Diss. ETH No. 22344

Modular Identification and Analysis of Biomolecular Networks

A thesis submitted to
ETH Zurich

to attain the degree of
Doctor of Sciences of ETH Zurich
(Dr. sc. ETH Zurich)

presented by
Moritz Lang
Dipl.-Ing. (Engineering Cybernetics), University of Stuttgart

born on February 22nd, 1984, *in* Darmstadt, Germany
citizen of the Federal Republic of Germany

accepted on the recommendation of
Prof. Dr. Jörg Stelling, examiner
Prof. Dr. Antonis Papachristodoulou, co-examiner
Prof. Dr. Mustafa Khammash, co-examiner

2014

Abstract

Modularization of mathematical models is a divide-and-conquer strategy, in which a large and complex biomolecular network is dissected into several, potentially overlapping sub-networks, the modules. Most available modularization approaches have focused on the identification of functional modules. Functional modules are sub-networks considered to fulfill specific tasks within the cell, or within their biomolecular network. In contrast, fewer methods have been developed to identify operational modules, sub-networks defined to assist in specific research tasks. The aim of an operational modularization is to simplify the identification and analysis of each individual module, as well as of the inter-module interactions, or to make it computationally less expensive than analyzing the whole, non-modularized network. In general, operational modularization methods are better suited for a specific research task if modules are not only defined by structural properties of the network, but also by the intended usage, as well as the available experimental data or limitations on experimental capabilities. In this thesis, we illustrate the applicability of the modularization framework as a generic problem solving strategy for systems identification and analysis tasks in systems biology.

If no or relatively little prior knowledge is available about a biomolecular network, black-box structural identification methods can detect the presence or absence of network features. These features restrict the topology of models adequately representing the network, thus, guiding further experimental discovery processes. Our first method concerns the detection of feedback and feedforward loops in biomolecular networks for which the structure is completely unknown. Systems either containing a feedback or a feedforward loop can show similar input-output dynamics. To discriminate between such loops, we induce system dynamics characteristic for either type of loop. Thereby, we utilize a relationship between the phase delay and the amplification of oscillatory signals by the network that is only found in nonlinear systems. Based on the results of our method, a first raw model of the network can be constructed from modules with partly characterized dynamical properties. These modules represent the feedback loop or the two arms of the feedforward loop, and the rest of the pathway.

If more knowledge of a biomolecular network is available, a core model of the network can be constructed containing all confirmed and characterized molecular interactions. However, for larger networks typically several potential model extensions exist, representing sets of hypothetical interactions which are not experimentally confirmed, yet. The core model together with each possible combination of such extensions define the set of candidate models of the network. The goal of model discrimination methods is to validate the existence of each model extension by determining the agreement of each candidate model with experimental data of the network dynamics. If several mutually compatible extensions exist, the application of most model discrimination approaches becomes computationally challenging due to the exponentially growing number of candidate models with the number of such extensions. Furthermore, the validation process for a given extension might be compromised due to inaccuracies in the core model, in the mathematical description of

the individual extensions, or due to other, already erroneously identified extensions. To decrease the number of candidate models as well as the dimensions of their parameter and state spaces, and to prevent error propagation as far as possible, we modularize the network such that each module contains exactly one hypothetical extension. We insulate the modules from each other such that they can be independently simulated and analyzed by using experimental data as virtual inputs for the modules. Our approach does not only significantly reduce the number of candidate models which have to be evaluated, but also the dimensions of their parameter and state spaces. These reductions in computational complexity allow to discriminate between models of larger networks with many potential extensions.

Our third modularization method is designed to simplify the parameter identification of biomolecular networks with known structure and reaction kinetics. Finding suitable parameter sets for general biomolecular networks with high-dimensional parameter spaces is computationally challenging, such that most available methods are essentially restricted to small- or medium size models. Similar to our model discrimination method, we define the modules such that they can be (partially) insulated by the experimental data. In the first phase of our algorithm, we solve independent parameter identification tasks for each module that operate only on the comparatively low-dimensional parameter space of the respective module. The first phase results in inconsistent dynamics of the species in the modules' interfaces. In the second phase, the consistency between the modules is iteratively established, while simultaneously adjusting the parameters to preserve optimality. The use of experimental data as virtual inputs in the first phase can be interpreted to entrain each module with dynamics similar to those of the real network. We believe that this entrainment is responsible for the apparently better convergence properties of our modular approach as compared to its non-modular counterpart, as observed in the parameter identification of two biomolecular networks.

Our last method concerns the analysis of the interactions between layers–functional modules defined by sets of reactions instead of species–in fully identified dynamic models of biomolecular signaling or metabolic pathways. In our approach, the strength and the dynamics of the interactions as well as the cooperativity between layers is defined strictly symmetrically, and only depends on the environment in which the interactions take place. Thus, our mathematical framework allows to unambiguously quantify the interactions of functional units also in biomolecular networks for which the notion of modules being downstream or upstream of one another is inadequate, for example due to feedback mechanisms. Furthermore, we deduce relationships between the dynamics of different groups of layers conceptually similar to notions in information and probability theory. These relationships allow us to analytically deduce the dynamics of certain sub-networks from known layer dynamics without requiring additional numerical integrations. The resulting reduction in computational requirements makes it possible to, for example, apply our layering framework to analyze and score complex metabolic engineering strategies. We demonstrate this capability by quantifying the dynamic interactions between elementary flux modes in a kinetic model of glycolysis in *Saccharomyces cerevisiae*.

Zusammenfassung

Basierend auf der Strategie des „Teilens und Herrschens“ werden bei der Modularisierung mathematischer Modelle große und komplexe biomolekulare Netzwerke in mehrere kleinere, gegebenenfalls überlappende Teilnetzwerke aufgeteilt, welche als Module bezeichnet werden. Die Mehrzahl der bisher entwickelten Modularisierungsansätze behandelt die Identifikation funktionaler Module. Diese stellen Teilnetzwerke dar, die für die Erfüllung spezifischer zellulärer Aufgaben zuständig sind, beziehungsweise spezifischer Aufgaben innerhalb ihres übergeordneten Netzwerkes. Demgegenüber wurden vergleichsweise wenige Methoden zur Identifikation zweckgebundener Module entwickelt, welche Teilnetzwerke darstellen, die nur als Zwischenschritt für die Beantwortung von spezifischen Forschungsfragen dienen. Das Ziel der zweckgebundenen Modularisierung besteht darin, dass die einzelnen Module, sowie deren wechselseitige Interaktion, einfacher oder weniger rechenintensiv identifiziert oder analysiert werden können als das komplette Netzwerk. Wenn bei der Definition eines solchen Modularisierungsansatzes nicht nur die Struktur des Netzwerkes, sondern auch die beabsichtigte Verwendung der Module sowie bereits vorhandene Messdaten oder technische Messbeschränkungen beachtet werden, kann im Allgemeinen davon ausgegangen werden, dass die resultierenden Module für die Beantwortung spezifischer Forschungsfragen besser geeignet sind. In dieser Doktorarbeit veranschaulichen wir die Anwendbarkeit dieses Modularisierungsprinzips als generische Problemlösungsstrategie im Kontext der Systembiologie anhand von Modularisierungsmethoden, die speziell für verschiedene Phasen der Systemidentifikation und -analyse entwickelt wurden.

Strukturelle Identifikationsmethoden, die keine Kenntnis über den inneren Aufbau eines biomolekularen Netzwerkes voraussetzen, bieten sich an um die Struktur weitgehend unbekannter Netzwerke topologisch zu beschränken und damit den weiteren experimentelle Forschungsprozess anzuleiten. Unsere erste Methode dient der strukturellen Identifikation und hat zum Ziel, Signalrück- und Vorwärtskopplungen in biomolekularen Netzwerken zu erkennen. Da Systeme, die entweder eine Rück- oder eine Vorwärtskopplung beinhalten, ähnliche Eingangs-Ausgangs-Beziehungen zeigen können, induzieren wir charakteristische Systemdynamiken, um zwischen diesen Kopplungsarten unterscheiden zu können. Dabei nutzen wir eine nur bei nichtlinearen Systemen gegebene Beziehung zwischen der Phasenverschiebung und der Verstärkung eines oszillatorischen Eingangssignals durch das Netzwerk aus. Basierend auf den Ergebnissen der Analyse kann ein erstes grobes Modell des Netzwerkes konstruiert werden. Dieses besteht aus mehreren Modulen mit partiell charakterisierten dynamischen Eigenschaften, welche die Rückkopplungsschleife oder die Arme der Vorwärtskopplung, sowie den Rest des Netzwerkes repräsentieren.

Bei biomolekularen Netzwerken mit bereits weitgehend bekanntem Aufbau kann ein mechanistisches Kernmodell des Netzwerkes konstruiert werden, welches alle bestätigten und charakterisierten molekularen Interaktionen umfasst. Jedoch existieren für größere Netzwerke oft weitere hypothetische Interaktionen, d.h. potentielle Modellerweiterungen. Das Ziel der Modelldiskriminierung besteht darin, die Übereinstimmung der Kandidatenmodelle, den um jede mögliche Kombination von hypo-

thetischen Interaktionen erweiterten Kernmodell, mit experimentellen Daten über das biomolekulare Netzwerk zu ermitteln, um so die Existenz der einzelnen Erweiterungen zu validieren. Da jedoch die Anzahl der Kandidatenmodelle exponentiell mit der Anzahl zueinander kompatibler Erweiterungen zunimmt, ist der mögliche Einsatzbereich der meisten etablierten Modelldiskriminierungsmethoden trotz der steigenden Leistung moderner Rechensysteme beschränkt. Erschwerend kommt hinzu, dass die Validierung einer Erweiterung zu fehlerhaften Ergebnissen führen kann, hervorgerufen durch Ungenauigkeiten im Kernmodell, der mathematischen Beschreibung einzelner Modellerweiterungen, oder durch bereits fehlerhaft detektierte andere Erweiterungen. Um die Anzahl sowie die Dimensionen der Parameter- und Zustandsräume der Kandidatenmodelle zu reduzieren und um Fehlerfortpflanzung bei der Validierung der einzelnen Erweiterungen zu beschränken, modularisieren wir das Netzwerk mit dem Ziel, dass jedes einzelne Modul genau eine hypothetische Modellerweiterung beinhaltet. Wir benutzen dann direkt experimentelle Messreihen als virtuelle Eingangssignale für die Module, um diese voneinander zu isolieren und eine getrennte Simulation und Analyse zu ermöglichen. Dieses Vorgehen reduziert nicht nur signifikant die Anzahl der Kandidatenmodelle, sondern auch die Dimension ihrer Parameter- und Zustandsräume, so dass eine effiziente Modelldiskriminierung auch bei größeren Modellen mit einer Vielzahl potenzieller Erweiterungen ermöglicht wird.

Unsere dritte Modularisierungsmethode dient der Parameteridentifikation von biomolekularen Netzwerken mit bekannter Struktur und Reaktionskinetiken. Da die Bestimmung geeigneter Parameterwerte für generelle biomolekulare Netzwerke mit hochdimensionalen Parameterräumen rechentechnisch anspruchsvoll ist, ist die Anwendung vieler moderner Parameteridentifikationsmethoden auf kleine bis mittelgroße Netzwerk beschränkt. Ähnlich wie bei unserer Modelldiskriminierungsmethode definieren wir die Module eines Netzwerkes in Hinblick auf vorhandene Messdaten, so dass diese zumindest partiell voneinander isoliert und damit abschnittsweise einzelnd identifizierbar sind. In der ersten Phase unseres Algorithmus lösen wir für jedes Modul getrennte Parameteridentifikationsprobleme, welche durch die vergleichsweise niedrigdimensionalen Parameterräume der jeweiligen Module rechentechnisch effizienter lösbar sind als die Parameteridentifikation des nicht-modularisierten Netzwerkes. Da dies jedoch zu inkompatiblen Dynamiken in den Modulschnittstellen führt, stellen wir in der zweiten Phase unseres Algorithmus iterativ die Konsistenz zwischen den Moduldynamiken her, während wir gleichzeitig die Parameterwerte zum Zwecke des Erhalts der Optimalität anpassen. Die Benutzung von experimentellen Messreihen als virtuelle Eingangssignale in der ersten Phase unseres Algorithmus kann dahingehend interpretiert werden, dass die einzelnen Module mit Dynamiken ähnlich derer im realen biomolekularen Netzwerk angeregt werden. Unserer Einschätzung nach ist diese Anregung der Grund für die besseren Konvergenzeigenschaften unserer modularen Parameteridentifikationsmethode im Vergleich zu ihrem nicht-modularen Gegenstück, welche wir anhand zweier biomolekularer Beispielnetzwerke beobachten konnten.

Unsere letzte Methode dient der Analyse und Quantifizierung der Interaktionen zwischen Modulen, welche Anhand von Reaktionen definiert sind („*layer*“), die gemeinsam eine zelluläre Aufgabe oder eine Aufgabe im übergeordneten Netzwerk erfüllen. Dabei definieren wir die dynamischen Interaktionen, deren Stärke, sowie die Kooperativität zwischen Modulen streng symmetrisch, so dass mit unserer Methode auch die Interaktionen zwischen funktionalen Modulen in biomolekularen Netzwerken analysiert werden können, für welche es durch verschiedene Arten der Rückkopplung nicht gerechtfertigt erscheint von einer sequentiellen Anordnung der Module auszugehen. Des Weiteren leiten wir Zusammenhänge zwischen den Dynamiken verschiedener Module her, die entfernt an bekannte Konzepte aus der Wahrscheinlichkeits- und Informationslehre erinnern. Diese Zusammenhänge ermöglichen uns analytisch die Dynamiken bestimmter Teilnetzwerke ohne erneute numerische In-

tegrationen aus den Dynamiken bekannter Module herzuleiten. Die aus der gezielten Ausnutzung solcher Zusammenhänge resultierende Reduktion der benötigten Rechenleistung ermöglicht es, unsere Methode zum Beispiel für die Analyse komplexer genetischer Eingriffe in den Metabolismus und deren Erfolgsbewertung einzusetzen. Dies veranschaulichen wir Anhand der Analyse der Interaktionen von durch elementare Flussmoden („*elementary flux modes*“) in der Glykolyse von *Saccharomyces cerevisiae* definierten funktionalen Modulen.