Die Analyse kleiner Moleküle mittels Methoden der Kombinatorik und des maschinellen Lernens[1]

Kai Dührkop[2]

Abstract: Massenspektrometrie ist eine Technik für die Analyse kleiner Moleküle im Hochdurchsatz. Aber sie liefert nur Informationen über die Masse der gemessenen Moleküle und, mittels Tandem-Massenspektrometrie, über die Massen der gemessenen Fragmente. Die automatisierte Auswertung von Massenspektren beschränkt sich oft auf die Suche in Spektrendatenbanken, so dass nur Moleküle derepliziert werden können, die bereits in einer solchen Datenbank gemessen wurden. In dieser Dissertation präsentieren wir zwei Methoden zur Beantwortung zweier zentraler Fragen: Was ist die Molekülformel eines gemessenen Ions? Und was ist seine Molekülstruktur? SIRIUS ist eine Methode der kombinatorischen Optimierung für die Annotation von Massenspektren und der Identifikation der Molekülformel. Dazu berechnet sie hypothetische Fragmentierungsbäume. Wir stellen ein neues Scoring Modell für die Berechnung von Fragmentierungsbäumen vor, welches die kombinatorische Optimierung als einen Maximum-a-posteriori-Schätzer auffasst. Dieses Modell ermöglicht es uns, Parameter und Hyperparameter des Scorings direkt aus den Daten abzuschätzen. Wir zeigen, dass dieses statistische Modell, dessen (Hyper)Parameter auf einem kleinen Datensatz geschätzt wurden, allgemeingültig für viele Datensätze und sogar für verschiedene Massenspektrometriegeräte ist. Wir stellen außerdem CSI:FingerID vor, eine Methode, die Kernel Support Vector Maschinen zur Vorhersage von molekularen Fingerabdrücken aus Tandem-Massenspektren nutzt. Diese vorhergesagten Fingerabdrücke können in Strukturdatenbanken gesucht werden. Dies ermöglicht erstmals die Aufklärung von Molekülstrukturen mittels Massenspektrometrie, ohne dabei auf den Abgleich von Massenspektren mittels Datenbanksuche angewiesen zu sein. Beide Methoden, SIRIUS und CSI:FingerID, sind als Kommandozeilenprogramm und als Benutzeroberfläche verfügbar. Die Vorhersage molekularer Fingerabdrücke ist als Webservice implementiert, der über eine Millionen Anfragen pro Monat erhält.

1 Einführung: Metabolomik und Massenspektrometrie

Die Bioinformatik beschäftigt sich mit der automatisierten Auswertung biologischer Daten. Im Fokus steht dabei üblicherweise die Genetik. Wir wissen heute, dass der Mensch mehr ist als nur die Summe seiner Gene und das viele Einflüsse, teils epigenetisch, teils umweltbedingt, auf uns und die Regulation unserer Gene einwirken. Unser Verhalten und unser Gesundheitszustand hängt nicht nur von den Genen und Proteinen, sondern von einer Vielzahl kleiner Moleküle ab. Diese werden mit der Nahrung, über die Haut, über Medikamente oder die Luft aufgenommen. Manche Stoffe werden sogar von Bakterien in unserem Darm erzeugt und wirken unmittelbar auf unseren Stoffwechsel [Co17]. Es wird geschätzt, dass jeder Mensch im Laufe seines Lebens zwei bis drei Millionen verschiedener kleiner Moleküle aufnimmt [IG07], und dass 80 bis 85 % aller Krankheiten mit kleinen

[1] Computational Methods for Small Molecule Identification

[2] Lehrstuhl Bioinformatik, Friedrich-Schiller-Universität, kai.duehrkop@uni-jena.de

Molekülen in Verbindung stehen [Up16]. Wie aktuell das Thema *kleine Moleküle* ist, zeigen die Diskussionen um Stickoxid und Feinstaub, oder um die krebserregende Wirkung des Herbizids Glyphosat.

Doch Moleküle sind nicht nur Auslöser von Krankheiten, sie können auch Krankheiten heilen. Insbesondere in Bakterien, Pilzen und Pflanzen dienen kleine Moleküle oft zur Verteidigung gegen Parasiten, Nahrungskonkurrenten und Krankheitserregern. Der bekannteste natürliche Wirkstoff ist das Antibiotikum Penicillin, der in bestimmten Schimmelpilzen gebildet wird, und dessen Entdeckung zu den bedeutendsten Entwicklungen der Medizin gehört. Seit 1970 wurden nur noch wenige neue Antibiotika entdeckt, weswegen die Weltgesundheitsorganisation 2014 eine Warnung über die *Post-antibiotische Ära* herausgab [Wo14]: Immer mehr Bakterien entwickeln Resistenzen gegen die bekannten Antibiotika. Dies macht die Suche nach neuen Antibiotika äußerst wichtig. Dabei suchen Forscher für gewöhnlich in Biomen, die noch wenig erforscht sind, wie zum Beispiel am Meeresboden [HF10, Am10]. Doch wie lassen sich potenzielle neue Antibiotika unter den tausenden von Molekülen, die in einer solchen Bodenprobe enthalten sind, aufspüren?

Mit einer ähnlichen Frage beschäftigt sich auch die Umweltforschung. Welche und wie viele potenziell giftige oder krebserregende Stoffe gelangen, zum Beispiel über das Trinkwasser, in unseren Körper? Wie weist man Stoffe, wie zum Beispiel Glyphosat, im Trinkwasser nach? Die wichtigste analytische Technik zum Nachweis kleiner Moleküle ist die Massenspektrometrie (MS). Man kann sich ein Massenspektrometer wie eine molekulare Waage vorstellen. Moleküle werden ionisiert (also mit einer Ladung versehen) und dann in einem elektrischen Feld beschleunigt. Dabei trennen sich schwere und leichte Moleküle auf, so dass die Masse (oder genauer, das Masse-Ladungs-Verhältnis m/z) jedes Moleküls bestimmt werden kann. Die Ausgabe ist ein MS Spektrum, bestehend aus einer Vielzahl von Peaks. Jeder Peak ist ein m/z Wert mit einer bestimmten Intensität, welche mit der Anzahl der gemessenen Ionen mit diesem Masse-Ladungs-Verhältnis korreliert. Hier liegt aber auch das Problem: Ein Massenspektrometer kann nur die Masse eines Moleküls bestimmen, nicht aber seine Zusammensetzung. Das bedeutet, dass man zwischen Molekülen mit gleicher Masse aber unterschiedlicher Struktur nicht unterscheiden kann.

Dieses Problem lässt sich mittels Tandem-Massenspektrometrie (MS/MS) lösen. Dabei wird zuerst ein MS gemessen und dann alle Ionen mit einem gewünschten m/z selektiert. Diese Ionen werden dann in einer Kollisionskammer in kleinere Bruchstücke fragmentiert. Diese Bruchstücke werden daraufhin in einem zweiten MS gemessen und ergeben das MS/MS Spektrum. Kleine Moleküle nehmen während der Ionisierung üblicherweise nur einen einzelnen Ladungsträger auf. Zerfällt ein solches einfach geladenes Ion in Bruchstücke, so kann nur ein Bruchstück den Ladungsträger enthalten, während die anderen Bruchstücke üblicherweise ungeladen sind. Da ein MS nur geladene Teile messen kann, gehen die ungeladenen Bruchstücke im Messvorgang verloren. Sie werden daher *Neutralverluste* genannt. Nur die geladenen Bruchstücke, *Fragmente* genannt, werden im zweiten MS gemessen.

Ein weiteres Problem ist, dass die genaue Fragmentierungsweise eines Moleküls nicht im Voraus bekannt ist. Überraschenderweise ist diese Fragestellung ein enorm schweres Problem und bis heute gibt es keine Methode die mit Sicherheit die Fragmentierung von Molekülen vorhersagen kann. Selbst quantenchemische Berechnungen liefern nur ungenaue

Ergebnisse [Sp18]. Stattdessen werden Datenbanken von MS/MS Messungen bekannter Moleküle angelegt. Durch die Suche in solch einer Datenbank lassen sich Stoffe identifizieren - aber eben nur solche, die zuvor gemessen und in einer Datenbank abgespeichert wurden. Die Folge ist, dass nur ein winziger Bruchteil der Moleküle, die in einer Probe gemessen werden, auch einem bekannten Stoff zugeordnet werden können [dDQ15, Up16].

Man unterteilt Experimente in der Massenspektrometrie grob in zwei Kategorien: Der gerichteten und ungerichteten Analyse. Bei der gerichteten sucht man gezielt nach bestimmten Stoffen in einer Probe (Beispiel: die Suche nach Drogen oder Dopingmitteln). Bei der ungerichteten Analyse versucht man alle Stoffe in einer Probe aufzuklären. Bei biologischen Proben spricht man dann auch von der Metabolomik, die, analog zur Genomik in der Genetik, die Gesamtheit aller Metabolite (also kleinen Moleküle) in einer biologischen Probe untersuchen will. Wie bereits erwähnt, arbeitet die Metabolomik heute weitestgehend mit Datenbanksuchen. Das heißt aber eben auch, dass nur Moleküle identifiziert werden können, die bereits bekannt sind und gemessen wurden. Für neue biologische Erkenntnisse braucht man jedoch Methoden, die auch völlig neue Stoffe identifizieren können. Hierfür sind in den letzten Jahren verschiedene Methoden entwickelt wurden, die üblicherweise nur noch eine Liste an Kandidatenstrukturen benötigen, aber nicht mehr eine Datenbank von tatsächlichen Messungen dieser Moleküle [HSB14, Bl18]. Im Laufe meiner Doktorarbeit haben wir an der Entwicklung zweier solcher Methoden gearbeitet: SIRIUS [BD16] und CSI:FingerID [Du15]. Dabei werden Techniken der kombinatorischen Optimierung, der Bayesschen Statistik und des maschinellen Lernens miteinander vereint. Zusammen ermöglichen beide Methoden die Strukturaufklärung kleiner Moleküle. In verschiedenen unabhängigen Evaluationen konnten sie alle anderen Methoden zur Strukturaufklärung mittels Massenspektrometrie deutlich schlagen [Du19]. section*Analyse der Summenformel mittels SIRIUS Die Masse eines Moleküls hängt nur von seiner Zusammensetzung, also den Atomen und ihren Elementen, nicht aber von der Anordnung der Atome oder den Bindungen ab. Eine solche Zusammensetzung kann in Form einer Summenformel dargestellt werden. So beschreibt die Summenformel $C_6H_{12}O_6$ ein Molekül mit 6 Kohlenstoff-, 12 Wasserstoff- und 6 Sauerstoffatomen. Der erste Schritt der Strukturaufklärung eines Moleküls liegt in der Bestimmung seiner Summenformel. Überraschenderweise ist bereits das ein schweres Problem: Obgleich moderne MS Geräte die Masse eines kleinen Moleküls auf die fünfte Nachkommastelle genau bestimmen können, gibt es doch, für große Massen, oft hunderte bis tausende Summenformeln die annähernd die gleiche Masse haben. Zur Unterscheidung dieser Summenformeln reicht die Masse allein nicht aus. Mit SIRIUS haben wir eine Software entwickelt, welche die Summenformel des Ions bestimmt. Dabei werden zwei Analyseverfahren kombiniert: Die Isotopenmusteranalyse und die Fragmentierungsanalyse.

Isotopen sind Atome mit gleicher Protonen- und Elektronenzahl (und damit von gleichem Element), aber unterschiedlicher Neutronenzahl. Sie unterscheiden sich also in ihrer Masse. Beispielsweise ist im Schnitt jedes 100te Kohlenstoffatom ein ^{13}C Isotop, während die anderen Kohlenstoffatome üblicherweise vom Typ ^{12}C sind. Entsprechend kann man jeder Summenformel nicht eine einzelne, sondern eine Verteilung von Massen zuordnen. Diese Verteilung lässt sich für eine Summenformel berechnen und mit dem gemessenen Isotopenmuster vergleichen. Wir haben ein Maximum Likelihood Verfahren entwickelt, wel-

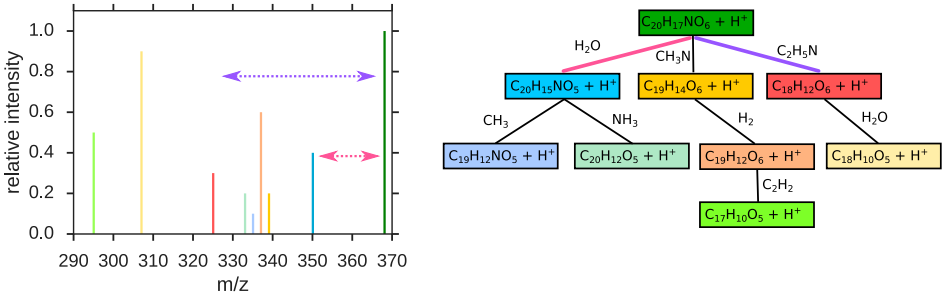

Abb. 1: Beispiel eines Fragmentierungsbaums (rechts) und dem zugehörigen MS/MS Spektrum (links) des Moleküls Bicucullin. Jeder Knoten im Baum ist die Erklärung eines Peaks im Spektrum. Jede Kante erklärt eine Fragmentierungsreaktion zwischen zwei Peaks im Spektrum.

ches einem Isotopenmuster die wahrscheinlichste Summenformel zuordnet [Bo09, Du19]. Isotopenmuster sind auch relevant, wenn man seltene Elemente wie Chlor oder Brom in einer Probe nachweisen will [Me16]. Dazu haben wir ein tiefes neuronales Netz entwickelt, welches solche seltenen Elemente aus einem Isotopenmuster vorhersagt [Du19].

Die Fragmentierungsanalyse berechnet Fragmentierungsbäume zur Aufklärung der Summenformel. Ein Fragmentierungsbaum beschreibt den Prozess der Fragmentierung als Baum, dessen Knoten die Summenformeln der Fragmente und dessen Kanten die Neutralverluste sind (Abbildung 1). Die Wurzel des Baums ist zugleich die Summenformel des gemessenen Ions. Zur Berechnung eines Fragmentierungsbaumes werden für alle Peaks im MS/MS Spektrum alle möglichen Summenformeln enumeriert. Dies ist über dynamische Programmierung effizient möglich [Bo08, Du13]. Danach wird ein gerichteter Fragmentierungsgraph erzeugt, dessen Knoten alle Summenformeln sind und der eine Kante zwischen je zwei Summenformeln enthält, die einander in einer Teilmengenbeziehung stehen. Jeder induzierte Teilbaum dieses Graphen ist eine mögliche Interpretation des Spektrums. Damit jedem Peak maximal eine Summenformel zugeordnet wird, färbt man den Graphen, so dass alle Knoten des gleichen Peaks die gleiche Farbe haben. Nun fordert man einen farbenfrohen Teilbaum, dessen Kanten- und Knotengewichte maximal sind. Als Gewichte wird ein beliebiges Scoring verwendet, welches wahrscheinlichen bzw. plausiblen Knoten und Kanten positive Gewichte, sowie den Unwahrscheinlichen negative Gewichte zuordnet. Das resultierende Problem „Maximaler Farbenfroher Teilbaum“ ist NP-schwer, kann aber in der Praxis mittels Integer Linearer Programmierung effizient gelöst werden [Ra13].

Die Fragmentierungsbaumberechnung geht auf die Arbeit von Rasche et al. [Ra11] zurück. Allerdings war das Scoring für die Kanten eher ad hoc gewählt und nur teilweise statistisch interpretierbar. Während meiner Doktorarbeit haben wir dieses algorithmische Problem in ein Maximum A Posteriori Problem umformuliert. Die Kantengewichte werden dabei zu *A Priori* Wahrscheinlichkeiten, die Knotengewichte zu *Likelihoods*. Dadurch bekommen die Kantengewichte eine statistische Interpretation. Darüber hinaus lassen sich diese Wahrscheinlichkeiten direkt aus echten Daten abschätzen: Für die A Priori Wahrscheinlichkeiten haben wir Wahrscheinlichkeitsverteilungen abgeschätzt und mittels Maximum

Likelihood Schätzern an gemessene Daten angepasst. Für die Likelihoods haben wir die Messfehler des MS Gerätes modelliert, sowie die Verteilung des Hintergrundrauschens.

Das resultierende neue Scoring wurde auf verschiedenen unabhängigen Datensätzen evaluiert. Die Identifikationsrate, also der Anteil korrekt bestimmter Summenformeln, stieg mit dem neuen Scoring um das doppelte auf 73,8 % an und schlägt damit nicht nur die vorherige Fragmentierungsbaumanalyse, sondern mit großem Abstand auch alle alternativen Methoden zur Summenformelbestimmung. Zusammen mit der Isotopenmusteranalyse erreicht SIRIUS Identifikationsraten von 86,36 %, 93,30 % und 93,75 % auf drei verschiedenen, unabhängigen Datensätzen. Neben dem neuen Scoring haben wur auch an Heuristiken und algorithmischen und technischen Optimierungen gearbeitet, was in einer 200fachen Beschleunigung der Fragmentierungsbaumberechnung resultierte [Du18, Wh15, Du19].

Strukturaufklärung mittels CSI:FingerID

Der zweite Teil meiner Doktorarbeit beschäftigt sich mit der Vorhersage von Molekülstrukturen mittels überwachtem maschinellem Lernen. Dabei stellt sich das Problem, dass sowohl MS Spektren als auch Moleküle keine einfachen numerischen Vektoren sind, wie sie normalerweise beim maschinellen Lernen als Eingabe und Ausgabe verwendet werden. Der erste Schritt für eine Strukturvorhersage aus MS Spektren liegt also darin, Spektren und Moleküle als Vektoren zu kodieren. Hier haben Markus Heinonen et al. Pionierarbeit geleistet [He12]: Sie kodieren Moleküle als binäre Vektoren, in denen jede Position für eine bestimmte Teilstruktur oder funktionelle Gruppe steht. Im Vektor steht eine 1, wenn das Molekül diese Teilstruktur enthält, ansonsten eine 0. Diese Form der Kodierung bezeichnet man auch als *Molekularen Fingerprint*. Da der Vektor binär ist, lässt sich die Strukturvorhersage als Menge von Klassifizierungsproblemen beschreiben: Für jede Teilstruktur wird ein Prediktor trainiert, der aus einem MS Spektrum das Vorhandensein dieser Struktur vorhersagt [He12]. Support Vektor Maschinen sind ein effizientes Verfahren zur Klassifizierung. Ihr Vorteil ist, dass sie nicht zwingend numerische Vektoren als Eingabe benötigen, sondern alternativ einen sogenannten Kernel: Dies ist eine Funktion, die das innere Produkt der Eingabevektoren berechnet. Beispielsweise muss das Spektrum nicht als Vektor kodiert werden, solange die Ähnlichkeit zweier Spektren direkt berechnet werden kann.

Auf diesen Grundlagen von Heinonen et al haben wir die Methode CSI:FingerID entwickelt [Sh14, Du15]. Statt nur das MS/MS Spektrum, verwenden wir Fragmentierungsbäume als Eingabe. Entsprechend haben wir Ähnlichkeitsfunktionen auf Fragmentierungsbäumen entwickelt. Beispielsweise eine Ähnlichkeitsfunktion, welche die Anzahl gemeinsamer Fragmente oder Neutralverluste zwischen zwei Bäumen zählt. Mittels dynamischer Programmierung lassen sich auch die Zahl gemeinsamer Pfade oder Teilbäume zählen. Besonders effektiv sind Ähnlichkeitsfunktionen, welche Eigenschaften auf den Molekülformeln bewerten, die aus der Masse des Peaks nicht ablesbar sind. Beispielsweise haben wir einen Kernel definiert, der aus der Summenformel eines Fragments die Zahl der kovalenten Atombindungen abschätzt. Ein solcher Kernel ist beispielsweise in der Lage, langkettige Moleküle von Molekülen mit vielen Ringen zu unterscheiden. Andere Ähnlichkeitsfunktionen suchen nach Teilsummenformeln, die zwei Fragmente gemeinsam haben, und die auf

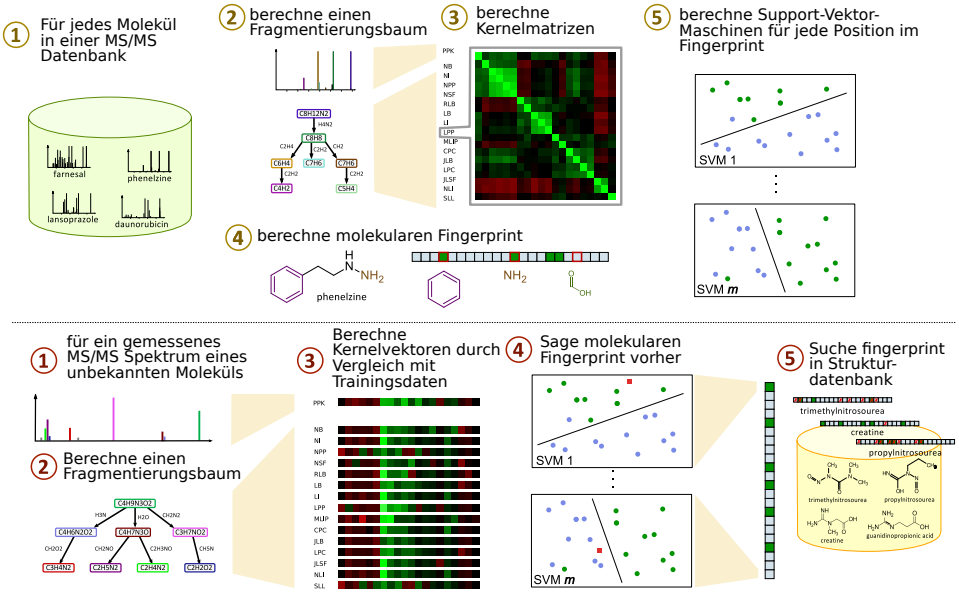

Abb. 2: Schematische Darstellung der Arbeitsweise von CSI:FingerID. In der Trainingsphase (Oben) wird das Maschinelle Lernverfahren auf einer Datenbank von Referenzspektren trainiert. In der Vorhersagephase (Unten) wird für ein unbekanntes Molekül dessen molekularer Fingerprint vorhergesagt und in einer Strukturdatenbank gesucht.

bestimmte Teilstrukturen hinweisen: So deutet das Vorhandensein der Summenformel H_3PO_4 auf eine Phosphatsäure im Molekül hin. Statt nur einer einzigen Ähnlichkeitsfunktion, definieren wir eine ganze Reihe von Kernels. Die 17 besten Kernels haben wir dann mittels multiplem Kernellernen kombiniert [Sh14]: Dabei werden die Kernels auf den Bäumen so linear kombiniert, dass die resultierende Funktion möglichst gut der Ähnlichkeitsfunktion auf den molekularen Fingerprints entspricht. Der resultierende Kernel hat eine deutlich verbesserte Qualität gegenüber dem ursprünglichen Kernel von Heinonen et al. Vor allem aber generalisiert er besser, wie sich auf unabhängigen Datensätzen gezeigt hat.

Um ein Molekül zu identifizieren, bestimmen wir erst seinen molekularen Fingerprint: Dabei kommen 2 937 Supportvektormaschinen zum Einsatz, die für jede Substruktur aus dem MS Spektrum und dem zugehörigen Fragmentierungsbaum die Wahrscheinlichkeit vorhersagen, das diese Struktur im gemessenen Molekül enthalten ist. Dieser vorhergesagte molekulare Fingerprint wird dann mittels Maximum Likelihood in einer Strukturdatenbank von Fingerprints gesucht. Siehe Abbildung 2 für den schematischen Aufbau von CSI:FingerID.

Wir haben CSI:FingerID auf verschiedenen Datensätzen evaluiert. Es ist konsistent und mit großem Abstand die beste Methode zur Identifizierung von Strukturen mittels Massenspektrometrie. In einer Evaluation mit Kreuzvalidierung konnte CSI:FingerID 3,3 Mal so viele Moleküle identifizieren wie die zweitbeste Methode. Die Identifikationsrate stieg

von 12,12 % auf 40,37 % an. Die Methode (zusammen mit ihrer Input-Output-Kernel-Regression Variante IOKR [Br16]) wurde auch zwei Mal blind in einem Wettbewerb, dem Critical Assessment of Small Molecule Identification (CASMI), evaluiert [Sc17]. In CASMI 2016 konnte die Methode 1,5 Mal so viele Moleküle wie die zweitbeste Methode identifizieren. In CASMI 2017 war der Abstand sogar noch größer - CSI:FingerID hatte 4,8 Mal so viele korrekte Identifikationen wie die nächstbeste Methode.

Fazit und Ausblick

Wir haben eine Kommandozeilenanwendung und grafische Nutzerschnittstelle für SIRIUS und CSI:FingerID implementiert. Die Integration mit CSI:FingerID geschieht über eine REST-Schnittstelle, so dass die Supportvektormaschinen auf unserem Server verbleiben und regelmäßig auf neuen Referenzdaten trainiert werden können. Unsere Methode wird mittlerweile weltweit eingesetzt und es werden Millionen Anfragen pro Monat an den CSI:FingerID Server gestellt. Verschiedene wichtige Massenspektrometrie Toolkits wie OpenMS [Rö16] und MZmine [Pl10], sowie die Global Natural Products Social Molecular Networking (GNPS) Datenbank [Wa16] integrieren SIRIUS mittlerweile in ihre Workflows. Auch wird unsere Methode zur Analyse kleiner Moleküle bereits von mehreren Pharmafirmen eingesetzt. CSI:FingerID ist in der Lage 40 % der MS/MS Spektren korrekt zu identifizieren. Schränkt man die Suche auf eine Datenbank biologischer Moleküle ein, steigt die Identifikationsrate sogar auf 75 % an. Dies macht eine qualitative Analyse im Hochdurchsatz überhaupt erst möglich.

Aufbauend auf SIRIUS und CSI:FingerID, entwickeln wir nun auch weitere Methoden, die nicht nur die Struktur, sondern auch abgeleitete Eigenschaften von Molekülen vorhersagen: Beispielsweise, ob ein Molekül zu einer bestimmten Klasse von Naturstoffen gehört. Oder ob ein gemessenes Molekül, ausgehend nur von seinem Massenspektrum, einem bekannten Wirkstoff oder Medikament ähnlich ist. Dies soll die Auswertung und Interpretation metabolomischer Daten sowie die Suche nach neuen Wirkstoffen vereinfachen und beschleunigen.

Literaturverzeichnis

[Am10] Aminov, Rustam: A brief history of the antibiotic era: lessons learned and challenges for the future. Front Microbiol, 1:134, 2010.

[BD16] Böcker, Sebastian; Duehrkop, Kai: Fragmentation trees reloaded. J Cheminform, 8:5, 2016.

[Bl18] Blaženović, Ivana; Kind, Tobias; Ji, Jian; Fiehn, Oliver: Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. Metabolites, 8(2), 2018.

[Bo08] Böcker, Sebastian; Lipták, Zsuzsanna; Martin, Marcel; Pervukhin, Anton; Sudek, Henner: DECOMP–from interpreting Mass Spectrometry peaks to solving the Money Changing Problem. Bioinformatics, 24(4):591–593, 2008.

[Bo09] Böcker, Sebastian; Letzel, Matthias; Lipták, Zsuzsanna; Pervukhin, Anton: SIRIUS: Decomposing isotope patterns for metabolite identification. Bioinformatics, 25(2):218–224, 2009.

[Br16] Brouard, Céline; Shen, Huibin; Dührkop, Kai; d'Alché-Buc, Florence; Böcker, Sebastian; Rousu, Juho: Fast metabolite identification with Input Output Kernel Regression. Bioinformatics, 32(12):i28–i36, 2016. Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2016).

[Co17] Cohen, Louis J; Esterhazy, Daria; Kim, Seong-Hwan; Lemetre, Christophe; Aguilar, Rhiannon R; Gordon, Emma A; Pickard, Amanda J; Cross, Justin R; Emiliano, Ana B; Han, Sun M; Chu, John; Vila-Farres, Xavier; Kaplitt, Jeremy; Rogoz, Aneta; Calle, Paula Y; Hunter, Craig; Bitok, J Kipchirchir; Brady, Sean F: Commensal bacteria make GPCR ligands that mimic human signalling molecules. Nature, 549(7670):48–53, 2017.

[dDQ15] da Silva, Ricardo R.; Dorrestein, Pieter C.; Quinn, Robert A.: Illuminating the dark matter in metabolomics. Proc Natl Acad Sci U S A, 112(41):12549–12550, 2015.

[Du13] Dührkop, Kai; Ludwig, Marcus; Meusel, Marvin; Böcker, Sebastian: Faster mass decomposition. In: Proc. of Workshop on Algorithms in Bioinformatics (WABI 2013). Jgg. 8126 in Lect Notes Comput Sci. Springer, Berlin, S. 45–58, 2013.

[Du15] Duehrkop, Kai; Shen, Huibin; Meusel, Marvin; Rousu, Juho; Böcker, Sebastian: Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc Natl Acad Sci U S A, 112(41):12580–12585, 2015.

[Du18] Duehrkop, Kai; Lataretu, Marie A.; White, W. Timothy J.; Böcker, Sebastian: Heuristic algorithms for the Maximum Colorful Subtree problem. In: Proc. of Workshop on Algorithms in Bioinformatics (WABI 2018). Jgg. 113 in Leibniz International Proceedings in Informatics (LIPIcs), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, S. 23:1–23:14, 2018.

[Du19] Duehrkop, Kai; Fleischauer, Markus; Ludwig, Marcus; Aksenov, Alexander A.; Melnik, Alexey V.; Meusel, Marvin; Dorrestein, Pieter C.; Rousu, Juho; Böcker, Sebastian: SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods, Marz 2019.

[He12] Heinonen, Markus; Shen, Huibin; Zamboni, Nicola; Rousu, Juho: Metabolite identification and molecular fingerprint prediction via machine learning. Bioinformatics, 28(18):2333–2341, 2012.

[HF10] Hughes, Chambers C.; Fenical, William: Antibacterials from the sea. Chem Eur J, 16(42):12512–12525, 2010.

[HSB14] Hufsky, Franziska; Scheubert, Kerstin; Böcker, Sebastian: New kids on the block: Novel informatics methods for natural product discovery. Nat Prod Rep, 31(6):807–817, 2014.

[IG07] Idle, Jeffrey R.; Gonzalez, Frank J.: Metabolomics. Cell Metab, 6(5):348–351, 2007.

[Me16] Meusel, Marvin; Hufsky, Franziska; Panter, Fabian; Krug, Daniel; Mueller, Rolf; Böcker, Sebastian: Predicting the presence of uncommon elements in unknown biomolecules from isotope patterns. Anal Chem, 88(15):7556–7566, 2016.

[Pl10] Pluskal, Tomáš; Castillo, Sandra; Villar-Briones, Alejandro; Oresic, Matej: MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinf, 11:395, 2010.

[Ra11] Rasche, Florian; Svatoš, Aleš; Maddula, Ravi Kumar; Boettcher, Christoph; Böcker, Sebastian: Computing fragmentation trees from tandem mass spectrometry data. Anal Chem, 83(4):1243–1251, 2011.

[Ra13] Rauf, Imran; Rasche, Florian; Nicolas, François; Böcker, Sebastian: Finding Maximum Colorful Subtrees in practice. J Comput Biol, 20(4):1–11, 2013.

[Rö16] Röst, Hannes L.; Sachsenberg, Timo; Aiche, Stephan; Bielow, Chris; Weisser, Hendrik; Aicheler, Fabian; Andreotti, Sandro; Ehrlich, Hans-Christian; Gutenbrunner, Petra; Kenar, Erhan; Liang, Xiao; Nahnsen, Sven; Nilse, Lars; Pfeuffer, Julianus; Rosenberger, George; Rurik, Marc; Schmitt, Uwe; Veit, Johannes; Walzer, Mathias; Wojnar, David; Wolski, Witold E.; Schilling, Oliver; Choudhary, Jyoti S.; Malmström, Lars; Aebersold, Ruedi; Reinert, Knut; Kohlbacher, Oliver: OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat Methods, 13(9):741–748, 2016.

[Sc17] Schymanski, Emma Louise; Ruttkies, Christoph; Krauss, Martin; Brouard, Céline; Kind, Tobias; Dührkop, Kai; Allen, Felicity Ruth; Vaniya, Arpana; Verdegem, Dries; Böcker, Sebastian; Rousu, Juho; Shen, Huibin; Tsugawa, Hiroshi; Sajed, Tanvir; Fiehn, Oliver; Ghesquière, Bart; Neumann, Steffen: Critical Assessment of Small Molecule Identification 2016: Automated Methods. J Cheminf, 9:22, 2017.

[Sh14] Shen, Huibin; Duehrkop, Kai; Böcker, Sebastian; Rousu, Juho: Metabolite Identification through Multiple Kernel Learning on Fragmentation Trees. Bioinformatics, 30(12):i157–i164, 2014. Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2014).

[Sp18] Spackman, Peter R.; Bohman, Bjoern; Karton, Amir; Jayatilaka, Dylan: Quantum chemical electron impact mass spectrum prediction for de novo structure elucidation: assessment against experimental reference data and comparison to competitive fragmentation modeling. Int J Quantum Chem, 118(2), 2018.

[Up16] Uppal, Karan; Walker, Douglas I.; Liu, Ken; Li, Shuzhao; Go, Young-Mi; Jones, Dean P.: Computational metabolomics: a framework for the million metabolome. Chem Res Toxicol, 29(12):1956–1975, 2016.

[Wa16] Wang, Mingxun et al.: Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol, 34(8):828–837, 2016.

[Wh15] White, W. Timothy J.; Beyer, Stephan; Duehrkop, Kai; Chimani, Markus; Böcker, Sebastian: Speedy Colorful Subtrees. In: Proc. of Computing and Combinatorics Conference (COCOON 2015). Jgg. 9198 in Lect Notes Comput Sci. Springer, Berlin, S. 310–322, 2015.

[Wo14] World Health Organization et al.: Antimicrobial resistance: global report on surveillance. World Health Organization, 2014.

Kai Dührkop wurde am 21.10.1988 in Rudolstadt geboren, ging bis 2007 auf das Friedrich-Froebel Gymnasium in Bad Blankenburg und studierte anschließend bis 2012 an der Friedrich-Schiller-Universität in Jena. Sein Studium schloss er mit einem Diplom in Bioinformatik ab. Die Diplomarbeit wurde zum Thema „A Sparse Dynamic Programming Algorithm for Fragmentation Tree Alignments“ verfasst. Von 2012 bis 2018 promovierte er am Lehrstuhl für Bioinformatik an der Friedrich-Schiller-Universität und schrieb seine Dissertation zum Thema „Computational Methods for Small Molecule Identification“. Seine Disputation fand am 20. September 2018 statt. Während seiner Promotion hat er 13 Publikationen verfasst, 7 davon als Erstautor. Er hielt wissenschaftliche Vorträge auf drei internationalen Konferenzen und besuchte zwei Dagstuhl- sowie ein Shonan-Seminar. Seit April 2019 arbeitet er im Rahmen eines wissenschaftlichen Austauschs an der University of California San Diego.