Methods and applications of fuzzy modularity in biological systems

Author(s):
Hinkley, Trevor

Publication Date:
2011

Permanent Link:

https://doi.org/10.3929/ethz-a-007052316 →

Diss. ETH No. 19986

Methods and Applications of Fuzzy Modularity in Biological Systems

A dissertation submitted to

ETH ZURICH

for the degree of

Doctor of Sciences

presented by

Trevor Hinkley

BSc, University of Nottingham. MSc, Imperial College London

Date of Birth: 16th April 1983

Citizen of The United Kingdom of Great Britain and Northern Ireland

Accepted under the recommendation of

Prof. Sebastian Bonhoeffer, examiner

Prof. Niko Beerenwinkel, co-examiner

2011

Summary

In this thesis, I introduce and apply a number of novel algorithms for the analysis of systems-level biological data.

Chapter 1 consists of an introduction to the broad concept of modularity. The extension of hierarchic modularity and its relevance to biological systems is examined.

In chapter 2, existing algorithms for the discovery of hierarchic modularity from graphically-represented datasets are examined. It is shown that the scaling parameters used in these algorithms are equivalent to the distance measures used in hierarchic clustering algorithms.

In chapter 3, an investigation is made into the mathematical properties of Markov-walk-based multiscale techniques when a partial eigendecomposition is used. Calculations for exact bounds on the optimization of these algorithms are derived.

In chapter 4, multiscale module-discovery techniques are extended into probabilistic, fuzzy algorithms that are true generalizations of the initial forms.

In chapter 5, a generalized kernel ridge regression (GKRR) is presented, with emphasis on its application to the calculation of epistatic fitness landscapes from probabilistic population-sequence data. In chapter 6, fuzzy modularity of protein-protein interaction networks of the organism *Saccharomyces Cerevisiae* is investigated in depth. The relations of a diverse array of biological data to discovered modules, and to a novel measure of protein fuzziness, are analysed. A speculative model of network growth is presented, based on these findings.

In chapter 7, the GKRR is applied to an extremely large dataset of sequences and fitness values of Human Immunodeficiency Virus 1, subtype B. A biological epistatic fitness landscape is derived, the first such landscape to be presented.

In chapter 8, prospects for future work, based on that presented here, is examined.

Zusammenfassung

In dieser Dissertation wende ich eine Reihe neuer Algorithmen für die Analyse Systembiologischer Daten an.

Kapitel 1 besteht aus einer Einführung in das Konzept der Modularität und aus einer Untersuchung der Erweiterung hierarchischer Modularität und ihre Relevanz für biologische Systeme.

In Kapitel 2 werden bestehende Algorithmen für die Entdeckung hierarchischer Modularität aus grafisch dargestellten Daten untersucht. Es wird gezeigt, dass die Skalierungsparameter dieser Algorithmen equivalent zu den Distanzmassen hierarchischer Clusteringalgorithmen sind.

In Kapitel 3 werden die mathematischen Eigenschaften Markovirrfahrt-basierenden Multiskalierung-Methoden untersucht, deren eine parzielle Eigenwertzerlegung zugrunde liegt. Es werden genaue Schranken für die Optimisierung dieser Algorithmen hergeleitet.

Kaptiel 4 erweitert multiskalierte Modulerkennungmethoden auf probabilistische unscharfe ("fuzzy") Methoden, welche wahre Generalisierungen der ursprünglichen Formen sind.

In Kapitel 5 wird eine generalisierte Kernel Ridge-Regression (GKRR) vorgezeigt, mit Gewichtung auf ihre Anwendung auf die Berechnung epistatischer Fitnesslandschaften aus probabilistischen Populationssequenzdaten.

In Kapitel 6 wird die unscharfe Modularität von Protein-Protein Interaktionsnetzwerken des Organismus *Saccharomyces Cerevisiae* im Detail untersucht. Die Verwandtschaft zwischen einer diversen Reihe biologischer Daten und entdeckten Modulen sowie eines neuen Masses der Proteinunschärfe wird untersucht. Ein spekulatives Modell von Netzwerkwachstum wird präsentiert, welches auf diesen Ergebnissen basiert.

In Kapitel 7 wird die GKRR Methode auf ein extrem grosses Datenset angewendet, welches aus Sequenzen und Fitnesswerten des Humanen Immundefizienz Virus 1, Subtyp B besteht. Eine biologische Epistase-Fitnesslandschaft

wird abgeleitet, welches die erste ihrer Art ist.

Kapitel 8 untersucht Möglichkeiten für weiterführende Arbeit, basierend auf den Ergebnissen, die hier gezeigt werden.