Berichte aus der Informatik

Stefan Bleuler

Search Heuristics for Module Identification from Biological High-Throughput Data

Shaker Verlag
Aachen 2008

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Zugl.: Zürich, ETH, Diss., 2007

Printed in Germany.

ISBN 978-3-8322-6928-9
ISSN 0945-0807

Shaker Verlag GmbH • P.O. BOX 101818 • D-52018 Aachen
Phone: 0049/2407/9596-0 • Telefax: 0049/2407/9596-9
Internet: www.shaker.de • e-mail: info@shaker.de

Search Heuristics for Module Identification from Biological High-Throughput Data

The advent of high-throughput measurement technologies in molecular biology enabled the determination of cellular parameters like the concentration of proteins, mRNA 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.