DISS. ETH NO. 17386

SEARCH HEURISTICS FOR MODULE IDENTIFICATION FROM BIOLOGICAL HIGH-THROUGHPUT DATA

A dissertation submitted to
ETH ZURICH

for the degree of
Doctor of Sciences

presented by

STEFAN BLEULER
Dipl. El.-Ing., ETH Zurich
born July 13, 1977
citizen of
Zollikon, ZH

accepted on the recommendation of
Prof. Dr. Eckart Zitzler, examiner
Prof. Dr. Peter Bühlmann, co-examiner

2007

Abstract

The advent of high-throughput measurement technologies in molecular biology enabled the determination of cellular parameters like the concentration of proteins, mRNAs or metabolites or the binding between molecules on a genome scale. The resulting data make new types of analyses possible which focus more on interactions between multiple elements such as genes, proteins or metabolites. A prominent type of analysis is to search for modules, i. e., groups of elements which exhibit similar properties in the measurements. The underlying assumption is that these similarities relate to common functions of the elements. While grouping alone does not explain the nature of specific interactions it often provides interesting hypotheses for further research or it can serve as preprocessing step for other types of analyses, e. g., the dimensionality of the data can be reduced by studying representatives for each module or by focusing on specific modules.

In most cases, such module identification tasks result in complex optimization problems many of which have been shown to be NP-hard. In the last few years, general module identification methods like k-means clustering or hierarchical clustering methods have been gradually adapted to the specifics of biological high-throughput data resulting amongst others in a number of so called biclustering algorithms. In contrast to standard clustering methods, biclustering algorithms do not require high similarity over all measurements but, taking gene expression as an example, they search for groups of genes which are similarly expressed over a subset of conditions. Despite this large advance, several important issues remained unsolved, such as the problems of integrating multiple data sets and different types of high-throughput measurements.

As a first step, this thesis confirms the usefulness of the basic biclustering approach in an extensive comparison of various existing heuristic biclustering approaches, a standard clustering method and a new exact algorithm based on a simple model. Building on these results, a flexible framework for biclustering is presented. The optimization algorithm consists of a hybridization of an evolutionary algorithm (EA) and a greedy local search. Thanks to the black-box scheme of the EA, this combination provides higher flexibility than most existing approaches. Building on this framework, the present thesis proposes approaches to three important open problems in module identification.

- In many biological studies several distinct gene expression data sets needs to be analyzed simultaneously. However, often measurement values are not directly comparable across data sets if they stem from different experiments, different labs or different measurement technologies. To address this problem, an approach for the joint bicluster

analysis of multiple expression data sets was developed. This allows to identify biclusters extending over multiple expression data sets even when measurement values are not directly comparable between the data sets.

- An even more challenging problem is the integration of multiple types of biological high-throughput data. A new data integration method is introduced which in contrast to existing approaches does not aggregate similarity measures on the different data sets but searches for a set of trade-off solution thereby visualizing potential conflicts between the information contained in the data sets.

- Often new measurement technologies require the development of new analysis methods. Thanks to the flexibility of the framework presented in this thesis it could be applied to extract information from a very recent type of measurements where only a few analysis methods exist, namely fluxome profiles. The resulting method is able to discriminate bacterial mutant strains based on their fluxome profiles.

Zusammenfassung

Moderne Messtechnologien in der Molekularbiologie ermöglichen es, verschiedene zelluläre Grössen wie die Konzentration von Proteinen, mRNA oder Metaboliten oder die Bindung zwischen Molekülen nicht nur für einzelne dieser Elemente sondern global zu bestimmen. Solche Daten erlauben neue Arten von Analysen, welche vermehrt die Interaktionen von verschiedenen Elementen z.B. verschiedenen Genen oder Proteinen untersuchen. Eine typische Analysemethode in dieser Kategorie ist die Modulidentifikation. Diese sucht nach Gruppen von Elementen, welche Ähnlichkeiten in Ihren Messwerten aufweisen. Dieser Strategie liegt die Annahme zu Grunde, dass solche Ähnlichkeiten auf eine gemeinsame Funktion hinweisen. Solche Gruppierungen erklären zwar die Art der Interaktionen nicht direkt, aber sie helfen Hypothesen darüber zu formulieren. Ausserdem können sie als Ausgangspunkt für weitere Analysen dienen. Beispielsweise kann die Dimension der Daten reduziert werden, indem nur die Interaktionen innerhalb einer Gruppe oder zwischen den Gruppen untersucht werden.

In den meisten Fällen resultieren solche Modulidentifikationen in komplexen Optimierungsproblemen und für viele davon wurde gezeigt, dass sie NP-schwer sind. In den letzten Jahren wurden allgemeine Methoden zur Modulidentifikation wie k-means Clustering oder Hierarchisches Clustering schrittweise den Anforderungen der biologischen Daten angepasst. Dabei wurden unter anderem so genannte Biclustering-Verfahren entwickelt, welche im Gegensatz zu klassischen Clustering-Methoden die Ähnlichkeiten nicht über alle Messpunkte bestimmen. Für das Beispiel von Genexpressionsdaten bedeutet dies, Gruppen von Genen zu suchen, welche über einen Teil der untersuchten Bedingungen ähnliche Profile aufweisen. Trotz dieser grossen Fortschritte sind einige wichtige Fragen offen geblieben. So ist es z.B. unklar wie man am besten eine Analyse von mehrere Genexpressions-Datensätzen vornimmt, oder wie man verschiedene Typen von Messungen kombinieren kann.

Als erster Schritt in dieser Dissertation, wurde die Nützlichkeit des grundsätzlichen Biclustering-Ansatzes mittels eines umfangreichen Vergleichs von mehreren existierenden, heuristischen Biclustering-Methoden, einem traditionellen Clustering-Verfahren und einem neuen exakten Algorithmus bestätigt. Basierend auf diesen Resultaten wurde ein flexibles Framework für Biclustering entwickelt, welches auf einer Kombination von einem Evolutionärem Algorithmus und einer Lokalen Suche basiert. Diese Kombination ermöglicht eine viel grössere Flexibilität in der Problemformulierung als bestehende Ansätze. Basierend auf diesem neuen Ansatz wurden drei wichtige offene Probleme im Bereich der Modulidentifikation angegangen.

- In vielen biologischen Studien müssen mehrere Genexpressionsdatensätze gemeinsam analysiert werden. Häufig können aber die Messwerte der verschiedenen Datensätze nicht direkt verglichen werden, wenn die Daten z.B. aus verschiedenen Experimenten, verschiedenen Labors oder verschiedenen Messverfahren stammen. Dazu wird in dieser Dissertation ein Verfahren vorgestellt, welches ein kombiniertes Biclustering von mehreren Datensätzen ermöglicht. Damit können Bicluster gefunden werden, welche sich über mehrere Datensätze erstrecken.

- Ein noch anspruchsvolleres Problem ist die Integration von verschiedenen Typen von Messungen. Dazu wurde ein Verfahren entwickelt, welches im Gegensatz zu bestehenden Methoden nicht die Ähnlichkeiten auf den verschiedenen Datensätzen in einem Ähnlichkeitsmass zusammenfasst, sondern eine so genannte Trade-off Front berechnet. Diese zeigt auf, in welchem Ausmass die verschiedenen Datentypen die gleiche Information enthalten.

- Häufig machen neue Messmethoden die Entwicklung von neuen Analysemethoden nötig. Dank der Flexibilität des hier vorgestellten Frameworks konnte dieses zur Analyse einer neuen Art von Messungen über den Zellstoffwechsel eingesetzt werden. Dabei entstand eine Methode, welche anhand dieser Fluxomprofile Ähnlichkeiten zwischen verschiedenen Bakterienstämmen identifiziert.