Biclustering for the Analysis of Global Regulatory Patterns in Large-Scale Gene Expression Data

DISS. ETH NO. 21797

Biclustering for the Analysis of Global Regulatory Patterns in Large-Scale Gene Expression Data

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by
OLIVER MORITZ VOGGENREITER
MSc Computational Biology & Bioinformatics, ETH Zurich

born on 25.04.1988
citizen of Germany

accepted on the recommendation of
Prof. Dr. Wilhelm Gruissem
Prof. Dr. Peter Bühlmann
Dr. Stefan Bleuler
Dr. Katja Bärenfaller

2014

Abstract

High-throughput technologies in the field of biology have increased the amount of data available in the public realm with regards to the relative quantities of biological agents such as proteins, genes, and transcripts under various conditions. This fast growth of the data has led to the application of new algorithms in order to come to grips with the increasing complexity of the biological samples. The large quantities of data promise to provide information about the relationships between various biological agents across many different testing conditions. This large variance in the origin of the data could potentially allow for the discovery and assessment of highly specific modules of regulation. Many different algorithms have been developed to find these modules, also known as biclusters, but there is a growing need for a solution that can run on data derived from thousands of experiments.

This thesis begins by providing an overview of some of the algorithms dealing with the discovery of biclusters in high-throughput data. We then take a closer look at one of these algorithms in order to identify potential improvements that could be made in the performance of the algorithm. Subsequently, an abstraction of the general problem of finding these biological modules is developed in order to provide the basis for a novel algorithm designed specifically for dealing with very large datasets. The thesis concludes by describing a tool which allows for the analysis and interpretation of the newly discovered modules.

The new algorithm is applied on a large dataset comprising the gene expression values of the small flowering plant, *Arabidopsis thaliana*, under thousands of different conditions. We perform an analysis of two of the biclusters discovered by the algorithm and find new interpretations for the results of a number of different publications.

Zusammenfassung

Moderne Messtechnologien in der Biologie haben die Menge der Daten über die relativen Werte von biologisch aktiven Komponenten (Proteine, Gene, und Transkripte) unter verschiedenen Bedingungen erhöht. Das schnelle Anwachsen der Daten führte dazu, dass neue Algorithmen entwickelt worden sind um die Komplexität der Biologischen Daten bewältigen zu können. Die große Menge der Daten verspricht weitaus mehr Informationen über die verschiedenen Beziehungen der Komponenten zu enthalten als bisher bekannt. Die vielen verschiedenen Testbedingungen unter der die Proben genommen würden erlauben viel genauere Module der Regulierung zu finden und zu analysieren. Viele verschiedene Algorithmen wurden entwickelt um diese Module, auch Bicluster genannt, zu finden. Um die heutige Menge an Daten zu bewältigen ist es jedoch nötig eine neue Lösung zu entwickeln.

Diese Dissertation beginnt mit einer Einführung zu bisherigen Bicluster Algorithmen. Danach wird eine dieser Methoden genauer analysiert um mögliche Verbesserungen in der Laufzeit zu entdecken. Anschließend wird eine Abstraktion der allgemeinen Problemformulierung "Entdeckung biologischer Module" entwickelt. Diese Abstraktion wird dann als Basis genutzt um einen spezialisierten Algorithmus zu entwichlen der auch grosse Datensätze auswerten kann. Das letzte Kapitel behandelt eine neue Applikation, welche die Analyse und Interpretation von neuen biologischen Modulen erleichtert.

Der neue Algorithmus wird auf einen großen Datensatz angewendet der die Gen-Expressions Werte des Modellorganismus *Arabidopsis thaliana* enthält. Dieser Datensatz besteht aus über tausend verschiedenen experimentellen Bedingungen. Zwei der gefundenen Bicluster werden anschließend analysiert und führen zu einer neuen Interpretation der Resultate mehrerer früherer Publikationen.