HUMBOLDT-UNIVERSITÄT ZU BERLIN

Exposé zur Diplomarbeit

„Entwurf und Implementierung einer integrierten Suche nach Metabolit- und Experimentdaten zur Unterstützung vergleichender Metabolite Profiling Experimente“

März 2007

Autor:
Georg Basler
Institut für Informatik
basler@informatik.hu-berlin.de

Betreuer:
Prof. Dr. Ulf Leser
Wissensmanagement in der Bioinformatik
Institut für Informatik

Prof. Dr. Joachim Selbig
Bioinformatik
Max Planck Institut für Molekulare Pflanzenphysiologie

1. Motivation

Bei der Gas-Chromatographie/Massenspektrometrie (GC/MS) werden von Stoffgemischen, insbesondere von biologischen Proben und Referenzsubstanzen, Chromatogramme erzeugt. Ein Chromatogramm ist ein dreidimensionales Signal aus den Zeiten, die die gefundenen Molekülfragmente zum Durchlaufen einer Trennsäule benötigen (Retentionszeiten), sowie zu jeder Retentionszeit einem (zweidimensionalen) Massenspektrum, welches sich aus dem Masse-Ladungsverhältnis und der Isotopenhäufigkeit der Fragmente zusammensetzt [1]. Durch den Vergleich der aus Pflanzen erhaltenen Spektren mit Spektren aus käuflich erwerbbaren Referenzsubstanzen ist eine Identifikation der in der Pflanze enthaltenen Metaboliten möglich [2]. Die so gewonnenen Daten werden in der „Golm Metabolome Database“ (GMD) abgelegt.

Informationen über das Experiment, wie die Spezies, aus der die biologische Probe stammt, oder die Umweltbedingungen, denen diese vor der GC/MS ausgesetzt war, werden momentan von dem für das Experiment Verantwortlichen lokal und in unstrukturierter Form abgelegt (z.B. in Textdokumenten). Dadurch sind die Ergebnisse nur umständlich reproduzierbar, und eine breit angelegte, vergleichende Suche nach Metaboliten aus Proben mit gemeinsamen Eigenschaften ist mit hohem manuellem Aufwand verbunden.

Zum derzeitigen Stand enthält die GMD noch keine Massenspektren, sondern lediglich die physikalischen Attribute einiger bei der GC/MS gemessenen Analyten. In Zukunft sollen jedoch auch die Spektren von Analyten und Referenzsubstanzen in die GMD aufgenommen und öffentlich verfügbar gemacht werden. Die Parameter eines Experiments, wie die untersuchte Spezies, Organe und Umweltbedingungen, sollen in ein kommerzielles „Laboratory Information Management System“[1] eingegeben werden. Das LIMS ist implementiert, jedoch noch nicht für die Aufnahme dieser Daten konfiguriert. Die Dokumentation und Skalierbarkeit des LIMS, insbesondere im Hinblick auf den automatisierten Zugriff, weisen deutliche Schwachstellen auf. Somit besteht zur Zeit keine Möglichkeit, die Experimentparameter automatisiert mit den zugehörigen Ergebnissen, insbesondere den Massenspektren, in Beziehung zu setzen.

2. Zielsetzung

Für jedes in Zukunft in der GMD abgelegte Massenspektrum soll eine eindeutige Referenz auf die im LIMS abgelegten Experimentparameter möglich sein. Diese Parameter sollen möglichst umfassend den Lebenszyklus und die an der ursprünglichen biologischen Probe vorgenommenen Präparationen erfassen. Dazu wird eine Erweiterung der dem LIMS zugrunde liegenden Oracle-Datenbank um Datenobjekte zur Aufnahme der Experimentparameter nötig sein.

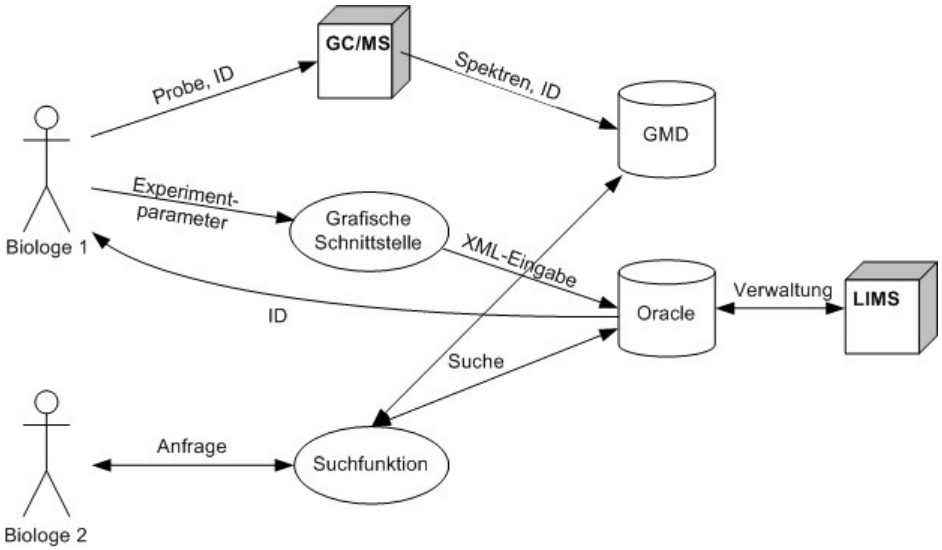
Auf den so ineinander in Beziehung gesetzten Spektren- und Experimentdaten soll eine Suchfunktion das Auffinden von Datensätzen mit möglichst beliebigen gemeinsamen Attributen ermöglichen. Darauf aufbauend soll untersucht werden, wie diese Suchfunktion performant implementiert werden kann.

3. Vorgehensweise

Um die strukturierte Aufnahme der Experimentparameter in die vom LIMS verwaltete Oracle-Datenbank zu ermöglichen, müssen zunächst Datenbanktabellen angelegt werden, die das biologische Experiment repräsentieren. Hierbei sollen zunächst nur die grundlegenden

[1] http://www.thermo.com/nautilus

Parameter, wie Spezies, Organ, Standortnummer und Identifikationsnummer angelegt werden; diese können später um detailliertere Informationen erweitert werden. Die Eingabe der Daten soll über eine grafische Benutzerschnittstelle erfolgen, welche eine starke Benutzerorientierung mit großer Flexibilität und Konfigurierbarkeit verbindet, um die Aufnahme der unterschiedlich komplexen Datensätze für den Benutzer zeitsparend zu gestalten. Wo möglich soll die Verwendung von standardisierten Begriffen, beispielsweise durch die „Gene Ontology“[2], vorgegeben werden. Um eine hohe Benutzerakzeptanz zu erreichen sollte im Vergleich zur bisherigen manuellen Dokumentation möglichst kein zusätzlicher Aufwand entstehen; dies kann beispielsweise durch die anschließende Ausgabe eines Datenblatts mit zusätzlichen für das Experiment relevanten Informationen, wie einer Historie der Gewächshausstandorte der Pflanze über ihre Lebensdauer, oder der Identifikationsnummer in Form eines Barcodes, erleichtert werden. Aus den Eingabedaten muss eine XML-Datei erzeugt werden, welche der (mangelhaft dokumentierten) Spezifikation des LIMS genügt.



Die Identifikationsnummer der Probe soll bei der GC/MS-Analyse übergeben und schließlich zusammen mit den Spektren in der GMD abgelegt werden (siehe Abbildung). Sie stellt das Schlüsselelement für die Beziehung zwischen Spektrum und zugehörigem Experiment dar. Daraufhin soll eine Suchfunktion entworfen werden, welche Suchanfragen sowohl an die Attribute der Analyten bzw. Spektren der GMD, als auch Experimentparameter der LIMS-Datenbank ermöglicht. Mit beiden Datenbanken soll parallel kommuniziert und so eine integrierte Sicht der Daten geschaffen werden. Vorstellbar sind Suchanfragen, die alle Analyten oder Metaboliten mit bestimmten gemeinsamen Experimentparametern als Ergebnis liefern, zum Beispiel alle identifizierten Metaboliten aus Lichtmangel ausgesetzten Blattzellen der Arabidopsis thaliana. Ebenso soll die Suche nach Experimenten ermöglicht werden, bei denen bestimmte Metaboliten gefunden wurden. Dadurch könnte eine wesentliche Aufgabe beim „Metabolite Profiling“, dem Auffinden von Mustern der Metabolitkonzentrationen in großen Mengen von Daten, weitgehend automatisiert werden [1].

Nachdem die grundlegende Suchfunktionalität vorhanden ist, soll auf einen der umgesetzten Aspekte näher eingegangen werden. So könnten Möglichkeiten der Datenorganisation von Spektren untersucht werden, wobei effiziente Ansätze zur Suche, wie die in [3] untersuchten Algorithmen, Ansätze zur Speicherung, oder des ähnlichkeitsbasierten Vergleichs von Spektren [4] miteinander verglichen, eine geeignete Methode ausgewählt und im Detail untersucht wird.

[2] http://www.geneontology.org

Literatur

1. Hummel, Jan et al. (im Druck): *The Golm Metabolome Database: a Database for GC-MS based Metabolite Profiling*. In: Hohmann, Stefan: Topics in Current Genetics: Nielsen, Jens/Jewett, Michael C.: Metabolomics. Berlin, Heidelberg, New York: Springer-Verlag.
2. Kopka, Joachim (2005): *Current challenges and developments in GC-MS based metabolite profiling technology*. In: Journal of Biotechnology 124 (2006). 312-322. PubMed: 16434119.
3. Stein, Stephen E./Scott, Donald R. (1994): *Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification*. In: Journal of the American Society for Mass Spectrometry 5 (1994). 859-866.
4. Hansen, Michael Edberg/Smedsgaard, Jørn (2004): *A New Matching Algorithm for High Resolution Mass Spectra*. In: Journal of the American Society for Mass Spectrometry 15 (2004). 1173-1180. PubMed: 15276164.