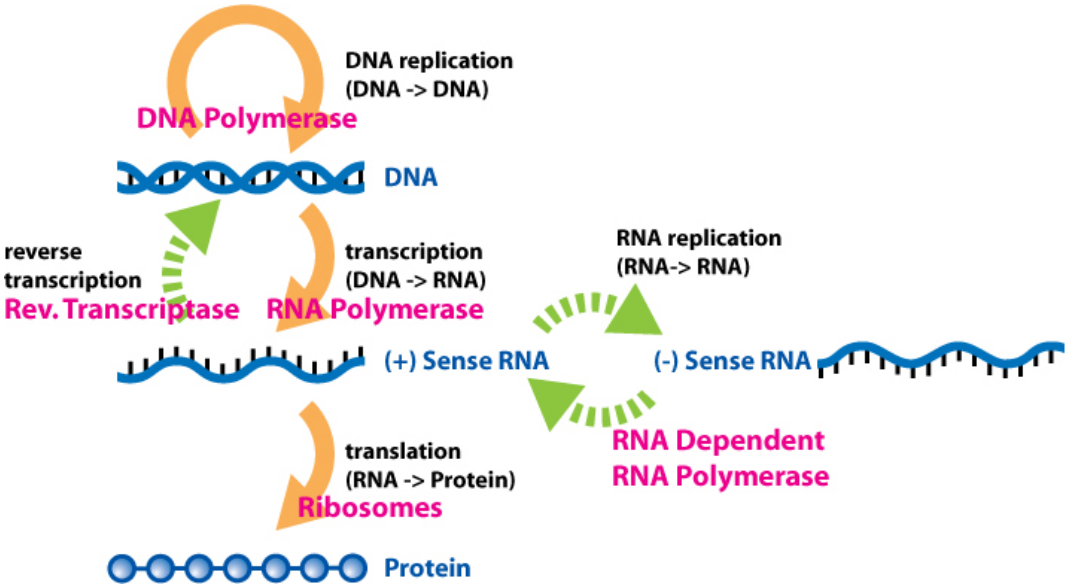


Figure 2.1: The central dogma of molecular biology (taken from Horspool, 2008). The biopolymers DNA, (+) Sense RNA, (-) Sense RNA, and proteins are depicted simplified in blue, processes translating sequential information are marked with arrows and black captions with the name of the process and the translated biopolymers, and the enzymes performing each process are marked in pink. Processes with continuous orange arrows are considered in the original formulation of the central dogma of molecular biology and processes with dashed green arrows have been added later.

Nowadays, the central dogma of molecular biology has been extended because besides DNA replication, transcription, and translation there are two additional ways in which sequential information can be transferred, namely reverse transcription for the translation of RNA to DNA and RNA replication for the translation of RNA to RNA. Gene expression is an active field of research and it has been shown that the expression of genes is regulated

at all stages. An introduction to the regulation of gene expression with emphasis on transcriptional initiation is given in subsection 2.1.2.

Enzymes are proteins which are essential for the catalysis of metabolic reactions for the transformation of small molecules into each other. Small molecules constitute the currently available building blocks of life in the cell including intermediates, storage molecules, and the basic modules for DNA, RNA, and proteins. An introduction to enzymes and small molecules is given in subsection 2.1.3

2.1.2 Regulation of gene expression and transcriptional initiation

The regulation of gene expression is vital in all living organisms as it adopts the gene expression towards cellular differentiation, environmental stimuli, and morphogenesis. This process is informally termed *gene regulation* and includes a wide range of mechanisms to fine–tune the produced amount of gene products. First described by Jacques Monod in 1961 for the lac operon in *Escherichia coli*, today's repertoire includes the transcriptional initiation, the processing of mRNAs, and post-translational modifications of proteins.

The transcriptional initiation is also known as *transcriptional regulation* and regulates the efficiency of the transcription from DNA to RNA. Hence, transcriptional initiation governs the number of produced mRNA copies in the cell which can be translated to polypeptides by translation. Important players of transcriptional initiation in procaryotes as well as eucaryotes are transcription factors (TFs) as illustrated in Figure 2.2.

TFs are proteins which show certain affinities to short DNA sequences. The degree of binding–affinity depends, amongst others, on the order of nucleobases in each DNA subsequence and DNA subsequences with a sufficient degree of binding–affinity are denoted transcription factor binding sites (TFBSs). The presence or absence of TFs at TFBSs alters the ability of the RNA polymerase to bind and translate a specific DNA region by transcription. The transcriptional regulation of genes encoding TFs results in gene regulatory networks which encode sophisticated programs of life.

The TFBSs of a specific TF show a specific base composition which depends on the TF. The pattern in this base composition is called *motif* and the discovery of motifs gives deep insights into gene regulatory networks. An introduction to the *de novo* motif discovery with phylogenetic footprinting is given in subsection 2.3.2.

2.1.3 Metabolism and small molecules

The term metabolism covers the conversion of small molecules by enzymes into each other (Pace, 2001). These small molecules are denoted metabolites and the assembly and de-

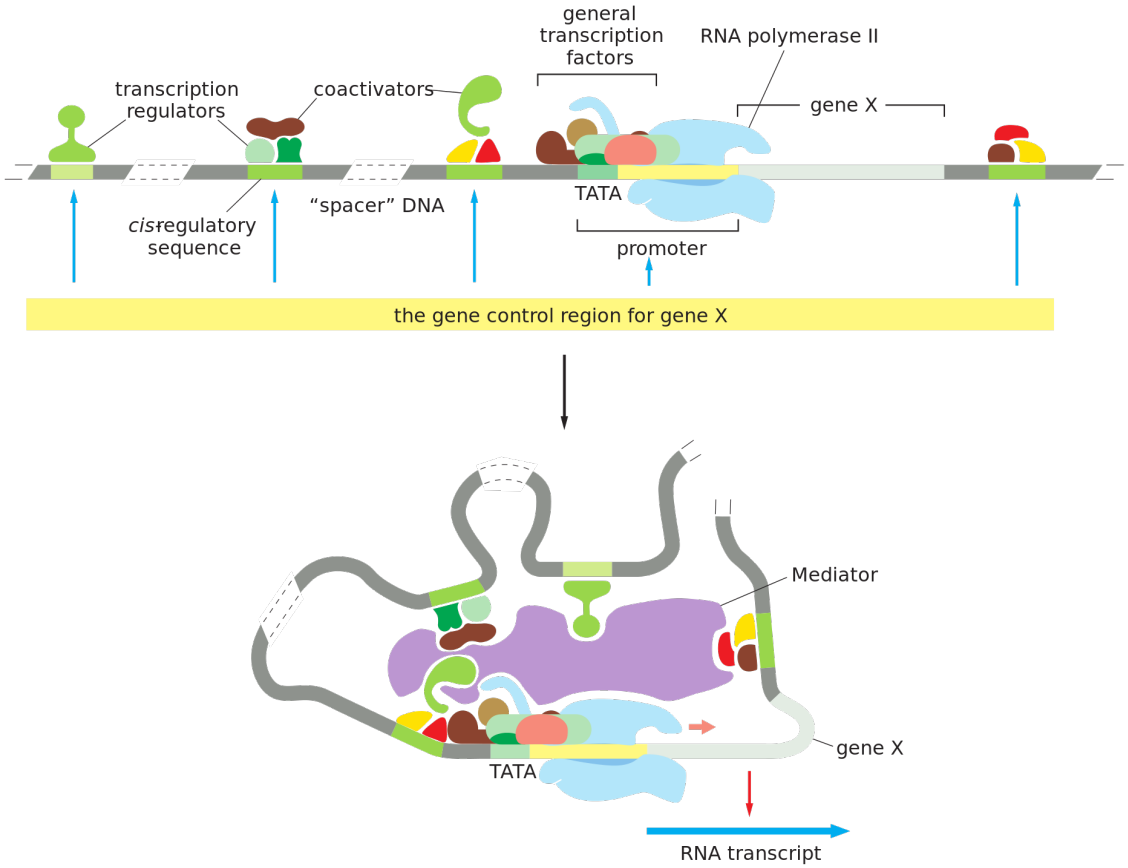


Figure 2.2: The transcriptional initiation of a typical eucaryotic gene (Figure Fig7-17 from Alberts et al., 2015). Upper part: The regulation of a representative gene "X" is governed by TFs (here referred to as general TF and transcription regulators) which bind to TFBSs (here referred to as *cis*–regulatory sequences) upstream and downstream of the gene "X". Lower part: The three–dimensional structure of the complex above is indicated. TFs and coactivators interact with different parts of the RNA polymerase II in order to start the transcription process.

pletion of metabolites is the basis of all processes in living organisms. Main tasks of the metabolism are the conversion of nutrients to ATP (the fundamental energy carrier of organisms), DNA, RNA, proteins, signaling molecules, and storage molecules as exemplified in Figure 2.3. These processes require a wealth of enzymes and intermediate metabolites.

The available set of enzymes encircles the range of possible metabolic reactions and results in a metabolic network (Ogata et al., 1999). The spread of the metabolic network is regulated by activation, deactivation, production, and depletion of enzymes depending on the developmental stage, tissue, and environmental stimuli. Metabolic networks are decom-

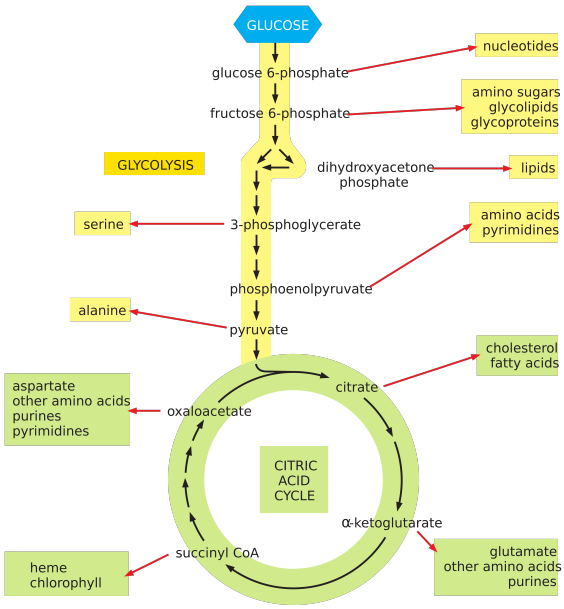


Figure 2.3: Glycolysis and the citric acid cycle are central pathways in the metabolism of most organisms (Figure Fig2-59 from Alberts et al., 2015). The glycolysis pathway and derived compound classes are indicated in yellow. The citric acid cycle and derived compound classes are indicated in green. Chemical compounds and compound classes are indicated in black and metabolic reactions are indicated with arrows, where black arrows indicate reactions of the glycolysis and the citric acid cycle and red arrows indicate outgoing metabolic pathways. The compound glucose is the input of the glycolysis and indicated in blue. The two pathways glycolysis and TCA process glucose for the production of energy and a big range of compounds.

posed somewhat arbitrarily to metabolic pathways for, amongst others, the allocation of ATP, the degradation of xenobiotics, and the production of secondary metabolites. The entirety of metabolites in the cell is called metabolome and the study of the metabolome is called *metabolomics*.

Metabolomics covers the research of metabolites using metabolite profiles to investigate the chemical fingerprints of different biological processes. An essential technology to prepare metabolite profiles is Mass Spectrometry (MS). An introduction to MS and MS data is given in subsection 2.3.3.

2.2 Computer science background

The manuscripts in this thesis touch different subjects in the field of computer science and in this section I give a brief introduction into the basics of these subjects. In subsection 2.2.1

I will introduce the Java programming language, in subsection 2.2.2 I will introduce the R programming language, in subsection 2.2.3 I will introduce databases, and in subsection 2.2.4 I will introduce data integration and ontologies.

2.2.1 The *Java* programming language

Java is a concurrent, class-based, and object-oriented computer programming language (Gosling et al., 2014). *Java* is compiled into standard bytecode that can run on any device with a *Java virtual machine* (*JVM*) in contrast to other common programming languages such as C++, Fortran, and Pascal which necessitate specific compilations for each operating system. *Java* is one of the most popular programming languages and frequently used for stand–alone applications, client–server web applications, and applets. There is a multitude of Java libraries for different communities such as *Colt* for physics, *cdk* for chemistry, and *BioJava* for bioinformatics (Wendykier et al., 2010; Steinbeck et al., 2003; Prlić et al., 2012). There is a rich set of free tools for the development of Java code such as the *eclipse* Integrated Development Environment (IDE), the *YourKit* profiler, and the unit test framework *JUnit*.

Java is used for the development of a multitude of bioinformatics tools such as Cytoscape, Ondex, and the CLC Genomics Workbench (Shannon et al., 2003; Köhler et al., 2006; Giotis et al., 2016). Such software is critical for the research in the life sciences as these implement different algorithms for the analysis of biological data. The *VANTED* system is a *Java*–based software to assist users in the analysis of wet lab data in context of biological networks. My colleagues and I contribute to the development of *VANTED* and *VANTED* add–ons in chapter 4 in sections 4.1 and 4.2.

2.2.2 The *R* programming language

R is a computer programming language for statistical computing and has broad capabilities for the preparation of publication–ready graphics (Morandat et al., 2012). *R* supports a command–line interface and is an interpreted programming language which is supported on all major operating systems. There is a multitude of standard functionality such as classical statistical tests, clustering, and multivariate analyses. The popular CRAN repository hosts *R* packages for diverse fields of application and in 12/2016 there are about $10,000$ freely available packages. There are many free tools for the development of *R* code such as the *RStudio* IDE, *devtools* for the management of packages, and *R–Forge* as a central platform for the development of *R* packages.

R is used for the development of a multitude of bioinformatics packages supporting import/export capabilities, reporting tools, and adapter for libraries from foreign program-

ming languages. Bioconductor is a free platform for open source software for the analysis of wet lab data in molecular biology and metabolomics and in 12/2016 there are more than 1,000 freely available packages (Gentleman et al., 2004). Popular *R* packages are *genefilter* for the filtering of genes from high–throughput data, *BioMart* for data integration, and *GenomicAlignments* for the handling of short genomic alignments (Bourgon et al., 2010; Durinck et al., 2005; Lawrence et al., 2013).

Many well established *R* packages are subject to continuous development to keep the packages up–to–date, to extend the functionality, and to improve the usability. First, my colleagues and I contribute to the development of the *R* package *DiffLogo* in chapter 5 in section 5.2 to enable the visual comparison of sequence motifs (see also 2.3.2). Second, my colleagues and I contribute to the development of the *R*-based web application *MetFamily* in chapter 6 in section 6.1 to enable researchers the analysis of MS data on the level of metabolite families (see also 2.3.3). Third, my colleagues and I contribute to the development of the *R* packages *xcms* and *CAMERA* in chapter 6 in section 6.2 to improve the extraction of isotope clusters from MS raw data (see also 2.3.3).

2.2.3 Databases

Databases have been invented in the 1960s for the persistent and structured storage and convenient query of a set of related data and databases are nowadays indispensable in industry and science. Access to the data is usually provided by a database management system (DBMS) which is a software for the support of diverse functionality such as database creation, database query, and database administration. Data which is managed by databases includes scientific information, customer accounts, and weather data.

There are relational databases and NoSQL ("not only SQL") databases. Relational databases are based on the relational model of data proposed by E. F. Codd in 1970 (Codd, 1970). The relational model is constituted by a set of linked tables to represent a part of the real world. With relational databases users are able to perform powerful queries using different query languages such as *Object Query Language* (*OQL*), *Structured Query Language* (*SQL*) and *XQuery*. NoSQL databases, in contrast, do not rely on the relational model of data in favor of response times, data transfer rates, and partition tolerance (Leavitt, 2010). NoSQL databases go back to the 1960s, but sophisticated developments root in the 21th century in times of big data and real–time web applications. Databases are an active field of development both in industry and science and future developments aim at new data models, concurrency control, and query optimization.

Databases in the life sciences are called *biological databases* and biological databases are popular for the structured storage and query of biological data such as genomic DNA sequences, 3D structures of proteins, and metabolite structures (Attwood et al., 2011).

The integration of biological data from databases exhibits a major challenge which is discussed in section 2.3.1.

2.2.4 Data integration and Ontologies

The continuously increasing availability of data and databases raises the problem to combine data from different sources to overcome the shortcomings of the individual data sources. The development of systems combining different databases began in the 1980s and the data integration emerged to an active field of development both in industry and science (J. M. Smith et al., 1986). The first concepts of data integration have been data warehouses which extract data from heterogeneous sources into a single view schema. Later, mediated schemas played a major role which involves the mapping of queries to the individual data sources (Lacroix, 2003). Current developments concern the semantic data integration tackling semantic conflicts in heterogeneous data sources such as different currencies, different data attributes, and ambiguous terms.

Ontology-based data integration is a well–established approach to resolve the aforementioned semantic conflicts. Ontologies allow to explicitly define the meaning of terms and their relations for a particular domain of interest in a form which is interpretable by both humans and computers (Sowa, 1995). A controlled vocabulary represents a set of terms without mutual relations and can be understood as a special case of ontology. Ontologies are applied in context of the semantic web, biomedical informatics, and library science. An introduction to the integration of biological data and ontologies in context of bioinformatics is given in subsection 2.3.1.

2.3 Bioinformatics background

The manuscripts in this thesis touch different subjects in the field of bioinformatics and in this section I give a brief introduction into the basics of these subjects. In subsection 2.3.1 I will introduce the integration of biological data, in subsection 2.3.2 I will introduce phylogenetic footprinting and phylogenetic shadowing, and in subsection 2.3.3 I will introduce MS.

2.3.1 Integration of biological data

Classical research in the life sciences apply a reductionistic approach which means that complex phenotypes are traced back to simple reasons. For instance, in case of diseases caused by mutations in a single gene such as hemophilia, sickle cell anemia, and cystic

fibrosis this approach works well. However, many problems in molecular biology such as cancer development, the increase of crop yield, and diabetes can not be reduced to such a direct cause and effect. Here, multiple factors need to be considered in frame of an integrative approach.

The field of bioinformatics faces thousands of biological databases and the integration of biological data from different sources is a prerequisite for more holistic analysis approaches (Hernandez et al., 2004). There are many systems for the integration of biological data such as *SRS*, *BioMoby*, and *KEGG* which enable life scientists to query biological data from heterogenous data sources (Etzold et al., 1996; Wilkinson et al., 2002; Ogata et al., 1999). A promising approach for the integration of biological data is the integration with biological networks. The *VANTED* system is designed to assist users in the analysis of wet lab data in context of biological networks (Junker et al., 2006). An analysis of user needs unraveled a set of missing features such as file import capabilities, mathematical methods for simulating metabolic pathways, and the support of standard graphical representations for biological entities. My colleagues and I contribute to the development of *VANTED* in chapter 4 in section 4.1 to support the analysis of biological data in context of biological networks.

Many ontologies have been designed to assist life science researchers in structuring a particular field of interest such as the *Gene Ontology* for the description of gene functions, *ChEBI* for the classification of small molecules, and the *NCBI taxonomy* for the classification of species (Ashburner et al., 2000; Hastings et al., 2013; Sayers et al., 2009). Such ontologies have been proved to support the integration of biological data and are the basis for prevalent analyses such as the enrichment analysis. However, the unified denomination of substances and meta data in biological data sets using ontologies and controlled vocabularies was not supported in *VANTED*. My colleagues and I contribute to the development of the *DBE2 information system* in chapter 4 in section 4.2 to tackle this problem.

2.3.2 Phylogenetic footprinting and phylogenetic shadowing

Phylogenetic footprinting is an approach for the identification of TFBS in DNA sequences for a set of species. In case of many closely related species this approach is called phylogenetic shadowing (Boffelli et al., 2003). The DNA sequences are preprocessed to a set of alignments, where each alignment comprises orthologous sequences from all considered species or a subset thereof. First applications of phylogenetic footprinting go back to Tagle et al. in 1988 and this approach has become increasingly applicable and lucrative as the number of published genomes increases rapidly (Robinson et al., 2011).

The key idea of phylogenetic footprinting is that functional sequences such as TFBSs are subject to negative selection during evolution. Hence, functional sequences should be less

prone to mutations compared to flanking sequences which leads to a higher degree of sequence conservation in these regions. The aim of phylogenetic footprinting is to identify these regions.

Chromatin immunoprecipitation with high–throughput DNA sequencing (*ChIP–seq*) is an essential technology for the identification of TFBSs in DNA sequences which correspond to a TF of interest (Johnson et al., 2007). However, *ChIP–seq* is subject to different sources of bias similar to many other techniques. Specifically, the binding–affinity bias leads to the prediction of distorted sequence motifs and the extraction of a biased sets of TFBSs. My colleagues and I develop a phylogenetic footprinting model which allows the estimation and the diminishment of the binding–affinity bias in *ChIP–seq* data in chapter 5 in section 5.1.

As the number of available algorithms and sequence data sets increases, it becomes a more and more important task to compare sequence motifs which have been extracted, e.g., using different algorithm parameters and sequence data from different experiments, tissues, and developmental stages (Colaert et al., 2009). These sequence motifs can be very similar to each other and can not be compared by eye. There is no tool for the comparison of sequence motifs which can handle arbitrary alphabets, which performs position–specific comparisons, and which supports a high degree of user–customization. My colleagues and I develop an appropriate tool called *DiffLogo* for the visual comparison of sequence motifs in chapter 5 in section 5.2.

2.3.3 Mass spectrometry

Mass spectrometry is an important technique to measure the mass of particles in a sample and early developments root in the late 19th century. The basis of modern mass spectrometry was laid by Arthur Jeffrey Dempster and Francis William Aston in the early 20th century (Dempster, 1918; Squires, 1998). Sector mass spectrometers have been important in the Manhatten Project to separate different isotopes of uranium and more recent developments have been awarded with multiple Nobel Prices.

Mass spectrometry is used in bioinformatics to measure the mass of thousands of small molecules and proteins in biological samples and hence represents a vital technique in both metabolomics and proteomics (Fenn et al., 1989). Measurements comprise mass spectra which contain noisy peaks for each measured molecule. For the purpose of structural elucidation these molecules are subjected to tandem mass spectrometry which involves the fragmentation of molecules in order to measure the resulting fragments in MS/MS spectra.

The analysis of mass spectra and MS/MS spectra is known as one of the major bottlenecks

in metabolomics and tools for the comprehensive analysis of MS data in a reasonable amount of time are urgently needed (Evans et al., 2009). A promising approach for the comprehensive analysis of MS data is to consider biochemically meaningful groups instead of each metabolite individually. These groups are denoted metabolite families and allow to capture biochemical patterns in a glut of measurable metabolites. My colleagues and I develop the tool *MetFamily* for the analysis of MS data on the level of metabolite families in chapter 6 in section 6.1. In addition, it is a big challenge to identify and quantify the individual metabolites in metabolite profiles and for this purpose my colleagues and I implement an approach for the extraction and validation of isotope clusters from MS data in chapter 6 in section 6.2.

2.4 Research objectives

In this chapter, I gave a brief introduction into the fields biology, computer science, and bioinformatics. In the field of bioinformatics I revealed six major challenges with considerable impact on current life science research. Two challenges are located in the field "Integration of biological data", two in the field "Phylogenetic footprinting", and two in the field "Data Processing and Interpretation of Mass Spectrometry Data".

In the field "Integration of biological data" my colleagues and I wish to improve the integration of biological data in context of biological networks using *VANTED* and to provide an ontology–assisted management of wet lab data in the *VANTED* system. In the field "Phylogenetic footprinting" my colleagues and I wish to detect and diminish the binding–affinity bias in *ChIP–seq* data and to provide a tool for the visual comparison of sequence motifs. In the field "Data Processing and Interpretation of Mass Spectrometry Data" my colleagues and I wish to improve the extraction of isotope clusters from MS raw data and to provide a freely available web application for the analysis of MS data on the level of metabolite families.

Peer-reviewed publications

This thesis is a cumulative thesis, accumulating research articles that have previously been published in peer-reviewed international journals and combining these to a thesis. The following list summarizes these publications. First authors are underlined and my name (Treutler né Mehlhorn) is marked in bold.

- Hendrik Rohn, Astrid Junker, Anja Hartmann, Eva Grafahrend-Belau, **Hendrik Treutler**, Matthias Klapperstück, Tobias Czauderna, Christian Klukas, and Falk Schreiber. VANTED v2: a framework for systems biology applications. *BMC systems biology*, 6(1):139+, November 2012. *doi:10.1186/1752-0509-6-139*

- **Hendrik Mehlhorn** and Falk Schreiber. DBE2 - management of experimental data for the VANTED system. *Journal of integrative bioinformatics*, 8(2)162+, July 2011. *doi:10.2390/biecoll-jib-2011-162*

- **Hendrik Treutler** and Steffen Neumann. Prediction, detection, and validation of isotope clusters in mass spectrometry data. *Metabolites*, 6(4):37+, October 2016. *doi:10.3390/metabo6040037*

- **Hendrik Treutler**, Hiroshi Tsugawa, Andrea Porzel, Karin Gorzolka, Alain Tissier, Steffen Neumann, and Gerd Ulrich U. Balcke. Discovering regulated metabolite families in untargeted metabolomics studies. *Analytical chemistry*, 88(16):8082-8090, August 2016. *doi:10.1021/acs.analchem.6b01569*

- Martin Nettling, **Hendrik Treutler**, Jesus Cerquides, and Ivo Grosse. Detecting and correcting the binding-affnity bias in ChIP-seq data using inter-species information. *BMC genomics*, 17(1):347+, May 2016. *doi:10.1186/s12864-016-2682-6*

- Martin Nettling, **Hendrik Treutler**, Jan Grau, Jens Keilwagen, Stefan Posch, and Ivo Grosse. DiffLogo: a comparative visualization of sequence motifs. *BMC bioinformatics*, 16(1):387+, November 2015. *doi:10.1186/s12859-015-0767-x*

I hereby declare that the copyright of the content of the articles Mehlhorn et al., 2011, Nettling, Treutler, Grau, et al., 2015, Nettling, Treutler, Cerquides, et al., 2016, Rohn et al., 2012, and Treutler and Neumann, 2016 is by the authors. These papers are available at:

- Mehlhorn et al., 2011:
  `http://journal.imbio.de/index.php?paper_id=162`

- Nettling, Treutler, Grau, et al., 2015:
  `https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0767-x`

- Nettling, Treutler, Cerquides, et al., 2016:
  https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-2682-6
- Rohn et al., 2012:
  https://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-6-139
- Treutler and Neumann, 2016:
  https://www.mdpi.com/2218-1989/6/4/37

I hereby declare that the copyright of the content of the article Treutler, Tsugawa, et al., 2016 is by the American Chemical Society (ACS). The paper is available at:

- https://pubs.acs.org/doi/abs/10.1021/acs.analchem.6b01569

3 Paper summary

This cumulative thesis covers six publications which tackle the six challenges highlighted in the introduction. As indicated in section 2.4, these publications touch three subject areas in the life sciences with two publications each, namely the "Integration of biological data" (Rohn et al., 2012; Mehlhorn et al., 2011), "Phylogenetic footprinting" (Nettling, Treutler, Cerquides, et al., 2016; Nettling, Treutler, Grau, et al., 2015), and "Data Processing and Interpretation of Mass Spectrometry Data" (Treutler and Neumann, 2016; Treutler, Tsugawa, et al., 2016). A rough classification with respect to the four omics of molecular biology, namely genomics, transcriptomics, proteomics, and metabolomics, is given in Figure 3.1.

3.1 The presented works in context

The presented publications have a potential impact on diverse research from all–over the world and on each other as each publication contributes one peace of a bigger puzzle, namely the understanding of biological systems in terms of systems biology. Systems biology aims at modelling the emergent properties of biological systems and developments in the fields "Integration of biological data", "Phylogenetic footprinting", and "Data Processing and Interpretation of Mass Spectrometry Data" are a prerequisite for this pursuit.

The subject area "Integration of biological data" aims at relating biological data of different types such as gene expression levels, protein modifications, and metabolite abundances. This is important in the life science research as data integration enables a combined data interpretation instead of interpreting each dataset individually. The subject area "Phylogenetic footprinting" is dedicated to the prediction of TFBSs in DNA sequences using sequence alignments from phylogenetically related species. These functional elements control an important step in the regulation of gene expression and the prediction of TFBSs by *de novo* motif prediction provides insights into gene regulatory networks. The subject area "Data Processing and Interpretation of Mass Spectrometry Data" involves the study of the small molecules in a tissue, organ, or organism and even these in single cells and

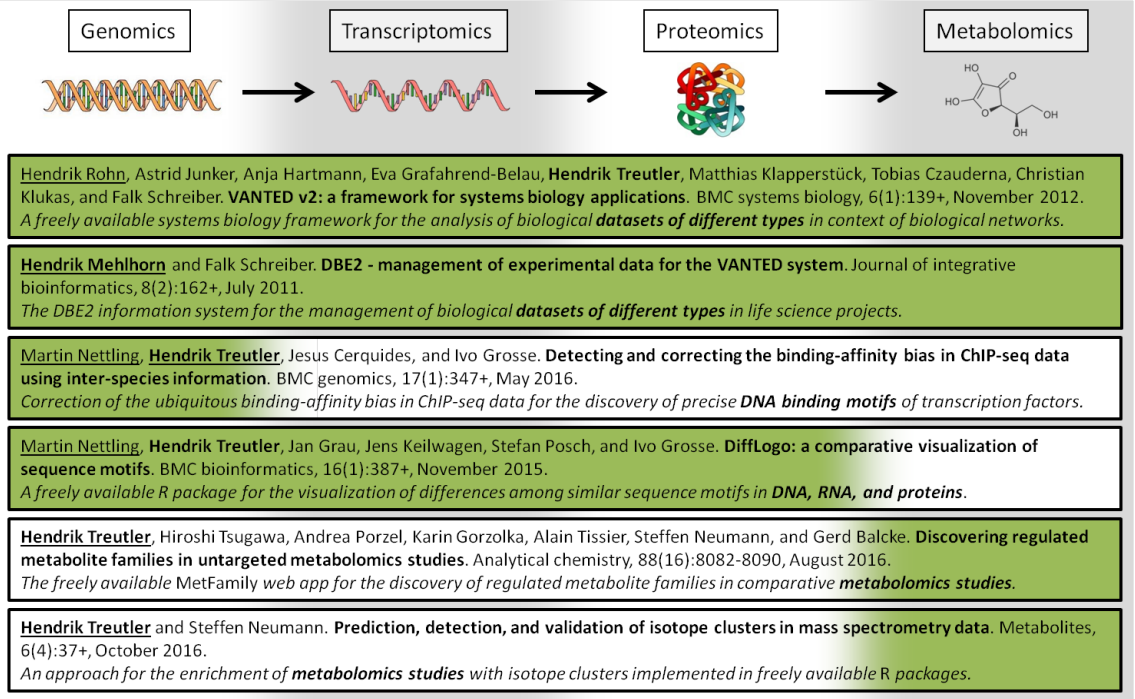


Figure 3.1: Six publications in context of four omics. The six publications of the dissertation at hand are shown in context of the four interconnected omics "Genomics", "Transciptomics", "Proteomics", and "Metabolomics" which consider the four major biological entities DNA, RNA, proteins, and metabolites respectively. Publication reference and summary is given in boxes for each publication. In the publication references authors with first–authorship are underlined and I am marked in bold. In the summaries the kind of data is marked in bold and the green sectors in the boxes indicate the appropriate kinds of omics.

exudates. Small molecules comprise metabolites, signalling molecules, and peptides and the measurement of small molecules in metabolite profiles using MS allows comprehensive snapshots of the physiology of the sample.

The analysis of biological datasets in context of biological networks is often a prerequisite in recent research and *VANTED* has proven to be a powerful framework for this task (Rohn et al., 2012). *VANTED* supports researchers in the analysis and visualization of biological data in context of biological networks to draw biological conclusions in the frame of the knowledge generation cycle in systems biology. The management of datasets in life science projects using the *DBE2 information system* decreases the technical hurdle for researchers and partners from industry to store, share, and analyze results from different experiments in an ontology–assisted way in *VANTED* (Mehlhorn et al., 2011).

An important type of data is the metabolite profile which provides a snapshot of the physiology of the sample of interest. The characterization of signals in metabolite profiles can be

improved in terms of quantification and identification using isotope clusters (Treutler and Neumann, 2016). The exhaustive extraction of isotope clusters from MS raw data using *xcms* and *CAMERA* facilitates the mapping of metabolite profiles on metabolic pathways using tools such as *VANTED*. In addition, an improved quantification of metabolites in metabolite profiles enables the determination of more precise effect sizes in comparative metabolomics studies using tools such as *MetFamily* (Treutler, Tsugawa, et al., 2016). *MetFamily* is designed for the discovery of regulated metabolite families which can provide strong biochemical hints towards differentially regulated metabolic pathways. In *VANTED* these metabolic pathways can be analyzed in context of the metabolite profile which can be stored and shared using the *DBE2 information system*. In addition, the regulation of enzymes in metabolic pathways can be essential for the interpretation of metabolite profiles and gene regulatory networks encode the entangled nature of this regulation.

DNA binding motifs of TFs are crucial for the elucidation of gene regulatory networks and *ChIP–seq* has become the major technology for the detection of these DNA binding motifs. Unfortunately, *ChIP–seq* data is distorted by the ubiquitous binding–affinity bias which leads to an artificial sharpening of the detected DNA binding motifs. The detection and correction of this bias is possible using a phylogenetic footprinting model which facilitates the construction of precise gene regulatory networks (Nettling, Treutler, Cerquides, et al., 2016). Gene regulatory networks are specific for the organism, tissue, and developmental process of interest which is reflected in slightly different DNA binding motifs. Often, such motif differences are not apparent using traditional approaches. *DiffLogo* is designed for the visual comparison of motifs in DNA, RNA, and proteins and allows the clustering of multiple motifs. This allows more mechanistic insights into the differences between gene regulatory networks under different conditions which might advance our understanding of complex biological processes (Nettling, Treutler, Grau, et al., 2015).

In summary, the interpretation of biological data in context of biological networks is a promising approach and tools such as *VANTED* are needed to fulfil this task. The *DBE information system* implements an ontology–assisted management of biological data in *VANTED* to facilitate the integration of biological data. In case of metabolite profiles, the precise quantification and identification of metabolites is a prerequisite for downstream analyses and data integration and continuously refined tools such as *xcms* and *CAMERA* approach this challenge. In addition, the web application *MetFamily* assists biologists and biochemists in the comparative analysis of metabolite profiles to gain biochemical hints such as the activation of certain enzymes and disordered metabolic pathways. *De novo* motif prediction with phylogenetic footprinting is a powerful approach for the construction of gene regulatory networks and tools implementing this approach need to incorporate certain sources of bias such as the binding–affinity bias. The extracted motifs differ under different conditions and researchers need tools such as *DiffLogo* to perform custom motif

comparisons in order to interpret findings, document work, and present results. Hence, the presented publications tackle different challenges in the life sciences and complement each other in order to gain a better understanding of biological processes in terms of systems biology.

In the next sections, I will summarize the objectives, methods, and results of the six publications in this thesis. In section 3.2 I will summarize two publications regarding the subject area "Integration of biological data", in section 3.3 I will summarize two publications regarding the subject area "Phylogenetic footprinting", and in section 3.4 I will summarize two publications regarding the subject area "Data Processing and Interpretation of Mass Spectrometry Data".

3.2 Integration of biological data

I will summarize the publication entitled "VANTED v2 – A framework for systems biology applications" in subsection 3.2.1 and I will summarize the publication entitled "DBE2 – Management of experimental data for the VANTED system" in subsection 3.2.2.

3.2.1 VANTED v2 – A framework for systems biology applications

The integration of biological data is often a requirement for the data analysis in life science projects. Here, data of different types such as genomics data, proteomics data, and metabolomics data is brought into context. My colleagues and I contributed to the development of the *VANTED* system which aims at the integration and interactive visualization of omics–data in context of biological networks such as metabolic pathways, gene regulatory networks, and signal transduction networks (Rohn et al., 2012). The mission of *VANTED* is to support biologists, biochemists, and bioinformaticians in the network–assisted interpretation of biological data sets and the preparation of publication–ready figures. Please find a reprint of this publication in chapter 4 in section 4.1.

Methods

The *VANTED* system was published in 2006 (Junker et al., 2006) and is designed as a framework for diverse functionality including statistical tests for data analysis, algorithms for graph layout, and interactive panels for the graphical user interface. Functional modules are encapsulated in plug–ins and add–ons which can be added during run–time. The code base of *VANTED* is written in *Java* and has been restructured in the frame of this publication to facilitate the maintainability and further developments.

VANTED supports seven main tasks, namely i) the import of experiment data and biological networks, ii) the visualization of experiment data in context of biological networks in a standardized graphical representation (Systems Biology Graphical Notation (SBGN)), iii) the integration of data from different types and different sources, iv) the simulation of mathematical models representing biological networks, v) the exploration of and interaction with experiment data in context of biological networks, vi) the topological analysis of biological networks and the statistical analysis of experiment data, and vii) the export of results to publication–ready figures and the export of network data and experiment data to computer–readable formats. The execution of these tasks requires diverse techniques regarding different programming languages, different concepts for user interaction, data management, and more.

Results, Discussion, and conclusions

VANTED has been widely used for the analysis of biological data in context of biological networks (Dongen et al., 2009; Grafahrend-Belau et al., 2009; Zurbriggen et al., 2009). In 12/2016, Google Scholar[1] reports 347 citations for the original paper from 2006 and 72 citations for the presented paper from 2012. An exemplary use case is the visualization of metabolite measurements and mRNA transcript abundances in the context of metabolic pathways from KEGG in SBGN. This use case integrates different types of experimental data with pathway information from KEGG in a standardized graphical representation which enables to ask new biological questions in the frame of life science projects. The *VANTED* system is open–source[2] and freely available at `http://www.vanted.org/`. *VANTED* is available for all major operating systems and includes data input templates for wet lab data, support for different biological network databases such as KEGG (Ogata et al., 1999), Biomodels (Le Novere et al., 2006), and MetaCrop (Schreiber et al., 2011), and further documentation.

Potential future developments concern the analysis of biological data in context of multiple biological networks from different domains such as protein–protein interaction networks, co–expression networks, and metabolic networks. Users will need efficient data handling strategies, additional algorithms, and an intuitive user interface for this purpose. A part of these requirements is currently under development.

[1] Google Scholar: `https://scholar.google.de/`
[2] VANTED source code hosted on BitBucket: `https://bitbucket.org/vanted-dev/vanted/`

3.2.2 DBE2 – Management of experimental data for the VANTED system

The management of datasets is crucial in most life science projects because wet lab data from different contributors needs to be exchanged in order to share and combine knowledge from different experiments. My colleagues and I developed the *DBE2 information system* (DBE2 = **D**atabase for **B**iological **E**xperiments **2**) for the management of experimental data for the *VANTED* system (Mehlhorn et al., 2011). The *DBE2 information system* is easy–to–use and supports user–rights management, ontology–controlled vocabularies, worldwide accessibility, and the opportunity to load, save, and edit the data. The mission of DBE2 is to support researchers in the sharing, ontology–assisted integration, and analysis of biological datasets in *VANTED*. Please find a reprint of this publication in chapter 4 in section 4.2.

Methods

We designed the *DBE2 information system* as a three-tier architecture consisting of a presentation tier interacting with the user, a logic tier implementing the exchange of data, and a data tier for the persistent storage of data.

We implemented the presentation tier in *Java* and denote it *DBE2 client*. The *DBE2 client* is a *VANTED* add–on for a seamless integration into *VANTED*. The *DBE2 client* supports a graphical user interface and users with just basic technical capabilities are able to load, save, share, and edit experiment data.

We implemented the logic tier as a *Java servlet* and denote it *DBE2 servlet*. The *DBE2 servlet* was installed at the IPK Gaterleben in 2010. The *DBE2 servlet* implements the download, save, and edit of experiment data including a user right management with user groups.

We implemented the data tier as a Oracle–database (version 11g) and denote it *DBE2 database*. The *DBE2 database* was installed at the IPK Gaterleben in 2010. The relational database schema of the *DBE2 database* represents data for user management, experiment data, and basis data. The user management data represents user accounts and associates user accounts with user groups in which users are able to share experiment data. The experiment data is organized in a tree–structure in four levels to represent data of different types. The four levels represent experiments, conditions, samples, and measurements where the first three levels comprise experiment meta data and the fourth level represents floating-point numbers, pictures, volumes, networks, and gradients. The basis data represents a controlled vocabulary of unified terms of general interest such as species names,

measurement units, and substance names. Hence, experiment data can be aggregated for certain species or measured substances which eases the integration of this data.

Names of chemical compounds and the corresponding taxonomy of compound classes are accessible from the ontology Chemical Entities of Biological Interest (CHEBI) (Hastings et al., 2013). Species names and the corresponding species taxonomy are accessible from the ontology *NEWT* UniProt Taxonomy Database (NEWT) (Phan et al., 2003). Both ontologies are retrieved in real time from the Ontology Lookup Service (OLS) which is a compendium of more than 150 life science ontologies accessible through an unified interface[1] (Côté et al., 2010).

Results, Discussion, and conclusions

The *DBE2 information system* was published in July 2011 and reported 43 registered users and 73 stored datasets. The *DBE2 information system* has been used several years in different life science projects with users in science and industry. An exemplary use case could be that a researcher from Australia uploads metabolite measurements to the *DBE2 database*, a project collaborator in the United States downloads this data, adds protein abundance measurements, and uploads the complemented data in turn, and a project collaborator in Germany downloads this data, integrates it into metabolic pathways, and draws biological conclusions. Personal interrogations with users indicated that the functionality of the *DBE2 information system* is satisfiable and that some adaptions of the user interface could improve the user experience. The *DBE2 information system* is currently unavailable for the *VANTED* system since *VANTED* of version 2.3 due to extensive code changes in the *VANTED* framework which break the code compatibility to the *DBE2 client*.

3.3 Phylogenetic footprinting

I will summarize the publication entitled "Detecting and correcting the binding–affinity bias in ChIP-seq data using inter–species information" in subsection 3.3.1 and I will summarize the publication entitled "DiffLogo: a comparative visualization of sequence motifs" in subsection 3.3.2.

[1] Ontology Lookup Service at EMBL-EBI: `https://www.ebi.ac.uk/ols/`

3.3.1 Detecting and correcting the binding–affinity bias in ChIP-seq data using inter–species information

The prediction of TFBSs and their motifs is essential for understanding transcriptional gene regulation. *ChIP–seq* has become the major technology to uncover genomic regions containing those binding sites, but motifs predicted by traditional computational approaches using these data are distorted by the ubiquitous binding–affinity bias. My colleagues and I contributed to the detection and correction of the binding–affinity bias in *ChIP–seq* data using inter-species information (Nettling, Treutler, Cerquides, et al., 2016). The objective of this work is the development of a phylogenetic footprinting model which is capable of verifing the presence of binding–affinity bias in *ChIP–seq* data and to extract corrected motifs from this data. Please find a reprint of this publication in chapter 5 in section 5.1.

Methods

We suppose that the binding–affinity bias can not be detected based only on *ChIP–seq* data from the reference species, but it can be detected using *ChIP–seq* data from the reference species and orthologous sequences from phylogenetically related species. The key idea is that the effect of the binding–affinity bias is strong in the reference species and that this effect decreased in phylogenetically related species due to mutations which allows the quantification of the binding–affinity bias. A toy example can be found in the manuscript in section "Using sequence information of phylogenetically related species to detect the binding–affinity bias".

We proposed a new phylogenetic footprinting model for the quantification and correction of the effect of the binding–affinity bias from alignments of TFBSs and flanking sequences. The *Java* implementation of this model is freely available[1]. We estimated the proposed phylogenetic footprinting model using a modified EM algorithm. The binding–affinity bias was quantified using estimates of a special parameter for the reference species which represents an inverse temperature derived from the Boltzmann distribution from thermodynamics. We measured the prediction accuracy of our model using the classification performance, namely 100 fold stratified repeated random sub-sampling validation. We used *DiffLogo* for the visualization of motif differences between traditional and corrected motifs (Nettling, Treutler, Grau, et al., 2015).

[1] *PhyFoo* source code on GitHub: `https://github.com/mgledi/PhyFoo`

Results, Discussion, and conclusions

We analyzed alignments of human *ChIP–seq* positive regions of five TFs with orthologous sequences from four other species. Based on these datasets we observed that the binding–affinity bias is reflected by an increased information content in motifs extracted from the reference species and that this increase of information content fades in other species proportional to the phylogenetic distance to the reference species. Hence, the binding–affinity bias leads to an artificial sharpening of the extracted motif and we illustrated this effect using *DiffLogo* (Nettling, Treutler, Grau, et al., 2015). We showed that correcting the binding–affinify bias improves motif prediction as quantified using the classification performance, i.e. the accuracy to discriminate motif–bearing alignments from non–motif–bearing alignments.

We suppose that present motifs in databases and literature are artificially sharpened versions of the true motif which potentially distorts our understanding of gene regulation. For example, corrected motifs can be used for the precise *in silico* prediction of binding sites, whereas motifs distorted by the binding–affinity bias would lead to imprecise *in silico* predictions. The refinement of motifs from databases and literature might lead to the prediction of novel binding sites, cis-regulatory modules, or gene-regulatory networks which might advance our understanding of transcriptional gene regulation as a whole.

3.3.2 DiffLogo: a comparative visualization of sequence motifs

The increasing amount of sequence data, motif prediction algorithms, and published motifs involves the demand and opportunity to compare different sequence motifs such as TFBSs, splice sites in pre–mRNAs, and phosphorylation sites in proteins. The comparison of sequence motifs using the *de facto* standard *sequence logo* (Schneider et al., 1990) is often difficult in case of highly similar motifs which have been extracted using different algorithms or from different samples. My colleagues and I developed *DiffLogo* for the intuitive visualization of subtle differences among similar sequence motifs such as binding patterns of one TF in different cell types, binding patterns of different TFs, and protein domains in different species (Nettling, Treutler, Grau, et al., 2015). The mission of *DiffLogo* is to support researchers in the comparative motifs analysis in order to interpret findings, document work, share knowledge, and present results. Please find a reprint of this publication in chapter 5 in section 5.2.

3. PAPER SUMMARY

Methods

The proposed graphical representation of motif differences is inspired by the well–known *sequence logo* and denoted *difference logo*. A *difference logo* of two motifs depicts for each motif position a stack of symbols representing DNA bases, RNA bases, or amino acids. The height of each stack is proportional to the difference between the symbol distributions in the respective motif position of both motifs.

By default, the Jensen–Shannon entropy is used for the quantification of position–specific distribution differences and the height of each base is proportional to the probability difference. However, the calculation of stack height and base height is customizable by the user with custom measures to enable the usage of *difference logos* in context of different biological questions. *DiffLogo* is implemented as an *R* package and users with moderate experience in *R* should be able to customize *DiffLogo* to user–specific demands.

The comparison of N motifs is achieved by a table with $N \times (N-1)$ pair–wise *difference logos*. The rows and columns of the table are ordered by a optimal leaf–ordered cluster tree to place similar motifs next to each other. The overall motif differences are displayed by a color–gradient of the background color of each *difference logo*.

Results, Discussion, and conclusions

Motifs from the domains genomics, transcriptomics, and proteomics can be compared using position weight matrices (PWMs) or sequence alignments. We compared DNA motifs from the human insulator CTCF from different cell lines using *DiffLogo* and successfully reproduced literature findings. In addition, we presented specific differences between F-box protein–protein binding domains from the three kingdoms "metazoa", "fungi", and "viridiplantae".

The proposed approach is implemented in the open–source *R* package *DiffLogo* and freely available at Bioconductor[1]. The package comprises example data, example code, and further documentation. In 2016, the *DiffLogo* package was downloaded approximately 150 times per month in average which suggests that a remarkable number of researchers can profit from the proposed approach.

A prerequisite for the usage of *DiffLogo* are basic capabilities in the *R* programming language, but we are aware that many researchers in the life sciences lack this requirement. In addition, the proposed version of *DiffLogo* does not properly align motifs with phase shifts or DNA motifs with reverse complements. Hence, my colleagues and I implemented

[1] *DiffLogo* source code on Bioconductor:
`https://www.bioconductor.org/packages/release/bioc/html/DiffLogo.html`

the web–server *WebDiffLogo* for the construction and visualization of multiple motif alignments[1][2].

3.4 Data Processing and Interpretation of Mass Spectrometry Data

I will summarize the publication entitled "Discovering Regulated Metabolite Families in Untargeted Metabolomics Studies" in subsection 3.4.1 and I will summarize the publication entitled "Prediction, Detection, and Validation of Isotope Clusters in Mass Spectrometry Data" in subsection 3.4.2.

3.4.1 Discovering Regulated Metabolite Families in Untargeted Metabolomics Studies

The elucidation of metabolic processes provides deep insights into complex processes in the cell and MS is a key technology for the identification and quantification of metabolites in biological samples. The measurement of fragment spectra is often crucial for the characterization of the molecular structure of metabolites. Usually, the structural elucidation of metabolites is done one–by–one and known as one of the major bottlenecks in metabolomics.

My colleagues and I developed an approach for the discovery of *metabolite families* from fragment spectra. Metabolite families are classes of biochemically related metabolites and the grouping of metabolites in metabolite families breaks down the wealth of measurable metabolites to meaningful classes. A subset of these metabolite families is differentially regulated in different samples and we denote these metabolite families *regulated metabolite families*. Regulated metabolite families provide a birds eye view in comparative metabolomics studies (Treutler, Tsugawa, et al., 2016). The proposed approach is implemented in a freely–available web app denoted *MetFamily*. The mission of *MetFamily* is to support researchers with limited technical capabilities in the comparative analysis of MS data on the level of metabolite families. Please find a reprint of this publication in chapter 6 in section 6.1.

[1] *WebDiffLogo* source code on GitHub:
`https://github.com/mgledi/DiffLogoUI`
[2] *WebDiffLogo* web page:
`http://difflogo.com`

3. PAPER SUMMARY

Methods

The input is a metabolite profile comprising the abundance of metabolites in different samples and a fragment library comprising fragment spectra of these metabolites. Hierarchical cluster analysis on fragment spectra is used to structurally relate metabolites for the detection of biochemically related metabolites denoted metabolite families. Principal component analysis on metabolite abundances is used for the detection of group–discriminating metabolites. The combination of both orthologous analyses allows the detection of metabolite families with differential regulation in different samples denoted regulated metabolite families.

The proposed approach is implemented in the easy–to–use web app *MetFamily*. It is recommended to obtain the metabolite profile and the fragment library from the raw data using the tool *MS-DIAL* (Tsugawa et al., 2015), but also other data sources have been successfully used as documented in the *MetFamily input specification*[1].

MetFamily is implemented as a Shiny web application in *R* version 3.3. The Shiny server hosting the Shiny web app is packaged in a docker container. Multi–user support is enabled by multiple docker instances which are orchestrated using *Docker Compose*.

Results, Discussion, and conclusions

The proposed approach was applied in a study comparing tomato trichomes with trichome–free leaves. We discovered two regulated metabolite families which were specific to trichomes, namely *branched chain acyl sugars* and *sesquiterpene glucosides*. Interestingly, the observed diversity of 73 acyl sugars in trichomes illustrates the low substrate specificity of BAHD acyltransferases which are upregulated in tomato glandular trichomes. Hence, we uncovered links between enzymatic promiscuity and organ-specific regulation of enzymes using *MetFamily*.

MetFamily moves biochemical questions in the centre of attention and preserves the user not to see the wood for the trees. The proposed analysis pipeline decreases the workload for the analysis from several days to a few hours which relaxes one of the major bottlenecks in metabolomics. The proposed approach is implemented in the easy–to–use web app *MetFamily* which is open–source[2] and freely available[3]. During data import there are user–customizable parameters which allow the adaption of *MetFamily* to data from different

[1] *MetFamily* input specification:
`http://pubs.acs.org/doi/suppl/10.1021/acs.analchem.6b01569/suppl_file/ac6b01569_si_004.pdf`
[2] *MetFamily* source code on GitHub: `https://github.com/Treutler/MetFamily`
[3] *MetFamily* web page: `http://msbi.ipb-halle.de/MetFamily/`

instruments. There is documentation for the functions of *MetFamily*[1].

Future developments concern the automated detection of metabolite families from fragment spectra to allow the halve–automated annotation of metabolite families in the *MetFamily* web app. Researchers without expert knowledge in the interpretation of fragment spectra would strongly benefit from this feature. In addition, a collaboration with the researchers behind the related tool *MS2LDA*[2][3] (Hooft et al., 2016) might be beneficial for both approaches could complement each other.

3.4.2 Prediction, Detection, and Validation of Isotope Clusters in Mass Spectrometry Data

The detection of isotope clusters allows an improved identification and quantification of metabolites in MS data. Isotope clusters enable precise predictions of the molecular formula of metabolites which is crucial for the structural elucidation. In addition, isotope clusters allow an improved quantification of metabolites and database searches with high precision. My colleagues and I developed an approach for the prediction, detection, and validation of isotope clusters in MS data (Treutler and Neumann, 2016). This approach can be applied to liquid chromatography–high resolution MS data and can easily extend existing analysis pipelines based on the frequently–used *R* packages *xcms* and *CAMERA*. The purpose of the proposed approach is an exhaustive extraction of reliable isotope clusters from MS data in order to improve the identification and quantification of metabolites. Please find a reprint of this publication in chapter 6 in section 6.2.

Methods

After measurement using mass spectrometers, feature detection algorithms perform peak picking to extract basic properties about peaks in the raw data such as mass–to–charge ratio, retention time, and peak height. A prerequisite for the prediction of isotope–signals in the raw data is a traditional peak picking which results in a set of peaks. The location of putative isotope peaks is predicted on basis of these peaks using chemical prior knowledge about the composition of isotope clusters. Additional isotope peaks are extracted in a second peak picking step using the feature detection algorithm *centWave* (Tautenhahn et al., 2008) given the predicted isotope peaks.

[1] *MetFamily* user guide:
`http://pubs.acs.org/doi/suppl/10.1021/acs.analchem.6b01569/suppl_file/ac6b01569_si_003.pdf`
[2] *MS2LDA* web page: `http://ms2lda.org/`
[3] *MS2LDA* source code on GitHub: `https://github.com/sdrogers/MS2LDA`

The subsequent detection of isotope clusters involves the arrangement of isotope peaks to putative isotope clusters. Putative isotope clusters are validated depending on the metabolite mass based on database statistics. These database statistics are compiled on compound databases such as ChEBI, KEGG, and PubChem to capture the typical distribution of isotope peaks from compounds of different mass ranges.

The proposed approach for the prediction, detection, and validation of isotope clusters is implemented in *R* version 3.3. The prediction and detection of isotope peaks has been integrated into the popular *R* package *xcms* (C. A. Smith et al., 2006) and the detection and validation of isotope clusters has been integrated into the commonly used *R* package *CAMERA* (Kuhl et al., 2011).

Results, Discussion, and conclusions

We extracted 37% more isotope peaks from measurements of *Arabidopsis thaliana* extracts. In a mix of standard compounds we predicted the correct molecular formula in 92% of the cases in the top three ranks. Statistics on different compound databases suggested that a mass–dependent validation of isotope clusters is more precise compared to existing approaches for the validation of isotope clusters.

The proposed approach is implemented in the open–source *R* packages *xcms* version 1.50.0 and *CAMERA* version 1.30.0 and is freely available at Bioconductor[1][2]. Both packages comprise example code and further documentation including code and documentation for the proposed approach. In 2016, the average number of package downloads per month was more than 1500 in case of the *xcms* package and more than 600 in case of the *CAMERA* package which suggests that a remarkable number of researchers can already profit from the proposed approach. In 12/2016, Google Scholar[3] reports 1693 citations for the *xcms* paper from 2006 and 205 citations for the *CAMERA* paper from 2011.

Future developments concern the adaption of the proposed approach for the prediction and detection of further related signals in MS data such as adducts and neutral losses. Especially the exhaustive extraction of signals from different adducts promises an improved detection of the neutral mass of metabolites which is a prerequisite for the identification of metabolites from MS data. Currently, the *xcms* package is restructured for version 3.0 to establish a flexible basis for future developments[4].

[1] *xcms* source code on Bioconductor:
`https://www.bioconductor.org/packages/release/bioc/html/xcms.html`
[2] *CAMERA* source code on Bioconductor:
`https://www.bioconductor.org/packages/release/bioc/html/CAMERA.html`
[3] Google scholar: `https://scholar.google.de/`
[4] *xcms 3.0* source code on GitHub: `https://github.com/Treutler/xcms/tree/xcms3`

Glossary

ChIP-seq (high-throughput) sequencing of immunoprecipitated chromatin.

DNA Deoxyribonucleic acid.

IDE Integrated Development Environment.

mRNA messenger RNA.

MS Mass Spectrometry.

RNA Ribonucleic acid.

SBGN Systems Biology Graphical Notation.

TF transcription factor.

TFBS transcription factor binding site.

References

Alberts, B., A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, and P. Walter (2015). *Molecular Biology of the Cell, Sixth Edition*. Garland Publishing. Garland Science.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25 (1), pp. 25–29.

Attwood, T. K., A. Gisel, N. E. Eriksson, and E. Bongcam-Rudloff (2011). "Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective". In: *Bioinformatics - Trends and Methodologies*. Ed. by Mahdavi. InTech. Chap. 1.

Berg, J.M., J.L. Tymoczko, and L. Stryer (2002). *Biochemistry, Fifth Edition*. W.H. Freeman.

Boffelli, Dario, Jon McAuliffe, Dmitriy Ovcharenko, Keith D. Lewis, Ivan Ovcharenko, Lior Pachter, and Edward M. Rubin (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science (New York, N.Y.)* 299 (5611), pp. 1391–1394.

Bourgon, Richard, Robert Gentleman, and Wolfgang Huber (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, 107 (21), pp. 9546–9551.

Chambers, Geoffrey K., Caitlin Curtis, Craig D. Millar, Leon Huynen, and David M. Lambert (2014). DNA fingerprinting in zoology: past, present, future. *Investigative genetics*, 5 (1), p. 3.

Codd, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. *Commun. ACM*, 13 (6), pp. 377–387.

Colaert, Niklaas, Kenny Helsens, Lennart Martens, Joel Vandekerckhove, and Kris Gevaert (2009). Improved visualization of protein consensus sequences by iceLogo. *Nature Methods*, 6 (11), pp. 786–787.

Côté, Richard, Florian Reisinger, Lennart Martens, Harald Barsnes, Juan Antonio Vizcaino, and Henning Hermjakob (2010). The ontology lookup service: bigger and better. *Nucleic acids research*, 38 (suppl 2), W155–W160.

Crick, Francis (1970). Central Dogma of Molecular Biology. *Nature*, 227 (5258), pp. 561–563.

Crick, Francis HC (1958). "On protein synthesis". In: *Symp Soc Exp Biol.* Vol. 12. 138-63, p. 8.

Darwin, Charles (1859). *On the origin of species.*

Dempster, A. J. (1918). A new Method of Positive Ray Analysis. *Phys. Rev.* 11, pp. 316–325.

Dongen, Joost T. van, Anja Fröhlich, Santiago J. Ramirez-Aguilar, Nicolas Schauer, Alisdair R. Fernie, Alexander Erban, Joachim Kopka, Jeremy Clark, Anke Langer, and Peter Geigenberger (2009). Transcript and metabolite profiling of the adaptive response to mild decreases in oxygen concentration in the roots of arabidopsis plants. *Annals of Botany*, 103 (2), pp. 269–280.

Durinck, Steffen, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics (Oxford, England)*, 21 (16), pp. 3439–3440.

Etzold, T., A. Ulyanov, and P. Argos (1996). SRS: information retrieval system for molecular biology data banks. *Methods in enzymology*, 266, pp. 114–128.

Evans, Anne M., Corey D. DeHaven, Tom Barrett, Matt Mitchell, and Eric Milgram (2009). Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Analytical chemistry*, 81 (16), pp. 6656–6667.

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

Gentleman, Robert C, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5 (10), p. 1.

Giotis, Efstathios S, Rebecca C Robey, Natalie G Skinner, Christopher D Tomlinson, Stephen Goodbourn, and Michael A Skinner (2016). Chicken interferome: avian interferon-stimulated genes identified by microarray and RNA-seq of primary chick embryo fibroblasts treated with a chicken type I interferon (IFN-α). *Veterinary Research*, 47 (1), p. 75.

Godfrey, M. D. and D. F. Hendry (1993). The Computer As Von Neumann Planned It. *IEEE Ann. Hist. Comput.* 15 (1), pp. 11–21.

Gosling, James, Bill Joy, Guy L. Steele, Gilad Bracha, and Alex Buckley (2014). *The Java Language Specification, Java SE 8 Edition*. 1st. Addison-Wesley Professional.

Grafahrend-Belau, Eva, Falk Schreiber, Dirk Koschützki, and Björn H. Junker (2009). Flux Balance Analysis of Barley Seeds: A Computational Approach to Study Systemic Properties of Central Metabolism. *Plant Physiology*, 149 (1), pp. 585–598.

Hastings, Janna, Paula de Matos, Adriano Dekker, Marcus Ennis, Bhavana Harsha, Namrata Kale, Venkatesh Muthukrishnan, Gareth Owen, Steve Turner, Mark Williams, and Christoph Steinbeck (2013). The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic acids research*, 41 (Database issue), pp. D456–D463.

Hernandez, Thomas and Subbarao Kambhampati (2004). Integration of Biological Sources: Current Systems and Challenges Ahead. *SIGMOD Rec.* 33 (3), pp. 51–60.

Hogeweg, Paulien (2011). The Roots of Bioinformatics in Theoretical Biology. *PLOS Computational Biology*, 7 (3), pp. 1–5.

Hooft, Justin Johan Jozias van der, Joe Wandy, Michael P Barrett, Karl EV Burgess, and Simon Rogers (2016). Topic modeling for untargeted substructure exploration in metabolomics. *Proceedings of the National Academy of Sciences*, 113 (48), pp. 13738–13743.

Horspool, Daniel (2008). *File:Extended Central Dogma with Enzymes.jpg*. An overview of the central dogma of molecular biochemistry with all unusual flows of information included (in green). URL: `https://commons.wikimedia.org/wiki/File:Extended_Central_Dogma_with_Enzymes.jpg`.

Johnson, David S., Ali Mortazavi, Richard M. Myers, and Barbara Wold (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.)* 316 (5830), pp. 1497–1502.

Junker, Björn H., Christian Klukas, and Falk Schreiber (2006). VANTED: a system for advanced data analysis and visualization in the context of biological networks. *BMC bioinformatics*, 7 (1), pp. 109+.

Kamionka, Mariusz (2011). Engineering of therapeutic proteins production in Escherichia coli. *Current pharmaceutical biotechnology*, 12 (2), pp. 268–274.

Köhler, Jacob, Jan Baumbach, Jan Taubert, Michael Specht, Andre Skusa, Alexander Rüegg, Chris Rawlings, Paul Verrier, and Stephan Philippi (2006). Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, 22 (11), pp. 1383–1390.

Kuhl, Carsten, Ralf Tautenhahn, Christoph Bottcher, Tony R Larson, and Steffen Neumann (2011). CAMERA: an integrated strategy for compound spectra extraction and

annotation of liquid chromatography/mass spectrometry data sets. *Analytical chemistry*, 84 (1), pp. 283–289.

Lacroix, Zoe (2003). *Bioinformatics: managing scientific data*. Academic Press.

Lawrence, Michael, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey (2013). Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology*, 9 (8), pp. 1–10.

Le Novere, Nicolas, Benjamin Bornstein, Alexander Broicher, Melanie Courtot, Marco Donizelli, Harish Dharuri, Lu Li, Herbert Sauro, Maria Schilstra, Bruce Shapiro, et al. (2006). BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic acids research*, 34 (suppl 1), pp. D689–D691.

Leavitt, Neal (2010). Will NoSQL Databases Live Up to Their Promise? *Computer*, 43 (2), pp. 12–14.

Mehlhorn, Hendrik and Falk Schreiber (2011). DBE2 - management of experimental data for the VANTED system. *Journal of integrative bioinformatics*, 8 (2), pp. 162+.

Morandat, Floréal, Brandon Hill, Leo Osvald, and Jan Vitek (2012). "Evaluating the Design of the R Language: Objects and Functions for Data Analysis". In: *Proceedings of the 26th European Conference on Object-Oriented Programming*. ECOOP'12. Beijing, China: Springer-Verlag, pp. 104–131.

Nettling, Martin, Hendrik Treutler, Jesus Cerquides, and Ivo Grosse (2016). Detecting and correcting the binding-affinity bias in ChIP-seq data using inter-species information. *BMC genomics*, 17 (1), pp. 347+.

Nettling, Martin, Hendrik Treutler, Jan Grau, Jens Keilwagen, Stefan Posch, and Ivo Grosse (2015). DiffLogo: a comparative visualization of sequence motifs. *BMC bioinformatics*, 16 (1), pp. 387+.

Ogata, Hiroyuki, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27 (1), pp. 29–34.

Pace, Norman R (2001). The universal nature of biochemistry. *Proceedings of the National Academy of Sciences*, 98 (3), pp. 805–808.

Phan, IQH, Sandrine F Pilbout, Wolfgang Fleischmann, and Amos Bairoch (2003). NEWT, a new taxonomy portal. *Nucleic Acids Research*, 31 (13), pp. 3822–3823.

Prlić, Andreas, Andrew Yates, Spencer E. Bliven, Peter W. Rose, Julius Jacobsen, Peter V. Troshin, Mark Chapman, Jianjiong Gao, Chuan Hock H. Koh, Sylvain Foisy, Richard Holland, Gediminas Rimsa, Michael L. Heuer, H. Brandstätter-Müller, Philip E. Bourne, and Scooter Willis (2012). BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics (Oxford, England)*, 28 (20), pp. 2693–2695.

Robinson, James T., Helga Thorvaldsdottir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov (2011). Integrative genomics viewer. *Nature Biotechnology*, 29 (1), pp. 24–26.

Rohn, Hendrik, Astrid Junker, Anja Hartmann, Eva Grafahrend-Belau, Hendrik Treutler, Matthias Klapperstück, Tobias Czauderna, Christian Klukas, and Falk Schreiber (2012). VANTED v2: a framework for systems biology applications. *BMC systems biology*, 6 (1), pp. 139+.

Sanger, F. and A. R. Coulson (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94 (3), pp. 441–448.

Sawicki, Mark P, Ghassan Samara, Michael Hurwitz, and Edward Passaro (1993). Human genome project. *The American journal of surgery*, 165 (2), pp. 258–264.

Sayers, Eric W., Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Michael Feolo, Lewis Y. Geer, Wolfgang Helmberg, Yuri Kapustin, David Landsman, David J. Lipman, Thomas L. Madden, Donna R. Maglott, Vadim Miller, Ilene Mizrachi, James Ostell, Kim D. Pruitt, Gregory D. Schuler, Edwin Sequeira, Stephen T. Sherry, Martin Shumway, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Tatiana A. Tatusova, Lukas Wagner, Eugene Yaschenko, and Jian Ye (2009). Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, 37 (Database issue), pp. D5–15.

Schneider, Thomas D and R Michael Stephens (1990). Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, 18 (20), pp. 6097–6100.

Schreiber, Falk, Christian Colmsee, Tobias Czauderna, Eva Grafahrend-Belau, Anja Hartmann, Astrid Junker, Björn H Junker, Matthias Klapperstück, Uwe Scholz, and Stephan Weise (2011). MetaCrop 2.0: managing and exploring information about crop plant metabolism. *Nucleic acids research*, 40 (D1), p. D1173.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13 (11), pp. 2498–2504.

Shockley, William (1952). Transistor electronics: Imperfections, unipolar and analog transistors. *Proceedings of the IRE*, 40 (11), pp. 1289–1313.

Smith, Colin A, Elizabeth J Want, Grace O'Maille, Ruben Abagyan, and Gary Siuzdak (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical chemistry*, 78 (3), pp. 779–787.

Smith, John M., Philip A. Bernstein, Umeshwar Dayal, Nathan Goodman, and Terry Landers (1986). "Multibase&Mdash;Integrating Heterogeneous Distributed Database Sys-

tems". In: *AFIPS Conference Proceedings; Vol. 55 1986 National Computer Conference*. Las Vegas, Nevada, USA: AFIPS Press, pp. 335–347.

Solomon, E.P., D.W. Martin, and L.R. Berg (2007). *Biology, Eighth Edition*. Cengage Learning.

Sowa, John F. (1995). Top-level ontological categories. *International Journal of Human-Computer Studies*, 43 (5-6), pp. 669–685.

Squires, Gordon (1998). Francis Aston and the mass spectrograph. *J. Chem. Soc., Dalton Trans.* (23), pp. 3893–3900.

Steinbeck, Christoph, Yongquan Han, Stefan Kuhn, Oliver Horlacher, Edgar Luttmann, and Egon Willighagen (2003). The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *Journal of chemical information and computer sciences*, 43 (2), pp. 493–500.

Sturgis, R.C. (2015). *The Mammals That Moved Mankind: A History of Beasts of Burden*. Bloomington, IN : Authorhouse 2015., p. 286.

Tautenhahn, Ralf, Christoph Böttcher, and Steffen Neumann (2008). Highly sensitive feature detection for high resolution LC/MS. *BMC bioinformatics*, 9 (1), p. 1.

Treutler, Hendrik and Steffen Neumann (2016). Prediction, Detection, and Validation of Isotope Clusters in Mass Spectrometry Data. *Metabolites*, 6 (4), p. 37.

Treutler, Hendrik, Hiroshi Tsugawa, Andrea Porzel, Karin Gorzolka, Alain Tissier, Steffen Neumann, and Gerd Ulrich U. Balcke (2016). Discovering Regulated Metabolite Families in Untargeted Metabolomics Studies. *Analytical chemistry*, 88 (16), pp. 8082–8090.

Tsugawa, Hiroshi, Tomas Cajka, Tobias Kind, Yan Ma, Brendan Higgins, Kazutaka Ikeda, Mitsuhiro Kanazawa, Jean VanderGheynst, Oliver Fiehn, and Masanori Arita (2015). MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature methods*, 12 (6), pp. 523–526.

Watson, J. D. and F. H. C. Crick (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171 (4356), pp. 737–738.

Weaver, Warren (1970). Molecular Biology: Origin of the Term. *Science*, 170 (3958), pp. 581–582.

Wendykier, Piotr and James G. Nagy (2010). Parallel Colt: A High-Performance Java Library for Scientific Computing and Image Processing. *ACM Trans. Math. Softw.* 37 (3), 31:1–31:22.

Wilkinson, Mark D. and Matthew Links (2002). BioMOBY: an open source biological web services proposal. *Briefings in bioinformatics*, 3 (4), pp. 331–341.

Zurbriggen, Matias D., Néstor Carrillo, Vanesa B. Tognetti, Michael Melzer, Martin Peisker, Bettina Hause, and Mohammad-Reza Hajirezaei (2009). Chloroplast-generated reactive oxygen species play a major role in localized cell death during the non-host interac-

tion between tobacco and Xanthomonas campestris pv. vesicatoria. *The Plant Journal*, 60 (6), pp. 962–973.

4 Integration of biological data

Publications presented in this thesis related to "Integration of biological data" are entitled "VANTED v2: a framework for systems biology applications" (Rohn et al., 2012) and "DBE2 – Management of experimental data for the VANTED system" (Mehlhorn et al., 2011).

4.1 VANTED v2: a framework for systems biology applications

In the following reference the first author is underlined and I am marked in bold.
Hendrik Rohn, Astrid Junker, Anja Hartmann, Eva Grafahrend-Belau, **Hendrik Treutler**, Matthias Klapperstück, Tobias Czauderna, Christian Klukas, and Falk Schreiber. VANTED v2: a framework for systems biology applications. *BMC systems biology*, 6(1):139+, November 2012. *doi:10.1186/1752-0509-6-139*

`https://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-6-139`

Rohn *et al. BMC Systems Biology* 2012, **6**:139
http://www.biomedcentral.com/1752-0509/6/139

BMC Systems Biology

SOFTWARE Open Access

VANTED v2: a framework for systems biology applications

Hendrik Rohn[1*], Astrid Junker[1], Anja Hartmann[1], Eva Grafahrend-Belau[1], Hendrik Treutler[1], Matthias Klapperstück[1], Tobias Czauderna[1], Christian Klukas[1] and Falk Schreiber[1,2,3]

Abstract

Background: Experimental datasets are becoming larger and increasingly complex, spanning different data domains, thereby expanding the requirements for respective tool support for their analysis. Networks provide a basis for the integration, analysis and visualization of multi-omics experimental datasets.

Results: Here we present VANTED (version 2), a framework for systems biology applications, which comprises a comprehensive set of seven main tasks. These range from network reconstruction, data visualization, integration of various data types, network simulation to data exploration combined with a manifold support of systems biology standards for visualization and data exchange. The offered set of functionalities is instantiated by combining several tasks in order to enable users to view and explore a comprehensive dataset from different perspectives. We describe the system as well as an exemplary workflow.

Conclusions: VANTED is a stand-alone framework which supports scientists during the data analysis and interpretation phase. It is available as a Java open source tool from http://www.vanted.org.

Keywords: Biological networks, Data visualization, Data integration, Data analysis, -Omics, Model simulation

Background

Systems biology comprises the iterative cycling between experimental (wet-lab) and computational (dry-lab) approaches with the aim of generating a holistic understanding of biological systems. The complexity and comprehensiveness of experimental datasets is exponentially increasing thereby elevating the requirements for respective tool support. This motivates the development of adequate software solutions supporting the analysis, integration and visualization of multiple large-scale datasets.

The reconstruction of different kinds of networks (e. g., metabolic, signaling, protein interaction and gene regulatory networks [1]) based on experimental datasets allows for the representation of the diverse nature of biological systems on a global scale. Networks provide the basis for qualitative and quantitative network analysis, for example, for structural analysis and simulation. Networks can furthermore be used for the integrated visualization of multi-omics experimental datasets. In combination with exploration functionalities and further data analysis steps such as correlation and clustering this is crucial for the gain of knowledge from large-scale datasets. New insights lead to the generation of new hypotheses giving feedback to the wet-lab, thereby closing the knowledge generation cycle in systems biology.

To deal with technical advances and the consequent increase of genome-wide datasets, a number of very diverse tools has been developed for network-centered visualization and analysis of experimental data [2,3]. A tool supporting every step of the knowledge generation cycle has to provide the following functionalities: (1) import of data and networks as well as (2) the export of data analysis results and visualizations in different standardized file formats to utilize existing resources, communicate findings and distribute new knowledge among researchers, (3) a variety of analytical methods to extract novel biological findings from large-scale datasets thereby reducing the complexity of the dataset, (4) data integration to combine data from multiple data domains and

*Correspondence: rohn@ipk-gatersleben.de
[1] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstr. 3, 06466 Gatersleben, Germany
Full list of author information is available at the end of the article

Rohn *et al. BMC Systems Biology* 2012, **6**:139
http://www.biomedcentral.com/1752-0509/6/139

support data analysis on a systems level and in the context of the 'global' expertise, (5) model simulation to analyze the dynamic behavior and function of biological systems, thereby elucidating potential targets of biotechnological usage, (6) visualization to ease the understanding of complex datasets and help to elucidate previously unknown functional relations and (7) exploration and interaction functionalities to support visual analysis of large scale datasets and to adapt visualizations according to individual purposes.

Here we present VANTED (version 2) (hereafter named VANTED), a framework for systems biology applications, which emerged from the initial VANTED version [4]. Based on the previously described functionalities it comprises a comprehensive set of tasks ranging from network reconstruction, data visualization, integration of various data types, network simulation to data exploration combined with a manifold support of systems biology standards for visualization and data exchange.

According to Figure 1 we will first introduce the seven main tasks of VANTED with a detailed explanation of various sub-tasks and indicate the possibilities for combining them in order to create systems biology workflows. In the second section an exemplary workflow is instantiated, demonstrating the combination of sub-tasks in order to explore a complex metabolite dataset. Finally, we discuss the benefits of the VANTED framework and describe potential future use cases and corresponding developments of the system.

Implementation

The initial VANTED framework was published in 2006 [4] and is widely used throughout the biologists community (see, for example, [5-11]). In the last years, the framework has been substantially extended and the structure has been changed by out-sourcing of sub-tasks from the VANTED core into add-ons, which are functional modules that can be added during run-time (see Table 1). Such modular approaches allow for a stable and easily maintainable framework core while enabling users to compose a set of functionalities according to individual purposes (see [12,13] for other examples). VANTED has been extended by several important technical improvements such as identifier enrichment for network elements, new input and output interfaces, self-organizing map clustering (SOM)[14], KEGG editor functionality [15] and many more. The new VANTED framework provides a diverse set of functionalities which support system biologists in visualizing and analyzing large-scale datasets (see Figure 1). These can be roughly categorized into seven main tasks, explained in the following sections and Table 1.

Import

Common network exchange formats are supported such as SBML [16], BioPAX [17], KGML [18], GML [19], DOT [20], SBGN-ML [21] and SIF [12] thereby enabling the exchange of data throughout the community. Various databases (e.g., KEGG [22]) provide network files which can be imported into VANTED via drag-and-drop.

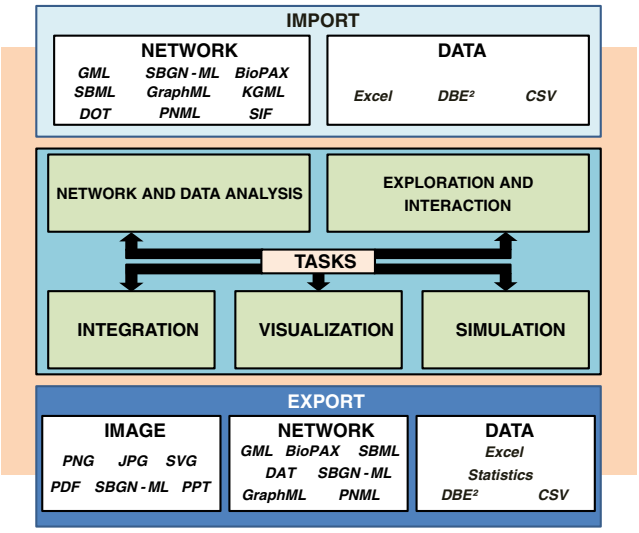


Figure 1 Overview of tasks supported by VANTED. After the initial import of network and experimental data, various tasks can be performed in a combinatorial fashion in order to instantiate a systems biology workflow. The export of results and visualizations is possible at each step of the workflow.

Rohn *et al. BMC Systems Biology* 2012, **6**:139
http://www.biomedcentral.com/1752-0509/6/139

Table 1 Summary of tasks supported by VANTED

Task	Sub-Tasks	Implemented in
Import	• networks (GML, GraphML, SBML, KGML, SIF, DOT, BioPAX, SBGN-ML, PNML)	core, METACROP add-on, DBE^2 add-on
	• experimental data (XLS, XLSX*, CSV)	
	• connection to experiment database DBE^2	
	• connection to network databases (MetaCrop, KEGG, RIMAS)	
Visualization	• charts (line, bar, pie, heat maps) on nodes and edges	core, SBGN-ED add-on, HIVE add-on, FLUXMAP add-on
	• automatic network layouts (e. g., Graphviz, force-directed, tree layout)	
	• SBGN support	
	• flux data support	
	• 3D visualization of networks and multimodal data	
Integration	• mapping of numerical or multimodal data	
	• mapping tables, identifier mapping	
	• linking other resources	core, HIVE add-on
Simulation	• constraint-based analysis	FBA-SIMVIS add-on,
	• Petri net analysis	PETRINET add-on
Exploration and interaction	• panning, zooming, collapsing, search, selection	
	• network exploration	
	• brushing, image exploration	core, GLIEP add-on, HIVE add-on
Analysis	• networks (centralities, shortest path, cycle detection, motifs)	core, CENTILIB add-on
	• statistics (correlation, clustering, t-test)	
	• enrichment analysis	
Export	• raster graphics (PNG, JPG), vector graphics (SVG, PDF, PPT, SBGN-ML)	core, DBE^2 add-on
	• interactive websites	
	• experimental data (XLS, XML, DBE^2)	
	• networks (GML, GraphML, DAT (Metatool), SBML, SBGN-ML, PNML)	

The first column comprises the task covered by the VANTED framework. The second column shortly summarizes sub-tasks. Underlined sub-tasks indicate new functionalities developed since the initial VANTED publication in 2006 [4]. The third column lists the modules of the VANTED framework (VANTED core, add-ons) that implement the described tasks.

VANTED is directly connected to the MetaCrop and the RIMAS databases. The MetaCrop database [23] contains manually curated information about metabolic pathways of major crop plants and corresponding networks in SBGN [24]. In addition to metabolic pathways the database comprises information about reaction kinetics and gene identifiers as well as related literature references. In order to filter, explore and import this information, the METACROP add-on provides seamless access [25]. Besides metabolic networks, gene regulatory networks of the RIMAS web portal [26] can be directly accessed. This information resource comprises SBGN-style networks about regulatory interactions during seed development of *Arabidopsis thaliana*.

The import of experimental data is preferably done by using XLS templates, which enable a structured import together with meta-data. Alternatively, plain text or CSV files may be used to import large datasets such as gene expression data, but require manual enrichment with meta-data. For unlimited accessibility, persistent storage

Rohn *et al. BMC Systems Biology* 2012, **6**:139
http://www.biomedcentral.com/1752-0509/6/139

and exchange of experimental data, the DBE^2 information system [27] is accessible via the DBE^2 add-on. The add-on utilizes ontologies from the Ontology Lookup Service [28] to unify terms such as compound names, species names and measurement units aiming at a facilitated data integration. As VANTED, DBE^2 supports different data types from numerical data to images, three-dimensional volumes and networks.

Visualization

Networks are represented as graphs composed of nodes and edges with fully customizable visual appearance. Numerous visual attributes such as the position, size, color and frame thickness of nodes as well as the color and thickness of edges and other visual attributes such as labels can be adapted according to individual purposes. In addition, a specialized set of node and edge shapes is provided, which build the basis for an SBGN compliant network visualization. SBGN-ED [29] enables VANTED to adapt networks for all SBGN languages in order to facilitate a standardized visual representation of biological entities. The visualization of such maps can be validated for syntactic and semantic correctness according to the SBGN specification.

Readable network layouts are important to improve the visual representation of networks. Besides the manual layout of network elements, automated graph layout algorithms are provided by calling the external Graphviz layouter API [30] or executing self-implemented layouters based on Tollis *et al*. [31] such as the force-directed layout, tree layout, circle layout, expression matrix layout, grid layout, subgraph layout and edge-routing algorithms. Further editing or improvement of automatic layouts can be done by manual curation using node merging and splitting algorithms. The latter is important for splitting frequently occurring nodes such as ATP or CO_2 in metabolic networks, thereby preventing edge-crossings throughout the network.

VANTED offers the integration of various datasets into network nodes and edges (data mapping) thereby enabling a network-based view on large-scale datasets. Options for visual representation of experimental data include shape and color coding of nodes and edges as well as more complex visualizations such as bar charts, pie charts, line charts and heat maps. Experimental factors of complex datasets such as time-resolution, varying genotypes and environmental conditions can be represented within one chart. Visualization of charts is performed by calling the JFREECHART library [32]. The FLUXMAP add-on [33] enables the visual representation of flux data by edge thickness adaptation. This supports the comparative visual analysis of complex flux distributions in an interactive way. Using the HIVE add-on [34] image-based data such as histological cross-sections, microscopy images, photographs and three-dimensional volume data such as NMR and CT data can be displayed in the network context based on a workspace approach and rendered using various 2D-, 3D- and network visualization functions.

Every shape, label, chart and even the selection are realized in VANTED as single Java Swing components placed in the graph window (for further technical details see [35]). Other commonly used libraries such as JUNG [36] render all graphics in a single component. VANTEDs approach is harder to implement, but scales better in terms of rendering speed and enables high flexibility in adapting and fine-tuning each component. The highly optimized CYTOSCAPE framework on the other hand scales very good, but does not enable comparable flexibility in terms of visualization of charts, shapes and other graphics.

In general, visualization is the most advanced feature of VANTED. Multiple options and functionalities enable users to generate appropriate visual representations thereby substantially facilitating the gain of knowledge compared to working with data tables. VANTED enables users to interact with up to 10k network elements, but the responsiveness depends on the visual complexity as complex charts, labels and other visualizations as well as high numbers of edge crossings may reduce this numbers considerably down to some thousand elements. For larger graphs, interaction may become unfeasible and algorithms such as automatic layouters consume a considerable amount of time.

Integration

Biological entities such as proteins, genes or metabolites are represented as nodes and any relation between such entities as node-connecting edges (e.g., regulation, interaction or conversion). Both network elements are attributed by technical properties such as visualization parameters (size, position, etc.) and properties related to their biological role. Each network element may contain links to other resources, usually represented as a hyperlink to any web-content such as a database entry. Nodes may link to other networks, enabling navigation and exploration of connected pathways (see also Section Exploration and interaction). Based on the present numerical attributes, for example, size, position and node degree, the user is able to compute new properties such as additional median values, which are stored as new element attributes and may be visualized or exported.

In VANTED, network elements are allowed to have several (alternative) identifiers. These identifiers provide the basis for data mapping which depends on common identifiers in network and experimental data. In case of different identifiers, synonyms have to be defined. For this mapping tables may be used to provide either additional labels for network elements or for biological entities in

Rohn *et al. BMC Systems Biology* 2012, **6**:139
http://www.biomedcentral.com/1752-0509/6/139

the experiment data. Mapping tables are simple XLS files, which list the existing names in the first column and additional names in the subsequent columns.

Simulation

Basis of the simulation task is the modeling capability of VANTED. Model reconstruction is based on a given network topology, which is manually created or imported from network files. Subsequently, model attributes such as stoichiometric coefficients, kinetic constants, firing rules and initial markings are added to the network or are already part of the import process (SBML files for example provide most attributes). So far, VANTED does not support the automated reconstruction of networks from external sources as described in [37].

These biological networks are finally transformed into mathematical models in order to analyze dynamic properties and behavioral attributes. The enrichment of metabolic networks with stoichiometric coefficients (represented by edge weights) and the definition of an optimization function is a prerequisite for the constraint-based network analysis. The FBA-SIMVIS [38] add-on enables VANTED to perform different techniques such as Flux Balance Analysis [39], Flux Variability Analysis [40], Robustness Analysis [41] and Knock-out Analysis. In combination with a dynamic and visual exploration of simulation results, this allows for the comprehensive analysis of metabolism in response to genetic or environmental perturbations. Metabolic networks can also be transformed into Petri nets [42], a second mathematical model, which is used for formal analysis and simulation of biological systems. The PETRINET [43] add-on enables VANTED to semi-automatically transform networks into valid Petri nets, simulate discrete and continuous Petri nets of varying complexity and analyze structural properties. Different visualization and interaction techniques such as brushing can be utilized in order to visually analyze P- and T-invariants, the reachability graph and varying markings of simulation steps.

Exploration and interaction

In terms of exploration of networks and data visualizations, VANTED supports standard interaction methods such as panning, zooming and overview+detail for selected network elements. The editing and rearrangement of network elements as well as the modification of attribute values and calculation of new attributes is possible in an interactive manner. Sophisticated selection and search functionalities provide the ability to find and explore network elements based on attribute values.

Furthermore, recurring entities in large networks or several networks may be linked in order to easily track interconnections between pathways. The GLIEP [44] add-on provides an interactive view for the exploration of interconnected networks by implementing a glyph visualization. Based on these glyphs the user is able to quickly switch between connected networks or to explore the overall interconnectivity using a focus+context technique. Furthermore, the HIVE add-on enables users to collapse networks into single nodes, thereby providing a clear representation of multiple (interconnected) networks. Connections between different networks are retained and link the network-overview nodes, which can be re-arranged or expanded according to user requirements.

On the basis of interaction events such as selection, brushing techniques [45] provide different views on visualized experimental data. The HIVE add-on enables users to explore and compare spatial distributions within a biological system by parallel visualization of segmented images and experimental values in the network view. Hovering over a segment in the image (e.g., corresponding to an organ) results in highlighting the respective measurement values in the network view. Furthermore it is possible to explore large numbers of images in the context of a network. If these images are related to a substance (e.g., GFP reporter expression for genes in a gene regulatory network), the user can integrate the respective images into the network nodes. If a number of nodes is selected, an image matrix is built up, spanning conditions, time points and replicate information. This matrix enables users to compare all images related to the selected nodes and to explore spatial patterns of different substances in the context of a biological network.

Further brushing techniques are provided by the PETRINET add-on for the analysis of Petri net properties such as invariants and the reachability graph. The user can move the mouse over nodes of the reachability graph, triggering the visualization of the respective state in the network visualization view.

Analysis

The analysis of network topology plays an important role for the understanding of interactions between biological entities. VANTED offers to compute several topological properties such as shortest paths between node pairs, network cycles and motifs. The detection of network motifs (such as feed-forward loops) is supported by the possibility to search for user-defined motifs which might be meaningful in the context of certain biological questions. The VANTED add-on CENTILIB [46] provides algorithms and methods for the computation and investigation of 17 different centralities in biological networks. Such centralities can be used for ranking of network nodes according to given criteria and for the detection of network hubs. Results of the centrality analysis can be explored and analyzed using a brushing-based approach.

The statistical evaluation of experimental datasets is a central part of data analysis. VANTED offers a series

Rohn *et al. BMC Systems Biology* 2012, **6**:139
http://www.biomedcentral.com/1752-0509/6/139

of tests for calculation of statistical parameters, for testing the normal distribution of datasets (David Quicktest [47]) and for outlier detection (Grubbs test). For the comparison of measurements with multiple conditions, several t-tests are available such as the unpaired t-test, the Welch-Satterthwaite t-test and the Mann-Whitney U-test with user-defined threshold settings for the calculated p-values. VANTED enables users to perform Pearson's and Spearman correlation analysis based on the mapped experimental data. Optional settings include a p-value threshold and the number of experiment conditions included in the analysis (see [4] for implementations details).

The calculation of clusters is a frequently used approach to categorize experimental data into functional or behavioral groups. For this task, VANTED supports self-organizing maps (SOM) [14]. A SOM is an artificial neural network, which is capable for the automated recognition of patterns within measurements and is well-suited for the categorization of time series data of biological entities. According to a user-defined number of target clusters, the SOM is trained and cluster attributes are automatically assigned to the network nodes. In addition such assignments can be done manually. The cluster sub-networks may then be independently laid out or colorized in order to visually catch clustered elements at a glance.

For gene expression data VANTED supports the computation and visualization of enrichments in the context of the GO [48] and the KEGG pathway [22] hierarchies. For example, for KEGG the procedure highlights classes of KEGG pathways in which the experimental data enriches significantly by assigning pie charts [49,50].

Export

VANTED provides a variety of file formats for data storage, publication and exchange. The GML and GraphML file formats are VANTED s native formats and accordingly support the storage of networks together with all related attributes such as layout information and the full set of mapped and integrated experimental data including the visualization options for mapped data. Additional information can be stored and exchanged as new attributes, e. g. a new custom attribute "myAttribute" enables to colorize all nodes with this attribute based on the respective attribute value. Such attributes can be created manually (e. g. cluster information and biological tags) or be the result of a computation (see [35] for further details).

For the exchange of data within the systems biology community, support for file formats such as DAT [51], SBGN-ML (provided by the SBGN-ED add-on) and BioPAX is implemented. VANTED additionally supports the SBML file format which allows for the storage and exchange of stoichiometric and kinetic models. When working with the PETRINET add-on, the Petri net and its configuration can be exchanged using the PNML file format. Experimental data which has been mapped onto a network can be extracted and exported using XLS sheets. The CSV format is supported for different kinds of node attributes as well as the export of analysis results such as correlation coefficients. All data types which are supported by VANTED (numerical data, images, three-dimensional volumes, networks) can be uploaded to the DBE^2 system for persistent data storage and exchange. Please note that VANTED usually serves as a data sink and the conversion between different file formats is not in the focus of the tool. Network topology (including labels) on the other hand is preserved in most cases.

Laid out networks can be exported to several graphic file formats, including raster images (PNG, JPG), as well as vector images (SVG, PDF, PPT). These file formats are well suited to be used as images in publications, presentations or as a basis for further graphical editing. Furthermore it is possible to export integrated networks as browseable and clickable images, embedded in HTML web sites. Those images can contain web-links to web resources or public databases. The publishing process of these web sites can be done in a semi-automatic fashion [52].

Results

The previously described tasks can be instantiated and combined in order to create manifold workflows supporting the interpretation of systems biology data. For demonstration purposes an exemplary workflow is executed with the VANTED framework, implementing the analysis of a comprehensive metabolic dataset taken from Sulpice *et al.* [53]. This dataset consists of measurements of enzyme activity data, metabolite data and different morphological parameters for a wide range of *Arabidopsis thaliana* ecotypes. In the following we focus on the first ecotype class A, which includes the most diverse ecotypes. The steps of the workflow are depicted in Figure 2 and the tutorial (Additional file 1).

Import

The import of enzyme activity data, metabolite data and morphological parameters of different *Arabidopsis thaliana* accessions from climate class A is realized using the VANTED XLS template (see Additional file 2). Experimental data may also be persistently stored in the DBE^2 database, enabling file sharing and on-click import of such experimental data into VANTED. In parallel to the import of the experimental data, 38 metabolic reference pathways are loaded from the MetaCrop database and merged into one SBGN network. Subsequently all reference pathways are assigned to their respective cellular location and the pathways in each subcellular compartment are connected to each other by merging identical metabolite

Rohn *et al. BMC Systems Biology* 2012, **6**:139
http://www.biomedcentral.com/1752-0509/6/139

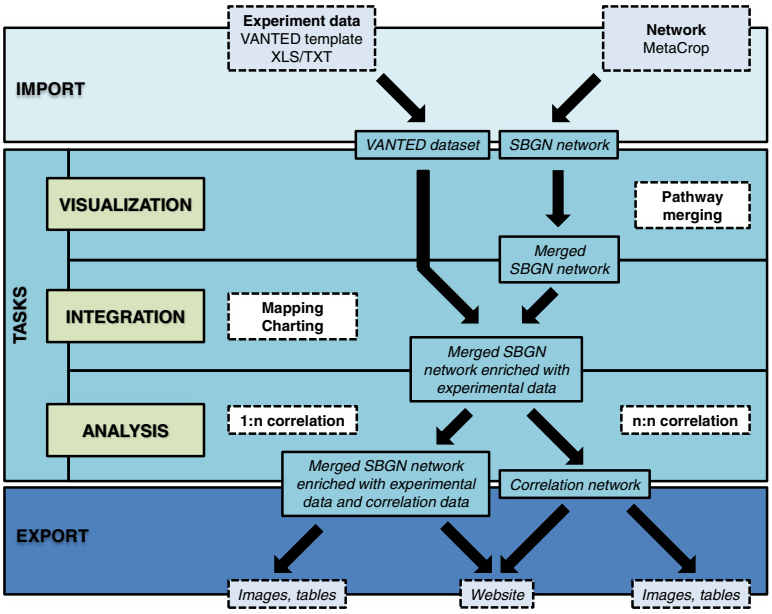


Figure 2 VANTED workflow for the exemplary use case. A complex metabolite dataset is imported into VANTED, integrated and visualized in the context of a large SBGN-style metabolic network. Based on data mapping, different kinds of correlation analyses are performed. The results of the workflow can be exported in various formats.

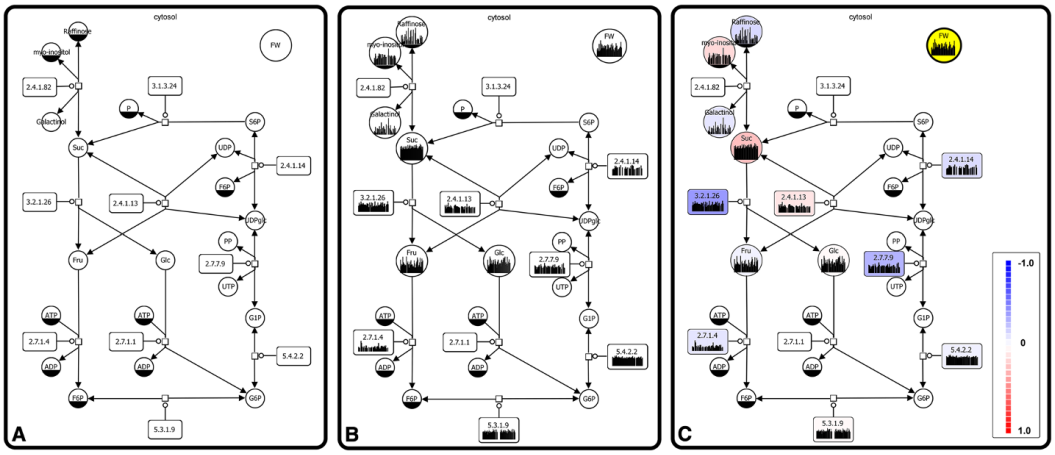


Figure 3 Visualization, integration and analysis of plant metabolic networks. (A) Metabolic network representing sugar metabolism in SBGN. A new node for the morphological parameter fresh weight (FW) was added to the network. **(B)** Integration of metabolic data into the network by visualization of corresponding charts inside the nodes. Metabolite concentrations are mapped to simple chemical glyphs whereas enzyme activity data is mapped to macromolecule glyphs. Bar charts display respective values for all *Arabidopsis thaliana* accessions of climate class A. **(C)** 1:n correlation analysis on mapped data for the detection of correlations between the morphological parameter FW and all other metabolic parameters. Correlation coefficients are visualized by color-coded nodes.

Rohn *et al. BMC Systems Biology* 2012, **6**:139
http://www.biomedcentral.com/1752-0509/6/139

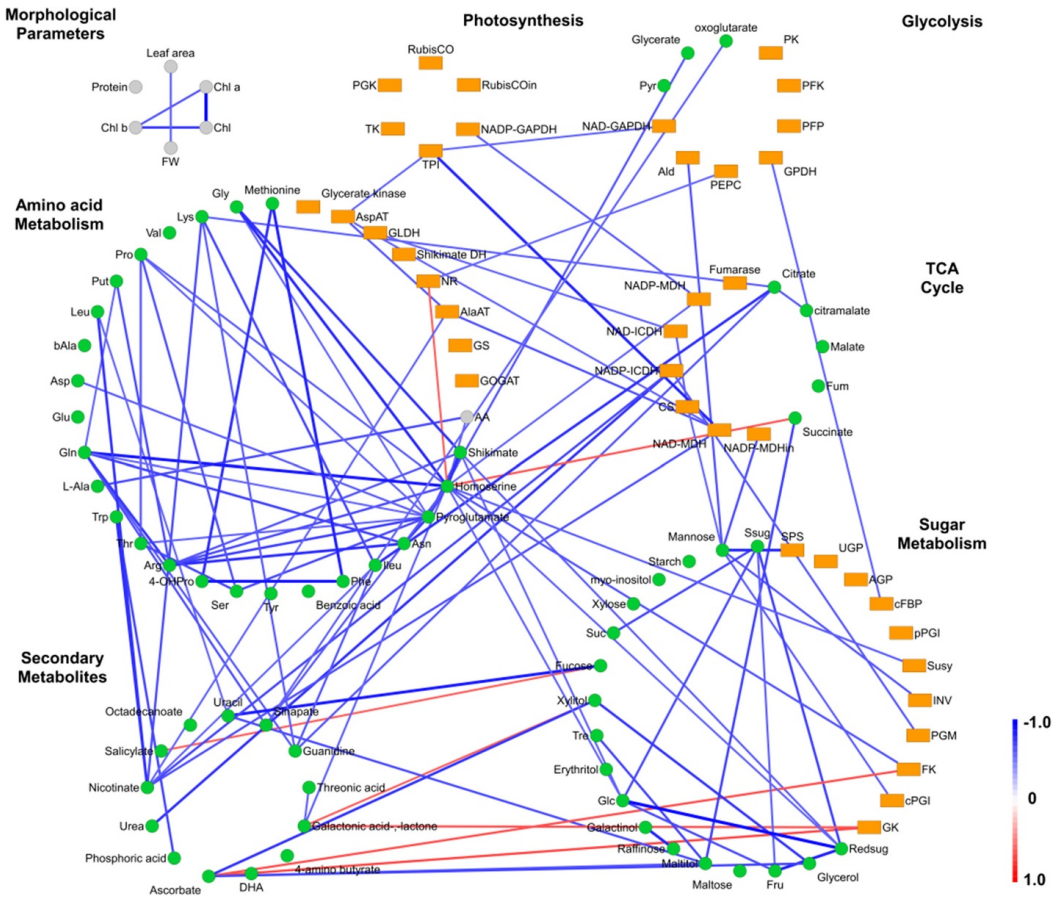


Figure 4 Correlation network for different pathways. Nodes representing metabolites (green), enzymes (orange) and morphological or other parameters (gray) are laid out as circles for each pathway. A n:n correlation was calculated, resulting in edges indicating a strong ($p \geq 0.95$) correlation, color-coded by the *r*-value. This visualization enables an overview about intra- and inter-pathway dependencies.

nodes. Finally a network layout is performed in order to optimize the edge routing and distance between nodes, resulting in the network which can be found in Additional file 3.

Visualization and integration

During data mapping, experimental data is integrated into the network by the visualization of corresponding charts inside the network nodes. To unify the identifiers in the network and the experimental dataset, a mapping table is used for the enrichment of network nodes with alternative identifiers (Figure 3a and Additional file 3). Subsequently, metabolite data is mapped to the nodes representing metabolites (simple chemical glyph) and enzyme activity data is mapped to nodes representing enzyme nodes (macromolecule glyph). New nodes for morphological parameters are added during the mapping process, as they are part of the experimental data, but do not occur in the network. The mapped experimental data is visually represented by bar charts inside the glyphs resulting in a data-enriched SBGN network (Figure 3b and Additional file 4).

Analysis

In order to identify similarities in the profiles of all accessions of climate class A, 1:n and n:n correlation analyses are performed. In case of the 1:n correlation analysis, the morphological parameter fresh weight (FW) is chosen as the target parameter and correlations were calculated to all other metabolic parameters in the network. Based

Rohn *et al. BMC Systems Biology* 2012, **6**:139
http://www.biomedcentral.com/1752-0509/6/139

on the resulting correlation coefficients network nodes are color-coded according to the correlation coefficient r (Figure 3c and Additional file 5). This visual representation of correlation results enables biologists to easily identify metabolic parameters with important influence on plant morphology at a global scale.

For the n:n correlation analysis, all metabolic parameters in the network are correlated with each other, including all metabolite and enzyme activity data as well as the data of morphological parameters. The resulting correlation values are visualized by generating new edges between correlating nodes. These edges are color-coded according to the negative (red) or positive (blue) correlations calculated with $p \geq 0.95$ and $|r| \geq 0.6$ Pearson's product-moment correlation. The resulting network is used to generate a correlation network at a pathway level, independent of the order of metabolic reactions within a pathway. Consequently, the metabolic dataset is used to generate new nodes in a network-independent manner which are then categorized according to the metabolic pathway (e.g., Glycolysis, TCA cycle) and laid out as pathway-specific circles (see Figure 4). During the n:n correlation analysis VANTED generates edges between nodes with data profiles of significant similarity thereby giving an overview about intra- and inter-pathway dependencies and allows for drawing conclusions about the interaction between single parameters. For example, the levels of amino acids show strong positive correlations among each other and with levels of TCA cycle intermediates, as these substances are precursors of the amino acids. This leads to the assumption that these mentioned parts of primary metabolism are stable throughout the different ecotypes. Secondary metabolites show strong negative correlations with enzymes of sugar metabolism among the considered *Arabidopsis thaliana* accessions. Variations of the levels of plant secondary metabolites are conceivable for accessions with different origin.

Discussion

The VANTED framework provides a rich variety of functionalities at the interface between data analysis, gain of knowledge out of large-scale datasets and the generation of feedback to the wet-lab part of the systems biology cycle. It supports both the fast and customizable visualization of networks and experimental data as well as the exploration, simulation and different kinds of data analysis. In contrast, most network-centered tools focus on a small subset of tasks (compare Table 2). For instance, OMIX provides high-quality and customizable network visualization but lacks analysis algorithms and direct connection to important databases. ONDEX focuses on the generation of large-scale biological networks from heterogeneous sources, but does not support charts and simulations. CELLDESIGNER is designed for the analysis of the dynamics of metabolic models, but does neither provide statistical analysis nor advanced interaction techniques. VANTED combines these features in one framework thereby reducing the use of several tools and tedious file exchanging procedures.

CYTOSCAPE is a widely used biological network analysis tool, which is the only competing tool providing all tasks in one system. Both tools cover a large portion of important systems biology tasks. CYTOSCAPE lacks some functions such as sophisticated charts and website export, but compared to VANTED provides additional functionality which is usually not in the focus of systems biology researchers, such as social graph topics. It has a big developer community which implemented a large number of plugins (over 150). Although the sheer number of extensions is quite impressive, the quality and complexity varies significantly. Many CYTOSCAPE plugins only provide simple functionalities such as the import of a certain file format, whereas others focus on very special applications which are not in the scope of the majority of potential users. In comparison to CYTOSCAPE, the VANTED add-on concept relies on a smaller set of add-ons each comprising a large set of functionalities which are necessary in order to perform a whole workflow. Many VANTED add-ons are able to interact with each other, thereby increasing the capabilities of the core tool. Examples for such combinations are the HIVE and the DBE^2 add-on, which together enable the persistent storage of volumetric and image data in the exchange database. Also the combination of FLUXMAP and SBGN-ED enables the visualization of flux data in SBGN networks. In summary, VANTED and CYTOSCAPE both enable the execution of various systems biology tasks within one tool. CYTOSCAPE provides a larger set of special sub-tasks with varying quality, whereas VANTED provides a small set of sub-tasks, which are optimized with regard to solving specific biological questions.

Conclusions

VANTED is a stand-alone framework which supports scientists during the data analysis and interpretation phase. This is achieved by integrating experimental data into biological networks and providing a rich variety of simulation, analysis and visualization functionalities. Manifold file exchange formats as well as connections to databases enable the examination of user data in the context of public resources. In comparison to other tools VANTED provides a large variety of functionalities, spanning most of the tasks during the analysis and visualization of large-scale datasets. The offered set of functionalities enables users to view and explore data from different perspectives, thereby facilitating the systemic analysis of a biological object. The support of various standards enables users to easily exchange files using well-established standard file

Table 2 Comparison of non-commercial tools for the network-centered visualization and analysis of biological data

Tasks	VANTED	CYTOSCAPE [12]	ONDEX [54]	OMIX [55]	CELL Designer [56]	PATHVISIO [57]	BIOUML [58]	VISANT [59]	PATHWAY Projector [60]	BINA [61]	MAPMAN [62]
Import											
networks	+	+	+	+	+	(+)	+	(+)	+	+	-
experimental data	+	+	+	+	(+)	+	+	-	+	(+)	+
connection to experiment database	+	+	+	-	(+)	-	-	-	-	+	-
connection to network databases	+	+	+	-	(+)	-	+	+	(+)	+	+
Visualization											
charts on nodes and edges	+	-	-	+	-	-	-	-	+	-	+
automatic network layouts	+	+	+	+	+	-	+	+	-	+	-
SBGN support	+	+	-	-	(+)	+	+	-	-	-	-
flux data support	+	+	-	+	+	-	-	-	(+)	+	-
3D visualization	+	(+)	-	+	-	-	-	-	-	-	-
Integration											
mapping of numerical or multimodal data	+	+	+	+	+	+	-	-	+	+	+
mapping tables, identifier mapping	+	+	+	-	-	(+)	+	-	+	(+)	-
linking other resources	+	+	+	+	+	+	+	+	+	-	+
Simulation											
constraint-based analysis	+	(+)	-	+	(+)	-	-	-	-	-	-
Petri net analysis	+	-	-	-	(+)	-	-	-	-	-	-
Exploration and interaction											
panning, zooming, collapsing, search and selection	+	+	+	+	(+)	(+)	+	+	+	+	(+)
network exploration	+	+	+	-	-	(+)	+	+	+	+	+
brushing, image exploration	+	-	-	-	-	-	-	-	-	-	-
Analysis											
networks	+	+	+	-	(+)	-	+	+	-	(+)	-
statistics	+	+	+	-	-	+	-	-	-	(+)	+
enrichment analysis	+	+	-	-	-	+	+	+	-	+	+
Export											
raster graphics, vector graphics	+	+	+	+	+	+	+	+	-	+	(+)
interactive websites	+	-	-	-	-	(+)	+	+	+	-	-
experimental data	+	+	-	-	+	-	+	-	+	+	-
networks	+	+	+	+	+	+	+	(+)	-	+	-

The first column comprises the sub-tasks of Table 1, which are covered by the respective tool. Please note that also add-ons and plugins of the respective system were evaluated. "-" no or inadequate support, "(+)" = partial support, "+" good support of the sub-task.

formats and allow for an accurate exchange of biological information using an unambiguous graphical representation (SBGN). To deal with future user requirements the VANTED system can be extended in a flexible way by using BeanShell and JRuby scripts or by writing new add-ons.

In the future we expect novel use cases to emerge for the VANTED framework, especially large datasets spanning multiple biological levels such as gene expression, protein activity, metabolite, flux and phenotypic data from one biological system [63]. Furthermore, the spatial resolution of the analyzed systems (e.g., compartmentation, tissues and organs) increases based on technological advances and enhanced quantity and quality of imaging techniques. Finally, mathematical models become more important for the understanding and prediction of complex behavior of biological systems.

Availability and requirements

- **Project Name:** VANTED
- **Project home page:** http://www.vanted.org
- **Operating system(s):** Platform independent (Java), the add-on FBASimVis will work on Windows computers only
- **Programming language:** Java 6/7
- **License:** GPL 2.0

Additional files

Additional file 1: Supplementary tutorial. ZIP file containing the data for recreating Figures 3 and 4. To guide the user, a PPT file is provided, which lists and describes all necessary steps to be performed in VANTED.

Additional file 2 : Filled experiment data template. VANTED template filled with metabolite data from Sulpice *et al.* [53], consisting of 64 metabolites, 37 enzymes and morphological parameters for 50 *Arabidopsis thaliana* ecotypes of climate class A. The file can be opened using MS Excel and imported into VANTED as an experiment dataset.

Additional file 3: Merged SBGN network. Large-scale metabolic network of plant primary metabolism in SBGN. The network has been created with VANTED based on merging different pathways downloaded from MetaCrop. This file serves as the basis for mapping experiment datasets and can be imported into VANTED as a network.

Additional file 4: Merged SBGN network enriched with experimental data. Enriched metabolic SBGN network after mapping additional file 2 onto additional file 3. Metabolite data of 50 *Arabidopsis thaliana* ecotypes is mapped to the network and visualized as bar charts inside the nodes. This file can be imported into VANTED as a network.

Additional file 5: Merged SBGN network enriched with experimental data and correlation data. Analysis of enriched metabolic SBGN network by performing a 1:n correlation between the morphological parameter fresh weight (FW) and all enriched network nodes. The correlation coefficient is visualized using a global color-code. This file can be imported into VANTED as a network.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
CK, HR and TC implemented the core. HR, HT, EGB, TC and MK implemented the add-ons. AJ, AH, EGB and HR developed the use case. FS supervised the project and gave conceptual advice. HR wrote the manuscript; all authors contributed to, read and approved the manuscript.

Acknowledgements
This work has been partly funded by BMBF (grants 0312706A, 3015426A, RUS 10/131) and DAAD (grant 54391720).

Author details
[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstr. 3, 06466 Gatersleben, Germany. [2]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Von-Seckendorff-Platz 1, 06120 Halle, Germany. [3]Clayton School of Information Technology, Monash University, Victoria 3800, Australia.

Received: 26 July 2012 Accepted: 1 November 2012
Published: 10 November 2012

References

1. Moreno-Risueno MA, Busch W, Benfey PN: **Omics meet networks - using systems approaches to infer regulatory networks in plants.** *Curr Opin Plant Biol* 2010, **13**(2):126–131.
2. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, Kohlbacher O, Neuweger H, Schneider R, Tenenbaum D, Gavin AC: **Visualization of omics data for systems biology.** *Nat Methods* 2010, **7:**S56–S68.
3. Suderman M, Hallett MT: **Tools for visually exploring biological networks.** *Bioinformatics* 2007, **23**(20):2651–2659.
4. Junker BH, Klukas C, Schreiber F: **VANTED: a system for advanced data analysis and visualization in the context of biological networks.** *BMC Bioinformatics* 2006, **7:**109. 1–13.
5. Bazzini AA, Manacorda CA, Tohge T, Conti G, Rodriguez MC, Nunes-Nesi A, Villanueva S, Fernie AR, Carrari F, Asurmendi S: **Metabolic and miRNA profiling of TMV infected plants reveals biphasic temporal changes.** *PLoS One* 2011, **6**(12):e28466.
6. Hofmann J, Ashry AENE, Anwar S, Erban A, Kopka J, Grundler F: **Metabolic profiling reveals local and systemic responses of host plants to nematode parasitism.** *Plant J* 2010, **62**(6):1058–1071.
7. Clauss K, von Roepenack-Lahaye E, Böttcher C, Roth MR, Welti R, Erban A, Kopka J, Scheel D, Milkowski C, Strack D: **Overexpression of sinapine esterase BnSCE3 in oilseed rape seeds triggers global changes in seed metabolism.** *Plant Physiol* 2011, **155**(3):1127–1145.
8. Kogel KH, Voll LM, Schäfer P, Jansen C, Wu Y, Langen G, Imani J, Hofmann J, Schmiedl A, Sonnewald S, von Wettstein D, Cook RJ, Sonnewald U: **Transcriptome and metabolome profiling of field-grown transgenic barley lack induced differences but show cultivar-specific variances.** *PNAS* 2010, **107**(14):6198–6203.
9. Riewe D, Grosman L, Zauber H, Wucke C, Fernie AR, Geigenberger P: **Metabolic and developmental adaptations of growing potato tubers in response to specific manipulations of the adenylate energy status.** *Plant Physiol* 2008, **146**(4):1579–1598.
10. van Dongen JT, Fröhlich A, Ramírez-Aguilar SJ, Schauer N, Fernie AR, Erban A, Kopka J, Clark J, Langer A, Geigenberger P: **Transcript and metabolite profiling of the adaptive response to mild decreases in oxygen concentration in the roots of arabidopsis plants.** *Ann Botany* 2009, **103**(2):269–280.
11. Gupta S, Maurya MR, Stephens DL, Dennis EA, Subramaniam S: **An integrated model of eicosanoid metabolism and signaling based on lipidomics flux analysis.** *Biophys J* 2009, **96**(11):4542–4551.
12. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13**(11):2498–2504.
13. Abramoff MD, Magelhaes PJ, Ram SJ: **Image Processing with ImageJ.** *Biophotonics International* 2004, **11:**36–42.
14. Kohonen T: **The Self-Organizing Map.** *Proc IEEE* 1990, **78:**1464–1480.
15. Klukas C, Schreiber F: **Dynamic exploration and editing of KEGG pathway diagrams.** *Bioinformatics* 2007, **23**(3):344–350.

16. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le Novere N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J: **The systems biology markup language (SBML): a medium for representation and exchange Of biochemical network models.** *Bioinformatics* 2003, **19**(4):524–531.
17. Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, Wu G, D'Eustachio P, Schaefer C, Luciano J, Schacherer F, Martinez-Flores I, Hu Z, Jimenez-Jacinto V, Joshi-Tope G, Kandasamy K, Lopez-Fuentes AC, Mi H, Pichler E, Rodchenkov I, Splendiani A, Tkachev S, Zucker J, Gopinath G, Rajasimha H, Ramakrishnan R, Shah I, Syed M, Anwar N, Babur O, Blinov M, Brauner E, Corwin D, Donaldson S, Gibbons F, Goldberg R, Hornbeck P, Luna A, Murray-Rust P, Neumann E, Ruebenacker O, Reubenacker O, Samwald M, van Iersel M, Wimalaratne S, Allen K, Braun B, Whirl-Carrillo M, Cheung KH, Dahlquist K, Finney A, Gillespie M, Glass E, Gong L, Haw R, Honig M, Hubaut O, Kane D, Krupa S, Kutmon M, Leonard J, Marks D, Merberg D, Petri V, Pico A, Ravenscroft D, Ren L, Shah N, Sunshine M, Tang R, Whaley R, Letovksy S, Buetow KH, Rzhetsky A, Schachter V, Sobral BS, Dogrusoz U, McWeeney S, Aladjem M, Birney E, Collado-Vides J, Goto S, Hucka M, Novere NL, Maltsev N, Pandey A, Thomas P, Wingender E, Karp PD, Sander C, Bader GD: **The BioPAX community standard for pathway data sharing.** *Nature Biotechnol* 2010, **28**(9):935–942.
18. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 2000, **28:**27–30.
19. Himsolt M: *GML: A portable Graph File Format*. University of Passau: Tech. rep.; 1996.
20. Ellson J, Gansner ER, Koutsofios E, North SC, Woodhull G: **Graphviz and dynagraph: static and dynamic graph drawing tools.** In *Graph Drawing Software*: Springer-Verlag; 2003:127–148.
21. van Iersel MP, Villeger AC, Czauderna T, Boyd SE, Bergmann FT, Luna A, Demir E, Sorokin A, Dogrusoz U, Matsuoka Y, Funahashi A, Aladjem MI, Mi H, Moodie SL, Kitano H, Novere NL, Schreiber F: **Software support for SBGN maps: SBGN-ML and LibSBGN.** *Bioinformatics* 2012, **28**(15):2016–2021.
22. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets. .** *Nucleic Acids Res* 2012, **40**(Database issue):D109—D114.
23. Schreiber F, Colmsee C, Czauderna T, Grafahrend-Belau E, Hartmann A, Junker A, Junker BH, Klapperstück M, Scholz U, Weise S: **MetaCrop 2.0: managing and exploring information about crop plant metabolism.** *Nucleic Acids Res* 2012, **40**(Database issue):D1173—D1177.
24. Le Novère, N, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegner K, Aladjem MI, Wimalaratne SM, Bergman FT, Gauges R, Ghazal P, Kawaji H, Li L, Matsuoka Y, Villéger A, Boyd SE, Calzone L, Courtot M, Dogrusoz U, Freeman TC, Funahashi A, Ghosh S, Jouraku A, Kim S, Kolpakov F, Luna A, Sahle S, Schmidt E, Watterson S, Wu G, Goryanin I, Kell DB, Sander C, Sauro H, Snoep JL, Kohn K, Kitano H: **The systems biology graphical notation.** *Nat Biotechnol* 2009, **27**(8):735–741.
25. Hippe K, Colmsee C, Czauderna T, Grafahrend-Belau E, Junker BH, Klukas C, Scholz U, Schreiber F, Weise S: **Novel developments of the MetaCrop information system for facilitating systems biological approaches.** *J Integrative Bioinf* 2010, **7**(3):125.
26. Junker A, Hartmann A, Schreiber F, Bäumlein H: **An engineer's view on regulation of seed development.** *Trends in Plant Science* 2010, **15**(6):303–307.
27. Mehlhorn H, Schreiber F: **DBE2- Management of experimental data for the VANTED system.** *J Integrative Bioinf* 2011, **8**(2):162.1–10.
28. Cote R, Jones P, Apweiler R, Hermjakob H: **The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries.** *BMC Bioinformatics* 2006, **7:**97.1–7.
29. Czauderna T, Klukas C, Schreiber F: **Editing, Validating, and Translating of SBGN Maps.** *Bioinformatics* 2010, **26**(18):2340–2341.
30. Ellson J, Gansner E, Koutsofios L, North S, Woodhull G, Short Description, Lucent Technologies: *Graphviz - open source graph drawing tools*, Lecture Notes in Computer Science: Springer-Verlag; 2001. 483–484.
31. Tollis IG, Di Battista G, Eades P, Tamassia R: *Graph Drawing: Algorithms for the Visualization of Graphs*: Prentice Hall; 1998.
32. Gilbert D, Morgner T: **JFreeChart, a free Java class library for generating charts.** [http://www.jfree.org/jfreechart]
33. Rohn H, Hartmann A, Junker A, Junker BH, Schreiber F: **FluxMap: a VANTED Add-on for the visual exploration of flux distributions in biological networks.** *BMC Syst Biol* 2012, **6:**33.1–9.
34. Rohn H, Klukas C, Schreiber F: **Creating views on integrated multidomain data.** *Bioinformatics* 2011, **27**(13):1839–1845.
35. Bachmaier C, Brandenburg FJ, Forster M, Raitner M, Holleis P: *Gravisto: Graph Visualization Toolkit*; 2004.
36. Madadhain J, Fisher D, Smyth P, White S, Boey Y: **Analysis and visualization of network data using JUNG.** *J Stat Software* 2005, **10:**1–35.
37. De RK, Tagore S: **Automated metabolic pathway reconstruction based on structural grammars.** *J Comput Sci Syst Biol* 2012, **5:**116–127.
38. Grafahrend-Belau E, Klukas C, Junker BH, Schreiber F: **FBASimViz: interactive visualization of constraint-based metabolic models.** *Bioinformatics* 2009, **25**(20):2755–2757.
39. Orth JD, Thiele I, Palsson BO: **What is flux balance analysis?** *Nature Biotechnol* 2010, **28**(3):245–248.
40. Mahadevan R, Schilling CH: **The effects of alternate optimal solutions in constraint-based genome-scale metabolic models.** *Metabolic Engineering* 2003, **5:**264–276.
41. Edwards JSuBP: **Robustness analysis of the Escherichia coli metabolic network.** *Biotechnol Progress* 2000, **16:**927–939.
42. Baldan P, Cocco N, Marin A, Simeoni M: **Petri nets for modelling metabolic pathways: a survey.** *Natural Computing* 2010, **9**(4):955–989.
43. Hartmann A, Rohn H, Pucknat K, Schreiber F: **Petri nets in VANTED: Simulation of Barley Seed Metabolism.** In *Proceedings of the 3rd International Workshop on Biological Processes & Petri Nets*; 2012:20–28.
44. Jusufi I, Klukas C, Kerren A, Schreiber F: **Guiding the interactive exploration of metabolic pathway interconnections.** *Information Visualization* 2012, **11**(2):136–150.
45. Martin AR, Ward MO: **High dimensional brushing for interactive exploration of multivariate data.** In *Proceedings on Visualization*; 1995:271–278.
46. Gräßler J, Koschützki D, Schreiber F: **CentiLib: comprehensive analysis and exploration of network centralities.** *Bioinformatics* 2012, **28**(8):1178–1179.
47. David H, Hartley H, Pearson E: **The distribution of the ratio, in a single, normal sample, of range to standard deviation.** *Biometrika* 1954, **41**(3–4):482–493.
48. The Gene Ontology Consortium: **The Gene Ontology project in 2008.** *Nucleic Acids Res* 2008, **36**(Database issue):D440–D444.
49. Klukas C, Schreiber F: **Integration of -omics data and networks for biomedical research.** *J Integrative Bioinf* 2010, **7**(2):112.1–6.
50. Sharbel TF, Voigt ML, Corral JM, Galla G, Kumlehn J, Klukas C, Schreiber F, Vogel H, Rotter B: **Apomictic and sexual ovules of Boechera display heterochronic global gene expression patterns.** *Plant Cell* 2010, **22**(3):655–671.
51. von Kamp A, Schuster S: **Metatool 5.0: fast and flexible elementary modes analysis.** *Bioinformatics* 2006, **22**(15):1930–1931.
52. Junker A, Rohn H, Czauderna T, Klukas C, Hartmann A, Schreiber F: **Creating interactive, web-based and data-enriched maps using the Systems Biology Graphical Notation.** *Nat Protocols* 2012, **7:**579–593.
53. Sulpice R, Trenkamp S, Steinfath M, Usadel B, Gibon Y, Witucka-Wall H, Pyl ET, Tschoep H, Steinhauser MC, Guenther M, Hoehne M, Rohwer JM, Altmann T, Fernie AR, Stitt M: **Network analysis of enzyme activities and metabolite levels and their relationship to biomass in a large panel of arabidopsis accessions.** *The Plant Cell Online* 2010, **22**(8):2872–2893.
54. Köhler J, Baumbach J, Taubert J, Specht M, Skusa A, Rüegg A, Rawlings C, Verrier P, Philippi S: **Graph-based analysis and visualization of experimental results with ONDEX.** *Bioinformatics* 2006, **22**(11):1383–1390.
55. Droste P, Miebach S, Niedenführ S, Wiechert W, Nöh K: **Visualizing multi-omics data in metabolic networks with the software Omix: a case study.** *Biosystems* 2011, **105**(2):154–161.
56. Funahashi A, Matsuoka Y, Jouraku A, Kitano H, Kikuchi N: **CellDesigner: a modeling tool for biochemical networks.** In *Proceedings of the 38th conference on Winter simulation*: Winter Simulation Conference; 2006:1707–1712.

57. van Iersel MP, Kelder T, Pico AR, Hanspers K, Coort S, Conklin BR, Evelo C: **Presenting and exploring biological pathways with PathVisio.** *BMC Bioinformatics* 2008, **9:**399.1–9.
58. Kolpakov FA: **BioUML- Framework for visual modeling and simulation of biological systems.** In *Proceedings of the International Conference on Bioinformatics of Genome Regulation and Structure*; 2002:130–133.
59. Hu Z, Hung JH, Wang Y, Chang YC, Huang CL, Huyck M, DeLisi C: **VisANT 3.5: Multi-scale network visualization, analysis and inference based on the gene ontology.** *Nucleic Acids Res* 2009, **37**(Web Server issue):W115—W121.
60. Kono N, Arakawa K, Ogawa R, Kido N, Oshita K, Ikegami K, Tamaki S, Tomita M: **Pathway projector: web-based zoomable pathway browser using KEGG atlas and Google Maps API.** *PLoS One* 2009, **4**(11):e7710.
61. Küntzer J, Backes C, Blum T, Gerasch A, Kaufmann M, Kohlbacher O, Lenhof HP: **BNDB - The Biochemical Network Database.** *BMC Bioinformatics* 2007, **8:**367.1–9.
62. Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY, Stitt M: **MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes.** *The Plant Journal* 2004, **37:**914–939.
63. Mochida K, Shinozaki K: **Advances in omics and bioinformatics tools for systems analyses of plant functions.** *Plant Cell Physiology* 2011, **52**(12):2017–2038.

doi:10.1186/1752-0509-6-139
Cite this article as: Rohn *et al.*: VANTED v2: a framework for systems biology applications. *BMC Systems Biology* 2012 **6**:139.

4.2 DBE2 – Management of experimental data for the VANTED system

In the following reference the first author is underlined and I am marked in bold.
Hendrik Mehlhorn and Falk Schreiber. DBE2 - management of experimental data for the VANTED system. *Journal of integrative bioinformatics*, 8(2):162+, July 2011. *doi:10.2390/biecoll-jib-2011-162*

`http://journal.imbio.de/index.php?paper_id=162`

Journal of Integrative Bioinformatics, 8(2):162, 2011 http://journal.imbio.de

DBE2 – Management of experimental data for the VANTED system

Hendrik Mehlhorn[1*], Falk Schreiber[1,2]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research, Corrensstr. 3, 06466 Gatersleben, Germany

[2]Martin-Luther-University Halle-Wittenberg, Institute of Computer Science, Von-Seckendorff-Platz 1, 06120 Halle, Germany

Summary

DBE2 is an information system for the management of biological experiment data from different data domains in a unified and simple way. It provides persistent data storage, worldwide accessibility of the data and the opportunity to load, save, modify, and annotate the data. It is seamlessly integrated in the VANTED system as an add-on, thereby extending the VANTED platform towards data management. DBE2 also utilizes controlled vocabulary from the Ontology Lookup Service to allow the management of terms such as substance names, species names, and measurement units, aiming at an eased data integration.

1 Introduction

High throughput phenotyping facilities and modern wet lab techniques such as GC/MS, multidimensional protein gels, and microarrays produce an continuously increasing amount of biological data sets. Each data set comprises a set of biological measurements as well as annotation data.

The data type of biological measurements is in many cases a simple decimal number. A decimal number may represent, for example, the concentration of a metabolite, the relative content of a messenger RNA, or the propotion of the expression levels of an enzyme under various conditions. Ordered sets of decimal numbers also represent one dimensional gradients, such as the concentration of a metabolite in a cell over time. Upcoming facilities enable the high throughput phenotyping of, for instance, plants, which yields a hugh amount of two dimensional images. Other techniques such as NMR or CT produce three dimensional volume data. The magnitude of biological measurement data necessitates data management systems in order to enable appropriate data analysis and data exchange.

Biological measurements arise in the context of certain experiment conditions and represent properties of specific biological entities (e.g. the concentration of a metabolite). This is reflected in the annotation data. Experiment conditions such as the availability of nutrients, water, and light, the time point of the measurement, the underlying genotype, and tissue constitute notable annotation data in praxis. Annotation data also comprises the names of measured biological entities as well as the unit of biological measurements. Standards such as PEDRo,

*To whom correspondence should be addressed. Email: mehlhorn@ipk-gatersleben.de

doi:10.2390/biecoll-jib-2011-162

Journal of Integrative Bioinformatics, 8(2):162, 2011 http://journal.imbio.de

MIAME, and ArMet [1, 2, 3] provide all necessary fields to reproduce the underlying biological experiment. These formats have been proposed for the standardization of annotation data, but the input demanded from the user is exhaustive which often prevents their usage. Fortunately bioinformatic analysis techniques such as flux balance analysis, correlation analysis, and network mapping, need just specific input parameters. The reduced specification demand of these techniques is an advantage which eases fast and appropriate data set explorations.

A complex biological problem often necessitates the examination of data sets from different data domains. Most existing databases do mainly only cover single data domains and are not uniformly addressable, which results in the need of data integration systems such as ONDEX or BridgeDB [4, 5].

The distribution of data sets also causes the problem of data annotation inconsistencies. For metabolite terms, protein names, or measurement units various formats or synonym relationships can be found. Ontologies are intended to structure the knowledge of an area of interest. Life science ontologies such as CHEBI (e.g. chemical compounds), NEWT (species taxonomy), and Gene Ontology (e.g. protein functions) are utilizable to overcome data annotation inconsistencies by strictly annotating the biological measurements using ontology terms [6].

The aim of the DBE2 information system (**D**atabase for **B**iological **E**xperiments **2**) is to manage biological data sets from different data domains in a unified and simple way. Only a small amount of annotation data is required, which is nevertheless appropriate for recent analysis techniques. The management and storage of data sets is consistent despite of the domain(s) the data arises from. The focus is on the integration of multimodal biological measurements of different data types such as zero dimensional decimal numbers, one dimensional gradients, two dimensional images, three dimensional volumes, and even biological networks being interesting in the context of the biological measurements.

The DBE2 information system is based on the DBE information system [7]. The DBE information system proved its usefulness for biologists by an easy usage, the permanent availability, and it's focus on the integration of data sets stored in the DBE database in a persistent and structured way.

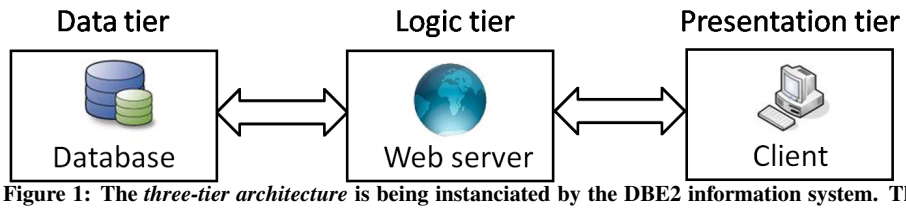


Figure 1: The *three-tier architecture* is being instanciated by the DBE2 information system. The *presentation tier* (the *DBE2 client*) calls the *logic tier* to download, upload, and edit biological data. The *logic tier* (the *DBE2 servlet*) employs the *data tier* (the *DBE2 database* and a file storage system) for the storage of experiment data and binary files. This happens using an underlying user management to manipulate or transfer data by defined queries.

There are several new developments in DBE2. The integration of further data domains and data types now enables a bigger area of application. Recent software engineering techniques helped restructuring the system, yielding a three-tier architecture (see Figure 1). The *DBE2 servlet* was introduced to implement all data accesses to the extended *DBE2 database* by queries and to facilitate the worldwide availability of the data. Big Files in biological data sets are stored in a *hierarchical storage management* by the *DBE2 servlet* to maintain the database performance.

doi:10.2390/biecoll-jib-2011-162

Journal of Integrative Bioinformatics, 8(2):162, 2011 http://journal.imbio.de

The *DBE2 client* in the shape of a VANTED [8] add-on now provides a graphical user interface to the DBE2 information system. Each data set access is controlled by defined servlet queries including an user account management. For the convenient integration of further clients a library is being supported which implements all *DBE2 servlet* calls in a functional way.

In addition the utilization of controlled vocabulary from the *Ontology Lookup Service* [9] raises the quality of annotation data such as species names, substance names, and measurement units. An adaption of certain ontology structures enables the organization of the underlying data sets in a reasonable and intuitive way.

This paper is organized as follows. The (1) *Introdution* conveys the area and the background of the DBE2 information system. Section (2) *DBE2 schema* discusses the representation of data sets in the *DBE2 database* and the whole system. A servlet enables the continuous and worldwide access to data sets, which is being introduced in Section (3) *DBE2 servlet*. The *DBE2 client* is designed as an graphical user interface to the system which is presented in Section (4) *DBE2 client*. This paper closes with a (5) *Discussion* to resume and discuss the presented content.

2 DBE2 schema

The DBE2 information system is designed to handle biological measurements of diverse data types and from various data domains such as metabolomics, proteomics, and phenomics. This happens in the context of the adjacent annotation data. A tree data structure of four levels represents data sets in an adequate and flexible way (see Figure 2).

A data set of biological measurements with it's annotation data is called an *experiment*. It contains meta information such as the experiment name, the coordinator, and the start date of the project. An *experiment* branches into a set of *condition*s as the experiment context of the measured data. Each *condition* represents the species, genotype, and variety as well as the treatment of the examined biological being. Each *condition* branches into a set of *sample*s, which specifiy the measurings of the underlying measured data in space and time. Measured data is allowed in the shape of (i) simple decimal numbers, (ii) pictures, and (iii) volumes which correspond to data of zero, two, and three dimensions. It is also possible to represent biological (iv) networks as well as one dimensional (v) gradients (by a set of ordered (i) decimal numbers). An abstract example is shown in Figure 2.

The resulting hierarchical tree structure is being implemented in the shape of relational tables in the *DBE2 database* as the *data tier* for a persistent and structured data storage as well as by a XML document schema for data exchange tasks. This enables dealing with experiment data of different types from various domains in a unified way using a corporate data structure.

The *DBE2 database* schema is implemented in an Oracle database (version 11g) and is shown in Figure 3. The database schema consists of four conceptual modules, namely the (i) *User management* module, the (ii) *Experiment data* module, the (iii) *Basis data* module, and the (iv) *Supplementary material* module. The (i) *User management* module provides a basic user right handling. Users need to possess a DBE2 account to store experiments in the *DBE2 database*. User accounts get organized via user groups. For every stored experiment a user group is defined with users having the right to access it. The experiment data is being stored in the (ii)

doi:10.2390/biecoll-jib-2011-162

Journal of Integrative Bioinformatics, 8(2):162, 2011 http://journal.imbio.de

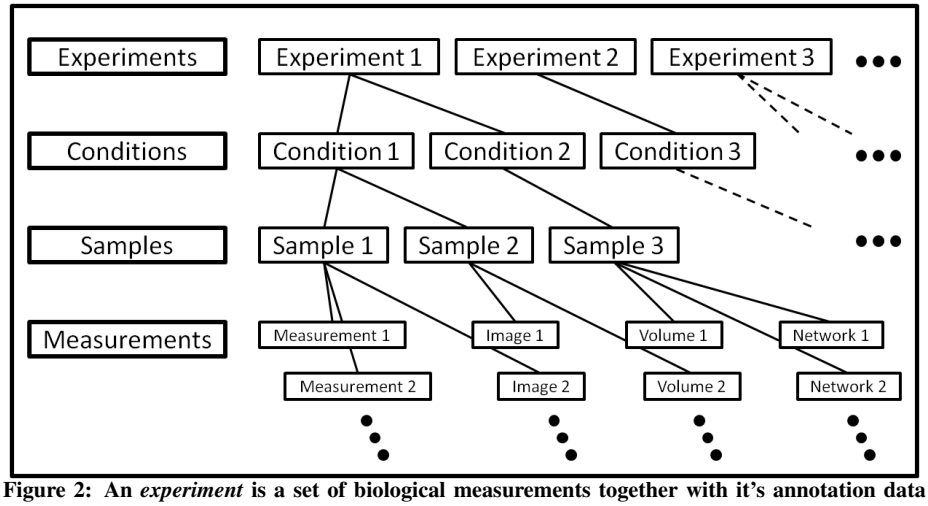


Figure 2: An *experiment* is a set of biological measurements together with it's annotation data and is represented by a tree structure in four levels. It branches into a set of *condition*s which represent the treatment and the genotype of the examined species. Each *condition* forks into a set of *sample*s, which provide information about the underlying measured data. Measured data is allowed to exhibit the shape of decimal numbers, images, volumes, and biological networks.

Experiment data module. This is done one to one according to the described hierarchical tree structure. A subset of the annotation data is being managed in the (iii) *Basis data* module. This module is designed to organize annotation data through controlled vocabulary of various areas such as species names, unit terms, and substance names. In addition the (iv) *Supplementary material* module enables to associate arbitrary supplementing files to any level of the *Experiment data* module as supplementary material. An important supplementing file could be a SDS gel image as the source of measured protein expression levels for instance.

Users of DBE are invited to supply annotation data to biological measurements according to the introduced hierarchical tree structure. Thus the support of information about the experiment in general, the species, the examined tissue, the time points of the biological measurements, and the name of the measured entities is necessary. This is sufficient for tasks such as the statistical comparison of data sets from varios differential treated breeding lines.

3 DBE2 servlet

Every request of the *DBE2 client* to the *DBE2 database* is implemented by the *DBE2 servlet*. The *DBE2 servlet* defines a set of queries, which builts an application programming interface (API). The encapsulation of *DBE2 database* transactions via a servlet as the *logic tier* assures worldwide data access and safe database manipulations. The *DBE2 servlet* is implemented as an Java HTTP servlet, which enables a dynamic treatment of all queries on the web server.

The *DBE2 servlet* implements the observance of the user right management. Every request associated to confidential user data sets contains parameters to authorize it. The handling of unexpected cases and errors includes exception reports and *DBE2 database* rollbacks. In this way the *DBE2 servlet* assures a consistent and safe data storage.

doi:10.2390/biecoll-jib-2011-162

Journal of Integrative Bioinformatics, 8(2):162, 2011 http://journal.imbio.de

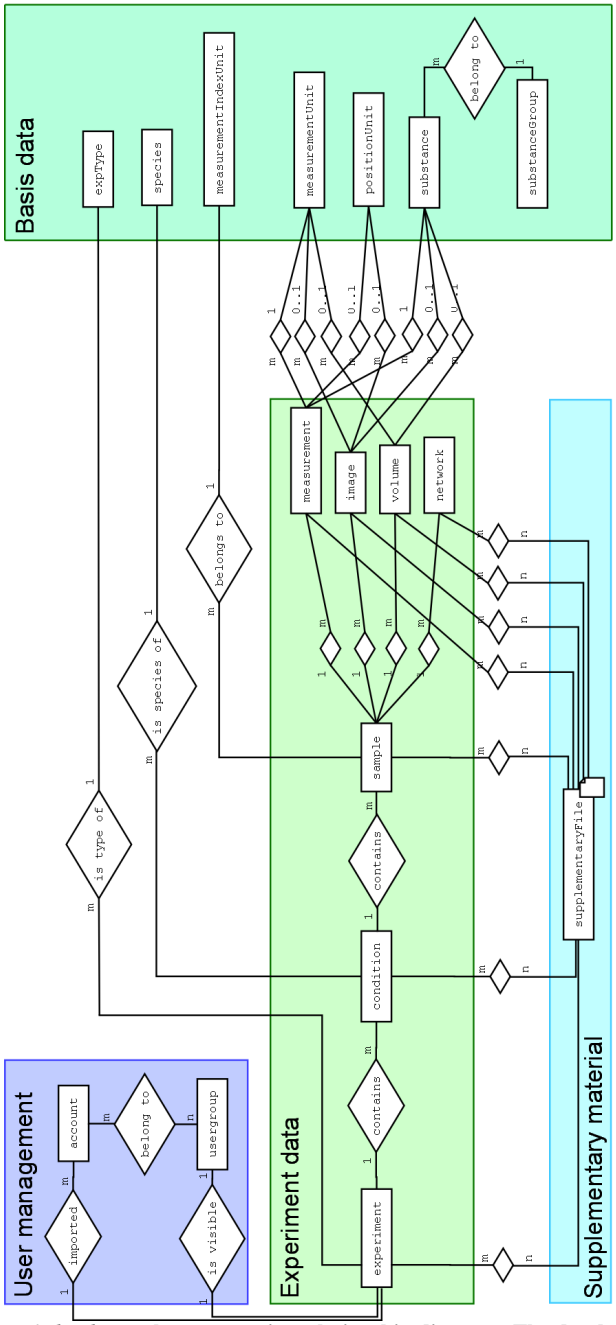


Figure 3: The *DBE2 database* schema as entity relationship diagram. The database schema comprises four modules: (i) *User management* module for user right handling, (ii) *Experiment data* module for experiment data storage, (iii) *Basis data* module for controlled vocabulary, and a (iv) *Supplementary material* module for the association of arbitrary files with entries in the *Experiment data* module.

doi:10.2390/biecoll-jib-2011-162

Journal of Integrative Bioinformatics, 8(2):162, 2011 http://journal.imbio.de

The queries provided by the *DBE2 servlet* enable an functional access to the database. Every client functionality regarding the *DBE2 database* is being realized by a certain set of *DBE2 servlet* queries. This prevents the client developer from communicating directly with the database via SQL queries, yielding a rather clean and safe development style.

The transfer of data from the *DBE2 client* to the *DBE2 servlet* and back works through streams, which afford the transfer of arbitrary sized data. For database performance reasons, binary files are stored using a *hierarchical storage management* (HSM), which realizes a compromise between data transfer performance and storage costs. This is implemented by the unpublished BFiler package in a transactional and simple way. Thus BFILEs (an Oracle SQL datatype) are stored in the *DBE2 database* as references to the physical file in the HSM instead of BLOBs (another Oracle SQL datatype which represents the whole file data).

4 DBE2 client

The *DBE2 client* supports a graphical user interface (GUI) to the DBE2 information system and represents an instance of the *presentation tier*. It executes all user actions regarding the *DBE2 database* via a set of *DBE2 servlet* queries. The *DBE2 client* is designed as a VANTED add-on, which smoothly enables the usage of analysis and visualization techniques of the VANTED system [8]. It is also possible to implement further clients for the usage of the system. Therefore a java library called the *DBE2 servlet client* is being supported to communicate with the *DBE2 servlet* in a functional and easy way.

The *DBE2 client* provides a user-friendly and easy way to upload, edit, and download *experiments* and to edit the annotation data of the experiment in the local XML format representation. In addition, the DBE2 information system administrator is allowed to manage terms in the *DBE2 database Basis data* module and to manage user rights by changing entries in the *DBE2 database User management* module. There is also a special public account which may be used by any user. This account enables users without an DBE2 account to download experiments which were explicitly approved for public access. See Figure 4 for an overview of all impotant *DBE2 client* use cases.

In the case of forbidden queries or unexpected cases the *DBE2 servlet* throws exceptions processed by the *DBE2 client*. In the case of known errors such as the injury of database table constraints, this happens either in way of offering an alternative proceeding or by descriptive messages. An example is that the name of *experiment*s has to be unique.

The usage of a controlled vocabulary for the standardization of annotation data supports easy mapping of data sets onto each other and onto biological networks. In case of the DBE2 information system the controlled vocabulary is being represented by the *Basis data* module in the *DBE2 database*. Whenever an user attempts to upload an *experiment* to the *DBE2 database*, the annotation data of the *experiment* has to be covered by the *Basis data* module. Uncovered annotation data has to be synchronized. Terms in the annotation data of the *experiment* which are missing in the *Basis data* module have to be added to expand the basis data pool or renamed to match the basis data pool. The *DBE2 client* offers the possibility to standardize annotation data by utilizing the *Ontology Lookup Service* (OLS). The OLS (*http://www.ebi.ac.uk/ontology-lookup/*) is a compendium of more than 70 life science ontologies retrievable through an unified interface. Users are able to search various ontologies for terms in the annotation data of experi-

doi:10.2390/biecoll-jib-2011-162

Journal of Integrative Bioinformatics, 8(2):162, 2011 http://journal.imbio.de

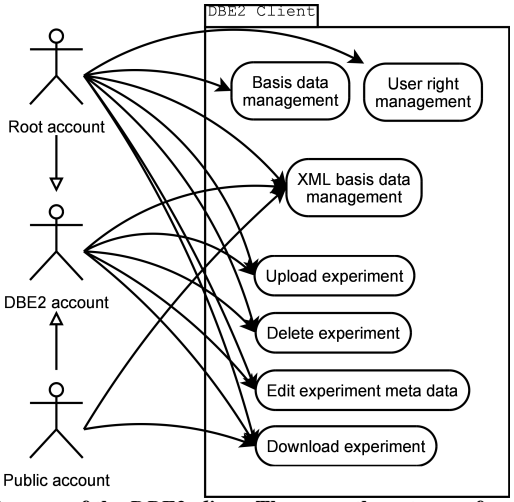


Figure 4: Use cases diagram of the *DBE2 client*. There are three types of accounts. The special (i) public account may used by everybody and is allowed to download particular experiments and to edit the basis data in the local XML representation of experiments. Users with a (ii) regular DBE2 account may upload, download, and edit experiments in addition. The DBE2 administrator uses the (iii) root account and is thus allowed to access all features including the management of *DBE2 database* basis data and the appointment of user rights.

ments. Species names and the according taxonomy tree are accessible from the ontology *NEWT UniProt Taxonomy Database* (short name: NEWT) [10]. Names of chemical compounds and the corresponding compound taxonomy are accessible from the ontology *Chemical Entities of Biological Interest* (short name: CHEBI) [6]. With the help of these ontologies, a part of the *DBE2 client* named *experiment manager* (see Figure 5), is able to represent the set of accessible *experiment*s in a hierarchical way according to their annotation data.

Since the *DBE2 client* is integrated seamlessly into VANTED, the DBE2 information system user is instantly able to use the functionalities of the VANTED system. These include the mapping of *experiment*s on biological networks, the arrangement of *experiment*s according to KEGG pathway hierarchies, and the corresponding visualization via a graph.

5 Discussion

In this paper the functionalities and the *three-tier architecture* of the DBE2 information system were presented. The DBE2 information system is designed for the management of biological measurements of various domains and types in an unified and easy way. The *DBE2 database* represents the *data tier* and stores biological data in a structured and persistent way. The *DBE2 servlet* represents the *logic tier* and realizes all data access and data manipulation in a functional and safe way. The *DBE2 client* represents the *presentation tier* and provides an easy to use GUI to the DBE2 information system.

Ontology support from the OLS aids the standardization of annotation data. The according controlled vocabulary eases the integrated analysis of biological measurements from various data sets. Term taxonomies supported by life science ontologies enable a logically structured

doi:10.2390/biecoll-jib-2011-162

Journal of Integrative Bioinformatics, 8(2):162, 2011 http://journal.imbio.de

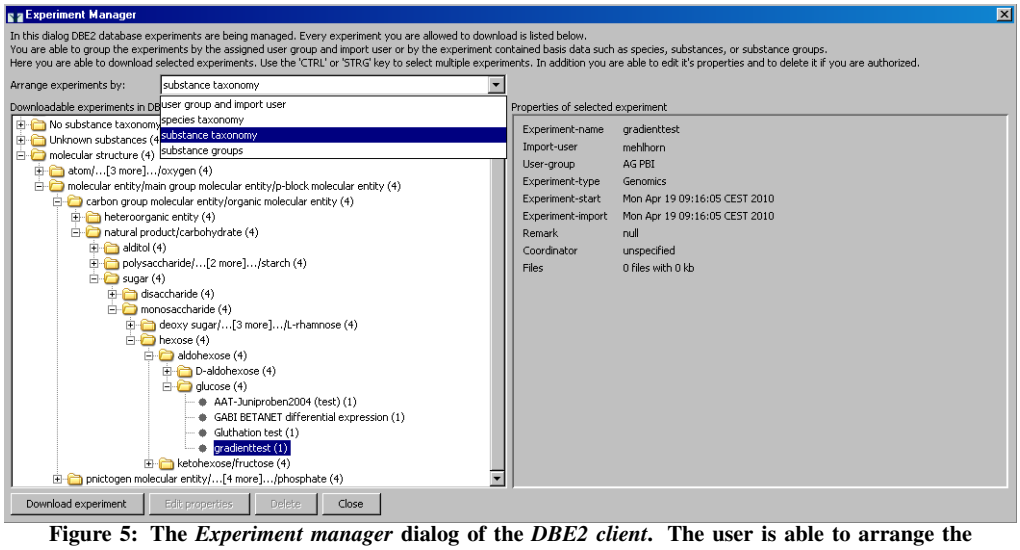


Figure 5: The *Experiment manager* dialog of the *DBE2 client*. The user is able to arrange the accessible *experiments* in four ways. The (i) user group arrangement helps to survey user rights to *experiments*. The (ii) substance taxonomy arrangement (shown), the (iii) species taxonomy arrangement, and the (iv) substance group arrangement enable an *experiment* overview according to the annotation data.

and intuitive survey over a big number of data sets.

Techniques from the seamlessly integrated VANTED system such as data mapping become even more powerful in the course of the consistent usage of ontology supported annotation data. For several use case examples of the VANTED system see [11, 12, 13].

The DBE2 information system is currently in use for three projects with users from the IPK as well as external users. There are 43 registered users and 73 experiments represented in the *DBE2 database*.

Future work concerns the support of additional features such as an extended ontology support for an enlarged data integration potential. In the course of a growing user community and usage of the DBE2 information system the identification and elimination of potential performance bottlenecks will be of great importance.

6 Availability and Requirements

- DBE2 information system: *DBE2 client* and *DBE2 servlet client*
  - DBE2 Web site: *http://www.vanted.org/addons/DBE2/index.html*
  - License: GNU General Public License
  - Programming language: Java version 1.5 or higher
  - Requirements: The *DBE2 client* requires the VANTED program

doi:10.2390/biecoll-jib-2011-162

Journal of Integrative Bioinformatics, 8(2):162, 2011 http://journal.imbio.de

- VANTED program
  - VANTED web site: *http://www.vanted.org/*
  - License: GNU General Public License
  - Operating system(s): Platform independent
  - Programming language: Java version 1.5 or higher
  - Requirements: Screen resolution of 1024 * 768 or higher, mouse, minimum 512 MB RAM recommended

Acknowledgements

We would like to thank Christian Klukas, who developed the initial DBE information system and contributed to the development of DBE2, Matthias Lange, Steffen Flemming, and Thomas Münch for their developmental support. In addition we would like to thank Hardy Rolletscheck for his valuable feedback on the usability of the *DBE2 client*. We also thank Tobias Czauderna, Hendrik Rohn, Anja Hartmann, Eva Grafahrend-Belau, Astrid Junker, and Matthias Klapperstück for fruitful discussions and testing. We thank Klaus Hippe for assistance in establishing the OLS support. The development of the DBE2 information system was supported in the frame of the project GABI Betanet (BMBF FKZ 0315054B).

References

[1] K. Garwood, T. McLaughlin, C. Garwood, S. Joens, N. Morrison, C. F. Taylor, K. Carroll, C. Evans, A. D. Whetton, S. Hart, D. Stead, Z. Yin, A. J. Brown, A. Hesketh, K. Chater, L. Hansson, M. Mewissen, P. Ghazal, J. Howard, K. S. Lilley, S. J. Gaskell, A. Brass, S. J. Hubbard, S. G. Oliver, and N. W. Paton. Pedro: a database for storing, searching and disseminating experimental proteomics data. *BMC Genomics*, 5(1):68, 2004.

[2] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics*, 29(4):365–371, 2001.

[3] H. Jenkins, N. Hardy, M. Beckmann, J. Draper, A. R. Smith, J. Taylor, O. Fiehn, R. Goodacre, R. J. Bino, R. Hall, J. Kopka, G. A. Lane, B. M. Lange, J. R. Liu, P. Mendes, B. J. Nikolau, S. G. Oliver, N. W. Paton, S. Rhee, U. Roessner-Tunali, K. Saito, J. Smedsgaard, L. W. Sumner, T. Wang, S. Walsh, E. S. Wurtele, and D. B. Kell. A proposed framework for the description of plant metabolomics experiments and their results. *Nature Biotechnology*, 22(12):1601–1606, 2004.

[4] J. Köhler, J. Baumbach, J. Taubert, M. Specht, A. Skusa, A. Rüegg, C. Rawlings, P. Verrier, and S. Philippi. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, 22(11):1383–1390, 2006.

doi:10.2390/biecoll-jib-2011-162

Journal of Integrative Bioinformatics, 8(2):162, 2011 http://journal.imbio.de

[5] M. van Iersel, A. Pico, T. Kelder, J. Gao, I. Ho, K. Hanspers, B. Conklin, and C. Evelo. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*, 11(1):5, 2010.

[6] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(suppl_1):D344–D350, 2008.

[7] L. Borisjuk, M.R. Hajirezaei, C. Klukas, H. Rolletschek, and F. Schreiber. Integrating data from biological experiments into metabolic networks with the DBE information system. *In Silico Biology*, 5(2):93–102, 2005.

[8] B. Junker, C. Klukas, and F. Schreiber. Vanted: A system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics*, 7(1):109.1–13, 2006.

[9] R. Cote, P. Jones, R. Apweiler, and H. Hermjakob. The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, 7(1):97, 2006.

[10] I. Q. Phan, S. F. Pilbout, W. Fleischmann, and A. Bairoch. NEWT, a new taxonomy portal. *Nucleic Acids Research*, 31(13):3822–3823, 2003.

[11] T. Czauderna, C. Klukas, and F. Schreiber. Editing, validating and translating of SBGN maps. *Bioinformatics*, 26(18):2340–2341, 2010.

[12] E. Grafahrend-Belau, C. Klukas, B. H. Junker, and F. Schreiber. FBA-SimVis: interactive visualization of constraint-based metabolic models. *Bioinformatics*, 25(20):2755–2757, 2009.

[13] C. Klukas and F. Schreiber. Integration of -omics data and networks for biomedical research with VANTED. *Journal of Integrative Bioinformatics*, 7(2):112, 2010.

doi:10.2390/biecoll-jib-2011-162

5 Phylogenetic footprinting

Publications presented in this thesis related to "Phylogenetic footprinting" are entitled "Detecting and correcting the binding-affinity bias in ChIP-seq data using inter-species information" (Nettling, Treutler, Cerquides, et al., 2016) and "DiffLogo: a comparative visualization of sequence motifs" (Nettling, Treutler, Grau, et al., 2015).

5.1 Detecting and correcting the binding-affinity bias in ChIP-seq data using inter-species information

In the following reference the first author is underlined and I am marked in bold.
Martin Nettling, **Hendrik Treutler**, Jesus Cerquides, and Ivo Grosse. Detecting and correcting the binding-affnity bias in ChIP-seq data using inter-species information. *BMC genomics*, 17(1):347+, May 2016. *doi:10.1186/s12864-016-2682-6*

`https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-2682-6`

Nettling *et al. BMC Genomics* (2016) 17:347
DOI 10.1186/s12864-016-2682-6

BMC Genomics

METHODOLOGY ARTICLE Open Access

Detecting and correcting the binding-affinity bias in ChIP-seq data using inter-species information

Martin Nettling[1*], Hendrik Treutler[2], Jesus Cerquides[3] and Ivo Grosse[1,4]

Abstract

Background: Transcriptional gene regulation is a fundamental process in nature, and the experimental and computational investigation of DNA binding motifs and their binding sites is a prerequisite for elucidating this process. ChIP-seq has become the major technology to uncover genomic regions containing those binding sites, but motifs predicted by traditional computational approaches using these data are distorted by a ubiquitous binding-affinity bias. Here, we present an approach for detecting and correcting this bias using inter-species information.

Results: We find that the binding-affinity bias caused by the ChIP-seq experiment in the reference species is stronger than the indirect binding-affinity bias in orthologous regions from phylogenetically related species. We use this difference to develop a phylogenetic footprinting model that is capable of detecting and correcting the binding-affinity bias. We find that this model improves motif prediction and that the corrected motifs are typically softer than those predicted by traditional approaches.

Conclusions: These findings indicate that motifs published in databases and in the literature are artificially sharpened compared to the native motifs. These findings also indicate that our current understanding of transcriptional gene regulation might be blurred, but that it is possible to advance this understanding by taking into account inter-species information available today and even more in the future.

Keywords: Binding-affinity bias, ChIP-seq, Phylogenetic footprinting, Evolution, Transcription factor binding sites, Gene regulation

Background

Predicting transcription factor binding sites and their motifs is essential for understanding transcriptional gene regulation and thus of importance in almost all areas of modern biology, medicine, and biodiversity research [1, 2]. Countless approaches exist for predicting motifs from these genomic regions [3–6], but predicting motifs from ChIP-seq data and similar experimental data is hampered by the contamination with false positive genomic regions as well as the enrichment of high-affinity binding sites [7–9].

The contamination with false positive genomic regions is caused by at least three reasons. First, the transcription factor or other DNA binding protein pulled down by immunoprecipitation may not bind directly to the binding site [10]. Second, ChIP-seq target regions may not contain a binding site due to experimental settings such as sequencing depth or DNA fragment length [11, 12]. Third, false positive regions may be predicted in the subsequent ChIP-seq data analysis due to never perfect analysis pipelines and too low signal cutoff thresholds [8]. These three effects may lead to the selection of false positive ChIP-seq regions that do not contain at least one binding site.

The enrichment of high-affinity binding sites is caused by at least two reasons. First, most antibodies have a preference of binding high-affinity binding sites with a higher probability than low-affinity binding sites, causing the set

*Correspondence: martin.nettling@informatik.uni-halle.de
[1] Institute of Computer Science, Martin Luther University, Halle (Saale), Germany
Full list of author information is available at the end of the article

of binding sites bound in the ChIP-seq experiment to be partially different from the set of binding sites bound in vivo [13, 14]. Second, true positive regions with low-affinity binding sites are rejected due to too high signal cutoff thresholds [5, 8]. These two effects may lead to an under-representation of low-affinity binding sites and an over-representation of high-affinity binding sites in ChIP-seq regions.

Taken together, the contamination with false positive genomic regions leads to the *contamination bias* [15] and thus to the prediction of artificially softened motifs, whereas the enrichment of sequences with high-affinity binding sites leads to the *binding-affinity bias* [16] and thus to the prediction of artificially sharpened motifs. Neglecting these effects leads to distorted motifs and could potentially affect all downstream analyses [17–20]. Existing approaches for predicting motifs are capable of detecting and correcting the contamination bias, which has been found to increase the quality of motif prediction considerably [8, 21, 22], and here we investigate the possibility of detecting and correcting the binding-affinity bias.

Detecting the binding-affinity bias seems impossible based on sequence data from one species alone, but it seems possible based on inter-species information. This is possible due to the fact that the binding-affinity bias is stronger in the target regions of the ChIP-seq experiment in the reference species than in orthologous regions of phylogenetically related species. This stronger binding-affinity bias yields more biased motifs in the reference species than in phylogenetically related species, and this difference may be used for detecting and potentially correcting the binding-affinity bias.

Phylogenetic footprinting models typically (i) take into account ChIP-seq data of only one species and (ii) do not take into account heterogeneous substitution rates among different DNA regions, heterotachious evolution of DNA regions, and loss-of-function mutations in binding sites. The consideration of (i) ChIP-seq data of more than one species and (ii) heterogeneity, heterotachy, and loss-of-function mutations are likely to improve both phylogenetic footprinting as well as the detection and correction of the binding-affinity bias, but in this work we investigate if the detection and correction of this bias is possible based on (i) ChIP-seq data of only one species and (ii) a simple phylogenetic footprinting model that neglects heterogeneity, heterotachy, and loss-of-function mutations.

We first investigate if the effect of observing more biased motifs in the reference species than in phylogenetically related species is measurable beyond statistical noise in target regions of five ChIP-seq data sets of human and in orthologous regions of monkey, dog, cow, and horse. We then develop a phylogenetic footprinting model that incorporates the binding-affinity bias, investigate if this model improves or deteriorates motif prediction compared to traditional models that do not incorporate it, and compare the motifs predicted with and without the correction of the binding-affinity bias.

Results and discussion

In subsection "Using sequence-information of phylogenetically related species to detect the binding-affinity bias", we describe the basic idea of how the binding-affinity bias could be detected based on inter-species information using a toy example. In the remaining subsections we perform three studies based on ChIP-seq data sets of five transcription factors and on multiple alignments of the human ChIP-seq target regions with orthologous regions from monkey, dog, cow, and horse. In subsection "Decrease of information contents in motifs from phylogenetically related species" we investigate if the effect of observing more biased motifs in the reference species than in phylogenetically related species is measurable in these five data sets. In subsection "Modeling the binding-affinity bias increases classification performance", we investigate if a correction of the binding-affinity bias leads to an improvement or a deterioration of the classification performance. In subsection "Modeling the binding-affinity bias leads to softened motifs", we compare the sequence motifs predicted with and without the correction of the binding-affinity bias.

Using sequence-information of phylogenetically related species to detect the binding-affinity bias

Detecting and correcting the binding-affinity bias might be possible because the binding-affinity bias inherent to the ChIP-seq experiment in the reference species (Fig. 1a) is stronger than the indirect binding-affinity bias in orthologous regions from phylogenetically related species. Under this assumption, the information content of the predicted motifs [23] should decrease with the phylogenetic distance from the reference species due to the increasing number of mutations.

To illustrate this idea, we present a toy example consisting of six binding sites from four phylogenetically related species in Fig. 1b and Table 1. In this toy example, we assume an exaggerated binding-affinity bias of three high-affinity binding sites captured by the ChIP-seq experiment and three low-affinity binding sites not captured by the ChIP-seq experiment. In real world applications the native motif is unknown and the motif predicted on the available data is biased to an unknown degree. In the presented toy example, however, the native motif is considered to be known so that the effect of the binding-affinity bias on the motifs of the reference species (species 1) and the phylogenetically related species (species 2, 3, and 4) can be illustrated.

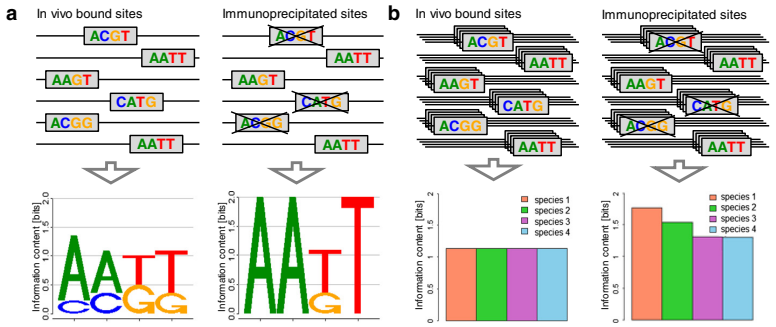


Fig. 1 Influence of binding-affinity bias on information content. **a** Binding-affinity bias in the reference species. The left column shows binding sites bound in vivo as well as the sequence logo. In the right column, enrichment of high-affinity binding sites by chromatin immunoprecipitation leads to a different motif with higher information content. **b** Binding-affinity bias in the reference species and three phylogenetically related species. The left column shows binding sites bound in vivo and the information content of the species–specific motifs. In the right column, the enrichment of high-affinity binding sites in the reference species and the other three species leads to different motifs with different information content in each species. The effect of this enrichment decreases with the phylogenetic distance from the reference species as reflected by decreasing information contents. Please find the sequences of all species in Table 1

Table 1 Influence of binding-affinity bias on information content. We illustrate the effect of binding-affinity bias with the given toy example of a ChIP-seq experiment for six binding sites in four species. Due to low binding-affinity, red binding sites are insufficiently bound. This results in the absence of red binding sites in the measured data which we denote binding-affinity bias. Binding sites with low binding-affinity typically comprise dissimilar bases in contrast to black binding sites with high affinity and common bases. The absence of red binding sites leads to a sharpening of the resulting motif, which we indicate using the information content. The information content without binding-affinity bias is equal in all species, whereas the information content with binding-affinity bias increases in all species. The vital point is that the effect of binding-affinity bias decreases with phylogenetic distance, which involves an increasing number of mutations. Please find a visualization of this toy example in Fig. 1b

	Species 1	Species 2	Species 3	Species 4
Binding site 1	A C G T	A C G T	A C T T	A A T T
Binding site 2	A A T T	A A T T	C A G T	A C G T
Binding site 3	A A G T	C A T G	A A G T	A A T G
Binding site 4	C A T G	A A G T	A C T G	A A G T
Binding site 5	A C G G	A C G G	A A G T	C A G T
Binding site 6	A A T T	A A T T	A A T G	A C T G
Number of mutations in all binding sites	0	6	9	14
Information content without binding-affinity bias	1.13	1.13	1.13	1.13
Information content with binding-affinity bias	1.77	1.54	1.31	1.31

The motif predicted from the three target regions containing high-affinity binding sites is strongly biased in reference species 1, and it is impossible to predict the native motif from only those three target regions. However, a shadow of this strong binding-affinity bias also exists in orthologous regions of species 2, 3, and 4, so the motifs predicted from these orthologous regions in species 2, 3, and 4 are biased, too. This bias in species 2, 3, and 4, however, is weaker than the bias in reference species 1, and this difference can be exploited for detecting and correcting the binding-affinity bias and for predicting the native motif from the three target regions of high-affinity binding sites in reference species 1 and their orthologous regions in species 2, 3, and 4.

Specifically, the binding-affinity bias introduced by the ChIP-seq experiment in reference species 1 causes a strong increase of the information content of the predicted motif (1.77 bit) compared to the native motif (1.13 bit). The shadow of the binding-affinity bias in species 2, 3, and 4 also causes an increase of the information contents of the motifs predicted in species 2 (1.54 bit), species 3 (1.31 bit), and species 4 (1.31 bit), but this increase in species 2, 3, and 4 is smaller than in reference species 1 (Table 1 and Fig. 1b). The increase of information content decreases with the number of observed mutations and thus the phylogenetic distance of species 2, 3, and 4 to reference species 1 in which the ChIP-seq experiment has been performed. Hence, the observation of a decreased information content of motifs predicted in orthologous regions of phylogenetically related species compared to the information content of the motif predicted in the

reference species could indicate the presence of a binding-affinity bias and possibly allow the correction of that bias.

Decrease of information contents in motifs from phylogenetically related species

We investigate this hypothesis on human ChIP-seq data of five transcription factors [10, 24] and multiple alignments of the human ChIP-seq target regions with orthologous regions from monkey, dog, cow, and horse [25] ("Data" Methods). We calculate the information contents of motifs from human (reference species), monkey, dog, cow, and horse for each of the five data sets ("Decrease of information contents in motifs from related species" Methods) and present the results in Fig. 2. We find for each of the five data sets that the information content of the motif from the reference species is significantly higher ($p < 1.83 \times 10^{-14}$, Wilcoxon Signed-Rank Test, Additional file 1: Table S1) compared to the information contents of the motifs from monkey, dog, cow, and horse.

Modeling the binding-affinity bias increases classification performance

Motivated by this observation, we develop a phylogenetic footprinting model capable of taking into account the contamination bias ($\mathcal{M}^{\mathrm{C}}_{-}$), the binding-affinity bias ($\mathcal{M}^{-}_{\mathrm{BA}}$), neither one or the other $\mathcal{M}^{-}_{-}$, or both ($\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$) ("Modeling the binding-affinity bias" Methods and Additional file 1: Section 1). In order to study to which degree these models are capable of modeling multiple alignments originating from ChIP-seq data, we consider the principle of parsimony [26], which states that the simplest of competing explanations is the most likely to be correct. As the new model $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ is more complex than the traditional model $\mathcal{M}^{\mathrm{C}}_{-}$, we should accept it only if it provides a more accurate representation of the data. A standard approach for measuring how accurately a model represents a data set is to measure its performance of classifying, in this case, motif-bearing and non-motif-bearing alignments, and a standard approach for measuring classification performance is stratified repeated random sub-sampling validation ("Measuring classification performance" Methods, Fig. 5).

Using this approach we measure the performance of the four models $\mathcal{M}^{-}_{-}$, $\mathcal{M}^{-}_{\mathrm{BA}}$, $\mathcal{M}^{\mathrm{C}}_{-}$, and $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ to classify each of the five data sets against the other four. Fig. 3a shows that $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ yields a higher classification performance than $\mathcal{M}^{\mathrm{C}}_{-}$ in all five data sets ($p < 2.3 \times 10^{-17}$, Wilcoxon Signed-Rank Test, Additional file 1: Table S2), indicating that the new model $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ is more realistic than the traditional model $\mathcal{M}^{\mathrm{C}}_{-}$. We also find that $\mathcal{M}^{-}_{\mathrm{BA}}$ yields a significantly higher classification performance than $\mathcal{M}^{\mathrm{C}}_{-}$ in all five data sets ($p < 1.8 \times 10^{-17}$, Wilcoxon Signed-Rank Test), which indicates that taking into account the binding-affinity bias has a larger impact on the classification performance than taking into account the contamination bias (Additional file 1: Figure S1, Figure S2, Figure S10, Figure S11, Figure S12, Figure S13, Figure S14, Figure S15 and Figure S16).

Modeling the binding-affinity bias leads to softened motifs

Next, we investigate the information contents of the corrected motifs predicted by models $\mathcal{M}^{-}_{\mathrm{BA}}$ and $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ that take into account the binding-affinity bias and the traditional motifs predicted by models $\mathcal{M}^{-}_{-}$ and $\mathcal{M}^{\mathrm{C}}_{-}$ that neglect this bias. Fig. 3b shows that the information contents of motifs predicted by $\mathcal{M}^{\mathrm{C}}_{-}$ are significantly higher than the information contents of motifs predicted by $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ ($p < 4.0 \times 10^{-18}$, Wilcoxon Signed-Rank Test). We also find that the information contents of motifs predicted by $\mathcal{M}^{-}_{-}$ are higher than the information contents of motifs predicted by $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ ($p < 4.0 \times 10^{-18}$, Wilcoxon Signed-Rank Test, Additional file 1: Table S4), stating that

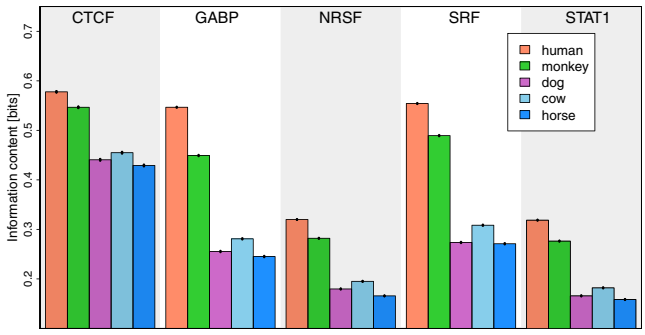


Fig. 2 Mean information content and standard error for motifs of five transcription factors in five species. The information content of motifs in the reference species (human) is significantly higher compared to the four phylogenetically related species ($p < 1.8 \times 10^{-14}$). The information content typically decreases with the phylogenetic distance from the reference species

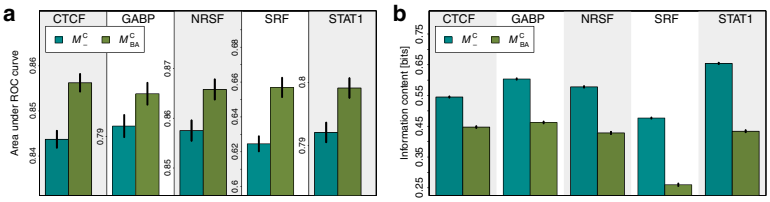


Fig. 3 Comparison of models $\mathcal{M}^{\mathrm{C}}_{-}$ and $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$. **a** Mean classification performance and standard error of the two models $\mathcal{M}^{\mathrm{C}}_{-}$ and $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ quantified by the area under the receiver operating characteristic curve. We find for each of the five data sets a significantly increased classification performance for $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ compared to $\mathcal{M}^{\mathrm{C}}_{-}$. Examples for ROC curves are shown in Additional file 1: Figure S10, Figure S11, Figure S12, Figure S13, Figure S14 and Figure S15. **b** Mean information content and standard error of the motifs predicted by the two models $\mathcal{M}^{\mathrm{C}}_{-}$ and $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$. We find for each of the five data sets a significantly decreased information content in motifs predicted by $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ compared to $\mathcal{M}^{\mathrm{C}}_{-}$ ($p < 4.0 \times 10^{-18}$)

the binding-affinity bias is stronger than the contamination bias. Equivalently, this states that the joint effect of both biases leads to an artificial sharpening of the motifs and an artificial overestimation of the binding affinities (Additional file 1: Figure S3, Figure S4, Figure S17, Figure S18).

Finally, we inspect the differences of the corrected motifs predicted by $\mathcal{M}^{-}_{\mathrm{BA}}$ and $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ and the traditional motifs predicted by $\mathcal{M}^{-}$ and $\mathcal{M}^{\mathrm{C}}_{-}$. Fig. 4 shows the differences between the base distributions of pairs of motifs for $\mathcal{M}^{\mathrm{C}}_{-}$ and $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ by difference logos ("Visualizing motif differences with DiffLogo" Methods). We find for each of the five data sets that the corrected motifs are softer than the traditional motifs distorted by the binding-affinity bias. Specifically, we find that the amount of decrease of the most abundant bases in the corrected motifs compared to the traditional motifs is roughly proportional to the base abundance, whereas the increase of the remaining bases is not proportional to the base abundance. Hence, the corrected motifs are not simply a uniformly softened version of the traditional motifs, but motifs with different degrees of dissimilarity at different positions (Additional file 1: Figure S5, Figure S6,Figure S7, Figure S8 and Figure S9).

Conclusions

We studied the possibility of detecting and correcting the binding-affinity bias in ChIP-seq data using interspecies information. We found that the fact that this bias is stronger in target regions of the reference species than its shadow in orthologous regions of phylogenetically related species enables the detection and correction of this bias. We proposed a phylogenetic footprinting model capable of taking into account the binding-affinity bias in addition to the contamination bias, and we applied this model and its three special cases that neglect one of the two biases or both to five ChIP-seq data sets. We found by stratified repeated random sub-sampling validation that taking into account the binding-affinity bias always improves motif prediction, that the motif binding-affinity bias leads to a distortion of motifs that is even stronger than the distortion caused by the contamination bias, and that the corrected motifs are typically softer than those predicted by traditional approaches. The comparison of corrected and traditional motifs showed small but noteworthy differences, suggesting that the refinement of traditional motifs from databases and from the literature might lead to the prediction of novel binding sites, *cis*-regulatory modules, or gene-regulatory networks and might thus advance our attempt of understanding transcriptional gene regulation as a whole.

Methods

In this section we describe "Decrease of information contents in motifs from related species" (i) the determination of the information contents of motifs in the reference species and phylogenetically related species, "Modeling the binding-affinity bias" (ii) the phylogenetic footprinting model that can take into account the binding-affinity bias, the contamination bias, neither one or the other, or both, "Measuring classification performance" (iii) the measurement of the classification performance of these four phylogenetic footprinting models using stratified repeated random sub-sampling validation, and "Visualizing motif differences with DiffLogo" (iv) the visualisation of differences between the corrected and the traditional motifs.

Decrease of information contents in motifs from related species

We determine the information content $I(P)$ of a motif P as described in [23]:

$$\begin{aligned} H_\ell(P) &= \log_2(|\mathcal{A}|) - \sum_{a\in\mathcal{A}} p_{\ell,a}\cdot\log_2(p_{\ell,a}) \\ I(P) &= \sum_{\ell=1}^{W} H_\ell(P), \end{aligned} \tag{1}$$

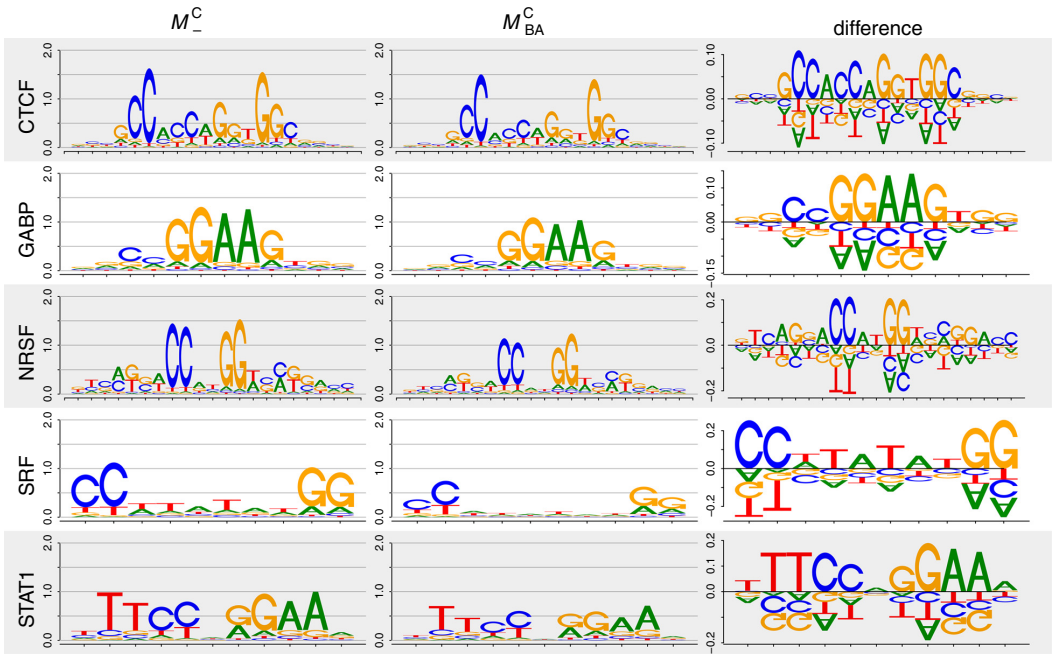


Fig. 4 Differences of motifs predicted by $\mathcal{M}^{C}_{-}$ and $\mathcal{M}^{C}_{BA}$. The height of the base stacks is quantified by the Jensen-Shannon divergence, where high base stacks represent high motif differences. We find significant motif differences exceeding 0.1 bit for all five data sets (Additional file 1: Figure S5, Figure S6, Figure S7, Figure S8 and Figure S9)

where $\mathcal{A} = A, C, G, T$ is the alphabet, $p_{\ell,a}$ is the probability of base a at position ℓ in motif P, and $H_\ell(P)$ denotes the information content of position ℓ in motif P.

We measure the information contents of motifs in five species using repeated random sub-sampling as follows. Initially, we choose one motif for each of the transcription factors CTCF, GABP, NRSF, SRF, and STAT1 from the JASPAR database, namely MA0139.1 for CTCF, MA0062.2 for GABP, MA0138.2 for NRSF, MA0083.2 for SRF, and MA0137.3 for STAT1 [27]. In the first step, we generate a test set from the set of positive alignments (Table 2) by removing randomly 200 alignments. In the second step, we predict for each transcription factor one binding site per target region in all target regions of the reference species (human) in the corresponding test data set, extract the predicted binding sites from the reference species as well as the binding sites at the same positions in the orthologous regions, and calculate for each species the information content of the resulting motif as specified above. We perform both steps 100 times and report the mean and standard error of the information content for each of the five species.

Modeling the binding-affinity bias

In this section we describe the probabilistic model for modeling the binding-affinity bias as a data generating process. A derivation of the log-likelihood for motif-bearing and non-motif-bearing alignments can be found in Additional file 1: Section 1.

Let O be the number of species. A data set comprises N independent multiple sequence alignments. We use X_n to refer to the n-th sequence alignment. Every alignment is formed by O sequences. The o-th

Table 2 Data set statistics for human ChIP-seq data. For each of the five transcription factors (TFs) CTCF, GABP, NRSF, SRF, and STAT1, we specify the (i) average length of transcription factor binding site (TFBS), the (ii) number of alignments, and the (iii) average length of alignments

TF	TFBS length	Number of alignments	Avg. length
CTCF	20 bp	467	213 bp
GABP	12 bp	451	236 bp
NRSF	21 bp	460	245 bp
SRF	12 bp	394	242 bp
STAT1	11 bp	360	244 bp

sequence is denoted by $X_n^{\cdot,o}$. By convention, the reference species (that in which the selection process has taken place) is species 1. Each sequence of alignment X_n is composed of L_n nucleotides. We denote by $X_n^{u,o}$ the u-th nucleotide of the o-th sequence of the n-th alignment. All nucleotides are presented by the set $\mathcal{A} = \{A, C, G, T\}$.

We assume the existence of a common ancestor of all of O species. The sequence of the common ancestor of the n-th alignment is a hidden variable Y_n, with Y_n^u representing its u-th nucleotide. The substitution probability that nucleotide Y_n^u is substituted by the nucleotide $X_n^{u,o}$ is denoted by the variable γ_o.

An alignment X_n may contain a binding site or not. This is denoted by the variable M_n. The length of the binding site is denoted by the variable W and the position of the binding site in alignment X_n is denoted by the variable ℓ_n.

The n-th alignment X_n is sampled as follows. The first decision to be made is whether or not the alignment contains a binding site. This is denoted by variable M_n which follows a Bernoulli distribution with parameter $1 - \alpha$. Thus, whenever variable M_n is equal to 1 (M_n^1), the alignment contains a binding site and when M_n is equal to 0 (M_n^0), it does not.

Thus, parameter α is the probability that alignment X_n contains no binding site. If α equals 0, the sampled data is uncontaminated, because all alignments contain a copy of the binding site. The larger the value of α, the higher the percentage of non motif-bearing alignments in the sampled data. A value of α equal to 1 models a data set where no binding sites are present.

Next we introduce the data generating process for non-motif-bearing alignments and later we explain that for motif-bearing alignments.

1. Sample the primordial sequence as follows: For each position u of the sequence sample nucleotide Y_n^u from the background equilibrium distribution π_0 independent of the previous nucleotides.
2. For each of the descent species $o \in \{1, \ldots, O\}$, sample its sequence given the primordial sequence as follows: To sample nucleotide u of the descent species o, we apply to nucleotide u of the primordial sequence the F81 [28] mutation model with the background equilibrium distribution π_0 and the substitution probability γ_o.

The generating process for motif-bearing sequences is slightly more complex, since it has to deal both with the generation of the binding site and with the selection process. First, we describe how to sample an alignment without taking into account the selection process. Second, we show how to modify this procedure so that the selection process is considered.

Sample a motif-bearing alignment X_n as follows:

1. Sample the start position of the binding site ℓ_n from the uniform distribution.
2. Sample the primordial sequence. For each position u of the sequence outside the binding site, we sample nucleotide Y_n^u from the background equilibrium distribution π_0. For each position u of the binding site, we sample nucleotide Y_n^u from the equilibrium distribution $\pi_{u-\ell_n+1}$.
3. For each of the descent species $o \in \{1, \ldots, O\}$, sample its sequence $X_n^{\cdot,o}$ as follows: For each position u of the descent species o outside the binding site, apply to nucleotide $X_n^{u,o}$ of the primordial sequence the F81 mutation model taking as equilibrium distribution π_0. For each position u of the descent species o inside the binding site, apply to nucleotide $X_n^{u,o}$ of the primordial sequence the F81 mutation model taking as equilibrium distribution $\pi_{u-\ell_n+1}$.

Finally, to model the selection process, we introduce the variable β. β is used to quantify the degree of the binding-affinity bias in the reference species. We assume that a transcription factor binds binding site B with a probability proportional to $p(B|\pi)^{\beta-1}$. As B occurs in vivo with probability $p(B|\pi)$, it occurs in the set of immunoprecipitated sequences with a probability proportional to $p(B|\pi) \cdot p(B|\pi)^{\beta-1} = p(B|\pi)^{\beta}$.

We can interpret the meaning of β as follows: If β is greater than one, low-affinity binding sites are more frequently rejected with respect to $p(B)$ and high-affinity binding sites are less frequently rejected with respect to $p(B)$. This leads to an under-representation of low-affinity binding sites and an over-representation of high-affinity binding sites in the ChIP-seq data set, thus modeling a data set that is affected by the binding-affinity bias. If β is equal to one, low-affinity binding sites are rejected as frequently as high-affinity binding sites, leading to a representative set of binding sites in the ChIP-seq data set, which is not affected by the binding-affinity bias.

Based on that selection model, sample a motif-bearing alignment that has passed the selection process as follows:

1. Sample a motif-bearing alignment disregarding the selection process following the procedure specified above.
2. Decide whether the alignment is accepted or rejected based on the probability of acceptance of the binding site found at the reference species. If the alignment is rejected, go to step 1.

Thus, we denote (i) the model with $\alpha = 0$ and $\beta = 1$ by $\mathcal{M}_{-}^{-}$, (ii) the model with with $\alpha > 0$ and $\beta = 1$ by

$\mathcal{M}^{C}_{-}$, (iii) the model with $\alpha = 0$ and $\beta > 1$ by $\mathcal{M}^{-}_{BA}$, and (iv) the model with $\alpha > 0$ and $\beta > 1$ $\mathcal{M}^{C}_{BA}$. $\mathcal{M}^{-}_{-}$ can neither handle the contamination bias nor the binding-affinity bias. $\mathcal{M}^{C}_{-}$ can only handle the contamination bias, but not the binding-affinity bias. $\mathcal{M}^{-}_{BA}$ can only handle the binding-affinity bias, but not the contamination bias. And $\mathcal{M}^{C}_{BA}$ can handle both the contamination bias and the binding-affinity bias.

We call $\mathcal{M}^{-}_{-}$, $\mathcal{M}^{C}_{-}$, $\mathcal{M}^{-}_{BA}$, and $\mathcal{M}^{C}_{BA}$ *foreground models*. For modeling the background alignments, we use the model with $\alpha = 1$ and $\beta = 1$, which we call *background model* and which we denote by $\mathcal{B}$.

Measuring classification performance

For measuring the classification performance of the four models $\mathcal{M}^{-}_{-}$, $\mathcal{M}^{-}_{BA}$, $\mathcal{M}^{C}_{-}$, and $\mathcal{M}^{C}_{BA}$ we perform stratified repeated random sub-sampling validation as illustrated in Fig. 5 using data sets of the five human transcription factors CTCF, GABP, NRSF, SRF, and STAT1 that have been used for benchmarking the phylogenetic footprinting program *MotEvo* [25].

In step 1, we generate two training sets and two disjoint test sets for each of the five transcription factors as follows. We randomly select 200 alignments from the set of alignments (Table 2) of a particular transcription factor as positive training set, and we choose the set of the remaining alignments as positive test set. We randomly select 500 alignments from the set of alignments of the four remaining transcription factors as negative training set and another disjoint set of 500 alignments as negative test set.

In step 2, we train a foreground model ($\mathcal{M}^{-}_{-}$, $\mathcal{M}^{-}_{BA}$, $\mathcal{M}^{C}_{-}$, or $\mathcal{M}^{C}_{BA}$) on the positive training set and a background model ($\mathcal{B}$) on the negative training set by expectation maximization [29] using a numerical optimization procedure in the maximization step.

We restart the expectation maximization algorithm, which is deterministic for a given data set and a given initialization, 150 times with different initializations and choose the foreground model and the background model with the maximum likelihood on the positive training data and the negative training data, respectively, for classification. We use a likelihood-ratio classifier of the two chosen foreground and background models, apply this classifier to the disjoint positive and negative test sets, and calculate the receiver operating characteristics curve, the precision recall curve, and the area under both curves as measures of classification performance.

We repeat both steps 100 times and determine (i) the mean area under the receiver operating characteristic curve and its standard error and (ii) the mean area under the precision recall curve and its standard error.

Data

The data used in this work originate from human ChIP-seq data of the five human transcription factors CTCF, GABP, NRSF, SRF, and STAT1, where the ChIP-seq data for GABP and SRF published in [10] are available from the QuEST web page [30], and the ChIP-seq data for CTCF, NRSF, and STAT1 published in [24] are available from the SISSRs web page [31]. All five data sets have been filtered for high-quality reads and mapped to a reference

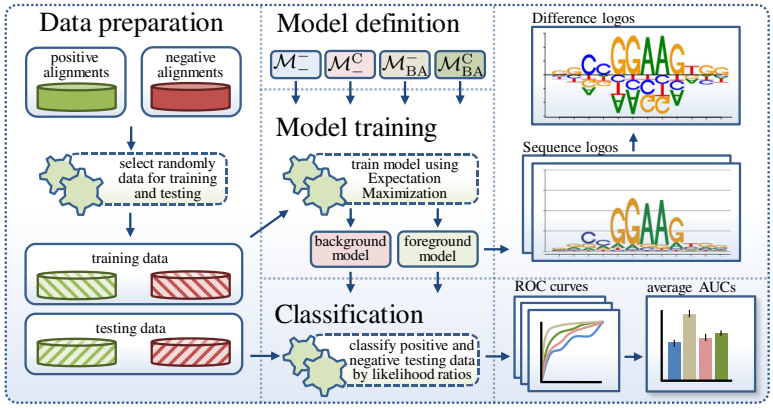


Fig. 5 Overview of the workflow presented in this manuscript. In the data preparation step, we randomly compile disjoint training data and testing data each with positive alignments and negative alignments for each of the transcription factors CTCF, GABP, NRSF, SRF, and STAT1. In the model training step, we train each of the four presented foreground models as well as a background model by expectation maximization with 150 restarts. We choose the foreground model and the background model with maximum likelihood, classify the testing data using a likelihood-ratio classifier, and extract different characteristics such as the ROC curve, the PR curve, the inverse temperature, and the inferred motif. We repeat the described procedure 100 times and calculate mean values and standard errors for several quantities such as the areas under the ROC curves or the PR curves

genome [10, 24], and peak calling has been performed by MACS [32]. Peaks have been extended or cropped to 400 bp, binding regions that potentially comprise more than one of the five transcription factors have been removed, and the 900 binding regions with the highest MACS score have been retained [25]. Orthologous regions from mouse, dog, cow, monkey, horse, and opossum have been extracted from the UCSC database [33], multiple alignments of these orthologous regions have been obtained using T-Coffee [34], and these multiple alignments are kindly provided by [25].

To prepare ungapped alignments from these gapped data sets of the five transcription factors CTCF, GABP, NRSF, SRF, and STAT1, we perform the following three steps. (i) Remove the species that cause the highest number of gaps in all alignments. Accordingly, we remove mouse and opossum and keep orthologous regions from human, monkey, cow, dog, and horse. (ii) Remove all columns in each of the alignments that contain at least one gap to obtain ungapped alignments. (iii) Remove all ungapped alignments that are shorter than 21 bp, which is the length of the longest motif (NRSF) in the performed studies. Table 2 shows details about the resulting data. All data are available as Additional file 2.

Visualizing motif differences with DiffLogo

We used the R package *DiffLogo* [35] to depict the differences between the predicted motifs of the models $\mathcal{M}_{-}^{-}$, $\mathcal{M}_{\mathrm{BA}}^{-}$, $\mathcal{M}_{-}^{\mathrm{C}}$, and $\mathcal{M}_{\mathrm{BA}}^{\mathrm{C}}$. DiffLogo is an open source software that is capable of depicting the differences between multiple motifs [35]. This is realized by visualizing all pairwise differences in an $N \times N$–grid with an empty diagonal. Each entry in the grid is called *difference logo*. The degree of difference of two motifs is calculated by the sum of all stack heights in the corresponding difference logo and is indicated by the background color from red (most dissimilar among all motif pairs) to green (most similar among all motif pairs). The individual sequence logos of the motifs are shown above the table.

A single difference logo depicts the position-specific differences between the base distributions of two sequence motifs. Differences are visualized using a stack of bases for each motif position. The height of each base stack is calculated by the Jensen-Shannon divergence, which is proportional to the degree of base distribution dissimilarity. The Jensen-Shannon divergence is zero if both base distributions are identical, increases with increasing difference of the two base distributions, and reaches a maximum of 2 bit if the two base distributions are maximally different, i.e., if two bases occur only in one of the two motifs each with a probability of 1/2 and the other two bases occur only in the other motif each with a probability of 1/2. The height of each base within a stack is given by the difference of abundance. Thus, the height of a base is proportional to the degree of differential symbol abundance. Bases with a positive height indicate a gain of abundance and bases with a negative height indicate a loss of abundance. The stack height in the positive direction must be equal to the stack height in the negative direction, because the sum of base abundance gain must be equal to the sum of base abundance loss.

Additional files

Additional file 1: Supplementary Methods, Results, Figures, and Examples. This file is structured in four sections.
In section 1, *Modeling the binding-affinity bias*, we describe how to determine the likelihood of non-motif-bearing and motif-bearing alignments modeling the contamination bias and the binding-affinity bias.
In section 2, *Example interpretation of difference logos*, we give an exemplary interpretation of some difference logos.
Section 3, *Supplementary Figures*, contains supplementary Figures S1-S18.
Section 4, *Supplementary Tables*, contains supplementary Tables S1-S10.
(PDF 3492 kb)

Additional file 2: Sequence data. This archive contains data files of gap-free alignments of the ChIP-seq positive regions for each of the transcription factors CTCF, GABP, NRSF, SRF, and STAT1 in FASTA format.
(ZIP 645 kb)

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
MN and IG developed the key idea. MN and JC developed the computational methods. MN and HT performed the studies. All authors wrote, read, and approved the final manuscript.

Acknowledgements
We thank Lothar Altschmied, Helmut Bäumlein, Sven-Erik Behrens, Karin Breunig, Jan Grau, Katrin Hoffmann, Robert Paxton, Patrice Peterson, and Marcel Quint for valuable discussions and DFG (grant no. GR3526/1), Gencat (2014 SGR 118), and Collectiveware (TIN2015-66863-C2-1-R) for financial support.

Author details
[1]Institute of Computer Science, Martin Luther University, Halle (Saale), Germany. [2]Leibniz Institute of Plant Biochemistry, Halle (Saale), Germany. [3]IIIA-CSIC, Campus UAB, Barcelona, Spain. [4]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany.

Received: 15 December 2015 Accepted: 28 April 2016
Published online: 10 May 2016

References

1. Nowrousian M. Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. Eukaryot Cell. 2010;9(9):1300–10.
2. Villar D, Flicek P, Odom DT. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. Nat Rev Genet. 2014;15(4):221–33.
3. Park PJ. Chip–seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10(10):669–80.
4. Furey TS. Chip–seq and beyond: new and improved methodologies to detect and characterize protein–dna interactions. Nat Rev Genet. 2012;13(12):840–52.
5. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. Chip-seq guidelines and practices of the encode and modencode consortia. Genome Res. 2012;22(9):1813–31.

6. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33(8):831–8. doi:10.1038/nbt.3300.
7. Hawkins J, Grant C, Noble WS, Bailey TL. Assessing phylogenetic motif models for predicting transcription factor binding sites. Bioinformatics. 2009;25(12):339–47.
8. Gomes AL, Abeel T, Peterson M, Azizi E, Lyubetskaya A, Carvalho L, Galagan J. Decoding chip-seq with a double-binding signal refines binding peaks to single-nucleotides and predicts cooperative interaction. Genome Res. 2014;24(10):1686–97.
9. Jain D, Baldi S, Zabel A, Straub T, Becker PB. Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments. Nucleic Acids Res. 2015;43(14):6959–68. doi:10.1093/nar/gkv637.
10. Valouev A, Johnson A, David S and. Sundquist, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods. 2008;5(9): 829–34.
11. Rye MB, Sætrom P, Drabløs F. A manually curated chip-seq benchmark demonstrates room for improvement in current peak-finder programs. Nucleic Acids Res. 2011;39(4):e25. doi:10.1093/nar/gkq1187.
12. Jung YL, Luquette LJ, Ho JWK, Ferrari F, Tolstorukov M, Minoda A, Issner R, Epstein CB, Karpen GH, Kuroda MI, Park PJ. Impact of sequencing depth in ChIP-seq experiments. Nucleic Acids Res. 2014;42(9): 178–4. doi:10.1093/nar/gku178.
13. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. Model-based analysis of chip-seq (macs). Genome Biol. 2008;9(9):137.
14. Nix DA, Courdy SJ, Boucher KM. Empirical methods for controlling false positives and estimating confidence in chip-seq peaks. BMC Bioinformatics. 2008;9(1):523.
15. Bailey TL, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, Madrigal P, Taslim C, Zhang J. Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput Biol. 2013;9(11):e1003326.
16. Håndstad T, Rye MB, Drabløs F, Sætrom P. A ChIP-Seq Benchmark Shows That Sequence Conservation Mainly Improves Detection of Strong Transcription Factor Binding Sites. PLoS ONE. 2011;6(4):18430. doi:10.1371/journal.pone.0018430.
17. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14(5):51.
18. Park D, Lee Y, Bhupindersingh G, Iyer VR. Widespread Misinterpretable ChIP-seq Bias in Yeast. PloS One. 2013;8(12):83506.
19. Teytelman L, Thurtle DM, Rine J, van Oudenaarden A. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. Proc Nat Acad Sci. 2013;110(46):18602–7.
20. Elliott JH, Grimshaw J, Altman R, Bero L, Goodman SN, Henry D, Macleod M, Tovey D, Tugwell P, White H, Sim I. Informatics: Make sense of health data. Nature. 2015;527:31–2.
21. Bailey TL, Elkan C. The value of prior knowledge in discovering motifs with MEME. Ismb. 1995;3:21–9.
22. Wilbanks EG, Facciotti MT. Evaluation of algorithm performance in chip-seq peak detection. PloS one. 2010;5(7):11471.
23. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. Information content of binding sites on nucleotide sequences. J Mol Biol. 1986;188(3):415–31.
24. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucl Acids Res. 2008;36(16):5221–31. doi:10.1093/nar/gkn488. http://nar.oxfordjournals.org/cgi/reprint/36/16/5221.pdf.
25. Arnold P, Erb I, Pachkov M, Molina N, van Nimwegen E. MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. Bioinformatics. 2012;28(4):487–94. doi:10.1093/bioinformatics/btr695.
26. Sober E. The principle of parsimony. Brit J Philos Sci. 1981;32(2):145–56.
27. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen C-YY, Chou A, Ienasescu H, Lim J, Shyr C, Tan G, Zhou M, Lenhard B, Sandelin A, Wasserman WW. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 2014;42(Database issue):142–7. doi:10.1093/nar/gkt997.
28. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981;17(6):368–76.
29. Lawrence CE, Reilly AA. An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins: Struct Funct Bioinformatics. 1990;7(1): 41–51.
30. Quantitative Enrichment of Sequence Tags: QuEST. http://mendel.stanford.edu/sidowlab/downloads/quest/. Accessed 29 Mar 2016.
31. ChIP-Seq Data Analysis: Identification of Protein–DNA Binding Sites with SISSRs Peak-Finder. http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/sissrs/. Accessed 29 Mar 2016.
32. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-seq (MACS). Genome Biol. 2008;9(9):137. doi:10.1186/gb-2008-9-9-r137.
33. Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, Kober KM, Miller W, Pedersen JS, Pohl A, Raney BJ, Rhead B, Rosenbloom KR, Smith KE, Stanke M, Thakkapallayil A, Trumbower H, Wang T, Zweig AS, Haussler D, Kent WJ. The UCSC genome browser database: 2008 update. Nucleic Acids Res. 2008;36(suppl 1):773–9. doi:10.1093/nar/gkm966.
34. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000;302(1): 205–17. doi:10.1006/jmbi.2000.4042.
35. Nettling M, Treutler H, Grau J, Keilwagen J, Posch S, Grosse I. DiffLogo: a comparative visualization of sequence motifs. BMC Bioinf. 2015;16(1):1.

5.2 DiffLogo: a comparative visualization of sequence motifs

In the following reference the first author is underlined and I am marked in bold.
Martin Nettling, **Hendrik Treutler**, Jan Grau, Jens Keilwagen, Stefan Posch, and Ivo Grosse. DiffLogo: a comparative visualization of sequence motifs. *BMC bioinformatics*, 16(1):387+, November 2015. *doi:10.1186/s12859-015-0767-x*

`https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0767-x`

Nettling *et al. BMC Bioinformatics* (2015) 16:387
DOI 10.1186/s12859-015-0767-x

BMC Bioinformatics

SOFTWARE Open Access

DiffLogo: a comparative visualization of sequence motifs

Martin Nettling[1*†], Hendrik Treutler[2†], Jan Grau[1], Jens Keilwagen[3], Stefan Posch[1] and Ivo Grosse[1,4]

Abstract

Background: For three decades, sequence logos are the *de facto* standard for the visualization of sequence motifs in biology and bioinformatics. Reasons for this success story are their simplicity and clarity. The number of inferred and published motifs grows with the number of data sets and motif extraction algorithms. Hence, it becomes more and more important to perceive differences between motifs. However, motif differences are hard to detect from individual sequence logos in case of multiple motifs for one transcription factor, highly similar binding motifs of different transcription factors, or multiple motifs for one protein domain.

Results: Here, we present *DiffLogo*, a freely available, extensible, and user-friendly R package for visualizing motif differences. *DiffLogo* is capable of showing differences between DNA motifs as well as protein motifs in a pair-wise manner resulting in publication-ready figures. In case of more than two motifs, *DiffLogo* is capable of visualizing pair-wise differences in a tabular form. Here, the motifs are ordered by similarity, and the difference logos are colored for clarity. We demonstrate the benefit of *DiffLogo* on CTCF motifs from different human cell lines, on E-box motifs of three basic helix-loop-helix transcription factors as examples for comparison of DNA motifs, and on F-box domains from three different families as example for comparison of protein motifs.

Conclusions: *DiffLogo* provides an intuitive visualization of motif differences. It enables the illustration and investigation of differences between highly similar motifs such as binding patterns of transcription factors for different cell types, treatments, and algorithmic approaches.

Keywords: Sequence analysis, Sequence logo, Sequence motif, Position weight matrix, Binding sites

Background

Biological polymer sequences encode information by the order of their monomers, i.e., bases or amino acids. Often specific parts of the polymer sequence are of particular interest, as they encode, for instance, the binding of transcription factors to specific binding sites [1, 2], the binding to micro-RNA-targets in mRNAs, splice donor sites and splice acceptor sites in pre-mRNAs [3, 4], the presence of phosphorylation sites in proteins, or the folding of specific protein domains [5]. The set of subsequences of one specific biological process are often represented as a sequence motif.

A sequence motif is a model, that represents the preference for the monomers based on a set of aligned biopolymer sequences. Sequence motifs are the result of pipelines comprising wet-lab experiments and motif prediction algorithms, and are frequently used as the basis of *in silico* predictions [6]. Thus, sequence motif are critical for research of a wide range of problems in biology and bioinformatics.

Considering a particular transcription factor, there are many pipelines that combine wet-lab experiments such as *HT-SELEX* [7, 8], *ChIP-Seq* [9] or *DNase-Seq footprinting* [10] with motif prediction algorithms such as *MEME* [2, 11], *ChIPMunk* [12], *POSMO* [13], or *Dimont* [14]. Wet-lab experiments differ in their experimental setup, e.g., ecotypes, cell types, developmental stage, time points, or treatment, and motif prediction algorithms differ in their mathematical theory and implementation details.

Visualizing the results of motif discovery is nowadays accomplished by sequence logos [15], the *de facto*

*Correspondence: martin.nettling@informatik.uni-halle.de
†Equal contributors
[1]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany
Full list of author information is available at the end of the article

standard for visualizing motifs in biology and bioinformatics. Sequence logos emerged as an essential tool for researchers to interpret findings, document work, share knowledge, and present results.

However, comparing multiple sequence logos by visual inspection is sometimes tricky. Differences between sequence logos of two unrelated transcription factors are usually obvious, whereas differences between sequence logos of the same transcription factor are often less obvious and rather hard to perceive as depicted in Fig. 1. Moreover, the results of motif discovery algorithms need to be compared against huge reference databases such as *JASPAR* [16] or *UniProbe* [17] or motifs from literature.

For this reason, the comparison of motifs is of primary interest. Several numerical measures including variants of Euclidean distance, Pearson correlation, and Jensen-Shannon divergence have been used to compare motifs [18–21]. These measures express the difference of motifs as a single number that can be easily utilized subsequently, e.g., for rankings or clustering algorithms. However, these measures lose the information of what exactly makes the difference between the motifs of interest. Hence, the comparison of multiple pairs of motifs can result in similar measures.

There are various tools for the analysis and visualization of motifs as summarized in Table 1. The R package *seqLogo* [22] is an implementation of sequence logos. In the context of motif comparison, sequence logos may be interpreted as a comparison of the input motif with a uniformly distributed motif. The web application *iceLogo* [23] extends this approach by comparing the input motif with a motif that follows the same background distribution at each motif position. Basically, *seqLogo* and *iceLogo* are designed for the presentation of single motifs. In contrast, the R package *MotifStack* [24] and the web application *STAMP* [25] are designed for the presentation of multiple motifs. Here, the input motifs are clustered and presented as sequence logos. Thus, the approach of both tools may be interpreted as multiple comparisons with a uniformly distributed motif. The web application *Two Sample Logo* [26] is capable of comparing two input motifs on the basis of probability theory. This comparison is performed for each motif position individually and results in a sophisticated motif comparison. Depending on the focus of each tool, the input format is a set of aligned sequences and/or a position frequency matrix or position weight matrix. In addition, some tools focus exclusively on DNA motifs, while others cover DNA, RNA, and protein motifs or even allow arbitrary alphabets. Table 1 summarizes tools and their capabilities. In section 4 of Additional file 1, we additionally provide comparative example plots generated by *seqLogo*, *iceLogo*, *STAMP*, *Two Sample Logo*, and *DiffLogo*.

We intend the pair-wise comparison of motifs and extend this idea towards the comparison of multiple motifs as follows.

We focus on the comparison of position-specific symbol distributions of two motifs. We neglect dependencies between different motif positions to reduce complexity. As suggested by the *sequence logo* approach, we intend to represent the characteristics of each motif position by the two properties stack height and symbol height within a stack. The stack height is to be proportional to the degree of distribution dissimilarity. The symbol height is to be proportional to the degree of differential symbol abundance.

We intend to compare three or more motifs on the basis of pair-wise motif comparisons. This comparison is to take into account all pair-wise motif comparisons, suggesting an arrangement in a grid with one row and one column for each motif and one cell for each motif comparison. Similar motifs are to be placed in nearby rows and columns, and the degree of similarity between all motifs is to become obvious at a glance analogous to heatmaps. The grid is to be complemented with a display of the individual sequence logos for further comparisons.

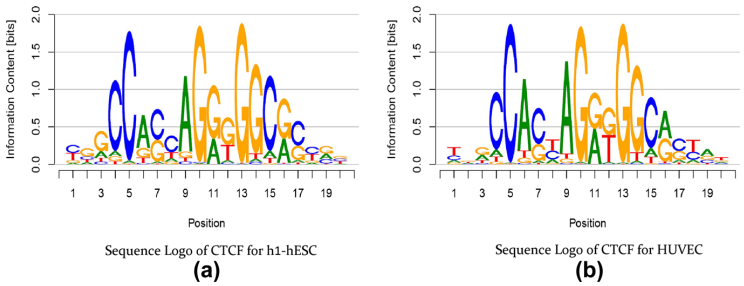


Fig. 1 Sequence logos of CTCF motifs from cell lines H1-hESC and HUVEC. The two sequence logos are highly similar in their conservation profile (height of stacks) and nucleotide preference at the individual motif positions

Table 1 Comparison of related tools. We compare six publicly available tools on the basis of five criteria

	Features				
Tools	Alphabet	Input format	Comparison	Clustering	Extensible
seqLogo	DNA	matrix	uniform	-	-
iceLogo	DNA/RNA, proteins	sequences	average	-	-
MotifStack	any	matrix	uniform	hclust	-
STAMP	DNA	sequences, matrix	uniform	UPGMA/SOTA	-
Two Sample Logo	DNA/RNA, proteins	sequences	position-specific	-	-
DiffLogo	any	sequences, matrix	position-specific	hclust, optimal leaf ordering	✓

In the first and second column, we examine the kind of supported input, in the third and forth column we examine the mode of action, and in the fifth column we examine whether the tool is extensible. For the criterion "alphabets" we summarize the supported biopolymers out of DNA, RNA, and proteins or arbitrary alphabets in case of "any". For the criterion "input format" we discriminate a set of "sequences" versus "matrix", which addresses at least one out of the formats position weight matrix (PWM), position frequency matrix (PFM), and position count matrix (PCM). For the criterion "comparison" we characterize the kind of distribution that is used for motif comparison ("uniform" is the uniform distribution, "average" is the average base distribution in a set of sequences, and "position-specific" is a position-specific distribution). For the criterion "clustering" we point out whether there is a clustering of motifs and which cluster-algorithm is used. For the criterion "extensible" we note whether the tool is extensible by the user

Implementation

In this section, we first define the used notation. We then briefly describe the classical sequence logo. Subsequently, we introduce the difference logo for the visualization of pair-wise motif differences. We discuss this new method and explore potential biological interpretations. Finally, we propose an approach for employing difference logos for the joint comparison of multiple motifs.

Basic notation and sequence logo

Consider a motif as an abstract description of a given set of aligned sequences of common length L from the alphabet $\mathcal{A}$. The relative frequency of symbol $a \in \mathcal{A}$ at position $\ell \in [1, L]$ corresponds to the (estimated) probability $p_{\ell,a}$. In case of two motifs, we use $p_{\ell,a}$ for the first motif and analogously $q_{\ell,a}$ for the second motif.

The well-known sequence logo visualizes a motif with a symbol stack for each position. We denote the height of the stack at position ℓ by H_ℓ and the height of symbol a within this stack by $H_{\ell,a}$. In the traditional sequence logo, H_ℓ and $H_{\ell,a}$ are defined by

$$H_\ell = \log_2(|\mathcal{A}|) - \sum_{a \in \mathcal{A}} p_{\ell,a} \cdot \log_2(p_{\ell,a}) \quad (1)$$

$$H_{\ell,a} = p_{\ell,a} \cdot H_\ell, \quad (2)$$

which states that the height of a stack at position ℓ reflects the degree of conservation at position ℓ quantified by the information content and that the height of each symbol at position ℓ is proportional to its frequency at position ℓ. Hence, the traditional sequence logo is an intuitive visualization of both (i) conserved motif positions and (ii) abundant bases.

The approach of *DiffLogo*

As specified earlier, we compare motifs per position. Similar to the sequence logo, we show a symbol stack for each position. We redefine the calculation of H_ℓ and use this measure as the total height of position ℓ reflecting the difference of the symbol distribution of both motifs at this position. We redefine the calculation of $H_{\ell,a}$ and use this measure as the height of a symbol within the stack at position ℓ. In the following, $H_{\ell,a}$ can be positive or negative. Symbols with positive values $H_{\ell,a}$ are plotted upward. Symbols with negative values $H_{\ell,a}$ are plotted downward.

Generally, there is a plethora of well-understood mathematical criteria that can be combined to define the height of a symbol stack and the relative heights of symbols within the stack such as probability differences, information divergences, distance measures, or entropies [27]. In the following, we present *DiffLogo* with the example of the Jensen-Shannon divergence for the calculation of H_ℓ and normalized probability differences for the calculation of $H_{\ell,a}$. We denote the combination of these two measures as weighted difference of probabilities.

Weighted difference of probabilities

We calculate the stack height for each motif position using the Jensen-Shannon divergence. The Jensen-Shannon divergence is a measure for the dissimilarity of two probability distributions based on information theory [28] (see Fig. 2). In contrast to other measures, the Jensen-Shannon divergence shows a comparable behavior when evaluating dissimilarities of distributions near the uniform distribution. The Jensen-Shannon divergence of two motifs at position ℓ is given by

$$H_\ell = \frac{1}{2} \sum_{a \in \mathcal{A}} p_{\ell,a} \log_2 \frac{p_{\ell,a}}{m_{\ell,a}} + \frac{1}{2} \sum_{a \in \mathcal{A}} q_{\ell,a} \log_2 \frac{q_{\ell,a}}{m_{\ell,a}}, \quad (3)$$

where $m_{\ell,a} = \frac{p_{\ell,a}+q_{\ell,a}}{2}$.

We define the height of each symbol by

$$H_{\ell,a} = r_{\ell,a} \cdot H_\ell, \quad (4)$$

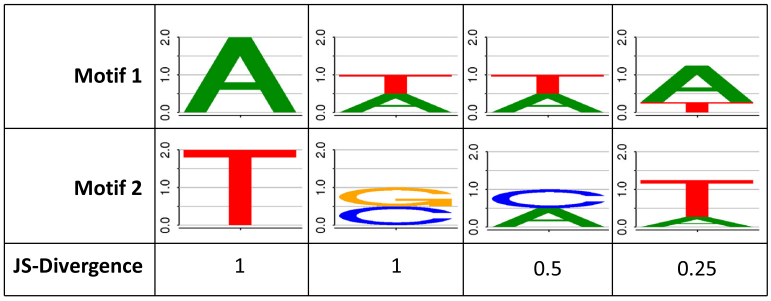


Fig. 2 Exemplary comparison of four DNA motifs of length one using the Jensen-Shannon divergence. Motif 1 and motif 2 are depicted as sequence logos. For each column, we compare the motif in the first row with the motif in the second row using the Jensen-Shannon divergence listed in the third row. In the first example we depict the case with only one base in each motif resulting in a maximal Jensen-Shannon divergence of 1 bit. In the second example we depict the case with two equally abundant bases both in motif 1 and motif 2 (both different) resulting again in a maximal Jensen-Shannon divergence of 1 bit. In the third example we depict the case with two equally abundant bases both in motif 1 and motif 2 (one equal and one different) resulting in a Jensen-Shannon divergence of 0.5 bit. In the fourth example we depict the case with two bases both in motif 1 and motif 2 (differentially abundant) resulting in a Jensen-Shannon divergence of 0.25 bit

where we define the weight $r_{\ell,a}$ as

$$r_{\ell,a} = \begin{cases} \frac{p_{\ell,a} - q_{\ell,a}}{\sum_{a' \in \mathcal{A}} |p_{\ell,a'} - q_{\ell,a'}|} & \text{if } p_\ell \neq q_\ell \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

$r_{\ell,a}$ is the probability difference of symbol a at position ℓ between two motifs normalized by the sum of absolute probability differences at this position. We use normalized probability differences as these are indicators for the gain or loss of symbol abundance and provide a view on the symbol distribution differences of both motifs. As a consequence, symbols less abundant in the second motif compared to the first motif are plotted upward, and symbols more abundant in the second motif compared to the first motif are plotted downward.

This representation emphasizes a high gain or loss of probability in co–occurrence with a high gain or loss of information content. The sum of the heights of symbols with a gain of probability and the sum of the heights of symbols with a loss of probability are equal at every position, because each gain of probability of one symbol implies a loss of probability of the remaining symbols. The advantage of this approach is that we are capable of seeing differences of position-specific symbol distributions and of seeing those symbols that are responsible for these differences by gaining or losing abundance.

Comparison of multiple motifs

According to the requirements formulated above, we propose a visualization for the joint comparison of $N \geq 3$ motifs given the measure H_ℓ as follows.

We plot the difference logos of all $N \times (N-1)$ motif pairs with a common ordinate scaling. We define a scalar dissimilarity value D for a pair of motifs as the sum of all stack heights in the corresponding difference logos,

$$D = \sum_{\ell=1}^{L} H_\ell. \quad (6)$$

We compute a motif order to group similar motifs. Here, we take the optimal leaf order of a hierarchical clustering of the motifs based on D (function *hclust* in R package *stats* and function *order.optimal* in R package *cba*). We arrange the difference logos ordered in an $N \times N$ grid with an empty diagonal. Difference logos opposing each other across the diagonal of the grid correspond to each other by an inversion of the ordinate. We visualize D with the background color of the corresponding difference logo using a color gradient from green (most similar among all pairwise comparisons) to red (most dissimilar). We outline the motif names above each column and left of each row. In addition, we allow the possibility of drawing the classic sequence logos and the cluster tree above the columns as auxiliary information.

The advantage of this approach is that we are capable of surveying the overall similarities and dissimilarities in the resulting difference logo grid. Greenish regions indicate similar motif groups and reddish rows and columns indicate less similar motifs. Given a region of interest, it is furthermore possible to comprehend the origins of dissimilarities from the individual difference logos and optionally the sequence logos.

R package

DiffLogo is written in R [29]. We provide the implementation as a ready-to-use R package. For symbol drawing, *DiffLogo* uses adapted methods from the package

seqLogo [22] in the software suite *bioconductor* [30]. *DiffLogo* allows the analysis of sequence motifs defined over arbitrary alphabets.

The core functions can be parameterized with functions for H_ℓ and $r_{\ell,a}$. Hence, the user is capable of combining different formulae for H_ℓ and $r_{\ell,a}$. We provide implementations of the Jensen-Shannon divergence and the normalized probability difference used for the difference logos presented in this manuscript. In addition, *DiffLogo* provides other implementations for H_ℓ and $r_{\ell,a}$ as alternatives. Exemplarily, we show the result of eight different combinations of measures for stack height and symbol height in Additional file 1: Tables S1 and S2. The *DiffLogo* package comprises example data, example code, and further documentation.

Results and discussion

In this section, we present three examples demonstrating the utility of *DiffLogo* in different applications. First, we examine differences in motifs of DNA binding sites of the same transcription factor from five different cell lines. Second, we examine differences in motifs of DNA binding sites of three different transcription factors with similar binding motifs. Third, we examine differences in motifs of a protein domain.

DNA motifs of same transcription factor

We consider sequence logos and difference logos of binding sites of the human insulator CTCF in different cell lines as obtained by motif discovery from ChIP-seq data [31] based on preprocessed ChIP-seq data from the ENCODE project. For CTCF motif inference, sequences with p-values smaller than 10^{-6} were selected. All data are freely available as Additional File of the original publication [31]. Since CTCF is a DNA-binding protein, the alphabet corresponds to the four nucleotides in this case.

In Fig. 1, we plot the sequence logos for two of these cell types, namely H1-hESC and HUVEC. Considering the sequence logos, both motifs look highly similar with regard to the conservation as well as the nucleotide preference of individual motif positions, and differences between both motifs are hard to perceive. Considering the corresponding difference logo in Fig. 3 (row 1, column 5 or row 5 column 1), however, we instantly see that indeed a large number of motif positions exhibits differences in nucleotide composition. We find the largest difference

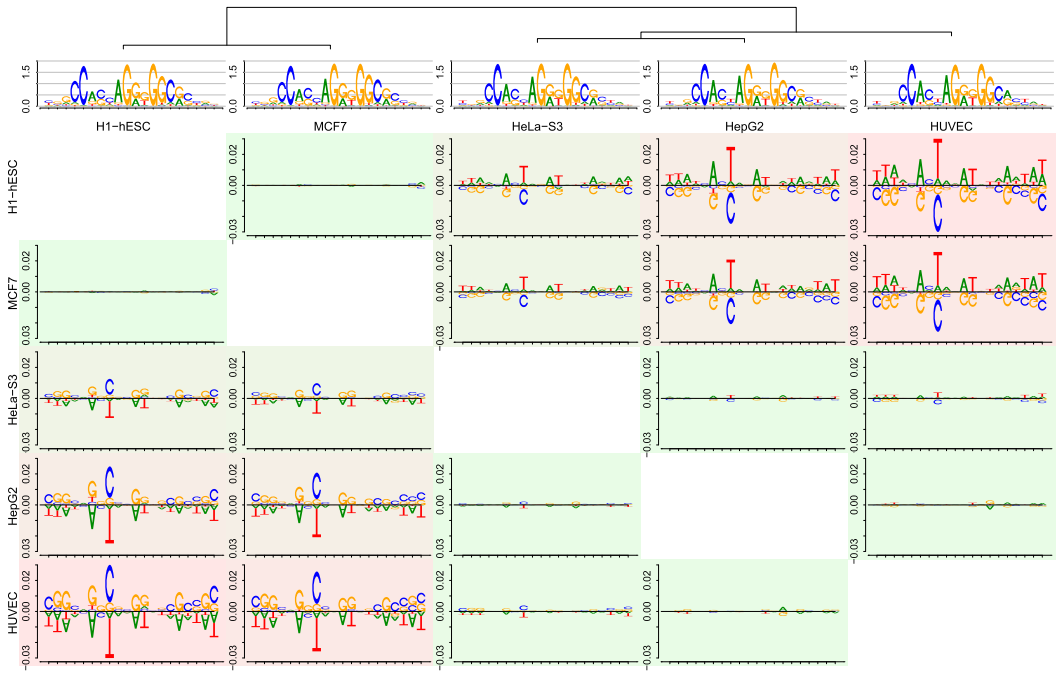


Fig. 3 Comparison of five DNA motifs using *DiffLogo*. Comparison of five CTCF motifs from cell lines H1-hESC, MCF7, HeLa-S3, HepG2, and HUVEC. We plot all pair-wise sequence logos and display the distance between each motif using the background color from green (similar) to red (dissimilar). We plot the sequence logos of each motif as well as the leaf-ordered cluster tree above. The motifs of H1-hESC and MCF7 are highly similar and substantially different from the other motifs, while the motifs of HeLe-S3, HepG2, and HUVEC are similar to each other as well. Due to leaf ordering, the difference between compared motifs increases with increasing distance from the main diagonal in the difference logo grid

according to the difference logo at position 8 of the motifs, where nucleotide C is more prevalent in cell type H1-hESC compared to HUVEC, whereas the opposite holds for nucleotide T. This difference is less visible in the sequence logos, even with hindsight from the difference logo, due to the low conservation at this position. Specifically, the probability of C increases from 0.35 (HUVEC) to 0.58 (H1-hESC), whereas the probability of T drops by a factor of 2 from 0.44 (HUVEC) to 0.21 (H1-hESC). Depending on the application, this difference at position 8 might have a decisive influence on the outcome of, e.g., *in silico* binding site prediction.

In the literature, several positions with substantial motif differences uncovered by *DiffLogo* are known to be related to CTCF binding affinity. For instance [32] show that "low occupancy" CTCF binding sites are enriched for C or G at position 18 compared to "high occupancy" sites, which in our case might indicate that the H1-hESC ChIP-seq data set contains a larger number of such "low occupancy" sites than the HUVEC data set.

In a large-scale study [33], CTCF core motifs are partitioned by the presence or absence of additional upstream and downstream motifs, where the greatest variations in the core motifs between partitions can be found at positions 1-3, 6, 8, 11, 12, 18, and 20, which cover those positions varying in the difference logo. Again, these partitions are related to binding affinity and occupancy of CTCF.

In summary, *DiffLogo* helps to identify several motif positions with substantial variation between cell types, known to be related to CTCF binding affinity and binding site occupancy.

In real-world applications, motifs for more than two cell types are often studied, which might render the pairwise comparison of difference logos a tedious task. We support such an evaluation across multiple cell types by a structured visualization of multiple difference logos as shown in Fig. 3. Here, we compare the pairwise difference logos of CTCF motifs from five cell types, namely H1-hESC, MCF7, HeLa-S3, HepG2, and HUVEC. The cluster tree and background color of the cells are based on numerical measures of motif differences (cf. Implementation) and guide us to the most notable differences between pairs of motifs. For instance, we observe from the tree and background colors that the motifs of H1-hESC and MCF7 are highly similar. The same holds true for the motifs of HeLa-S3, HepG2, and HUVEC, whereas motifs show substantial differences between these two groups. To further facilitate the visual comparison of multiple motifs, we leaf-order the cluster tree such that neighboring motifs are as similar as possible. Due to this ordering, the difference between motif pairs increases with increasing distance from the main diagonal of the difference logo grid. For instance, the topology of the clustering would allow to invert the order of the three leaves under the right sub-tree in Fig. 3, which, however, would bring the quite dissimilar motifs of HUVEC and MCF7 in direct neighborhood. From Fig. 3, we also observe that the two motifs of H1-hESC and HUVEC are the most dissimilar ones among the motifs studied. A visualization of all nine available motifs can be found in Additional file 1: Figure S1.

DNA motifs of different transcription factors

We demonstrate the utility of *DiffLogo* for motifs derived from binding assays for the human transcription factors Max, Myc, and Mad (Mxi1) from Mordelet *et al*. [34]. These three basic helix-loop-helix transcription factors are members of a regulatory network of transcription factors that controls cell proliferation, differentiation, and cell death. Each transcription factor binds to different sets of target sites, regulates different sets of genes, and thus plays a distinct role in human cells. However, Myc, Max, and Mad have almost identical PWMs, which all correspond to an E-box motif with consensus sequence CACGTG.

The PWMs considered here have been derived from probe sequences and corresponding binding intensities of *in-vitro* genomic context protein-binding microarrays [34]. The exact binding sites within the probe sequences are predicted by the de-novo motif discovery tool Dimont [14] using Slim models [35]. For each of the three transcription factors, the top 1,000 predicted binding sites are used to generate the corresponding PWM.

In Fig. 4, we plot the sequence logos and difference logos of Myc, Max, and Mad. We observe from the sequence logos that the binding motifs are almost identical. Considering the difference logos, we observe that the six core nucleotides are conserved in the motifs of all three transcription factors. We find the largest differences between the motif of Max and the motifs of Myc and Mad. In case of Max and Myc, we find a Jensen-Shannon divergence greater than 0.01 bit at positions 11, 12, 22, and 26. In case of Max and Mad, we find a Jensen-Shannon divergence greater than 0.01 bit at positions 3, 12, 22, and 25. In both cases, we mainly find more purine (adenine and guanine) in the motif of Max than in the motifs of Myc and Mad.

Protein motifs

As a third example, we demonstrate the utility of *DiffLogo* using the F-box domain, which plays a role in protein-protein binding. The complete F-box domain in this example is 48 amino acids long [36]. Here, we investigate the middle section from the 12th to the 35th amino acid.

In Fig. 5, we plot the sequence logos and difference logos of F-box domains from the three kingdoms metazoa, fungi, and viridiplantae. We observe from the cluster

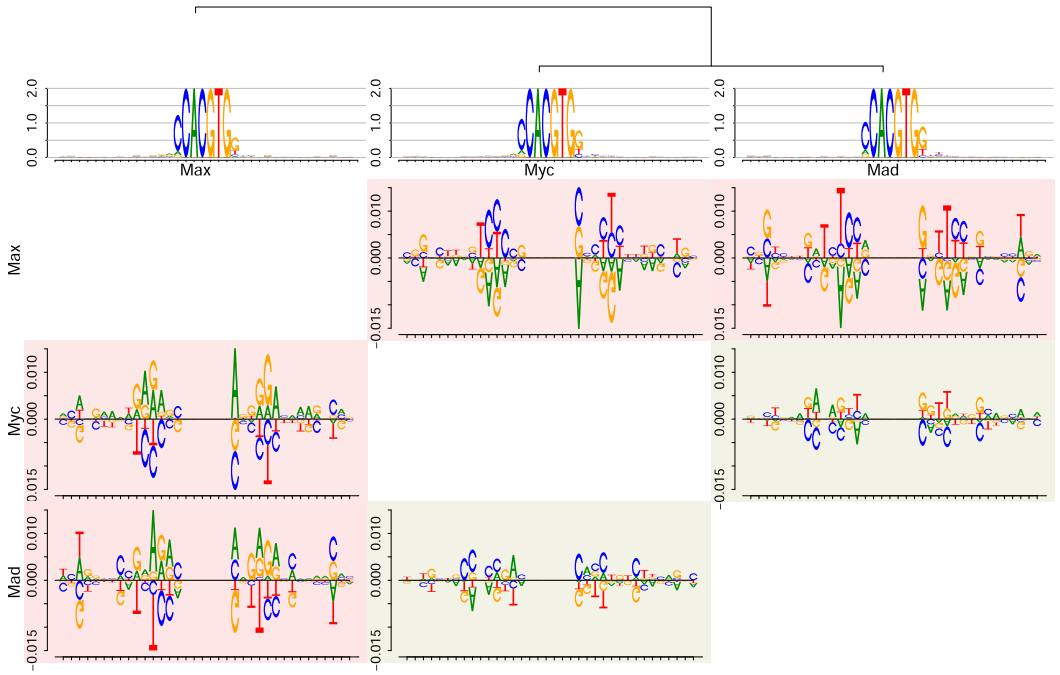


Fig. 4 Comparison of E-Box motifs of Max, Myc, and Mad using *DiffLogo*. We plot all pair-wise difference logos and display the distance between each motif using the background color from green (similar) to red (dissimilar). We plot the sequence logos of each motif as well as the leaf-ordered cluster tree above. The motifs of the transcription factors Myc and Mad are more similar to each other than to the motif Max. The six core nucleotides with consensus sequence CACGTG are conserved in the motifs of all three transcription factors and, hence, are not visible in the difference logos

tree and the background colors that the motifs of metazoa and fungi are highly similar, whereas motifs of this group show substantial differences to viridiplantae. The largest difference can be seen between motifs of metazoa and viridiplantae.

When comparing metazoa and fungi with viridiplantae, *DiffLogo* identifies positions 6, 17, and 22 with high values of the Jensen-Shannon divergence. The differences at positions 6 and 22 could be expected from the differences of the sequence logos, whereas the differences at position 17 are not immediately obvious from them. At position 6 the abundance of arginine (R) in viridiplantae is 0.54 and thus more than 10 times higher than in fungi and 12 times higher than in metazoa. At position 22 tryptophane (W) is highly abundant in viridiplantae and 4 and 3.4 times more abundant than in metazoa and fungi. At position 17 the most noticeable differences in viridiplantae to fungi and metazoa can be seen for amino acid cysteine (C), valine (V), alanine (A), and serine (S). The overall abundance increases from 0.13 in metazoa and 0.12 in fungi to 0.64 in viridiplantae. In contrast, the abundance of arginine (R), glutamine (Q), and lysine (K) is only 0.044 in viridiplantae and 0.44 in metazoa and fungi. A visualization of the full F-Box domain from four kingdoms can be found in Additional file 1: Figure S2.

Conclusion

We present *DiffLogo*, an easy-to-use tool for a fast and efficient comparison of motifs. *DiffLogo* may be applied by users with only basic knowledge in R and is highly configurable and extensible for advanced users. We introduce weighted differences of probabilities to emphasize large differences in position-specific symbol distributions. We present visual comparisons of multiple motifs stemming from motifs of one transcription factor in different cell types, different transcription factors with similar binding motifs, and species-specific protein domains. Figures generated by *DiffLogo* enable the identification of overall motif groups and of sources of dissimilarity. Using *DiffLogo*, it is easily possible to compare motifs from different sources, so *DiffLogo* facilitates decision making, knowledge sharing, and the presentation of results. We make *DiffLogo* freely available in an extensible, ready-to-use R package including examples and documentation. *DiffLogo* is part of *Bioconductor*.

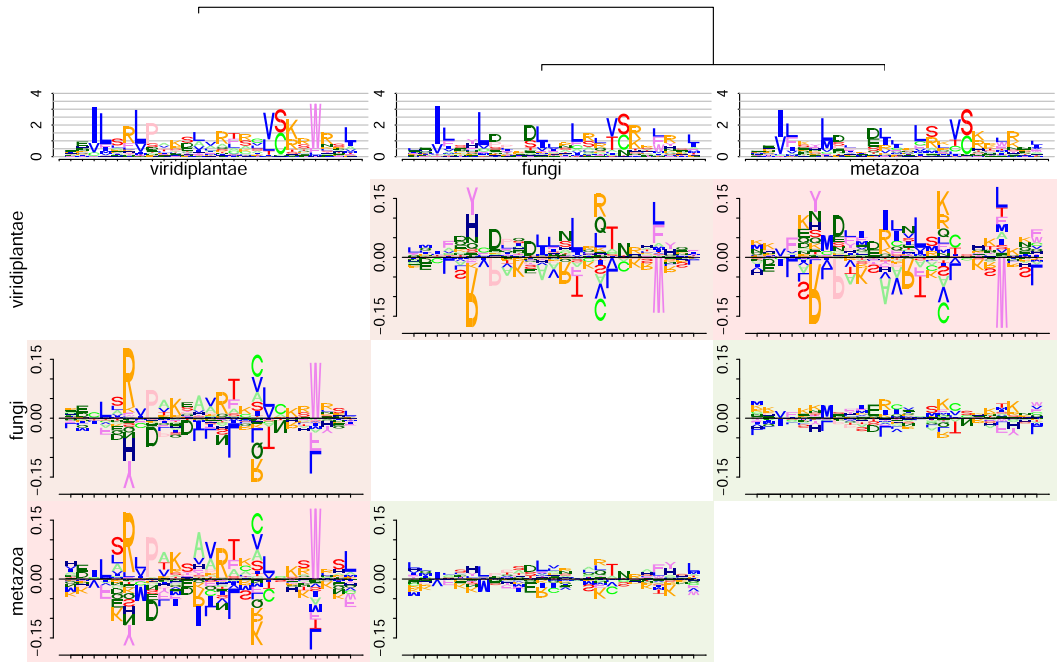


Fig. 5 Comparison of three F-box domain motifs using *DiffLogo*. We compare the F-box domains from the kingdoms metazoa, fungi, and viridiplantae and plot all pair-wise difference logos and display the distance between each motif using the background color from green (similar) to red (dissimilar). We plot the sequence logos of each motif as well as the leaf-ordered cluster tree above. The motifs of metazoa and fungi are highly similar. All other pairwise comparisons show substantial differences

Availability and requirements

Project name: DiffLogo
Project home page: http://github.com/mgledi/DiffLogo
Availability: http://bioconductor.org/packages/DiffLogo
Operating system(s): Platform independent
Programming language: R
Other requirements: Installation of R 1.8.0 or higher
License: LGPL ($\geq$ 2)
Any restrictions to use by non-academics: None

Additional file

Additional file 1: Supplementary Methods, Results, Figures, and Examples. This file is structured in four sections. Section 1, *Additional examples*, contains Figures S1 and S2. Figure S1 shows a *DiffLogo* grid for nine CTCF motifs. Figure S2 shows a *DiffLogo* grid for four F-box domain motifs. In section 2, *CTCF with and without clustering*, we show in detail the impact of clustering and optimal leaf ordering for a *DiffLogo* grid of nine CTCF motifs. In section 3, *Alternative combinations of stack heights and symbol weights*, we first describe the mathematical background of four implementations of H_ℓ and two implementations of $r_{\ell,a}$. Afterwards, we show the result of the eight possible combinations in Tables S1 and S2 on two sequence motifs. In section 4, *Tool comparison*, we compare *DiffLogo* with the five tools *seqLogo*, *iceLogo*, *MotifStack*, *STAMP*, and *Two Sample Logo*.

From the set of nine CTCF motifs we selected the pair of motifs with the highest similarity according to the Jensen-Shannon divergence (GM12878 and K562) and the pair of motifs with the lowest similarity according to the Jensen-Shannon divergence (H1-hESC and HUVEC) for the comparison of the five different tools. (PDF 8775 kb)

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
MN conceived the idea. MN, HT, JK, JG, SP, and IG developed the idea and the computational methods. MN and HT implemented and tested *DiffLogo*. All of the authors read and approved the final version of the manuscript.

Acknowledgements
We thank Karin Breunig, Jesus Cerquides, Ralf Eggeling, and Martin Porsch for valuable discussions and contributing data and DFG (grant no. GR3526/1) for financial support.

Author details
[1]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany. [2]Leibniz Institute of Plant Biochemistry, Halle (Saale), Germany. [3]Institute for Biosafety in Plant Biotechnology, Julius Kühn-Institut (JKI), Federal Research Centre for Cultivated Plants, Quedlinburg, Germany. [4]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany.

Received: 10 April 2015 Accepted: 8 October 2015
Published online: 17 November 2015

References

1. Staden R. Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res. 1984;12:505–19.
2. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology. San Diego: Department of Computer Science and Engineering, University of California; 1994.
3. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997;268(1):78–94.
4. Yeo G, Burge CB. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. J Comput Biol. 2004;11(2–3):377–94.
5. Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, et al. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res. 2010;38(Database issue):161–6. doi:10.1093/nar/gkp885.
6. Elnitski L, Jin VX, Farnham PJ, Jones SJM. Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques. Genome Res. 2006;16:4140006.
7. Zhao Y, Granas D, Stormo GD. Inferring binding energies from selected binding sites. PLoS Comput Biol. 2009;5(12):1000590.
8. Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, et al. Multiplexed massively parallel selex for characterization of human transcription factor binding specificities. Genome Res. 2010;20(6):861–73.
9. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007;316(5830):1497–502.
10. Galas DJ, Schmitz A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res. 1978;5(9):3157–170. doi:10.1093/nar/5.9.3157.
11. Bailey TL, Williams N, Misleh C, Li WW. Meme: discovering and analyzing dna and protein sequence motifs. Nucleic Acids Res. 2006;34(Web-Server-Issue):369–73.
12. Kulakovskiy IV, Boeva VA, Favorov AV, Makeev VJ. Deep and wide digging for binding motifs in chip-seq data. Bioinforma. 2010;26(20):2622–23.
13. Ma X, Kulkarni A, Zhang Z, Xuan Z, Serfling R, Zhang MQ. A highly efficient and effective motif discovery method for chip-seq/chip-chip data using positional information. Nucleic Acids Res. 2012;40(7):50.
14. Grau J, Posch S, Grosse I, Keilwagen J. A general approach for discriminative de novo motif discovery from high-throughput data. Nucleic Acids Res. 2013;41(21):197.
15. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990;18(20):6097–100.
16. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. Jaspar: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004;32(Database issue):91–4.
17. Newburger DE, Bulyk ML. Uniprobe: an online database of protein binding microarray data on protein–dna interactions. Nucleic Acids Res. 2009;37(suppl 1):77–82.
18. Hughes JD, Estep PW, Tavazoie S, Church GM. Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae1. J Mol Biol. 2000;296(5):1205–14. doi:10.1006/jmbi.2000.3519.
19. Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B. Computational detection of cis -regulatory modules. Bioinformatics. 2003;19(suppl 2):5–14. doi:10.1093/bioinformatics/btg1052.
20. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, et al. Transcriptional regulatory code of a eukaryotic genome. Nature. 2004;431(7004):99–104. doi:10.1038/nature02800.
21. Linhart C, Halperin Y, Shamir R. Transcription factor and microrna motif discovery: The amadeus platform and a compendium of metazoan target sets. Genome Research. 2008;18(7):1180–9. doi:10.1101/gr.076117.108.
22. Bembom O. SeqLogo: Sequence logos for DNA sequence alignments. 2015. http://www.bioconductor.org/packages/release/bioc/html/seqLogo.html, accessed 2015.03.05.
23. Colaert N, Helsens K, Martens L, Vandekerckhove J, Gevaert K. Improved visualization of protein consensus sequences by iceLogo. Nat Meth. 2009;6(11):786–7. doi:10.1038/nmeth1109-786.
24. Jianhong Ou LJZ. MotifStack: Plot Stacked Logos for Single or Multiple DNA, RNA and Amino Acid sequence. http://www.bioconductor.org/packages/release/bioc/html/motifStack.html. Accessed on 13 Feb 2015.
25. Mahony S, Benos PV. STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res. 2007;35(Web Server issue):272–58. doi:10.1093/nar/gkm272.
26. Vacic V, Iakoucheva LM, Radivojac P. Two sample logo: a graphical representation of the differences between two sets of sequence alignments. Bioinforma. 2006;22(12):1536–7. doi:10.1093/bioinformatics/btl151.
27. Ali SM, Silvey SD. A general class of coefficients of divergence of one distribution from another. J R Stat Soc Series B (Methodological). 1966;28(1):131–42.
28. Lin J. Divergence measures based on the Shannon entropy. Inf Theory, IEEE Trans on. 1991;37(1):145–51. doi:10.1109/18.61115.
29. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2013. http://www.R-project.org/.
30. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome biology. 2004;5(10):80–16. doi:10.1186/gb-2004-5-10-r80.
31. Eggeling R, Gohr A, Keilwagen J, Mohr M, Posch S, Smith AD, et al. On the value of intra-motif dependencies of human insulator protein ctcf. PLoS ONE. 2014;9(1):85629. doi:10.1371/journal.pone.0085629.
32. Plasschaert RN, Vigneau S, Tempera I, Gupta R, Maksimoska J, Everett L, et al. CTCF binding site sequence differences are associated with unique regulatory and functional trends during embryonic stem cell differentiation. Nucleic acids research. 2014;42(2):774–89. doi:10.1093/nar/gkt910.
33. Nakahashi H, Kwon K-RKR, Resch W, Vian L, Dose M, Stavreva D, et al. A genome-wide map of CTCF multivalency redefines the CTCF code. Cell reports. 2013;3(5):1678–89. doi:10.1016/j.celrep.2013.04.024.
34. Mordelet F, Horton J, Hartemink AJ, Engelhardt BE, Gordân R. Stability selection for regression-based models of transcription factor-DNA binding specificity. Bioinforma. 2013;29(13):117–25. doi:10.1093/bioinformatics/btt221.
35. Keilwagen J, Grau J. Varying levels of complexity in transcription factor binding motifs. Nucleic Acids Res. 2015;43(18):e119.
36. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42(D1):222–30. doi:10.1093/nar/gkt1223.

6 Data Processing and Interpretation of Mass Spectrometry Data

Publications presented in this thesis related to "Data Processing and Interpretation of Mass Spectrometry Data" are entitled "Discovering Regulated Metabolite Families in Untargeted Metabolomics Studies" (Treutler, Tsugawa, et al., 2016) and "Prediction, Detection, and Validation of Isotope Clusters in Mass Spectrometry Data" (Treutler and Neumann, 2016).

6.1 Discovering Regulated Metabolite Families in Untargeted Metabolomics Studies

In the following reference the first author is underlined and I am marked in bold.
Hendrik Treutler, Hiroshi Tsugawa, Andrea Porzel, Karin Gorzolka, Alain Tissier, Steffen Neumann, and Gerd Ulrich U. Balcke. Discovering regulated metabolite families in untargeted metabolomics studies. *Analytical chemistry*, 88(16):8082-8090, August 2016. *doi:10.1021/acs.analchem.6b01569*

`https://pubs.acs.org/doi/abs/10.1021/acs.analchem.6b01569`

analytical chemistry

Article

pubs.acs.org/ac

Discovering Regulated Metabolite Families in Untargeted Metabolomics Studies

Hendrik Treutler,[‡] Hiroshi Tsugawa,[∥] Andrea Porzel,[§] Karin Gorzolka,[‡] Alain Tissier,[†] Steffen Neumann,[‡] and Gerd Ulrich Balcke*[,†]

[†]Leibniz Institute of Plant Biochemistry, Department of Cell and Metabolic Biology, Weinberg 3, D-06120 Halle/Saale, Germany

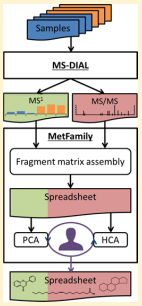
[‡]Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, D-06120 Halle/Saale, Germany

[§]Leibniz Institute of Plant Biochemistry, Department of Bioorganic Chemistry, Weinberg 3, D-06120 Halle/Saale, Germany

[∥]RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa 230-0045, Japan

Supporting Information

ABSTRACT: The identification of metabolites by mass spectrometry constitutes a major bottleneck which considerably limits the throughput of metabolomics studies in biomedical or plant research. Here, we present a novel approach to analyze metabolomics data from untargeted, data-independent LC-MS/MS measurements. By integrated analysis of MS^1 abundances and MS/MS spectra, the identification of regulated metabolite families is achieved. This approach offers a global view on metabolic regulation in comparative metabolomics. We implemented our approach in the web application "MetFamily", which is freely available at http://msbi.ipb-halle.de/MetFamily/. MetFamily provides a dynamic link between the patterns based on MS^1-signal intensity and the corresponding structural similarity at the MS/MS level. Structurally related metabolites are annotated as metabolite families based on a hierarchical cluster analysis of measured MS/MS spectra. Joint examination with principal component analysis of MS^1 patterns, where this annotation is preserved in the loadings, facilitates the interpretation of comparative metabolomics data at the level of metabolite families. As a proof of concept, we identified two trichome-specific metabolite families from wild-type tomato *Solanum habrochaites* LA1777 in a fully unsupervised manner and validated our findings based on earlier publications and with NMR.



INTRODUCTION

Metabolomics experiments provide small molecule measurements from biological samples in a broad range of applications including cancer research, drug development, and plant science.[1−5] Mass spectrometry (MS) coupled to liquid chromatography (LC) is an essential analytical technology to acquire a snapshot of the metabolic state of a sample. On the basis of untargeted MS measurements, it is possible to measure thousands of detectable signals as MS^1 features per chromatographic run and to acquire signal profiles of small molecules based on retention time (RT), accurate mass-to-charge ratio (m/z), and abundance.[6] Univariate or multivariate statistical analysis is then applied to signal profiles of different sample groups to detect MS^1 features that are group-discriminating or of interest based on the experimental design.

Hints for the structural characterization or even identification of MS^1 features are obtained from tandem MS measurements (MS/MS), where the metabolites undergo fragmentation resulting in MS/MS spectra. MS/MS spectra can be collected by data-dependent acquisition (DDA) or in data-independent acquisition (DIA) mode, requiring a trade-off between dwell time and spectral purity.[7,8] Using DIA, it is possible to collect thousands of MS^1 features from a single LC run as well as the associated MS/MS spectra.[9] However, in most studies, the identity of the vast majority of MS^1 features is unknown. Structure elucidation of each individual MS^1 feature from complex biological samples, e.g., by NMR and interpretation of MS/MS spectra, is currently out of reach. Thus, the biochemical relation between MS^1 features remains largely unexplained.

Group-discriminating MS^1 features are often structurally related, e.g., if particular metabolic pathways are differentially regulated as a consequence of disease,[10] stress,[11] genetic manipulation,[12] or in the case of organ-specific accumulation of structurally related metabolites.[13] Structurally related metabolites often exhibit latent similarity in their MS/MS spectra in which characteristic fragmentation patterns arise from common functional groups or structural features. For instance, upon negative mode ionization and collision-induced dissociation (CID), adenylated metabolites such as adenyl nucleotides, CoA esters, and NAD cofactors form a fragment ion of m/z 134.0472 Da ($C_5N_5H_4^-$), which corresponds to the mass of the purine core element. Under the same conditions, glucosides often form a fragment ion of m/z 161.0455 Da ($C_6H_9O_5^-$), characteristic of the hexose side-chain. Thus, on the basis of existing information, precursor ions showing these character-

Received: April 21, 2016
Accepted: July 24, 2016

DOI: 10.1021/acs.analchem.6b01569
Anal. Chem. XXXX, XXX, XXX−XXX

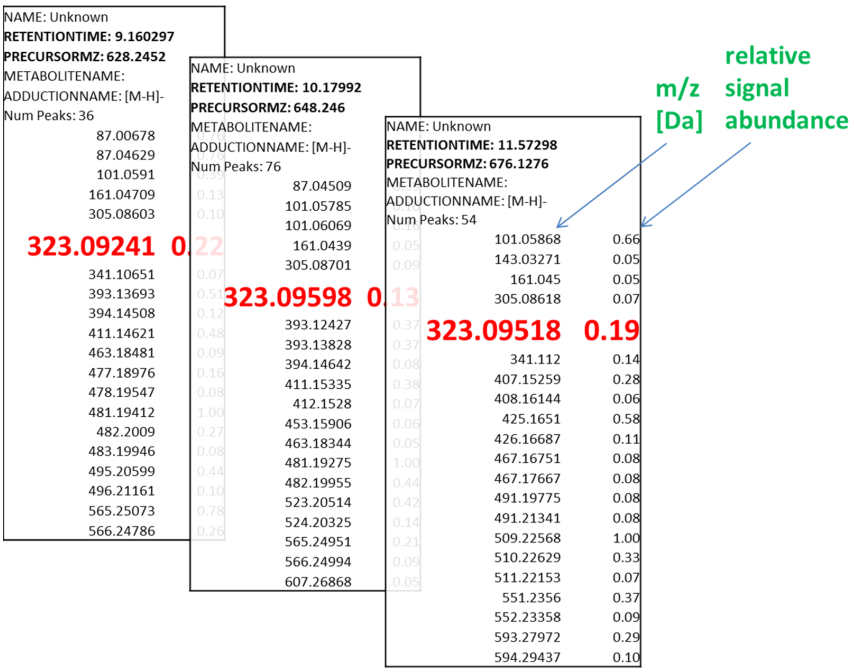


Figure 1. MS/MS library format before upload into MetFamily.

istic fragments could be grouped together as metabolites sharing common structural features, or *metabolite families*. However, even pre-existing MS/MS information characteristic of certain metabolite families is sparse. Hence, novel approaches that turn MS^1- and MS/MS-features into interpretable information within a reasonable amount of time are urgently needed. These approaches should be able to relate MS^1 abundances to latent similarity at the MS/MS spectral level.

Recently, several studies reported on the organization of hundreds of MS^1 features by molecular networking depicting relationships between structurally related molecules based on their spectral similarity.[14−17] However, an explicit assignment of MS^1 features to similarity clusters and the source of structural similarity between up- or downregulated MS^1 features was not apparent. Previously, Wagner et al. used GC-MS data for hierarchical cluster analysis (HCA) to arrange known and structurally related metabolites.[18] Using HCA, it was possible to identify structural classes among 59 metabolites. Rasche et al. described FT-BLAST[19] to compare spectra and computationally derived fragmentation trees, revealing clusters of structurally closely related compounds. However, neither Wagner et al. nor Rasche et al. considered the abundance of MS^1 features in different samples.

Inspired by the idea to comprehensively analyze molecular networks and to explicitly group MS^1 features, we performed HCA across hundreds of MS/MS spectra obtained from glandular trichomes of wild-type tomato *Solanum habrochaites* LA1777. Glandular trichomes of vascular plants such as tomato are metabolic factories producing a plethora of secondary metabolites involved in plant defense and the communication with its environment.[13,20] We considered characteristic fragments prevalent in MS/MS similarity clusters to assign MS^1 features to certain trichome-specific metabolite families. In addition, we applied principal component analysis (PCA) to metabolite profiles for the discovery of group-discriminating MS^1 features and combined the information on metabolite families obtained from HCA (MS/MS feature similarity) with the PCA loadings (sample-specific MS^1 abundance). This combination of statistical analyses of MS^1 feature abundances and MS/MS structural annotations can not only speed-up the individual analysis steps, but allows us to address new questions, such as the discovery of group-discriminating metabolite families with biochemical relevance. Here, we exemplarily selected two metabolite families being produced by tomato glandular trichomes which play important roles in the plant defense against herbivores, namely the branched chain acyl sugars[21−24] and the sesquiterpene glucosides which are potentially poisonous to plant herbivores.[25,26] We implemented the proposed methodology in the Open Source web application "MetFamily" and made our approach freely available (accessible via http://msbi.ipb-halle.de/MetFamily/).

MATERIALS AND METHODS

Fragment Matrix Assembly. MetFamily processes a metabolite profile of a set of MS^1 features together with an MS/MS library comprising MS/MS spectra for these MS^1 features. We obtain both data sets as output of MS-DIAL,[9] where the metabolite profile contains extracted m/z/retention time features from MS^1 scans with the corresponding feature abundances (Data S-1 of the Supporting Information, SI) and the MS/MS library contains deconvoluted MS/MS spectra of the MS^1 features with relative intensities of the fragment ions (Figure 1, Data S-2). Instead of MS-DIAL, other tools can produce similar input data as described in Note S-3. Upon data

DOI: 10.1021/acs.analchem.6b01569
Anal. Chem. XXXX, XXX, XXX−XXX

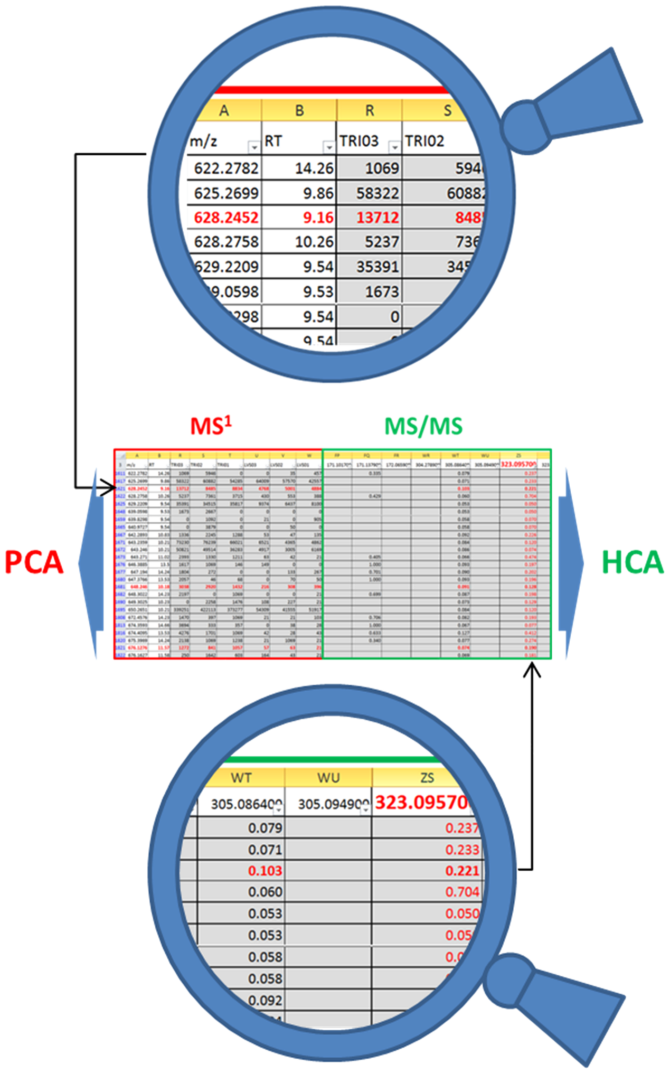


Figure 2. Combined data matrix after data preprocessing by MetFamily. The quantification part (red, left) contains the MS^1 features (rows; precursor ions) and the MS^1 abundances in individual samples. In the fragment part (green, right), the column headers are the mean of binned MS/MS features (m/z or neutral loss) from the MS/MS library. Upper zoom: m/z; retention time of feature (628.2452; 9.16) and its respective peak heights in two trichome samples. Lower zoom: relative MS/MS intensities of fragment ion m/z 323.09570 Da. Arrows to the left and to the right: MS^1 abundances are analyzed using PCA and MS/MS spectra are analyzed using HCA.

import, MetFamily aligns all MS/MS spectra with a user-defined m/z error to create the *fragment matrix* as shown in Figure 2, where the relative intensity of unique MS/MS fragments is associated with the corresponding MS^1 feature (i.e., precursor ion) and its MS^1 abundance in individual samples (Data S-3). For our showcase, this preprocessing step takes one or 2 min. The fragment matrix is assembled as follows.

First, we process the set of all fragments. Here, we remove fragments with an intensity below a user-defined noise threshold. We normalize fragment intensities within each MS/MS spectrum to a maximum of 1 (base peak). In addition, we add one neutral loss (NL) for each fragment by calculating the mass of the neutral loss as the difference of fragment m/z and precursor m/z in MS^1 (intentionally a negative m/z value). The intensity of the NLs is chosen equal to the intensity of the corresponding fragment. In this manuscript, we treat fragments and NLs equally by denoting both as fragments.

Second, we align the individual MS/MS spectra (Figures 1 and 2). Here, we match fragments from different MS/MS spectra with similar m/z and merge these to *fragment groups* of unique m/z. We call the mean of all fragment m/z's of one fragment group the *fragment group mean*. For the alignment of the individual MS/MS spectra, we use an efficient algorithm

DOI: 10.1021/acs.analchem.6b01569
Anal. Chem. XXXX, XXX, XXX−XXX

implemented in the R package *xcms*[31] (version 1.44.0). This algorithm avoids the usage of fixed m/z bins with a heuristic approach that groups fragments with similar m/z and decomposes contiguous fragment groups using hierarchical clustering. Here, a fragment m/z matches a fragment group, if the following:

$$|m - m_{\text{group}}| \leq mz\text{Abs}_{\text{MS/MS}} + m \times mz\text{PPM}_{\text{MS/MS}}/1E6$$

where m is the fragment m/z, m_{group} is the fragment group mean, $mz\text{Abs}_{\text{MS/MS}}$ is a parameter representing the absolute m/z error, and $mz\text{PPM}_{\text{MS/MS}}$ is representing the relative m/z error in ppm (parts per million). See Table S-3 for a summary of user-customizable parameters. After fragment group assembly, we remove fragment groups which correspond to isotopic ions. Specifically, we detect fragment groups with a m/z difference of 1.0033 Da (regarding the fragment group means ± m/z error) which correspond to ^{13}C isotopes. Third, we create the fragment matrix with one row for each unique MS^1 precursor and columns of fragment groups (Figure 2). We register the intensity of each fragment in the row and column of the corresponding MS^1 feature and fragment group, respectively. For each MS^1 feature, we generate an ID given by "m/z/retention time" in MS^1.

Finally, we add the set of MS^1 abundances in all samples and other annotations to each row resulting in a combined data matrix. The combined data matrix represents the data basis for subsequent analyses and can be examined in a spreadsheet program for complementing analyses (Figure 2 and Data S-3).

MS^1/MS/MS Combined Data Analysis. A principal component analysis (PCA) for the set of m MS^1 features in n samples is performed as follows. Given the m by n matrix of scaled MS^1 abundances, we calculate the scores and the loadings. Here, MetFamily supports the scaling functions *log$_2$ transformation*, *Pareto scaling*, *Centering*, and *Autoscaling*.[27] The scores comprise one data point per sample and reflect differences between samples. The loadings comprise one data point per MS^1 feature and emphasize MS^1 features with differential abundance between samples.

We perform a hierarchical cluster analysis (HCA) on MS/MS spectra of a set of MS^1 precursor features as follows. We calculate the distance matrix of pairwise dissimilarities between the MS/MS spectra of all MS^1 features. Here, we provide different distance functions to score common and distinct fragments. Specifically, we recommend the distance function 'Jaccard (intensity-weighted)', which sums the intensities of common and disjoint fragments:

$$f(s_i, s_j) = 1 - \frac{sum(map(s_i \cap s_j))}{sum(map(s_i \cup s_j))}$$

where s_i and s_j are the fragments in the MS/MS spectrum of MS^1 feature i and j. To suppress noise and emphasize the importance of intense fragments, *map* discretizes the intensities of the fragments as follows. Intensities smaller than 0.2 are mapped to 0.01, intensities greater or equal than 0.2 and smaller than 0.4 are mapped to 0.2, and intensities greater or equal than 0.4 are mapped to 1. Given the distance matrix, we calculate a hierarchical cluster dendrogram where each cluster of MS^1 features represents a putative metabolite family.

For each cluster of MS/MS spectra, we calculate the *cluster-discriminating power* for prevalent fragments as follows. For each fragment present in more than 50% of the MS/MS spectra in a cluster, we measure the ability of this fragment to discriminate spectra in the cluster from spectra outside the cluster as

$$\text{cdp}(f_{k,l}) = \frac{p_{\text{in}} - p_{\text{out}}}{n}$$

where $f_{k,l}$ is the l-th fragment of the k-th cluster, p_{in} is the number of MS/MS spectra in the k-th cluster containing the fragment $f_{k,l}$, p_{out} is the number of MS/MS spectra outside the k-th cluster containing the fragment $f_{k,l}$, and n is the total number of MS/MS spectra in the k-th cluster. If $p_{\text{out}} > p_{\text{in}}$, then we define $\text{cdp}(f_{k,l}) = 0$. The cluster-discriminating power of a fragment is in the range from zero to one, and a fragment with a cluster-discriminating power close to one indicates a very specific fragment.

Clusters containing fragments with a cluster-discriminating power close to one indicate metabolite families. Currently, the annotation of metabolite families based on characteristic MS/MS fragments is performed by a mass spectrometry expert who manually evaluates the hierarchy of putative metabolite families and labels a set of clusters with functional and/or structural annotations based on characteristic fragment patterns. Each MS^1 feature can be labeled with one annotation, i.e., membership in a metabolite family.

Plant Growth and Harvest. *Solanum habrochaites* LA1777 was grown on soil in a greenhouse (65% humidity, light intensity: 165 μmol s^{-1} mm^2, 21−24 °C, 16 h light period) and watered with tap water every 2 days. The plant material was harvested 12 weeks after germination during the light phase in the early afternoon. For trichome harvest, tomato leaves were put on the hand palm (using gloves) and trichomes were quickly brushed off the leaves by a 2 cm broad paint brush which was dipped in liquid nitrogen. The frozen trichomes were collected in a mortar filled with liquid nitrogen. Trichomes from 15 plant leaves were pooled under cryogenic conditions and further purified by sieving through steel sieves of 150 μm mesh width (Retsch, Hahn, Germany). After removal of trichomes, the plant leaves were immediately quenched in liquid nitrogen. Pooled leaves were ground in a mortar under liquid nitrogen conditions. After evaporation of all liquid nitrogen during storage at −80 °C leaves and trichomes were lyophilized overnight and stored in a deep freezer until extraction.

Metabolite Extraction. Using wall-reinforced cryo-tubes of 1.6 mL volume (Precellys Steel Kit 2.8 mm, Peqlab Biotechnologie GmbH, Erlangen, Germany) filled with 5 steel beads (3 mm), 25 mg aliquots of dry leaf or trichome powder was suspended in 900 μL dichloromethane/ethanol (−80 °C). Then, 200 μL of 50 mM aqueous ammonium formate/formic acid buffer (0 °C, pH 3) was added to each vial, and two rounds of cell rupture/metabolite extraction were conducted by FastPrep bead beating (60 s, speed 5.5 m/s, first round −80 °C, second round room temperature, FastPrep24 instrument with cryo adapter, MP Biomedicals LLC, Santa Ana, CA, U.S.A.). After phase separation by centrifugation at 20 000g (2 min, 0 °C) the aqueous phase was removed, and 600 μL of the organic phase was collected. Following, 500 uL tetrahydrofuran (THF) was added to exhaustively extract hydrophobic metabolites and the Fastprep and centrifugation were repeated accordingly. The THF supernatant was combined with the first organic phase extract and dried in a stream of nitrogen gas. The dried extract was resuspended in 150 μL 75% methanol (aqueous) and filtered over 0.2 μm PVDF.

DOI: 10.1021/acs.analchem.6b01569
Anal. Chem. XXXX, XXX, XXX−XXX

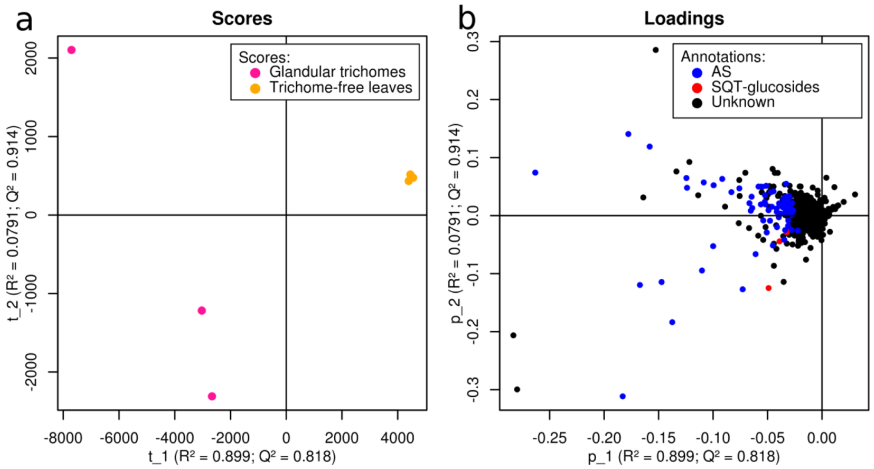


Figure 3. Principal component analysis of metabolite extracts of glandular trichomes and leaves of *Solanum habrochaites* LA1777. Comparison of 2585 MS^1 features from TOF-MS measurements ($n = 6$). (a) scores and (b) loadings with annotations. The PCA loadings with annotations indicate a predominant enrichment of acyl sugars in glandular trichomes. AS: acyl sugars, SQT-glucosides: sesquiterpene glucosides, and Unknown: Not characterized here.

Analytical Conditions for Liquid Chromatography and Mass Spectrometry. 0.5 μL methanolic extract was injected into an Acquity-UPLC (Waters Inc.) and separated on a Nucleoshell RP18 (150 mm × 2 mm × 2.7 μm; Macherey & Nagel, Düren, Germany) at 40 °C. The mobile phase A was 0.33 mM ammonium formate with 0.66 mM formic acid in water; mobile phase B was acetonitrile. The gradient was 0 min, 5% B; 2 min, 5% B; 19 min, 95% B; 21 min, 95% B; 21.1 min, 5% B; and 24 min, 5% B. The column flow rate was 0.4 mL/min, the autosampler temperature was 4 °C.

ESI-(−)-Mass Spectrometry was performed on an AB Sciex TripleTOF 5600 system (Q-TOF) equipped with a DuoSpray ion source. All analyses were performed at the high sensitivity mode for both TOF MS^1 and product ion scan. The mass calibration was automatically performed every 20 injections using an APCI calibrant solution via a calibration delivery system (CDS). The instrument (TripleTOF 5600, Sciex, Toronto, Canada) was configured to simultaneously acquire high resolution MS/MS spectra for all MS^1 features (sequential window acquisition of all theoretical fragment-ion spectra, SWATH)[28] (Figure S-1). The SWATH parameters were MS^1 accumulation time, 150 ms; MS^2 accumulation time, 20 ms; collision energy, −45 V; collision energy spread, 35 V; cycle time, 1160 ms; Q1 window, 25 Da; mass range, m/z 65−1250. The other parameters were curtain gas, 35; ion source gas 1, 60; ion source gas 2, 70; temperature, 600 °C; ion spray voltage floating, −4.5 kV; declustering potential, 35 V.

Raw Data Processing. After measurement, raw data of triplicate trichome and trichome-free leaf material was converted from the vendor file format (in our case *.wiff) into the common file format of Reifycs Inc. (Analysis Base File format *.abf) using the freely available Reifycs ABF converter (http://www.reifycs.com/AbfConverter/index.html).This process took about 1 min per sample. After conversion, the freely available MS-Dial software was used for feature detection, ion species annotation, compound spectra extraction, and peak alignment between samples.[9] Data processing by MS-Dial using the parameters in Table S-1 took about 30 min. Data processing by MetFamily using the parameters in Table S-2 took 1 min.

Notably, neither the use of SWATH-triggered CID fragmentation nor the use of MS-Dial are prerequisite to run MetFamily. Any data independent or data dependent acquisition to collect MS/MS spectra and other peak picking and deconvolution software can alternatively be used.[29−32] In that case, their output has to be provided as a text file containing the peak intensities and a msp-type spectral library which are formatted as exemplified in Data S-1 and Figure 1, and described in Note S-3. However, as unique feature, MS-Dial jointly deconvolutes MS^1 and MS/MS features and automatically predicts the precursor ion when DIA was applied. Via the Reifycs ABF converter, MS-DIAL accepts all of major MS vendor-formats as well as the common mzML data and is applicable to either DIA or DDA MS/MS fragmentation methods.

Substance Purification. Since NMR requires purified analytes in the upper μm range, 1 kg of LA1777 leaf material was surface-extracted with methanol for 2 h. After evaporation, a methanolic concentrate of this extract was produced and injected into a LC system in 100 μL increments. For peak separation using semipreparative HPLC and an analysis by mass spectrometry (1260 Infinity system, Agilent), a full scan between 200 and 800 m/z was performed after negative electrospray ionization (ion source: API-ES, gas temperature: 350 °C, drying gas 10 mL/min, nebulizer pressure 35 psig, capillary voltage 4500 V). For HPLC, a XTerra prep MS C18 column (5 μm × 7.8 mm × 150 mm; Waters) was used and run at a flow rate of 6 mL/min at 25 °C. Solvent A was 0.3 mM ammonium formate acidified with formic acid to pH 6.2. Solvent B was acetonitrile. Gradient conditions were: 0−5 min 5% B; 5−87 min linear gradient to 95% B; 87−88 min 95% B; and 88−90 min 5% B. For fractionation, m/z 605.5, 737.5, and 751.5 triggered the selective collection. A makeup pump that transferred an aliquot of the eluate to the mass analyzer was set to 0.5 mL/min 50% A - 50% B. Subsequently, all collected fractions were dried by lyophilization prior to NMR analysis.

DOI: 10.1021/acs.analchem.6b01569
Anal. Chem. XXXX, XXX, XXX−XXX

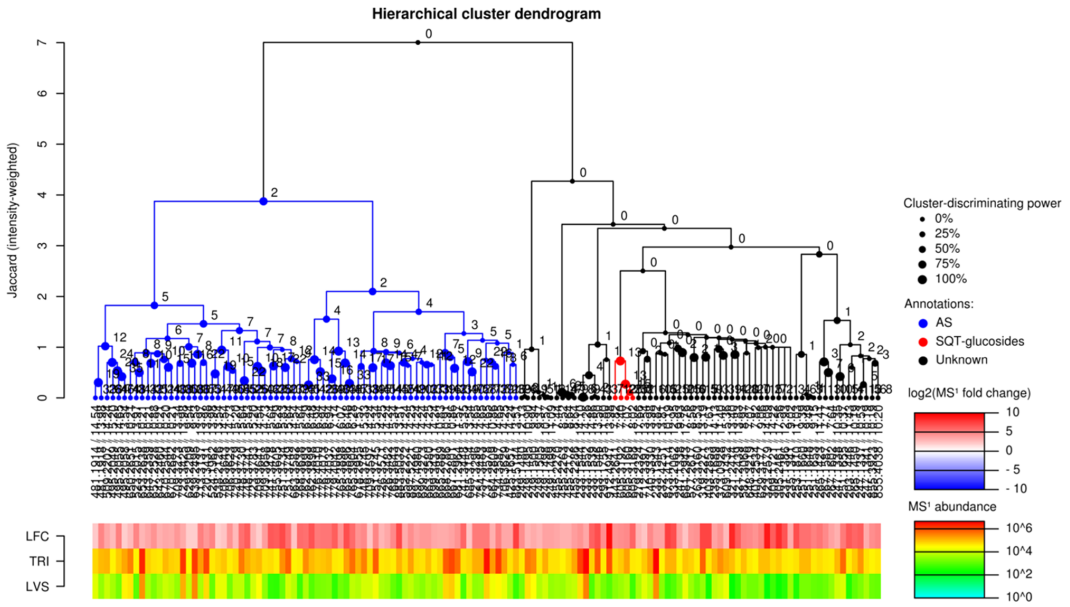


Figure 4. Hierarchical cluster analysis of 135 trichome-specific MS^1 features using the corresponding MS/MS spectra obtained from organic extracts of *S. habrochaites* LA1777. For comparison of the groups trichomes versus leaf focusing on trichome-specific features, the set of 2585 MS^1 features was filtered using an MS^1 abundance threshold of 20 000 counts and a $\log_2$-fold change (LFC) of two. The heatmap below depicts the LFC and the absolute MS^1 abundance in glandular trichomes (TRI) and trichome-free leaves (LVS), respectively. The 135 filtered MS^1 features clearly segregated into two main signal-clusters which in turn further segregated into signal-clusters with different levels of similarity between MS/MS spectra. Specifically, we identified a cluster of 73 short branched chain acyl sugars (AS, in blue) and a cluster of four sesquiterpene glucosides (SQT-glucosides, in red) on the basis of a set of characteristic fragments which were prevalent in both clusters (see legend "Annotations" on the right). Both signal-clusters show characteristic fragments with a cluster-discriminating power of 80% and more (size of the branch nodes, see legend "Cluster-discriminating power" on the right). 58 trichome-specific MS^1 features partially showed further clusters, but remained uncharacterized in this study (Unknown in black).

Analytical Conditions for NMR. NMR spectra were recorded on an Agilent/Varian VNMRS 600 NMR spectrometer operating at a proton NMR frequency of 599.83 MHz using a 5 mm inverse detection cryoprobe. 2D NMR spectra were recorded using standard pulse sequences (gDQCOSY, zTOCSY, gHSQCAD, gHMBCAD) implemented in Agilent (Varian) VNMRJ 4.2A (CHEMPACK 7.1) spectrometer software. A TOCSY mixing time of 80 ms was used. HSQC experiments were run with multiplicity editing and optimized for $^1J_{CH}$ = 146 Hz. HMBC experiments were optimized for a long-range coupling constant of 8 Hz; a 2-step $^1J_{CH}$ filter was used (130−165 Hz). Proton and carbon chemical shifts are referenced to internal TMS (0 ppm).

RESULTS AND DISCUSSION

As a proof of concept, we applied MS signal profiles to compare the metabolism of a special plant organ in tomato, the glandular trichomes, to tomato leaves. Plant glandular trichomes are secretory cells that protrude from the epidermis of many vascular plants. As "metabolic factories", they produce important drugs such as the antimalaria artemisinin or compounds known to be involved in plant defense.[20,33] Here, we used *Solanum habrochaites* LA1777, a wild type tomato accession with a rich profile of secondary metabolites produced in the glandular trichomes.[34] We used six UPLC-(−)ESI-SWATH-MS/MS runs of triplicate trichome and trichome-free leaf extracts (cf. Materials and Methods). However, MetFamily is applicable to a larger number of samples and sample groups. We used MS-DIAL[9] for data preprocessing and exported (i) a signal profile with MS^1 features and (ii) a spectral library with deconvoluted MS/MS spectra extracted from the raw data (Data S-1 and Data S-2). Using the software MetFamily, we aligned the MS/MS spectra of the spectral library resulting in a novel fragment matrix structure, and we fused this fragment matrix with the matching set of MS^1 features from the six individual samples to a single matrix (cf. Materials and Methods, Figures 1 and 2, Data S-3, Table S-3).

MetFamily provides options to perform principal component analyses (PCA). Here, we performed a PCA on 2585 MS^1 features detected in glandular trichomes or leaves of LA1777 using Pareto-scaled data. In our example, PC1 shows a clear separation between trichomes and leaves with R^2 = 0.90, Q^2 = 0.82 and a large number of MS^1 features more abundant in glandular trichomes (Figure 3A,B). A Scree plot on additional principal components is provided in Figure S-2. Up to this point, all data have been acquired in a fully untargeted manner and traditionally this is where group-discriminating MS^1 features would be subjected to tedious manual structure elucidation. In our approach, we amended the loadings plot of the PCA (Figure 3B) with a set of structural annotations based on characteristic MS/MS fragments which we identified in different signal-clusters using HCA (Figure 4). Using MetFamily, we performed a hierarchical cluster analysis (HCA) on MS/MS spectra of the fragment matrix (Data S-

DOI: 10.1021/acs.analchem.6b01569
Anal. Chem. XXXX, XXX, XXX−XXX

6. DATA PROCESSING AND INTERPRETATION OF MASS SPECTROMETRY DATA

3). For PCA as well as for HCA, MetFamily allows the usage of thresholds for the MS^1 abundance of individual MS^1 features (average of all samples) and for the $\log_2$-fold change between the average MS^1 abundance of two sample groups. Since we were interested in abundant trichome-specific metabolites, we retained 135 MS^1 features in the HCA with MS^1 abundances ≥20 000 counts and a $\log_2$-fold change ≥2 comparing trichomes versus leaf. After hierarchical cluster analysis, the resulting dendrogram indicated a clear segregation into two main clades with internal spectral similarity (Figure 4).

The first signal-cluster contained 73 MS^1 features which correspond to short branched chain acyl sugars[21] (AS, blue in Figure 4). The structural similarities among members of this clade was supported by prevalent fragment ions 87.0451 Da (theoretical mass for $C_4H_7O_2^-$ is 87.0452) and 101.0603 Da (theoretical mass for $C_5H_9O_2^-$ is 101.0608), which are indicative for short branched acyl groups. These acyl moieties were esterified to sucrose as reflected by the fragments 323.0957 Da (theoretical mass for $C_{12}H_{19}O_{10}^-$ is 323.0984; sucrose-H_2O—H^-) and 305.0864 Da (theoretical mass for $C_{12}H_{17}O_9^-$ is 305.0878; sucrose-$2H_2O$—$H]^-$). MS/MS fragmentation patterns and NMR analysis of two selected MS^1 features of this clade ([m/z; RT]: [737.3578; 14.65] and [751.3749; 15.64]) confirmed the membership to the metabolite family of short branched chain acyl sugars (Figures S-3, S-4, S-7, S-8, S-11–S-14, S-15–S-19 and Tables S-5, S-6). Our NMR analysis revealed that the feature [737.3578; 14.65] comprised an isomeric mixture of isobutyl, isopentyl, and anteisobutyl acyl moieties, which were not resolvable using our chromatography. MS/MS fragmentation and NMR of various AS have been thoroughly studied earlier by Ghosh et al., where compounds selected here for analysis were annotated as acylsucrose S4:21[2] (theoretical m/z:737.36012 Da (formate adduct-H)) and acylsucrose S4:22[6] (theoretical m/z: 751.37577 Da (formate adduct-H)), respectively.[21]

The second signal-cluster contained a group of four MS^1 features which correspond to sesquiterpene glycosides (SQT-glucosides, red in Figure 4). The structural similarities among members of this clade was supported by three prevalent fragment ions: m/z 401.2548 Da (theoretical mass for $C_{21}H_{37}O_7^-$ is 401.2545), 563.3051 Da (theoretical mass for $C_{27}H_{47}O_{12}^-$ is 563.3073), and 605.3176 Da (theoretical mass for $C_{29}H_{49}O_{13}^-$ is 605.3179) (Figure S-5). Recently, Ekanayaka et al. identified a novel class of trichome-specific sesquiterpene glucosides from *S. habrochaites* using these fragment ions and elucidated the structures of purified representatives by NMR.[26] In our study, CID fragmentation and preparative isolation of MS^1 feature [605.3160; 7.07, an abundant in-source fragment] with subsequent NMR confirmed the structure of 12-*O*-(6″-*O*-malonyl-*β*-D-glucopyranosyl-(1 → 2)-*β*-D-glucopyranosyl)-cam-pherenane-2-endo,12-diol, a member of the novel sesquiter-pene glucoside metabolite family (Figures S-3–S-6, S-9, S-10 and Tables S-3, S-4).

After annotation of both metabolite families, the correspond-ing MS^1 features are highlighted by their color-code in the PCA loadings (Figure 3B). In our case, it was evident that the representatives of both metabolite families were enriched in glandular trichomes, indicating a trichome-specific upregulation of short branched chain acyl sugars and sesquiterpene glucosides. Please note that the hierarchical cluster dendrogram comprised more clades with internal spectral similarity, but we concentrated on the short branched chain acyl sugars and sesquiterpene glucosides whose structures were confirmed by NMR. A detailed workflow exemplified here is given in Figure S-1, and the full showcase protocol is given in Note S-1. A general user guide for MetFamily is given in Note S-2.

Additional Features of MetFamily. MetFamily also supports semitargeted analyses. In this case, sets of MS^1 features can be selected by certain fragment masses, neutral losses, or combinations thereof within a user-defined mass error in ppm as filter criteria. Using this option, only selected MS^1 features are considered in subsequent PCA or HCA calculations and the data analysis is consequently constrained to selected metabolite families. For example, to isolate only glycosylated MS^1 features from all data the user can specify a fragment ion of m/z 161.0455 Da ($C_6H_9O_5^-$) from MS/MS spectra in negative mode and can then focus on the regulation of enzymatic glycosylations in a biological context (for details, see the MetFamily user guide in Note S-2). When we applied this filter with a mass error of 25 ppm, we obtained 568 MS^1 features from our example data, presumably containing a hexose as a structural moiety. In addition, it is possible to search MS^1 features with certain fragments or neutral losses postanalysis. The corresponding MS^1 features can then be jointly visualized in the PCA loadings and the hierarchical cluster dendrogram.

It is possible to export different kinds of results from MetFamily. Selected sets of precursor ions can be exported and, e.g., reloaded into the original MS data acquisition software. Further, it is possible to export both the hierarchical cluster dendrogram and the PCA plots as publication-ready high quality images. The set of parameters used for the initial data import can be exported and imported. Finally, it is possible to export the whole project (including all annotations and color codes) to enable the user to share the project or to continue the data analysis at a later time (Data S-4).

■ CONCLUSIONS

The web application "MetFamily" presented here constitutes a novel approach to analyze metabolomics data from untargeted, data-independent LC-MS/MS measurements. Rather than relying on the time-consuming structure identification of individual metabolites, MetFamily assists in the interpretation of complex metabolomics data by identifying metabolite families through patterns in MS/MS. These are generated by similarity clustering of associated MS/MS spectra and can be annotated with names and colors. After preprocessing of LC-MS/MS raw data, MetFamily performs a joint data analysis of MS^1 abundances and MS/MS spectra in which the annotation of metabolite families facilitates the interpretation of com-parative data sets. Structure elucidation at the metabolite level can be performed afterward in a much more focused way. As a proof of concept, we identified two trichome-specific metabolite families from wild type *Solanum habrochaites* LA1777 in a fully unsupervised manner and validated our findings based on earlier publications and with NMR. The plethora of identified trichome-specific acyl sucroses correlates with upregulation of acyltransferases of the BAHD family in tomato glandular trichomes (Schilmiller 2012). In addition, the size of the clade "acyl sugar" is related to a low substrate specificity of BAHD acyltransferases, illustrating that MetFam-ily can uncover links between enzymatic promiscuity and organ-specific regulation of enzymes.

Using the proposed approach, it is now possible to obtain a comprehensive overview of data sets containing thousands of mass features within a reasonable amount of time. Thus, by

DOI: 10.1021/acs.analchem.6b01569
Anal. Chem. XXXX, XXX, XXX–XXX

providing a dynamic link between structural similarity at the MS/MS level (HCA) and the corresponding MS^1-signal intensity-based patterns (PCA) we bridge the gap between raw data and structural information. Moreover, using MetFamily, precursor ions can now be filtered via combinations of fragment ions and neutral losses, permitting the selection of metabolite families based on characteristic fragmentation patterns.

While traditional compound identification is based on the comparison of MS/MS spectra (or electron impact MS spectra) with reference spectra from known compounds, future developments should exploit spectral patterns of MS/MS features being characteristic of certain metabolite families. Public knowledge on such characteristic fragment ions or neutral losses, e.g., based on metabolite families, can assist mass spectrometry specialists in the elucidation of unknown features and will open new perspectives in life science.

AVAILABILITY

Project name: MetFamily
Source code: https://github.com/Treutler/MetFamily
Availability: http://msbi.ipb-halle.de/MetFamily/
Operating system(s): Platform independent
Programming language: R
Other requirements: Installation of R 3.2.2 or higher; License: GPL 3
Any restrictions to use by nonacademics: None

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.6b01569.

General work flow as flowchart (Figure S-1) Scree plot of the first five principle components of Figure 3 (Figure S-2). Exported parameter file of MS-DIAL (Table S-1) and exported parameter file from MetFamily and parameter explanation (Tables S-2, S-3). Structure elucidation of three selected MS^1 features (Figures S-3−S-19, Tables S-4−S-6). Metabolite profile of the showcase (Data S-1), MS/MS library of the showcase (Data S-2), matrix of the showcase (Data S-3), and annotated MetFamily project file (Data S-4) (PDF)
Protocol for the presented showcase (PDF)
user guide for MetFamily (PDF)
Detailed specification of MetFamily input files (PDF)
(TXT)
(ZIP)
(ZIP)
(ZIP)

AUTHOR INFORMATION

Corresponding Author

*E-mail: gerd.balcke@ipb-halle.de (G.U.B.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We would like to thank Anja Ehrlich for the preparative isolation of metabolites, Anja Henning and Nick Bergau for trichome harvest and data acquisition. We thank Prof. Masanori Arita and K.G. for fruitful discussions and review of the manuscript.

REFERENCES

(1) Jorge, T. F.; Rodrigues, J. A.; Caldana, C.; Schmidt, R.; van Dongen, J. T.; Thomas-Oates, J.; Antonio, C. *Mass Spectrom. Rev.* **2016**, *35*, 620.
(2) Tonoli, D.; Varesio, E.; Hopfgartner, G. *Chimia* **2012**, *66*, 218−222.
(3) Wishart, D. S.; Mandal, R.; Stanislaus, A.; Ramirez-Gaona, M. *Metabolites* **2016**, *6*.
(4) Suhre, K.; Shin, S. Y.; Petersen, A. K.; Mohney, R. P.; Meredith, D.; Wagele, B.; Altmaier, E.; CardioGram; Deloukas, P.; Erdmann, J.; Grundberg, E.; Hammond, C. J.; de Angelis, M. H.; Kastenmuller, G.; Kottgen, A.; Kronenberg, F.; Mangino, M.; Meisinger, C.; Meitinger, T.; Mewes, H. W.; Milburn, M. V.; Prehn, C.; Raffler, J.; Ried, J. S.; Romisch-Margl, W.; Samani, N. J.; Small, K. S.; Wichmann, H. E.; Zhai, G.; Illig, T.; Spector, T. D.; Adamski, J.; Soranzo, N.; Gieger, C. *Nature* **2011**, *477*, 54−60.
(5) Fiehn, O. *Plant Mol. Biol.* **2002**, *48*, 155−171.
(6) Fernie, A. R.; Trethewey, R. N.; Krotzky, A. J.; Willmitzer, L. *Nat. Rev. Mol. Cell Biol.* **2004**, *5*, 763−769.
(7) Roemmelt, A. T.; Steuer, A. E.; Poetzsch, M.; Kraemer, T. *Anal. Chem.* **2014**, *86*, 11742−11749.
(8) Zhu, X.; Chen, Y.; Subramanian, R. *Anal. Chem.* **2014**, *86*, 1202−1209.
(9) Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M. *Nat. Methods* **2015**, *12*, 523−526.
(10) Ferslew, B. C.; Xie, G.; Johnston, C. K.; Su, M.; Stewart, P. W.; Jia, W.; Brouwer, K. L.; Sidney Barritt, A. t. *Dig. Dis. Sci.* **2015**, *60*, 3318−3328.
(11) Kaling, M.; Kanawati, B.; Ghirardo, A.; Albert, A.; Winkler, J. B.; Heller, W.; Barta, C.; Loreto, F.; Schmitt-Kopplin, P.; Schnitzler, J. P. *Plant, Cell Environ.* **2015**, *38*, 892−904.
(12) Qu, G.; Quan, S.; Mondol, P.; Xu, J.; Zhang, D.; Shi, J. *J. Integr. Plant Biol.* **2014**, *56*, 849−863.
(13) Glas, J. J.; Schimmel, B. C. J.; Alba, J. M.; Escobar-Bravo, R.; Schuurink, R. C.; Kant, M. R. *Int. J. Mol. Sci.* **2012**, *13*, 17077−17103.
(14) Watrous, J.; Roach, P.; Alexandrov, T.; Heath, B. S.; Yang, J. Y.; Kersten, R. D.; van der Voort, M.; Pogliano, K.; Gross, H.; Raaijmakers, J. M.; Moore, B. S.; Laskin, J.; Bandeira, N.; Dorrestein, P. C. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, E1743−1752.
(15) Garg, N.; Kapono, C. A.; Lim, Y. W.; Koyama, N.; Vermeij, M. J. A.; Conrad, D.; Rohwer, F.; Dorrestein, P. C. *Int. J. Mass Spectrom.* **2015**, *377*, 719−727.
(16) Nguyen, D. D.; Wu, C. H.; Moree, W. J.; Lamsa, A.; Medema, M. H.; Zhao, X. L.; Gavilan, R. G.; Aparicio, M.; Atencio, L.; Jackson, C.; Ballesteros, J.; Sanchez, J.; Watrous, J. D.; Phelan, V. V.; van de Wiel, C.; Kersten, R. D.; Mehnaz, S.; De Mot, R.; Shank, E. A.; Charusanti, P.; Nagarajan, H.; Duggan, B. M.; Moore, B. S.; Bandeira, N.; Palsson, B. O.; Pogliano, K.; Gutierrez, M.; Dorrestein, P. C. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, E2611−E2620.
(17) Li, D.; Baldwin, I. T.; Gaquerel, E. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, E4147−4155.
(18) Wagner, C.; Sefkow, M.; Kopka, J. *Phytochemistry* **2003**, *62*, 887−900.
(19) Rasche, F.; Scheubert, K.; Hufsky, F.; Zichner, T.; Kai, M.; Svatos, A.; Bocker, S. *Anal. Chem.* **2012**, *84*, 3417−3426.
(20) Tissier, A. *Plant J.* **2012**, *70*, 51−68.
(21) Ghosh, B.; Westbrook, T. C.; Jones, A. D. *Metabolomics* **2014**, *10*, 496−507.
(22) Kim, J.; Kang, K.; Gonzales-Vigil, E.; Shi, F.; Jones, A. D.; Barry, C. S.; Last, R. L. *Plant Physiol.* **2012**, *160*, 1854−1870.
(23) Schilmiller, A.; Shi, F.; Kim, J.; Charbonneau, A. L.; Holmes, D.; Jones, A. D.; Last, R. L. *Plant J.* **2010**, *62*, 391−403.
(24) Schilmiller, A. L.; Charbonneau, A. L.; Last, R. L. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 16377−16382.

DOI: 10.1021/acs.analchem.6b01569
Anal. Chem. XXXX, XXX, XXX−XXX

(25) Ekanayaka, E. A.; Celiz, M. D.; Jones, A. D. *Plant Physiol.* **2015**, *167*, 1221.
(26) Ekanayaka, E. A. P.; Li, C.; Jones, A. D. *Phytochemistry* **2014**, *98*, 223−231.
(27) van den Berg, R. A.; Hoefsloot, H. C.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J. *BMC Genomics* **2006**, *7*, 142.
(28) Gillet, L. C.; Navarro, P.; Tate, S.; Rost, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. *Mol. Cell. Proteomics* **2012**, *11*, 016717.
(29) Lommen, A. *Anal. Chem.* **2009**, *81*, 3079−3086.
(30) Lommen, A.; Kools, H. J. *Metabolomics* **2012**, *8*, 719−726.
(31) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779−787.
(32) Kuhl, C.; Tautenhahn, R.; Bottcher, C.; Larson, T. R.; Neumann, S. *Anal. Chem.* **2012**, *84*, 283−289.
(33) Paddon, C. J.; Westfall, P. J.; Pitera, D. J.; Benjamin, K.; Fisher, K.; McPhee, D.; Leavell, M. D.; Tai, A.; Main, A.; Eng, D.; Polichuk, D. R.; Teoh, K. H.; Reed, D. W.; Treynor, T.; Lenihan, J.; Fleck, M.; Bajad, S.; Dang, G.; Dengrove, D.; Diola, D.; Dorin, G.; Ellens, K. W.; Fickes, S.; Galazzo, J.; Gaucher, S. P.; Geistlinger, T.; Henry, R.; Hepp, M.; Horning, T.; Iqbal, T.; Jiang, H.; Kizer, L.; Lieu, B.; Melis, D.; Moss, N.; Regentin, R.; Secrest, S.; Tsuruta, H.; Vazquez, R.; Westblade, L. F.; Xu, L.; Yu, M.; Zhang, Y.; Zhao, L.; Lievense, J.; Covello, P. S.; Keasling, J. D.; Reiling, K. K.; Renninger, N. S.; Newman, J. D. *Nature* **2013**, *496*, 528−532.
(34) McDowell, E. T.; Kapteyn, J.; Schmidt, A.; Li, C.; Kang, J. H.; Descour, A.; Shi, F.; Larson, M.; Schilmiller, A.; An, L. L.; Jones, A. D.; Pichersky, E.; Soderlund, C. A.; Gang, D. R. *Plant Physiol.* **2011**, *155*, 524−539.

DOI: 10.1021/acs.analchem.6b01569
Anal. Chem. XXXX, XXX, XXX−XXX

6.2 Prediction, Detection, and Validation of Isotope Clusters in Mass Spectrometry Data

In the following reference the first author is underlined and I am marked in bold.
Hendrik Treutler and Steffen Neumann. Prediction, detection, and validation of isotope clusters in mass spectrometry data. *Metabolites*, 6(4):37+, October 2016. *doi:10.3390/metabo6040037*

`https://www.mdpi.com/2218-1989/6/4/37`

metabolites

MDPI

Article

Prediction, Detection, and Validation of Isotope Clusters in Mass Spectrometry Data

Hendrik Treutler [1,2,*] and Steffen Neumann [1]

[1] Department of Stress and Developmental Biology, Leibniz Institute for Plant Biochemistry, Weinberg 3, Halle 06120, Germany; steffen.neumann@ipb-halle.de

[2] Institute of Computer Science, Martin-Luther-University Halle-Wittenberg, Von-Seckendorff-Platz 1, Halle 06120, Germany

* Correspondence: hendrik.treutler@ipb-halle.de; Tel.: +49-345-5582-1472

Academic Editor: Peter D. Karp
Received: 31 August 2016; Accepted: 14 October 2016; Published: 20 October 2016

Abstract: Mass spectrometry is a key analytical platform for metabolomics. The precise quantification and identification of small molecules is a prerequisite for elucidating the metabolism and the detection, validation, and evaluation of isotope clusters in LC-MS data is important for this task. Here, we present an approach for the improved detection of isotope clusters using chemical prior knowledge and the validation of detected isotope clusters depending on the substance mass using database statistics. We find remarkable improvements regarding the number of detected isotope clusters and are able to predict the correct molecular formula in the top three ranks in 92% of the cases. We make our methodology freely available as part of the Bioconductor packages *xcms* version 1.50.0 and *CAMERA* version 1.30.0.

Keywords: isotope cluster; software; raw data

1. Introduction

The elucidation of the metabolism provides deep insights into complex processes in the cell such as responses to nutrition deficiency, pathogen exposure, and drought stress in plants or the implications of mutations, age, and tissue development in animals. Mass spectrometry is a key technology for the identification and quantification of metabolites in biological samples. After measurement using mass spectrometers, feature detection algorithms extract basic properties about peaks in the raw data such as retention time and peak height. The set of properties describing single peaks are called *features* and the exhaustive extraction of features is a prerequisite for downstream analyses such as metabolite identification and quantitative comparisons between samples.

The feature detection algorithm *centWave* in the R package *xcms* version 1.50.0 [1] adapts the following procedure. First, a set of *regions of interest* (ROIs) is identified in the ROI identification step, where ROIs are two-dimensional intervals in the mass-to-charge (m/z) dimension and the retention time dimension containing potential signals. The set of ROIs is examined in the ROI examination step in order to validate, localize, and quantify features. In the ROI identification step, a heuristic method is applied to the raw data to substantially reduce the processing time of the more computationally intensive ROI examination step. This heuristic method aims at a high specificity at the cost of sensitivity, especially in case of features with a low signal-to-noise ratio. Consequently, potentially important features in the raw data are not detected and the information behind these features cannot be used in downstream analyses.

Most chemical elements are present in different variants called isotopes. Though chemically almost equivalent, the isotopes of a particular chemical element differ in mass and are thus well distinguishable using mass spectrometry. The isotopes of each element have a known natural

abundance and the distribution of isotopes across all atoms of a molecule results in a set of related signals. The features extracted from these signals are called *isotopologue features* and the set of all isotopologue features from one analyte is called *isotope cluster* also known as isotope pattern. Unfortunately, many of these signals are below the detection limit which results in the underestimation of isotopologue features.

Based on isotope clusters, it is possible to determine the charge state, abundance, and elemental composition of the measured ion with high precision. The arrangement of isotopologue features to isotope clusters leads to a considerable reduction of data complexity facilitating the interpretation of data sets. It has been demonstrated that the analysis of isotope clusters leads to an increased confidence and precision of comparative analyses [2]. Isotope clusters from precursor ions and tandem mass spectrometry are pivotal for the determination of the molecular formula using software like SIRIUS [3], Rdisop [4], and others [5–12]. The molecular formula strongly facilitates the identification of molecules known as a major bottleneck in metabolomics [13,14] and has been demonstrated metabolome-scale [15]. There are approaches in metabolomics and proteomics which use isotope clusters to improve peak picking [16–18]. In addition, isotope clusters have been used as a valuable source for the assessment of the data quality [19] and for database searches with high precision [20].

The detection of isotope clusters is usually performed after peak picking by consideration of coeluting features separated by certain distances in the m/z dimension. However, a validation of putative isotope clusters in terms of the removal of leading peaks from hydrogen–losses and the decomposition of overlapping isotope clusters into individual isotope clusters is usually lacking in case of small molecules. The deconvolution of overlapping isotope clusters has been described in case of peptides and proteins, for isotope dilution experiments, and in case of substances with known molecular formula [17,21,22].

Aiming at the exhaustive detection and precise validation of isotope clusters, we propose the following approach for liquid chromatography–high resolution mass spectrometry data. We predict new ROIs for putative isotope peaks based on previously detected features and implement this approach in combination with the *centWave* algorithm as part of the R package *xcms* version 1.50.0 [23]. We validate putative isotope clusters depending on the mass of the substance based on database statistics and implement this approach as part of the R package *CAMERA* version 1.30.0 [24].

For evaluation purposes, we apply the modified *centWave* algorithm to different sets of mass spectrometry raw data and detect and validate isotope clusters as proposed. We evaluate the results using various performance measures and find remarkable improvements regarding the number of detected isotope clusters. The extended R packages *xcms* and *CAMERA* are available at Bioconductor [25].

2. Results

We demonstrate the performance of our approach for an enhanced isotope cluster detection and validation. First, we describe the workflow which includes our approach; Second, we evaluate the proposed targeted peak picking with predicted isotope ROIs compared to peak picking with random ROIs and traditional peak picking on basis of various performance measures; Third, we evaluate the proposed isotope detection routine with mass–specific isotope cluster validation compared to several isotope detection routines on basis of various performance measures; Fourth, we present the isotope ratio quantiles which are used for the validation of isotope clusters; Fifth, we exemplify the proposed isotope detection routine with and without mass–specific isotope cluster validation on six example substances.

2.1. Workflow of the Approach

We integrated the proposed methodology into an untargeted workflow which extracts annotated peak tables from LC-MS raw data as summarized in Figure 1. The user supplies the LC-MS raw data files in a *xcms*-supported format, namely one of AIA/ANDI NetCDF, mzXML, mzData, or mzML.

Metabolites **2016**, *6*, 37

The workflow incorporates one function from the R package *xcms* [23], one function from the R package *CAMERA* [24], and two new function as follows.

First, we perform peak picking without any prior knowledge which we denote as *traditional peak picking*. Here, we use the *centWave* algorithm [1] which applies a heuristic for the detection of ROIs (ROI identification step). Given the set of detected ROIs, chromatographic peaks are extracted using continuous wavelet transformation (ROI examination step). This step results in a peak table with one row for each detected feature and one column for each feature property such as m/z, retention time, integrated peak area, and signal-to-noise ratio.

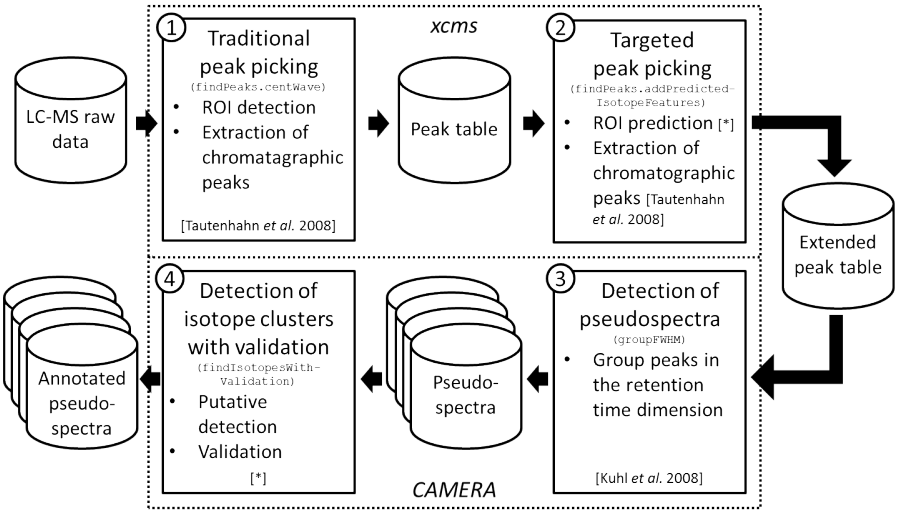


Figure 1. Workflow of the proposed approach. We depict data sets with cylinders, algorithms with continuous rectangles, and R packages with dotted rectangles. Each algorithm rectangle comprises the step number (top left corner), the purpose of the algorithm (heading), the R function name (monospace font), algorithm steps (itemized), and a reference for the algorithm or the individual algorithm steps (in square brackets, asterisk stands for this manuscript). ① The workflow starts with traditional peak picking on LC-MS raw data to extract a peak table comprising features; ② This peak table is extended by a targeted peak picking which targets on isotope features; ③ The extended peak table is split into putative compound spectra denoted pseudospectra; ④ The detection and validation of isotope clusters is performed on each pseudospectrum resulting in annotated pseudospectra.

Second, we perform the proposed targeted peak picking as described in Section 4.1. Here, a set of isotope ROIs is predicted on basis of the previously extracted peak table. Given the set of predicted isotope ROIs, chromatographic peaks are extracted using continuous wavelet transformation (ROI examination step). Notably, this ROI examination step is identical to the ROI examination step in the traditional peak picking step with the exception that we use relaxed peak picking parameters this time. This step results in an extended peak table which is enriched with features corresponding to isotope isotope peaks as demonstrated in the second results section.

Third, we extract *pseudospectra* from the extended peak table [24]. This step aims at the extraction of compound spectra on basis of the retention times, but multiple coeluting compounds are potentially assigned to the same spectrum which is the reason for the usage of the term pseudospectrum. In case of multiple raw data files a retention time correction (*xmcs* function `retcor`) can be advisable prior to the extraction of pseudospectra. This step results in a set of pseudospectra. Each pseudospectrum is a peak table comprising all properties of a subset of the features from the extended peak table.

Fourth, we detect isotope clusters in each pseudospectrum using the proposed isotope detection routine with mass–specific isotope cluster validation as described in Section 4.2. Here, putative isotope clusters are detected and putative isotope clusters are validated based on database statistics as demonstrated in the third results section. This step results in a set of annotated pseudospectra, i.e., the given set of pseudospectra enriched with isotope annotations.

The presented workflow is implemented exemplarily in the vignette IsotopeDetectionVignette in R package *CAMERA* in version 1.30.0. In addition the R package *CAMERA* supports a number of further analyses given the set of annotated pseudospectra. This includes, amongst others, the annotation of adducts and neutral losses, the filling of missing values, and the combination of results from opposite ion modes.

2.2. Targeted Peak Picking Using Predicted Isotope ROIs

We examine whether the proposed prediction of isotope ROIs in combination with the *centWave* algorithm increases the number of detected isotope peaks. To verify the specificity of the predicted isotope ROIs to isotopes, we compare predicted isotope ROIs with the same number of random ROIs denoted *noise ROIs*. In addition, we compare our approach to the unmodified *centWave* algorithm with different signal-to-noise thresholds snthr. We evaluate our approach based on a dilution series experiment with 40 LC-MS measurements. These data sets comprise both strong and weak signals and constitute the basis to test the detection of weak signals like isotope peaks.

We evaluate the performance of predicted isotope ROIs detected with different relaxed signal-to-noise thresholds `snthr'` as described in Section 4.1 on 40 LC-MS measurements described in Section 4.4. We quantify the performance using the performance measures (i) number of detected peaks; (ii) number of detected isotope peaks; (iii) number of detected isotope clusters; (iv) *isotope coverage*; and (v) Peak Picking Score (*PPS*). The isotope coverage is the ratio between the number of detected isotope peaks and the number of detected peaks. The isotope coverage ranges from 0 to 1, where 0 means that no isotope clusters have been detected and 1 means that all peaks are part of isotope clusters. A higher isotope coverage indicates a higher peak picking quality as exploited in [19]. The PPS was proposed in [19] for the quantification of the peak picking quality and implemented in the R package *IPO*. The PPS is defined as the ratio between the number of reliable peaks squared and the number of non–reliable peaks. The number of reliable peaks is defined as the number of peaks in isotope clusters which are detected in the *IPO* package by a custom isotope detection routine. The number of non–reliable peaks is defined as the number of peaks which are not in a isotope cluster although it is to be expected based on different criteria. We compute each performance measure as a function of the relaxed signal-to-noise threshold `snthr'` $\in \{100, 95, ..., 5\}\% * \text{snthr}$, where $\text{snthr} = 25$ is the signal-to-noise threshold used in the traditional peak picking step.

In Figure 2 we show the performance of the traditional peak picking in combination with targeted peak picking with isotope ROIs as well as traditional peak picking in combination with targeted peak picking with noise ROIs for varying signal-to-noise threshold `snthr'`. In addition, we show the performance of traditional peak picking with varying signal-to-noise threshold `snthr`. In case of predicted isotope ROIs, all five measures increase with decreasing `snthr'`. The isotope coverage appears to saturate for a relaxed signal-to-noise threshold `snthr'` of approximately 6.25. For this threshold, we find in case of predicted isotope ROIs an average increase of approximately +10% peaks, +37.6% isotope peaks, +33.5% isotope clusters, +25.2% isotope coverage, and +102.8% PPS in contrast to noise ROIs, suggesting an isotope-specific improvement of peak picking. More specifically, 20 isotope clusters could be extended and 37 isotope clusters could be newly detected. In addition, we find that the PPS decreases for a relaxed signal-to-noise threshold `snthr'` lower than 5. This finding confirms the general observation that peak picking with a too low signal-to-noise threshold results in unreliable peaks and is therefore not advisable. We also tested the performance of traditional peak picking with varying signal-to-noise threshold `snthr` and find that the number of peaks more

Metabolites **2016**, *6*, 37

than doubles. However, the proportion of low–intensity peaks which are not part of isotope clusters increases disproportionately and there is no specificity for isotope peaks.

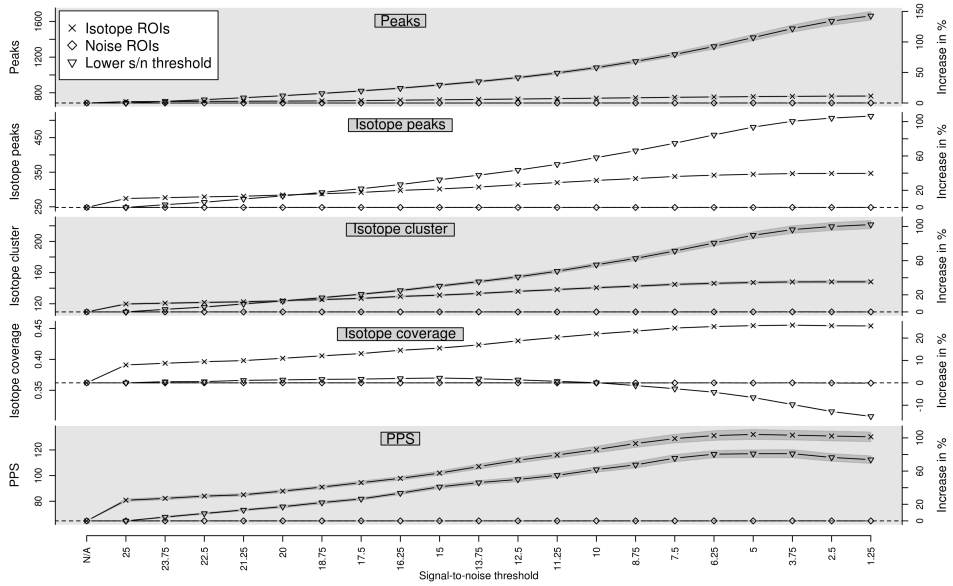


Figure 2. Evaluation of predicted isotope ROIs for varying relaxed signal-to-noise threshold `snthr'`. We show the mean (solid line) and the standard error of the mean (SEM, interval in dark grey) of the performance measures (**i**) number of detected peaks; (**ii**) number of detected isotope peaks; (**iii**) number of detected isotope clusters; (**iv**) isotope coverage; and (**v**) Peak Picking Score (PPS). In case of isotope ROIs and noise ROIs, we plot the performance of each measure without additional ROIs in the first column ("N/A") as reference value (horizontal dashed line) and in the subsequent columns with additional ROIs for decreasing relaxed signal-to-noise threshold `snthr'`. In case of "Lower S/N threshold", we plot the performance of each measure for decreasing signal-to-noise threshold `snthr` without additional ROIs. All four measures increase for predicted isotope ROIs with decreasing signal-to-noise threshold `snthr'` in contrast to noise ROIs.

2.3. Isotope Cluster Detection and Validation

There is a multitude of isotope detection routines for the recognition of isotope clusters. These detect coeluting features which are separated by certain distances in the m/z dimension and group these features to isotope clusters. However, a validation of detected isotope clusters is typically based on simple *ad hoc* rules. There are at least four cases for which the validation of isotope clusters can be beneficial as shown in Figure 3.

First, valid isotope clusters can be verified which strengthens the trust in the data; Second, multiple coeluting substances with mass differences of a few dalton can result in isobaric ion species and thus in overlapping isotope clusters [26]. These are potentially misinterpreted as a single isotope cluster affecting downstream analyses. This necessitates the deconvolution of the overlapping isotope cluster into at least two valid isotope clusters; Third, substances can be affected by hydrogen loss as reported in [27] and exploited in [28]. This leads to mass differences similar to isotope peaks ($\text{mass}(^1\text{H}) = 1.008 \approx 1.0034 = \text{mass}(^{13}\text{C}) - \text{mass}(^{12}\text{C})$) and results in a small trailing peak which is potentially misinterpreted as monoisotopic peak of the putative isotope cluster. This may result in the assumption of a wrong monoisotopic mass and may even lead to the rejection of the entire isotope cluster on the basis of failed intensity-checks [24]. Although this small trailing peak corresponds to

Metabolites **2016**, *6*, 37

the same substance, it needs to be removed from the isotope cluster in order to allow more precise molecular formula predictions. Fourth, the intensity of small peaks is systematically underestimated by some mass spectrometers which leads to distorted ratios between different isotope peaks as reported previously [3]. This intensity bias would lead to distorted molecular formula predictions and the removal of these underestimated peaks from the isotope cluster allows more precise molecular formula predictions.

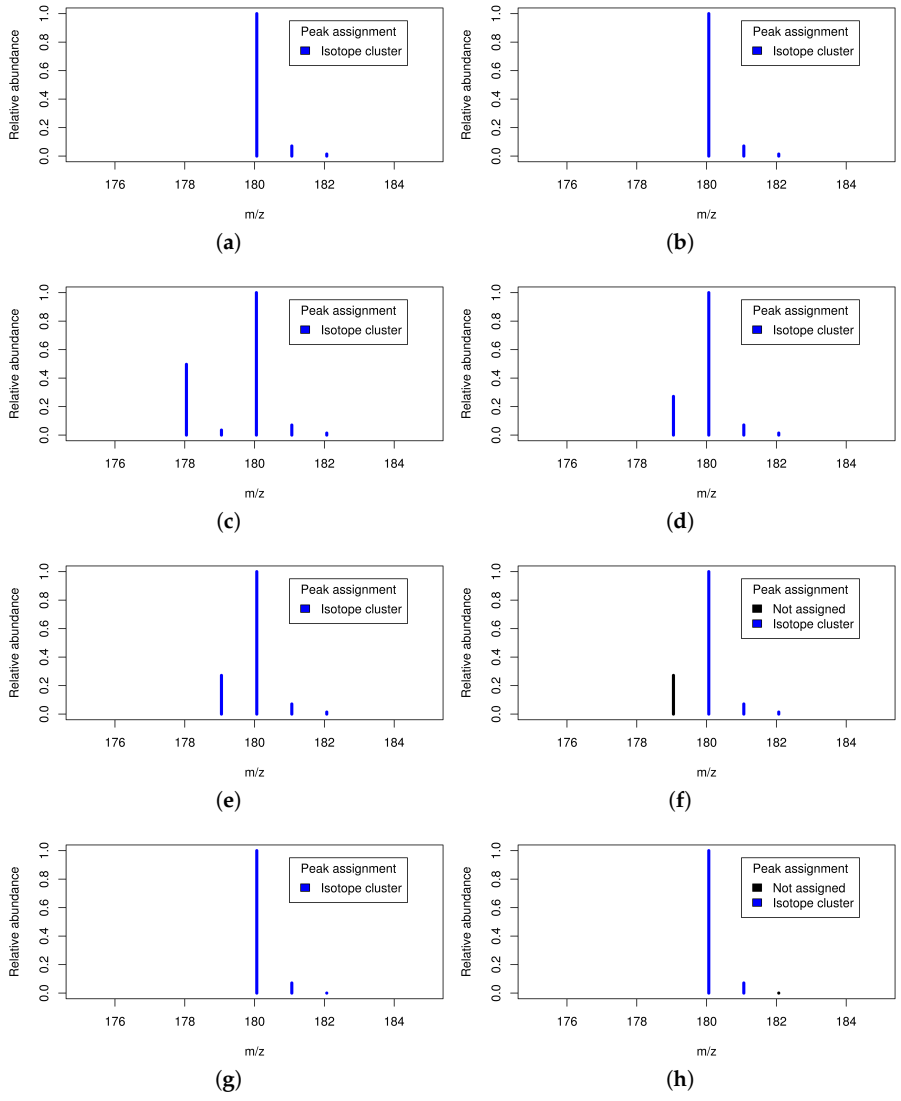


Figure 3. Four cases necessitating the validation of putative isotope clusters. Figure 3**a,b**: Valid isotope cluster without and with isotope cluster validation; Figure 3**c,d**: Two overlapping isotope clusters without and with isotope cluster validation; Figure 3**e,f**: Hydrogen loss without and with isotope cluster validation; Figure 3**g,h**: Underestimated small peak without and with isotope cluster validation.

We compare the proposed isotope detection routine with mass–specific isotope cluster validation (IDR_{NewVal}) with the isotope detection routine without isotope cluster validation ($IDR_{NewNoVal}$), the isotope detection routine implemented in the *AStream* package ($IDR_{AStream}$) [29], the isotope detection routine implemented in the *CAMERA* package (IDR_{CAMERA}) [24], and the isotope detection routine implemented in the *mzMatch* package ($IDR_{mzMatch}$) [30]. The isotope detection routines from *AStream*, *CAMERA*, and *mzMatch* apply different requirements for the validation of isotope clusters. In $IDR_{AStream}$ it is required that the abundance of the monoisotopic peak, the first isotope peak, and the second isotope peak decreases strictly, which corresponds to a ratio <1 between consecutive isotope peaks. In IDR_{CAMERA} it is required that the ratio of the monoisotopic peak to the first isotopic peak is within an interval which is given by the ratios of the monoisotopic peak to the first isotopic peak of a substance consisting exactly one carbon atom and a substance consisting exactly $mass_{mono}/mass(^{12}C)$ carbon atoms, where $mass_{mono}$ is the assumed monoisotopic mass of the substance. In $IDR_{mzMatch}$ it is required that isotope peaks show a high correlation regarding coelution.

We evaluate the performance of the isotope cluster detection and validation described in Section 4.2 on a dilution series experiment with 40 LC-MS measurements described in Section 4.4. We quantify the performance using the performance measures (i) number of detected peaks; (ii) number of detected isotope peaks; (iii) number of detected isotope clusters; and (iv) isotope coverage, i.e., the proportion of detected isotope peaks versus all detected peaks. We compute each performance measure without predicted isotope ROIs as well as with predicted isotope ROIs for a relaxed signal-to-noise threshold `snthr'` of 6.25. We present the results with predicted isotope ROIs relative to the results without predicted isotope ROIs in Figure 4. These results are a subset of the results in Figure A1 in the Appendix A where we present the results for varying relaxed signal-to-noise threshold `snthr'`. We relate the results to the quality of the predicted molecular formulas presented in the Appendix B on a gold standard of 11 data sets with known content.

In Figure 4 we show the performance measures for IDR_{NewVal}, $IDR_{NewNoVal}$, $IDR_{AStream}$, IDR_{CAMERA}, and $IDR_{mzMatch}$. We find that all four measures increase with predicted isotope ROIs in case of all isotope detection routines. $IDR_{NewNoVal}$ detects the most isotopes which reflects the fact that there are no constraints regarding the shape of the isotope cluster. This indicates that a certain proportion of the detected isotope clusters might be invalid. We point out, that this highly sensitive algorithm can be useful in case of substances containing uncommon elements such as Cl, Br, Se, or B as scrutinized in [31]. $IDR_{mzMatch}$ detects by far the lowest number of isotopes which reflects that this algorithm requires a high degree of correlation between isotope peaks resulting in a high specificity at the cost of sensitivity. $IDR_{NewNoVal}$ and $IDR_{mzMatch}$ show the lowest number of correctly predicted molecular formulas as shown in Appendix B. We find comparable results for $IDR_{AStream}$, IDR_{CAMERA}, and IDR_{NewVal}. Also the numbers of correctly predicted molecular formulas are similar as shown in Appendix B. Interestingly, IDR_{NewVal} showed the highest number of correctly predicted molecular formulas and was also able to rank the highest number of correct molecular formulas to the first three ranks. Remarkably, in case of 85% to 92% of all tested ions the detected isotope clusters from all isotope detection routines with or without predicted isotope ROIs were sufficient for the prediction of the correct molecular formula to the first three ranks. This finding states, that the prediction of molecular formulas from isotope clusters works well in general and hence it is challenging to improve upon.

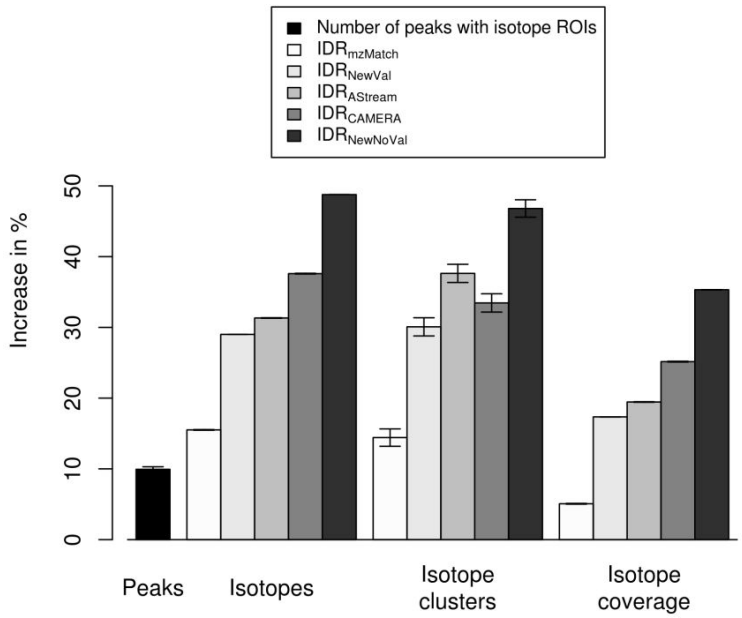


Figure 4. Evaluation of predicted isotope ROIs in combination with different isotope detection routines for a relaxed signal-to-noise threshold `snthr` of 6.25. We plot the increase of the mean and the standard error of the mean (SEM, error bars) of the performance measures (**i**) number of detected peaks; (**ii**) number of detected isotope peaks; (**iii**) number of detected isotope clusters; and (**iv**) isotope coverage relative to the performance of the *CAMERA* isotope detection routine without predicted isotope ROIs. All four measures increase with predicted isotope ROIs.

2.4. Isotope Cluster Statistics

We examine the compounds of the publicly available databases ChEBI [32], KEGG [33], KNApSAcK [34], LIPID MAPS [35], and PubChem [36] in order to compute mass–specific confidence intervals for the abundance–ratio of the monoisotopic peak to the first to fifth isotope peak as described in Section 4.3. For each database and each isotope peak, we compute multiple quantiles in order to define confidence intervals with different confidence levels. We validate isotope clusters on basis of mass–specific confidence intervals of peak abundance–ratios as described in Section 4.2.

We exemplarily examine the interval size and magnitude of the computed confidence intervals of isotope ratios. A small interval size indicates a small range of observed isotope ratios for the analyzed substances and allows a precise definition of valid isotope ratios, whereas a large interval size indicates a diverse range of observed isotope ratios for the analyzed substances and requires a loose definition of valid isotope ratios. If the interval size and magnitude of the computed confidence intervals depends on the mass range, then mass–specific confidence intervals can increase the specificity of isotope cluster validation.

See Figure 5 for the 95% confidence interval of the ratios of the monoisotopic peak to the first; second, and third isotope peak for the database KEGG with a mass window size of 50 dalton. The ratio of the monoisotopic peak to the first isotope peak depends on the abundance of the first isotope peak, which is dominated by the proportion of ^{13}C. This results in a relatively narrow confidence interval, because the variation of the number of carbon atoms is limited within a 50 dalton mass window. The ratio of the monoisotopic peak to the second isotope peak depends on the abundance of the second isotope peak, which is dominated by the proportion of ^{13}C and ^{34}S. The 97.5%-quantile and the 50%-quantile are higher compared to the case of the first isotope peak because the second isotope

peak has typically a lower abundance than the first isotope peak. In contrast, the 2.5%-quantile is smaller compared to the case of the first isotope peak because a subset of compounds comprises at least one sulfur (partially also chlorine or bromine) with a high abundance of ^{34}S (or ^{37}Cl, ^{81}Br) causing a relatively high abundance of the second isotope peak and thus a small ratio of the monoisotopic peak to the second isotope peak. This results in a relatively large confidence interval. The ratio of the monoisotopic peak to the third isotope peak mainly depends on the abundance of the third isotope peak, which is dominated by the proportion of ^{13}C and ^{34}S (and ^{37}Cl, ^{81}Br). This results in a relatively large confidence interval analogous to the case of the second isotope peak. The quantiles are higher compared to the case of the second isotope peak because the third isotope peak has typically a lower abundance compared to the second isotope peak. We find that the magnitude of the quantiles substantially depends on the mass of the substances. Specifically, the quantiles are typically inversely proportional to the substance mass. For example, in case of the mass interval 200 to 250 dalton versus the mass interval 800 to 850 dalton the 50%-quantiles deviate by a factor of 3.5 in case of the ratio of the monoisotopic peak to the first isotope peak, by a factor of 8.4 in case of the ratio of the monoisotopic peak to the second isotope peak, and by a factor of 25.6 in case of the ratio of the monoisotopic peak to the third isotope peak. This finding suggests that mass–specific confidence intervals can indeed increase the specificity of isotope cluster validation. See Figure C1 in Appendix C for an overview of all computed quantiles and the resulting symmetric confidence intervals of the ratio of the monoisotopic peak to the first isotope peak for the database PubChem with a mass window size of 50 dalton.

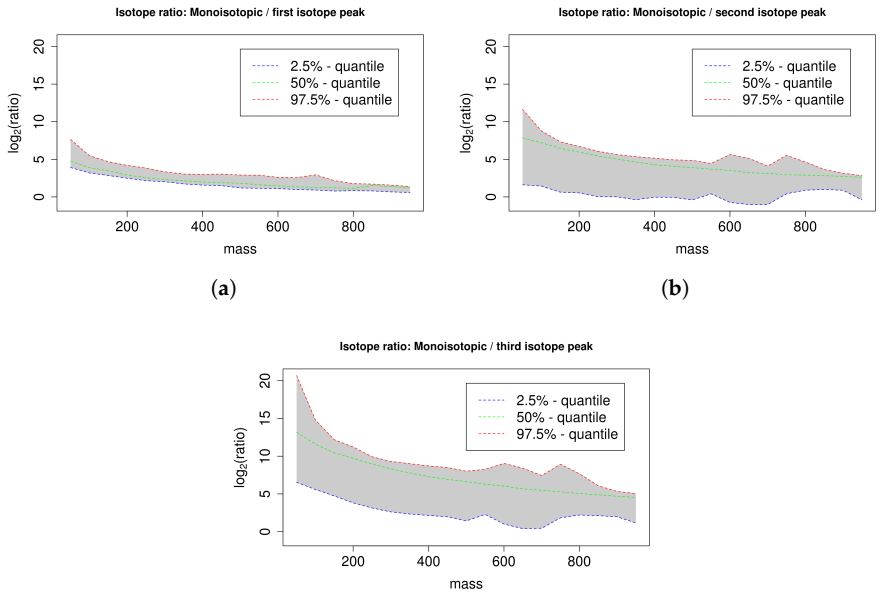


Figure 5. 95% confidence interval of the ratio of the monoisotopic peak to the first (**a**), second (**b**), and third isotopic peak (**c**) of all compounds in KEGG for different compound masses arranged in mass windows of size 50 dalton. We plot the 50%-quantile in green, the 2.5%-quantile in blue, and the 97.5%-quantile in red and we emphasize the enclosed 95% confidence interval in grey. The ratios decrease with increasing compound mass reflecting the increasing proportion of isotopic atoms.

2.5. Exemplary Isotope Cluster Detection

We exemplify the detection of isotope clusters for selected substances to demonstrate the proposed isotope detection routine without isotope cluster validation $IDR_{NewNoVal}$ and the isotope detection routine with mass–specific isotope cluster validation IDR_{NewVal}. We simulate the mass and relative intensity of the monoisotopic peak and the first five isotope peaks of six substances with enviPat [37] in centroid mode with a resolution of 10,000, namely (i) aspartic acid which has a low mass and comprises only the elements CHNO (see Table 1 for details); (ii) cysteine which has a low mass and comprises sulfur; (iii) chloramphenicol which has a low mass and comprises chlorine; (iv) digoxigenin monodigitoxoside which has a medium mass and comprises only the elements CHNO; (v) 2-Chloro-2′-deoxyadenosine-5′-triphosphate which has a medium mass and comprises chlorine; and (vi) autoinducer-2 which has a low mass and contains boron. The isotopic fine structure of these substances is not detectable at this resolution and hence each simulated peak is a mixture of multiple peaks from the isotopic fine structure. We only include isotope peaks with an abundance of at least 0.01% of the abundance of the monoisotopic peak which results in isotope clusters of size 4, 5, 6, 6, 6, and 6 respectively.

For each isotope cluster, we calculate the minimal absolute mass error Δm^{abs} in units of dalton and the minimal relative mass error Δm^{ppm} in units of PPM which are required for a successful isotope cluster detection. The incorporation of a mass error is necessary because the mass differences between individual isotope peaks depend on the elemental composition and hence deviates from the default mass difference of ^{13}C isotopes. It is possible to use only one of both parameters or a combination of both parameters to enable the detection of isotope clusters (see Equation (2) in Section 4.2).

We merge all six isotope clusters resulting in a single synthetic spectrum comprising 33 peaks. We apply the isotope detection routines $IDR_{NewNoVal}$ and IDR_{NewVal} as described in Section 4.2 to the synthetic spectrum. We evaluate whether the isotope detection routines are able to assemble the original isotope clusters.

In Table 1 we show the results. We find that $IDR_{NewNoVal}$ is able to detect all six isotope clusters provided that a sufficiently large mass error is set (e.g., $\Delta m^{abs} = 0.01$). In case of a smaller mass error (e.g., $\Delta m^{abs} = 0.005$) we find that isotope clusters become split at isotope peaks which are dominated by the isotopes of sulfur, chlorine, or boron, i.e., the second isotope peak of substance (ii); the second and fourth isotope peak of substance (iii); the second isotope peak of substance (v); and the first isotope peak of substance (vi). We find that $IDR_{NewVal}$is able to validate all but one isotope cluster. The first peak of the boron-containing substance (vi) is not included in the isotope cluster, because the abundance of this peak is too small relative to the space of biological substances of this mass. Hence, the excluded peak is assumed to be a potential hydrogen-loss. However, this isotope cluster can be correctly identified without validation or with specialized approaches [31].

Table 1. Isotope cluster detection exemplified for six substances. We show the substance name, the sum formula, the mass of the monoisotopic peak and the first five isotope peaks (rounded to five digits), the mass difference to the monoisotopic peak (Δm, rounded to five digits), the relative peak intensity (Int., normalized to 100 and rounded to two digits), the absolute m/z error Δm^{abs} and the relative m/z error in ppm Δm^{ppm} for a successful isotope cluster detection (Δm^{abs} is rounded to five digits and Δm^{ppm} is rounded to one digit), whether the isotope cluster assignment using the isotope detection routine without isotope cluster validation $IDR_{NewNoVal}$ is successful or not (No val., "+"/"−"), and whether the isotope cluster assignment using the isotope detection routine with mass–specific isotope cluster validation IDR_{NewVal} is successful or not (Val., "+"/"−"). $IDR_{NewNoVal}$ is able to detect the isotope clusters of all substances and IDR_{NewVal} successfully validates the isotope clusters of all but one substance.

Substance Name	Sum Formula	Mass	Δm	Int.	Δm^{abs}	Δm^{ppm}	No Val.	Val.
Aspartic acid	$C_4H_7NO_4$	133.037508		100.00	0.00191	14.3	+	+
		134.040468	1.00296	4.96			+	+
		135.041918	2.00441	0.93			+	+
		136.044728	3.00722	0.04			+	+
Cysteine	$C_3H_7NO_2S$	121.019749		100.00	0.00895	73.9	+	+
		122.021976	1.00223	4.59			+	+
		123.016385	1.99664	5.05			+	+
		124.019165	2.99942	0.19			+	+
		125.018404	3.99866	0.03			+	+
Chloramphenicol	$C_{11}H_{12}Cl_2N_2O_5$	322.012327		100.00	0.00913	28.4	+	+
		323.015369	1.00304	13.00			+	+
		324.009595	1.99727	66.20			+	+
		325.012562	3.00024	8.53			+	+
		326.007250	3.99492	11.54			+	+
		327.010016	4.99769	1.45			+	+
Digoxigenin monodigitoxoside	$C_{29}H_{44}O_8$	520.303618		100.00	0.00078	1.5	+	+
		521.307027	1.00341	32.24			+	+
		522.309803	2.00619	6.70			+	+
		523.312531	3.00891	1.04			+	+
		524.315166	4.01155	0.13			+	+
		525.317742	5.01412	0.01			+	+
2-Chloro-2′-deoxyadenosine-5′-triphosphate	$C_{10}H_{15}ClN_5O_{12}P_3$	524.961858		100.00	0.00817	15.6	+	+
		525.964411	1.00255	13.30			+	+
		526.959596	1.99774	35.41			+	+
		527.962023	3.00017	4.63			+	+
		528.963673	4.00182	1.11			+	+
		529.966017	5.00416	0.12			+	+
Autoinducer-2	$C_5H_{10}BO_7$	192.055590		24.37	0.00689	35.9	+	−
		193.052059	0.99647	100.00			+	+
		194.055706	2.00012	6.13			+	+
		195.056530	3.00094	1.59			+	+
		196.059851	4.00426	0.09			+	+
		197.060963	5.00537	0.01			+	+

3. Discussion

Aiming at the exhaustive detection and precise validation of isotope clusters we propose an additional targeted peak picking step with predicted isotope ROIs and the mass–specific validation of putative isotope clusters based on database statistics. Compromising between peak reliability and

exhaustive detection we use a relaxed signal-to-noise of 6.25 threshold for predicted isotope ROIs and achieve an increase of +37.6% isotope peaks and +102.8% PPS. We use this relaxed signal-to-noise threshold by default in the freely available implementation of this algorithms in the R package xcms. The targeted peak picking with predicted isotope ROIs can easily be adapted in other tools such as *MZmine2* [38], *apLCMS* [39], and related approaches [40]. The validation of putative isotope clusters in combination with predicted isotope ROIs results in the highest number of correctly predicted molecular formulas and also the highest number of correct molecular formulas among the first three ranks. However, the ranks of correctly predicted molecular formulas were robust with respect to different approaches for peak picking and isotope cluster detection and it is challenging to improve upon. We exemplify the use of the proposed isotope detection routine with and without mass–specific isotope cluster validation and find that it is possible to detect substances with and without biologically unusual elements using an absolute mass error of 0.01 dalton. Consequently, we use this absolute mass error by default in the freely available implementation of these algorithms in the R package *CAMERA*.

The enhanced isotope cluster detection and validation presented in this work could improve the accuracy of substance quantification. All isotope peaks of one isotope cluster originate from the same substance and we point out that the consideration of a greater number of features from a certain substance—although small and noisy—reduces the technical variance in the data. In turn, this would enhance the precision and yield of comparative analyses, because a reduced data variance would not only improve calculated fold changes but would enable the statistically valid detection of smaller effect sizes. The slight improvement in molecular formula prediction could affect a considerable number of substances in case of metabolome-scale metabolite identification studies. Especially in untargeted metabolomics reliable hints for metabolite identification are urgently needed.

4. Materials and Methods

We present the methodology of the proposed approach and the used data for evaluation. Specifically, we describe (i) the targeted peak picking with predicted isotope ROIs; (ii) the detection and mass–specific validation of isotope clusters; (iii) the computation of isotope ratio quantiles; and (iv) two sets of mass spectrometry raw data.

4.1. Targeted Peak Picking with Predicted Isotope ROIs

A requirement for the prediction of isotope ROIs is a set of peaks that have been detected previously. This initial peak picking can be accomplished by one of the numerous peak picker which are available [1,18,38]. In untargeted approaches, these peak picker typically do not use any prior knowledge and we refer to this kind of peak picking as *traditional peak picking*. We propose the following approach for the targeted detection of isotope peaks. This approach is designed for liquid chromatography–high resolution mass spectrometry data and does not consider the isotopic fine structure available with ultrahigh resolution mass spectrometry.

Given a set of detected peaks from traditional peak picking, a maximum charge $Z = 3$, and a maximum number of isotopes $I = 5$ we predict putative isotope ROIs as follows. For each charge state $z \in \{1, ..., Z\}$ and for each isotope number $i \in \{1, ..., I\}$, we compute the theoretical m/z distance to the monoisotopic peak

$$d_{z,i} = \frac{i * \Delta m}{z}, \quad (1)$$

where Δm = mass(^{13}C) – mass(^{12}C) $\approx$ 1.003355. We use Δm as an approximation for the mass difference between successive peaks in isotope clusters because the isotopic nuclide ^{13}C has usually the largest impact on isotope clusters in biological samples. Other isotopic nuclides such as ^{15}N, ^{18}O, and ^{34}S cause isotope peaks with mass differences which can only be discriminated from ^{13}C-isotope peaks using mass spectrometers with resolution above 40,000 (in case of ions with an m/z of 500 dalton). For each peak detected by traditional peak picking we predict for each charge state z and for each isotope number i one putative isotope ROI. Each putative isotope ROI is composed of the retention time

Metabolites **2016**, *6*, 37

interval of the detected peak and the m/z interval of the detected peak shifted by $d_{z,i}$ as exemplified in Figure 6. An additional targeted peak picking is performed based on the set of predicted isotope ROIs using a relaxed signal-to-noise threshold $\texttt{snthr'} = \texttt{snthr} * r/100$, where `snthr` is the signal-to-noise threshold for traditional peak picking and $r \in \{100, 95, ..., 5\}$. Subsequently, the peak table from traditional peak picking and the peak table from the targeted peak picking on basis of putative isotope ROIs are merged and redundant peaks are removed.

For control purposes, we generate a set of noise ROIs given the set of predicted isotope ROIs as follows. To approximate the distribution of the predicted isotope ROIs in the m/z dimension and the retention time (RT) dimension, we calculate the minimum and maximum m/z and RT of the predicted isotope ROIs and use a uniform distribution in the calculated intervals of both dimensions. To approximate the distribution of peak widths in m/z and RT we calculate a histogram of peak widths in m/z relative to the peak m/z and a histogram of peak widths in RT. For each predicted isotope ROI we sample one new noise ROI which m/z and RT is uniformly drawn within the calculated ranges in m/z and RT and which peak width in m/z and RT is drawn from the calculated histograms. Subsequently, targeted peak picking is applied to the set of noise ROIs using a relaxed signal-to-noise threshold `snthr'` analog to predicted isotope ROIs and the results from traditional peak picking and targeted peak picking on basis of noise ROIs are merged as before.

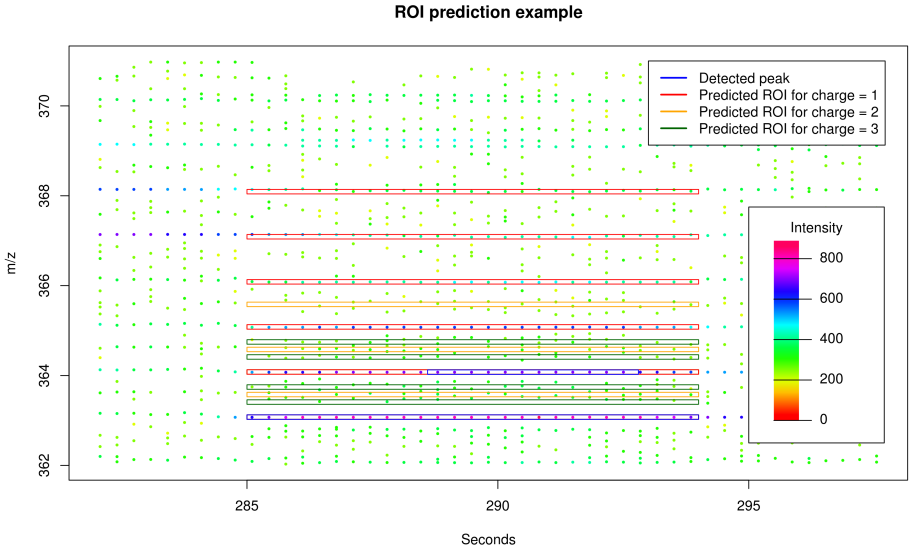


Figure 6. Exemplary section of LC-MS raw data. We mark two detected peaks from traditional peak picking in blue and 12 predicted isotope ROIs in red, orange, and green calculated on basis of the (monoisotopic) peak (apex $m/z \approx 363.075$ dalton / retention time ≈ 291 seconds) given a maximum isotope number $I = 5$ and a maximum charge state $Z = 3$. Via prediction of isotope ROIs, we are able to expand the region of the already detected first isotope peak and to encompass the signals of the second, third, fourth, and fifth isotope peak. Here, the subsequent peak picking procedure will not find relevant signals for the predicted isotope ROIs corresponding to the charge states 2 (orange) and 3 (green) and will reject these accordingly.

4.2. Detection and Mass–Specific Validation of Isotope Clusters

We propose an approach for the detection and validation of isotope clusters in liquid chromatography–high resolution mass spectrometry data which does not resolve the isotopic fine

structure. In this approach we detect putative isotope clusters based on characteristic distances in the m/z dimension. We validate putative isotope clusters depending on the substance mass and we refer to this validation as *mass–specific validation*. We detect and validate isotope clusters given a set of coeluting features, a maximum charge $Z = 3$, a relative m/z error in ppm Δm^{ppm}, and an absolute m/z error Δm^{abs} as follows.

First, we detect putative isotope clusters. For each charge state $z \in [1, Z]$, we mark all pairs of peaks (p_1, p_2) for which

$$\delta_{z,p_1,p_2} = ||mass(p_1) - mass(p_2)| - \Delta m/z| \leq \max\left(\frac{mass(p_1) * \Delta m^{ppm}}{10^6}, \Delta m^{abs}\right) \quad (2)$$

holds, where $\Delta m = \text{mass}(^{13}C) - \text{mass}(^{12}C) \approx 1.003355$ is the expected distance between two isotope peaks (cf. Section 4.1). For each charge state and for each peak p, we compute all putative isotope clusters $(p_1, p_2, ..., p_n)$ for which $\delta_{c,p',p''}$ holds for each successive pair of peaks (p', p''). We retain the putative isotope cluster with the maximum number of peaks and remove the peaks of this putative isotope cluster from the set of available peaks. We iteratively perform the last steps with the remaining peaks until there are no putative isotope clusters with at least two peaks left.

Second, we validate the set of putative isotope clusters which have been extracted previously depending on the monoisotopic mass. See Figure 3 for four cases which necessitate the following validation of putative isotope clusters. For each putative isotope cluster $(p_1, p_2, ..., p_n)$ we examine the second to last peak $p' \in (p_2, ..., p_n)$. For each peak p' we compute the ratio of the abundance of the monoisotopic peak p_1 and the abundance of peak p'. Specifically, we compute the minimum and maximum ratio considering that the abundance estimates of both peaks are affected by the ubiquitous noise using an estimate of the signal-to-noise ratio of both peaks. If the computed interval of ratios does not overlap with the 99% confidence interval derived from the KEGG database for the current monoisotopic mass (mass window size 50) we split the putative isotope cluster. In this case we turn the peak p' into the new monoisotopic peak resulting in a new putative isotope cluster $(p', ..., p_n)$ which is validated as well. We retain all putative isotope clusters which comprise at least two peaks and consider these as validated isotope clusters.

4.3. Isotope Ratio Quantiles

We perform isotope statistics for each of the databases ChEBI, KEGG, KNApSAcK, LIPID MAPS, and PubChem as follows [32–36]. We iterate all compounds, compute the exact mass and the theoretical isotope cluster from the molecular formula, and record the ratio of the monoisotopic peak to the first to fifth isotope peak. We group all compounds by the exact mass in consecutive mass windows for each of the mass window sizes 10, 25, 50, 100, and 250 dalton to support different compromises between mass specificity and quantile robustness. For each mass window size, each mass window, and each isotope peak (1st–5th) we compute the isotope ratio for several p-quantiles, where $p \in \{5.0 \times 10^{-6}$, 0.999995, 1.0×10^{-5}, 0.99999, 5.0×10^{-5}, 0.99995, 1.0×10^{-4}, 0.9999, 5.0×10^{-4}, 0.9995, 0.001, 0.999, 0.005, 0.995, 0.01, 0.99, 0.025, 0.975, 0.05, 0.95, 0.1, 0.9, 0.5$\}$. For each mass window size and each isotope peak we record the isotope ratio in a matrix with one row for each p-quantile and one column for each mass window. We encapsulate the resulting data for each database, each mass window size, and each isotope peak in an R object of class *S4* named `compoundQuantiles`. This implementation supports a simple API for convenient retrieval of the data (see documentation of package *CAMERA* version 1.50.0 for details). Based on this implementation, it is also possible to compute isotope ratios amongst isotope peaks, e.g., the confidence interval of the isotope ratio between the third isotope peak and the fifth isotope peak for a given mass range.

Metabolites **2016**, *6*, 37

4.4. Data Sets

4.4.1. MM48

We perform a case study based on a gold standard data set comprising 11 LC-MS measurements (UPLC-ESI-QTOF-MS, positive mode) each of a solution of 48 known reference substances denoted as *MM48*. The raw data is available in MetaboLights [41] accession MTBLS381 in Supplementary Materials link. This set of compounds was also used in [24] and the measurements have been deposited in MetaboLights accession MTBLS188. We compile a ground truth of detectable ions as follows. First, we assume a set of three expected ions ($[M]^+$, $[M + H]^+$, $[M + Na]^+$) as well as isotope peaks up to the fifth isotope peak (i.e., $[M + 1]^+$, $[M + 2]^+$, $[M + 3]^+$, $[M + 4]^+$, and $[M + 5]^+$ in case of the $[M]^+$ ion) for each compound and calculate the exact mass of these 18 molecular formulas (three ions each with an isotope cluster with six peaks); Second, we check the abundance of these ions in the 11 data sets and define all ions with a peak area of at least 1000 counts within a retention time interval of at most five seconds as measurable ions constituting the ground truth. Considering the set of ions which are measurable in at least six of 11 data sets, we detect 72 monoisotopic ions (see Figure 7), 63 isotope clusters with at least two ions, and 190 ions in total.

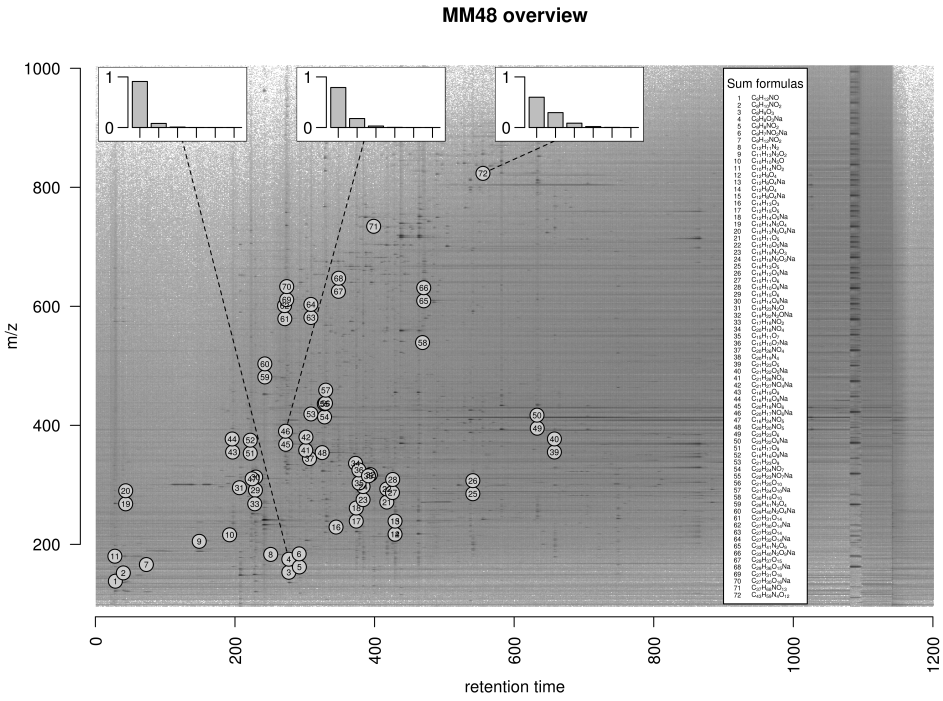


Figure 7. Overview of monoisotopic measurable ions in the MM48 data set. We plot the logarithmic raw data intensities in the dimensions mass-to-charge ratio (m/z) and retention time and mark the location of 72 monoisotopic ions which are measurable in at least six of eleven data sets. In case of three ions with exact mass 175.037, 390.095, and 823.413 dalton, we exemplarily plot the theoretical relative intensities of the monoisotopic peak and the first to fifth isotope peak in the insets at the top. The set of measurable ions spans a huge range in both dimensions with different isotope clusters constituting a diverse basis for validation purposes.

4.4.2. Dilution Series

We perform a case study based on 40 LC-MS measurements (UPLC-ESI-QTOF-MS, positive mode), which is a subset of the data used in [24] and is available from the MetaboLights repository with accession MTBLS188. This set of measurements is composed of a dilution series varying the ratio of solution and leaf sample. Specifically, the ratio of solution and leaf sample is 0:100, 25:75, 50:50, and 75:25 in 10 data sets each. This experimental design implies a diverse range of cases in the data regarding the signal-to-noise ratio of peaks and constitutes the basis to test the detection of weak signals like isotope peaks.

5. Conclusions

We implemented the targeted peak picking with predicted isotope ROIs in combination with the *centWave* algorithm as part of the R package *xcms* in version 1.50.0 (functions `findPeaks.centWaveWithPredictedIsotopeROIs` and `findPeaks.addPredictedIsotopeFeatures`). We implemented the mass–specific validation of putative isotope clusters as part of the R package *CAMERA* in version 1.30.0 (function `findIsotopesWithValidation`).

Supplementary Materials: The following are available online at www.ebi.ac.uk/metabolights/MTBLS381, 11 MM48 raw data files used for performance evaluation in the manuscript.

Acknowledgments: The open access fee was funded by the Deutsche Forschungsgemeinschaft (DFG funding No. NE 1396/5-1). The authors would like to thank Carsten Kuhl and Christoph Böttcher for providing the UPLC-ESI-QTOF-MS data and Sarah Scharfenberg for refining the manuscript.

Author Contributions: Hendrik Treutler, and Steffen Neumann conceived and designed the methodology; Hendrik Treutler performed the case studies; Hendrik Treutler wrote the paper. Hendrik Treutler, and Steffen Neumann read and approved the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Isotope Cluster Detection and Validation: Extended Results

We compare the proposed isotope detection routine with mass–specific isotope cluster validation (IDR_{NewVal}) against the isotope detection routine without isotope cluster validation ($IDR_{NewNoVal}$), the isotope detection routine implemented in the *AStream* package ($IDR_{AStream}$) [29], the isotope detection routine implemented in the *CAMERA* package (IDR_{CAMERA}) [24], and the isotope detection routine implemented in the *mzMatch* package ($IDR_{mzMatch}$) [30].

We evaluate the performance of the isotope cluster detection and validation described in Section 4.2 on a dilution series experiment with 40 LC-MS measurements described in Section 4.4. We quantify the performance using the performance measures (i) number of detected peaks; (ii) number of detected isotope peaks; (iii) number of detected isotope clusters; and (iv) isotope coverage, i.e., the ratio of the number of detected isotope peaks and the number of all detected peaks. We compute each performance measure as a function of the relaxed signal-to-noise threshold $\texttt{snthr'} \in \{100, 95, ..., 5\}\ \% * \texttt{snthr}$, where $\texttt{snthr} = 25$ is the signal-to-noise threshold of the traditional peak picking step. In the Section 2.3 we show an excerpt of these results, i.e., we present the results for each isotope detection routine with predicted isotope ROIs relative to the results of IDR_{CAMERA} without predicted isotope ROIs in Figure 4.

In Figure A1 we show the performance measures for IDR_{NewVal}, $IDR_{NewNoVal}$, $IDR_{AStream}$, IDR_{CAMERA}, and $IDR_{mzMatch}$.

Metabolites **2016**, *6*, 37

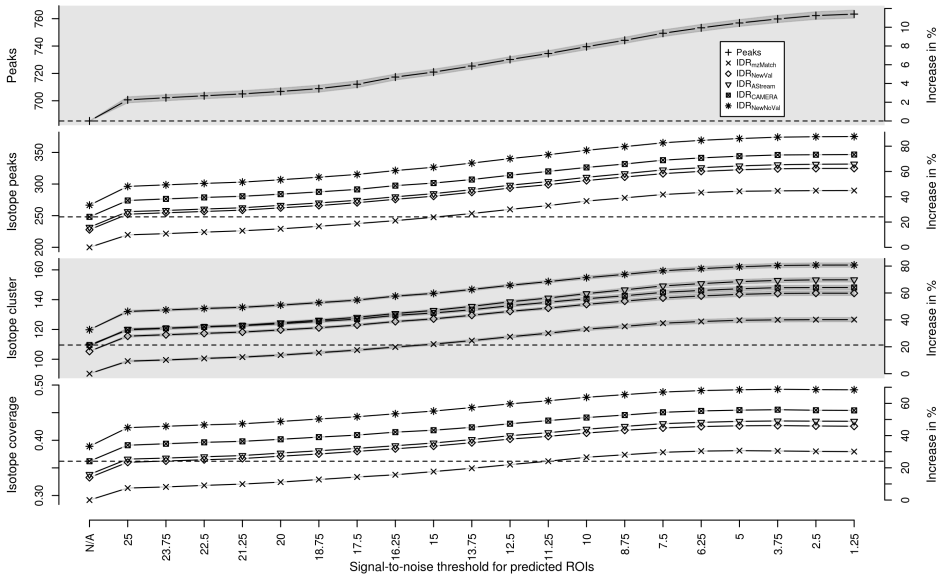


Figure A1. Evaluation of predicted isotope ROIs in combination with validated isotope clusters for varying relaxed signal-to-noise threshold `snthr'`. We plot the mean (solid line) and the standard error of the mean (SEM, interval in dark grey) of the performance measures (**i**) number of detected peaks; (**ii**) number of detected isotope peaks; (**iii**) number of detected isotope clusters; and (**iv**) isotope coverage. We plot the performance of each measure without additional ROIs in the first column ("N/A") as reference value (horizontal dashed line). All four measures of all isotope detection routines increase with decreasing signal-to-noise threshold `snthr'`.

Appendix B. Prediction of Molecular Formulas From Isotope Clusters

In order to study to which degree the proposed approach is capable of improving the detection and validation of isotope clusters, we test the quality of predicted molecular formulas. The prediction of molecular formulas is an important step towards the identification of substances and can be done automatically on the basis of isotope clusters. We use 11 LC-MS measurements with 48 known compounds and select a set of 72 ions. We predict for each ion a list of ranked molecular formula candidates using SIRIUS and evaluate the rank of the correct molecular formula [3].

We evaluate the performance of predicted isotope ROIs described in Section 4.1 and the isotope detection routine with mass–specific isotope cluster validation described in Section 4.2 on 11 LC-MS measurements of known compounds described in Section 4.4 using predicted molecular formulas from SIRIUS as described in the Appendix D.4. We quantify the performance using the number of compounds with a certain rank averaged over all measurements. If the proposed approaches increase the quality of detected isotope clusters, then the rank of the predicted molecular formulas should decrease and be ranked first in the ideal case. We compare different combinations of two peak picking approaches and five isotope detection routines, namely (iA) the traditional peak picking and (iB) the traditional peak picking in combination with targeted peak picking with predicted isotope ROIs (see Section 4.1) and (iiA) the isotope detection algorithm from *AStream*; (iiB) the isotope detection algorithm from *mzMatch*; (iiC) the isotope detection algorithm from *CAMERA*; (iiD) the proposed isotope detection algorithm without isotope cluster validation; and (iiE) the proposed isotope detection algorithm with mass–specific isotope cluster validation resulting in ten combinations of algorithms (see Section 4.2 and the Appendix D). In Table B1 we show the ranks of the predicted molecular formulas for ten algorithms averaged over 11 data sets.

Table B1. Molecular formula prediction from isotope clusters. Using SIRIUS we predict molecular formulas from isotope clusters which have been detected using different algorithms. In the first column we indicate whether we use targeted peak picking with predicted isotope ROIs ('+') or not ('−') and in the second column we indicate the isotope detection algorithm ($IDR_{AStream}$ for the algorithm implemented in R package *AStream*, IDR_{CAMERA} for the algorithm implemented in R package *CAMERA*, $IDR_{mzMatch}$ for the algorithm implemented in R package *mzMatch*, $IDR_{NewNoVal}$ for the proposed isotope detection algorithm without isotope cluster validation, and IDR_{NewVal} for the proposed isotope detection algorithm with mass–specific isotope cluster validation). We specify the number of ions with a molecular formula on rank 1, on rank 2, on rank 3, between rank 4 and rank 10, on a rank above 10, the number of ions which molecular formula is not among the top 1000 candidates ('No rank'), and the number of ions which have not been detected during peak picking ('No peak'). We arranged the isotope detection algorithms by the number of ions with molecular formula on rank 1.

Predicted Isotope ROIs	Isotope Detection Algorithm	Rank 1	Rank 2	Rank 3	3 < Rank $\leq$ 10	Rank > 10	No Rank	No Peak
−	$IDR_{mzMatch}$	48.82	11.55	1.18	3.36	0	4.64	2.45
+	$IDR_{mzMatch}$	48.18	12	1.18	3.36	0	4.82	2.45
−	$IDR_{NewNoVal}$	49.09	10.91	0.91	1.55	0	7.09	2.45
+	$IDR_{NewNoVal}$	49.36	11.18	0.73	1.64	0	6.73	2.36
−	$IDR_{AStream}$	52.82	11.27	1.09	1.82	0	2.55	2.45
+	$IDR_{AStream}$	53.27	11.55	0.55	1.91	0	2.36	2.36
−	IDR_{CAMERA}	53.73	10.27	0.82	1.55	0	3.18	2.45
+	IDR_{CAMERA}	52.82	11	0.64	1.64	0	3.55	2.36
−	IDR_{NewVal}	53.82	11.09	1	1.55	0	2.09	2.45
+	IDR_{NewVal}	54.09	11.36	0.73	1.64	0	1.82	2.36

Appendix C. Isotope Cluster Statistics: Full Quantile Set for PubChem

In Figure C1 we depict all computed quantiles and the resulting symmetric confidence intervals of the isotope ratio of the monoisotopic peak to the first isotope peak for the database PubChem with a mass window size equal to 50 dalton. See Section 4.3 for a detailed description of the database statistics.

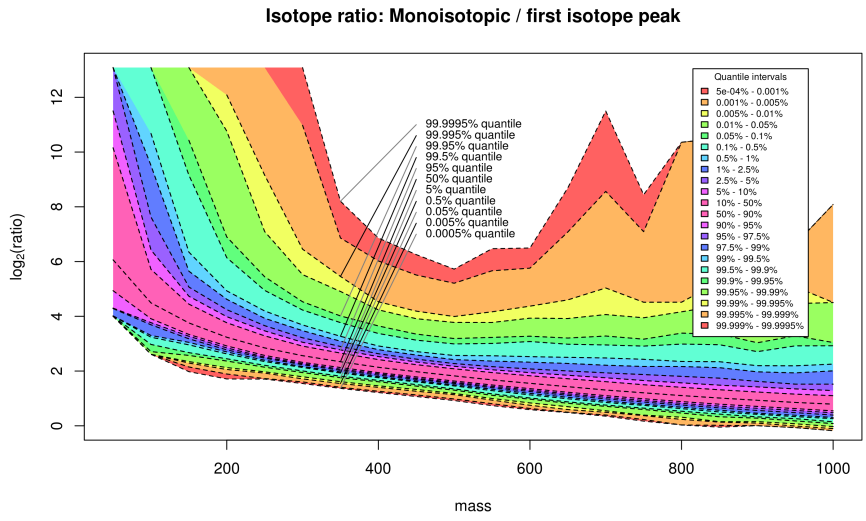


Figure C1. The full set of 23 quantiles of the monoisotopic peak versus the first isotopic peak for the PubChem database for different compound masses arranged in mass windows of size 50 dalton. We emphasize the enclosed confidence intervals with different colors.

Appendix D. Software Versions and Processing Parameters

Tools versions, used functions, and parameters of *xcms/CAMERA*, *AStream*, *mzMatch*, and *SIRIUS* are given subsequently.

Appendix D.1. xcms/CAMERA

We use the R package *xcms* version 1.44.0 [23] and the R package *CAMERA* version 1.27.0 [24] for peak picking using *centWave* [1], the grouping of features into pseudospectra, and the detection of isotope clusters. We processed the raw data of each LC-MS measurement individually as follows. We performed peak picking with the *centWave* algorithm with parameters `peakwidth` $= (5, 12)$, `prefilter` $= (2, 200)$, `ppm` $= 10$, and `snthr` $= 25$. We use a signal-to-noise ratio of 25, because it has been shown that this ratio yields reliable molecular formula predictions from mass spectrometry data [42]. Subsequently, we group detected peaks by retention time into pseudospectra-groups using function *groupFWHM* with `perfwhm` $= 1$ and standard parameters and detect isotope clusters using function *findIsotopes* with `intensityValue` $=$ 'intb' and standard parameters.

Appendix D.2. AStream

We use the R package *AStream* version 2.0 [29] for the detection of isotope clusters. We import the peaks which have been detected using *xcms* into the *AStream* datalist structure. We apply the function `data.norm` with the parameters `mz.tol` $= 0.005$ (the mean m/z error for `ppm` $= 10$ as used in *xcms* and *mzMatch*) and we detect isotope clusters using function `isotope.search` with the parameter `mz.tol` $= 0.005$. In a postprocessing step we remove contradictory isotope annotations, i.e., if (i) peak B is annotated as [M + 1] isotope peak of peak A and (ii) peak C is annotated as [M + 2] isotope peak of peak A and (iii) peak C is annotated as [M + 1] isotope peak of peak B; then we remove annotation (iii).

Appendix D.3. mzMatch

We use the R package *mzmatch.R* version 2.0-13 [30] for the detection of isotope clusters. We import the peaks which have been detected using *xcms* via the peakML file format used by *mzMatch* using the function `PeakML.xcms.write.SingleMeasurement` with the parameters `writeRejected` $=$ TRUE, `ppm` $= 10$, `addscans` $= 0$, and `ApodisationFilter` $=$ FALSE. We convert this data using function `mzmatch.ipeak.Combine` and we detect isotope clusters using function `mzmatch.ipeak.sort.RelatedPeaks` with the parameters `ppm` $= 10$ and `rtwindow` $= 50$. In a postprocessing step we remove all isotope clusters with gaps, i.e., the isotope cluster with monoisotopic peak [M] and isotope peak [M + 2] without the [M + 1] isotope peak is considered non-evaluable and removed from the output. Approximately 10% of the isotope annotations are removed in this way.

Appendix D.4. Prediction of Molecular Formulas Using SIRIUS

We predict ranked candidate lists from isotope clusters using command–line SIRIUS [3] version 3.1.3. We use the parameters `-elements` $= CHNOPS$, `-isotope` $=$ *score*, `-candidates` $= 1000$, `-ppm-max` $= 10$, and `-profile` $= qtof$ and give the ion species (`-ion`), the monoisotopic m/z (`-mz`), the (m/z, intensity) pairs (`into` intensity from *xcms*; `-ms1`), and an empty MS/MS spectrum (`-ms2`) as input. We rank the resulting candidate lists according to the `tree` score and select the rank of the correct molecular formula.

References

1. Tautenhahn, R.; Böttcher, C.; Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinform.* **2008**, *9*, 504, doi:10.1186/1471-2105-9-504.
2. Trutschel, D.; Schmidt, S.; Grosse, I.; Neumann, S. Joint Analysis of Dependent Features within Compound Spectra Can Improve Detection of Differential Features. *Front. Bioeng. Biotechnol.* **2015**, *3*, doi:10.3389/fbioe.2015.00129.

3. Böcker, S.; Letzel, M.C.; Lipták, Z.; Pervukhin, A. SIRIUS: Decomposing isotope patterns for metabolite identification. *Bioinformatics* **2009**, *25*, 218–224.
4. Dührkop, K.; Hufsky, F.; Böcker, S. Molecular Formula Identification Using Isotope Pattern Analysis and Calculation of Fragmentation Trees. *Mass Spectrom.* **2014**, *3*, doi:10.5702/massspectrometry.S0037
5. Stoll, N.; Schmidt, E.; Thurow, K. Isotope pattern evaluation for the reduction of elemental compositions assigned to high-resolution mass spectral data from electrospray ionization fourier transform ion cyclotron resonance mass spectrometry. *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 1692–1699.
6. Kind, T.; Fiehn, O. Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinform.* **2006**, *7*, 234, doi:10.1186/1471-2105-7-234.
7. Zhang, J.; Gao, W.; Cai, J.; He, S.; Zeng, R.; Chen, R. Predicting molecular formulas of fragment ions with isotope patterns in tandem mass spectra. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2005**, *2*, 217–230.
8. Ipsen, A.; Want, E.J.; Ebbels, T.M.D. Construction of Confidence Regions for Isotopic Abundance Patterns in LC/MS Data Sets for Rigorous Determination of Molecular Formulas. *Anal. Chem.* **2010**, *82*, 7319–7328.
9. Pluskal, T.; Uehara, T.; Yanagida, M. Highly accurate chemical formula prediction tool utilizing high-resolution mass spectra, MS/MS fragmentation, heuristic rules, and isotope pattern matching. *Anal. Chem.* **2012**, *84*, 4396–4403.
10. Jarussophon, S.; Acoca, S.; Gao, J.M.; Deprez, C.; Kiyota, T.; Draghici, C.; Purisima, E.; Konishi, Y. Automated molecular formula determination by tandem mass spectrometry (MS/MS). *Analyst* **2009**, *134*, 690–700.
11. Meringer, M.; Reinker, S.; Zhang, J.; Muller, A. MS/MS Data Improves Automated Determination of Molecular Formulas by Mass Spectrometry. *MATCH Commun. Math. Comput. Chem.* **2011**, *2011*, 259–290.
12. Snider, R.K. Efficient calculation of exact mass isotopic distributions. *J. Am. Soc. Mass Spectrom.* **2007**, *18*, 1511–1515.
13. McLafferty, F.W.; Turecek, F. Interpretation of Mass Spectra, 4th ed. *J. Chem. Educ.* **1994**, *71*, doi:10.1021/ed071pA54.5.
14. Clendinen, C.S.; Stupp, G.S.; Ajredini, R.; Lee-McMullen, B.; Beecher, C.; Edison, A.S. An overview of methods using (13)C for improved compound identification in metabolomics and natural products. *Front. Plant Sci.* **2015**, *6*, doi:10.3389/fpls.2015.00611.
15. Daly, R.; Rogers, S.; Wandy, J.; Jankevics, A.; Burgess, K.E.; Breitling, R. MetAssign: Probabilistic annotation of metabolites from LC-MS data using a Bayesian clustering approach. *Bioinformatics* **2014**, *30*, 2764–2771.
16. Hussong, R.; Tholey, A.; Hildebrandt, A. Efficient Analysis of Mass Spectrometry Data Using the Isotope Wavelet. In Proceedings of the 3rd International Symposium on Computational Life Science (COMPLIFE 2007), Utrecht, The Netherlands, 4–5 October 2007; Volume 940, pp. 139–149.
17. Slawski, M.; Hussong, R.; Tholey, A.; Jakoby, T.; Gregorius, B.; Hildebrandt, A.; Hein, M. Isotope pattern deconvolution for peptide mass spectrometry by non-negative least squares/least absolute deviation template matching. *BMC Bioinform.* **2012**, *13*, doi:10.1186/1471-2105-13-291.
18. Kenar, E.; Franken, H.; Forcisi, S.; Wörmann, K.; Häring, H.U.U.; Lehmann, R.; Schmitt-Kopplin, P.; Zell, A.; Kohlbacher, O. Automated label-free quantification of metabolites from liquid chromatography-mass spectrometry data. *Mol. Cell. Proteom. MCP* **2014**, *13*, 348–359.
19. Libiseller, G.; Dvorzak, M.; Kleb, U.; Gander, E.; Eisenberg, T.; Madeo, F.; Neumann, S.; Trausinger, G.; Sinner, F.; Pieber, T.; et al. IPO: A tool for automated optimization of XCMS parameters. *BMC Bioinform.* **2015**, *16*, doi:10.1186/s12859-015-0562-8.
20. Ojanperä, S.; Pelander, A.; Pelzing, M.; Krebs, I.; Vuori, E.; Ojanperä, I. Isotopic pattern and accurate mass determination in urine drug screening by liquid chromatography/time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **2006**, *20*, 1161–1167.
21. Fabregat-Cabello, N.; Sancho, J.V.; Vidal, A.; González, F.V.; Roig-Navarro, A.F.F. Development and validation of a liquid chromatography isotope dilution mass spectrometry method for the reliable quantification of alkylphenols in environmental water samples by isotope pattern deconvolution. *J. Chromatogr. A* **2014**, *1328*, 43–51.
22. Haimi, P.; Uphoff, A.; Hermansson, M.; Somerharju, P. Software tools for analysis of mass spectrometric lipidome data. *Anal. Chem.* **2006**, *78*, 8324–8331.
23. Smith, C.A.; Want, E.J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **2006**, *78*, 779–787.

24. Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T.R.; Neumann, S. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.* **2012**, *84*, 283–289.
25. Gentleman, R.C.; Carey, V.J.; Bates, D.M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; et al. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **2004**, *5*, R80, doi:10.1186/gb-2004-5-10-r80.
26. Meija, J.; Caruso, J.A. Deconvolution of isobaric interferences in mass spectra. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 654–658.
27. Johnstone, R.A.W.; Rose, M.E. *Mass Spectrometry for Chemists and Biochemists*, 2nd ed.; Cambridge University Press: Cambridge, UK, 1996.
28. Yamagaki, T.; Watanabe, T. Hydrogen radical removal causes complex overlapping isotope patterns of aromatic carboxylic acids in negative-ion matrix-assisted laser desorption/ionization mass spectrometry. *Mass Spectrom.* **2012**, *1*, doi:10.5702/massspectrometry.A0005.
29. Alonso, A.; Julià, A.; Beltran, A.; Vinaixa, M.; Díaz, M.; Ibañez, L.; Correig, X.; Marsal, S. AStream: An R package for annotating LC/MS metabolomic data. *Bioinformatics* **2011**, *27*, 1339–1340.
30. Scheltema, R.A.; Jankevics, A.; Jansen, R.C.; Swertz, M.A.; Breitling, R. PeakML/mzMatch: A File Format, Java Library, R Library, and Tool-Chain for Mass Spectrometry Data Analysis. *Anal. Chem.* **2011**, *83*, 2786–2793.
31. Meusel, M.; Hufsky, F.; Panter, F.; Krug, D.; Müller, R.; Böcker, S. Predicting the Presence of Uncommon Elements in Unknown Biomolecules from Isotope Patterns. *Anal. Chem.* **2016**, *88*, 7556–7566.
32. Degtyarenko, K.; de Matos, P.; Ennis, M.; Hastings, J.; Zbinden, M.; McNaught, A.; Alcántara, R.; Darsow, M.; Guedj, M.; Ashburner, M. ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **2008**, *36*, D344–D350.
33. Ogata, H.; Goto, S.; Sato, K.; Fujibuchi, W.; Bono, H.; Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **1999**, *27*, 29–34.
34. Afendi, F.M.M.; Okada, T.; Yamazaki, M.; Hirai-Morita, A.; Nakamura, Y.; Nakamura, K.; Ikeda, S.; Takahashi, H.; Altaf-Ul-Amin, M.; Darusman, L.K.; et al. KNApSAcK family databases: Integrated metabolite-plant species databases for multifaceted plant research. *Plant Cell Physiol.* **2012**, *53*, doi:10.1093/pcp/pcr165.
35. Sud, M.; Fahy, E.; Cotter, D.; Brown, A.; Dennis, E.A.; Glass, C.K.; Merrill, A.H.; Murphy, R.C.; Raetz, C.R.; Russell, D.W.; et al. LMSD: LIPID MAPS structure database. *Nucleic Acids Res.* **2007**, *35*, D527–D532.
36. Kim, S.; Thiessen, P.A.; Bolton, E.E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B.A.; et al. PubChem Substance and Compound databases. *Nucleic Acids Res.* **2015**, *44*, D1202–D1213.
37. Loos, M.; Gerber, C.; Corona, F.; Hollender, J.; Singer, H. Accelerated Isotope Fine Structure Calculation Using Pruned Transition Trees. *Anal. Chem.* **2015**, *87*, 5738–5744.
38. Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinform.* **2010**, *11*, doi:10.1186/1471-2105-11-395.
39. Yu, T.; Park, Y.; Johnson, J.M.; Jones, D.P. apLCMS—Adaptive processing of high-resolution LC/MS data. *Bioinformatics* **2009**, *25*, 1930–1936.
40. Woldegebriel, M.; Vivó-Truyols, G. Probabilistic Model for Untargeted Peak Detection in LC–MS Using Bayesian Statistics. *Anal. Chem.* **2015**, *87*, 7345–7355.
41. Haug, K.; Salek, R.M.; Conesa, P.; Hastings, J.; de Matos, P.; Rijnbeek, M.; Mahendraker, T.; Williams, M.; Neumann, S.; Rocca-Serra, P.; et al. MetaboLights—An open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* **2013**, *41*, D781–D786.
42. Koch, B.P.; Dittmar, T.; Witt, M.; Kattner, G. Fundamentals of Molecular Formula Assignment to Ultrahigh Resolution Mass Data of Natural Organic Matter. *Anal. Chem.* **2007**, *79*, 1758–1763.

Eidesstattliche Erklärung / Declaration under Oath

This work was conducted from 03/2009 to 08/2012 under the supervision of Prof. Dr. Falk Schreiber at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben and from 09/2014 to 12/2016 under the supervision of Dr. Steffen Neumann at the Leibniz Institute of Plant Biochemistry (IPB) in Halle and Prof. Dr. Ivo Grosse at the Martin Luther University Halle–Wittenberg.

I declare under penalty of perjury that this thesis is my own work entirely and has been written without any help from other people. I used only the sources mentioned and included all the citations correctly both in word or content. This thesis has not previously been presented in identical or similar form to any other German or foreign examination board.

Halle (Saale), 26.01.2017 Hendrik Treutler

Hendrik Treutler

Researcher

Lerchenfeldstrasse 1
06110
Halle (Saale)
+49 (176) 222 83 968
+49 (345) 279 87 556
hendrik.treutler@gmail.com

Personal matters

birth 03/10/85 (né Mehlhorn)

family status Married, three children

Education

02/2009 **Diploma**, *Martin-Luther-University Halle–Wittenberg*, Halle (Saale), *Diploma in bioinformatics*.

06/2003 **Abitur**, *Dr.–Wilhelm–André–Gymnasium*, Chemnitz.

Diploma thesis

title "*Implementation von iterativen Algorithmen zur Motivvorhersage mit Centroiden*"

supervisor Prof. Dr. Ivo Grosse

description *De novo* motif prediction with centroid solution outperforms traditional approaches

Experience

Vocational

09/2014–today **Research associate**, *Leibniz Institute of Plant Biochemistry (IPB)*, Halle (Saale).
Ph.D. student in the field of mass spectrometry.
Selected subjects:

- Established the web application *MetFamily* for the discovery of substance classes which are characteristic for certain plant samples.
- Developed and implemented an approach for the extraction of isotope clusters from mass spectrometry data in the bioconductor packages *xcms* and *CAMERA*.
- Developed and implemented an approach for the comparison of sequence motifs and the visualization of motif differences in the bioconductor package *DiffLogo*.

03/2009–08/2012 **Research associate**, *Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)*, Stadt Seeland (OT Gatersleben).
Ph.D. student in the field of plant bioinformatics.
Selected subjects:

- Established an information system for the management, sharing, and ontology–assisted unification of wet lab data in the frame of research projects between partners from science and industry.
- Developed approaches for the integration of biological networks using sets of related identifiers.

Miscellaneous

07/2008–08/2009 **Programmer**, *ICON Genetics*, Halle (Saale).
Industrial activity for the development of a GUI–based tool for the classification and quantification of antibody mRNAs

Languages

German **First language**

English **Business fluent**

Publications*

Hendrik Mehlhorn, Matthias Lange, Uwe Scholz, and Falk Schreiber. Extraction and prediction of biomedical database identifier using neural networks towards data network construction. In M. D. Lytras P. Ordez de Pablos and R. Tennyson, editors, *Cases on Open-Linked Data and Semantic Web Applications*, pages 58–83. Information Science Reference (an imprint of IGI Global), January 2013.

Hendrik Rohn, Astrid Junker, Anja Hartmann, Eva Grafahrend-Belau, **Hendrik Treutler**, Matthias Klapperstück, Tobias Czauderna, Christian Klukas, and Falk Schreiber. VANTED v2: a framework for systems biology applications. *BMC systems biology*, 6(1):139+, November 2012.

Hendrik Mehlhorn, Matthias Lange, Uwe Scholz, and Falk Schreiber. IDPredictor: predict database links in biomedical database. *Journal of integrative bioinformatics*, 9(2), June 2012.

Hendrik Mehlhorn and Falk Schreiber. DBE2 - management of experimental data for the VANTED system. *Journal of integrative bioinformatics*, 8(2), July 2011.

Hendrik Mehlhorn and Falk Schreiber. TransID – the Flexible Identifier Mapping Service. In *Proceedings of the 9th International Symposium on Integrative Bioinformatics 2013*, pages p. 112–121, March 2013.

Hendrik Treutler and Steffen Neumann. Prediction, detection, and validation of isotope clusters in mass spectrometry data. *Metabolites*, 6(4), October 2016.

Hendrik Treutler, Hiroshi Tsugawa, Andrea Porzel, Karin Gorzolka, Alain Tissier, Steffen Neumann, and Gerd Ulrich U. Balcke. Discovering regulated metabolite families in untargeted metabolomics studies. *Analytical chemistry*, 88(16):8082–8090, August 2016.

Martin Nettling, **Hendrik Treutler**, Jesus Cerquides, and Ivo Grosse. Detecting and correcting the binding-affinity bias in ChIP-seq data using inter-species information. *BMC genomics*, 17(1), May 2016.

Martin Nettling, **Hendrik Treutler**, Jan Grau, Jens Keilwagen, Stefan Posch, and Ivo Grosse. DiffLogo: a comparative visualization of sequence motifs. *BMC bioinformatics*, 16:387+, November 2015.

* Hendrik Treutler (né Mehlhorn) is marked in bold, first authors are underlined

Open source software

- **DiffLogo**: Visualization of differences among sequence motifs in R
  `https://github.com/mgledi/DiffLogo`
- **xcms**: Processing of mass spectrometry data in R
  `https://github.com/sneumann/xcms`
- **CAMERA**: Annotation and analysis of mass spectrometry data in R
  `https://github.com/sneumann/CAMERA`
- **MetFamily**: Discovery of regulated metabolite families from mass spectrometry data in R
  `https://github.com/Treutler/MetFamily`
- **samplingDataCRT**: Sampling data for SWD and other study designs in R
  `https://github.com/trutscheld/samplingData`

Halle (Saale), 24.01.2017 Hendrik Treutler