Explorative Projekte im NGFN

Das Nationale Genomforschungsnetz (NGFN) ist als kooperative Netzwerkstruktur organisiert. Wissenschaftler aus sehr unterschiedlichen Fachrichtungen bringen ihr jeweiliges Know-How und die von ihnen angewandten Technologien in gemeinsamen Projekten zusammen. Neben den bereits bekannten NGFN-Komponenten der krankheitsorientierten Genomnetze und den Systematisch-Methodischen Plattformen, die auf Kernbereich- und Plattformtechnologien der ersten Förderphase aufbauen, wird es in Zukunft mit den Explorativen Projekten (EP) ein neues Strukturelement geben, in dem innovative Ideen zu methodischen Neu- und Weiterentwicklungen oder Ansätze zur Erforschung inhaltlich neuer krankheitsorientierter Gebiete bearbeitet werden. Diese Explorativen Projekte gewährleisten, dass auch Forschungsideen unterstützt werden, die sich noch in einem sehr frühen Stadium befinden. Ihre Ergebnisse sollen neue Technolgien und Anwendungsfelder für die Humangenomforschung erschließen und direkt in das NGFN einfließen, wo sie umgesetzt werden können. Einen ersten Einblick in diesen neuen Bereich geben die folgenden drei Artikel.

Netzwerke unter der Lupe

Systembiologie – Inferenz regulatorischer Netzwerke basierend auf Genexpressionsdaten
Christian Spieth, Nora Speer und Andreas Zell

Im Bereich der molekularen Genetik und der Systembiologie wurden in den letzten Jahren große Erkenntnisgewinne erzielt. Die dabei verwendeten Verfahren sind großteils noch nicht fest etabliert und die zur Verfügung stehenden Analysetools sind noch nicht auf die Art und Menge der biologischen Daten ausgelegt, welche mit modernen High-Throughput Techniken erzeugt werden können. Die meisten Algorithmen zur Analyse von Genexpressionsdaten, die bisher entwickelt wurden, ignorieren im Auswertungsprozess bestehendes biologisches Wissen über genregulatorische Mechanismen oder ontologische Klassifikation einzelner Gene. Das vorhandene Wissen wird oft erst im zweiten Analyseschritt hinzugezogen, wenn es zur eigentlichen Interpretation der gefundenen Cluster oder Klassifikationsergebnisse kommt. Dadurch bleibt eine Vielzahl an Informationen in den Daten, wie zum Beispiel die zeitliche Abhängigkeit der Expressionslevel, unberücksichtigt. Aus diesem Grund ist es notwendig, eine neue Klasse von Analysemethoden zu entwickeln, die regulatorische Netzwerke und deren quantitative Parameter direkt aus den Experimentdaten bestimmen können und im Gegensatz zu herkömmlichen Analysetechniken vorhandenes biologisches und medizinisches Wissen über regulatorische Mechanismen bereits bei der Analyse nutzen. Ziel des hier vorgestellten Explorativen Projekts zur automatisierten Inferenz von regulatorischen Netzwerken ist deshalb die Entwicklung von computergestützten Verfahren und Algorithmen, mit denen sich regulatorische Systeme, wie zum Beispiel genregulatorische Netzwerke oder metabolische Pathways, automatisch auf Experimentdaten basierend rekonstruieren lassen.

Interferenzprobleme

Die Inferenz regulatorischer Netze ist sehr schwierig, da das zugrunde liegende Problem auf Grund geringer Datenmengen im Verhältnis zur Anzahl der beteiligten Systemkomponenten hochgradig unterbestimmt ist. Konzeptionell gesehen besteht das Inferenzproblem in der Auswahl eines geeigneten mathematischen Modells, welches die regulatorischen Prozesse innerhalb der Zelle abbildet, und der Bestimmung der zugehörigen Modellparameter.

Modelle – mehr oder weniger real

Abhängig von der zur Verfügung stehenden Rechnerkapazität wurden in der Vergangenheit bereits einige mehr oder weniger realistische Modelle zur Simulation von regulatorischen Netzwerken vorgeschlagen.

Die einfachsten Modelle sind Random Boolean Networks, welche den Systemzustand auf boolesche, also wahrheitsbedingte Zustände reduzieren. In dieser Art der Modellierung kann ein Gen dementsprechend entweder „an" oder „aus" sein, d.h. exprimiert werden oder nicht. Obwohl diese Modelle effizient simuliert werden können und bereits eine Vielzahl von dynamischen Eigenschaften von Netzwerken an ihnen untersucht werden konnte, sind sie nicht geeignet für eine realistische Abbildung biologischer Vorgänge in einer Zelle, weil sie wie erwähnt nur zwei diskrete Zustände erlauben. Qualitative Modelle erweitern zwar das Modell der booleschen Netzwerke von lediglich zwei Zuständen auf n Zustände, diskretisieren jedoch immer noch die möglichen Zustände des Netzwerkes. Lediglich quantitative Modelle können die Biologie zufrieden stellend darstellen, da sie, statt diskretisierten Zuständen, kontinuierliche Übergänge zwischen verschiedenen Zuständen erlauben. Simulatoren für quantitative regulatorische Prozesse können in zwei Kategorien unterschieden werden. Zum einen deterministische Modelle, welche Metabolite durch eine Stoffkonzentration charakterisieren und eine gleichmäßige Verteilung des Stoffes innerhalb des Reaktionsvolumens annehmen.

In diesem Fall wird die Veränderung der Stoffkonzentration durch gewöhnliche Differentialgleichungen beschrieben. Zum anderen stochastische Modelle, bei denen einzelne Moleküle betrachtet werden, die sich zufällig im Reaktionsvolumen bewegen und miteinander reagieren. Die Anzahl der einzelnen Moleküle wird hier durch eine stochastische Zeitreihe beschrieben.

Parameterbestimmung

Bei der Bestimmung der Parameter wird versucht, das mathematische Modell so zu parametrisieren, dass es die experimentellen Daten am besten abbildet, d.h. bei der Simulation des Modells eine möglichst ähnliche Dynamik erzeugt. Methoden zur Inferenz regulatorischer Netzwerke hängen immer von den zugrunde liegenden Modellen ab. Zur eigentlichen Bestimmung der Parameter der eingesetzten mathematischen Modelle werden unter anderem direkte analytische Methoden und Gleichungslöser oder Evolutionäre Algorithmen (EA) eingesetzt. Letztere sind dafür bekannt, dass sie sich sehr gut für Parameteroptimierung eignen. EAs sind Mitglieder der Familie der stochastischen Suchalgorithmen und basieren auf dem Darwinschen Prinzip der Evolution mit Mutation und Selektion. Basierend auf EAs wurden in der Vergangenheit bereits einige Inferenzstrategien zur Parameterbestimmung an unserem Lehrstuhl entwickelt.

Verschiedene Phasen

Der Ablauf des gesamten Inferenzprozesses ist in verschiedene Teilphasen unterteilt und gliedert sich wie folgt:

i) Datensammlung: Für die spätere Inferenz werden Zeitverläufe der Expressionslevel von biologischen Prozessen benötigt. Diese können durch High-Throughput-Methoden, wie zum Beispiel DNA Microarrays, gesammelt werden. Weiterhin können auch metabolische Daten zur Modellierung von Signalwegen oder zur Identifizierung von Parametern in metabolischen Systemen für die Inferenz verwendet werden.

ii) Datenfilterung: Um die Komplexität und Größe der Datensätze zu reduzieren, werden in einem ersten Schritt diejenigen Daten, die statistisch gesehen nicht an dem zu untersuchenden biologischen Prozess teilnehmen, herausgefiltert. Dazu werden zum Beispiel die Gene, deren Expressionssignal unter einer festgelegten Schranke liegt, entfernt, da die Expressionsstärke im Rauschen des Microarrays liegt. Oder es werden die Gene gefiltert, deren Zeitverlauf sich nicht signifikant ändert und die daher vermutlich nicht an den Regulationsmechanismen beteiligt sind.

iii) Clustering: Zur weiteren Reduzierung der Dimensionalität des Inferenzproblems werden ähnliche Expressionsprofile geclustert. Nach Anwendung eines Clusterverfahrens kann das dadurch gefundene Clusterzentrum als Expressionsprofil angesehen werden, welches die anderen Profile im Cluster repräsentiert und dadurch die Dimensionalität nochmals verringert.

Die grundlegende Idee zur Reduktion der Dimensionalität durch Clusterverfahren ist, dass co-exprimierte Gene sehr wahrscheinlich auch co-reguliert sind. Diese Annahme trifft jedoch nicht notwendigerweise für alle Gene innerhalb eines Clusters zu. Deshalb ist es nicht ausreichend, nur mathematische Clusterverfahren zu verwenden. Dadurch, dass bekanntes biologisches Wissen in den Clusterprozess miteinfließt, können Clusteralgorithmen für Genexpressionsdaten verbessert werden. Dabei sind insbesondere Genfunktionen bereits bekannter co-regulierter Gene, Interaktionen von Genprodukten, zelluläre Lokalisation (z.B. Mitochondrien oder Ribosomen) oder bekannte Zusammenhänge in metabolischen oder genregulatorischen Netzwerken von großem Interesse.

Von unserem Lehrstuhl wurde bereits eine Clustertechnik vorgestellt, die biologisches und, im Speziellen, funktionelles Vorwissen basierend auf Gene Ontology (GO) nutzt, um bessere Cluster zu finden. Die Einbringung von Vorwissen wird dadurch möglich, dass die Informationen aus GO maschinell leicht weiter verarbeitet werden können und diese Ontologie in allen öffentlichen Datenbanken zu den verwendeten Standards zählt. Das neue Clusterverfahren kann auf der einen Seite von Biologen dazu verwendet werden, die zu untersuchenden Gene funktionell zu klassifizieren. Auf der anderen Seite bieten wir damit eine Möglichkeit, sowohl die Genexpressionswerte als auch die funktionellen Annotationen zu nutzen. Dadurch wird die Interpretation der gesammelten experimentellen Daten erleichtert, da coregulierte Gene innerhalb ihres biologischen Kontextes betrachtet werden. Das vorgestellte Verfahren kann somit als eigenständiges Expertensystem verwendet werden oder dazu benutzt werden, die Dimensionalität der Expressionsdaten sinnvoll zu reduzieren.

iv) Inferenz Abstrahiert kann das Verhalten einer Zelle als genregulatorisches Netzwerk von n Genen gesehen werden. Jedes Gen g_i produziert bei der genetischen Expression eine gewisse Menge an biochemischen Molekülen x_i und beeinflusst somit die Konzentrationen aller

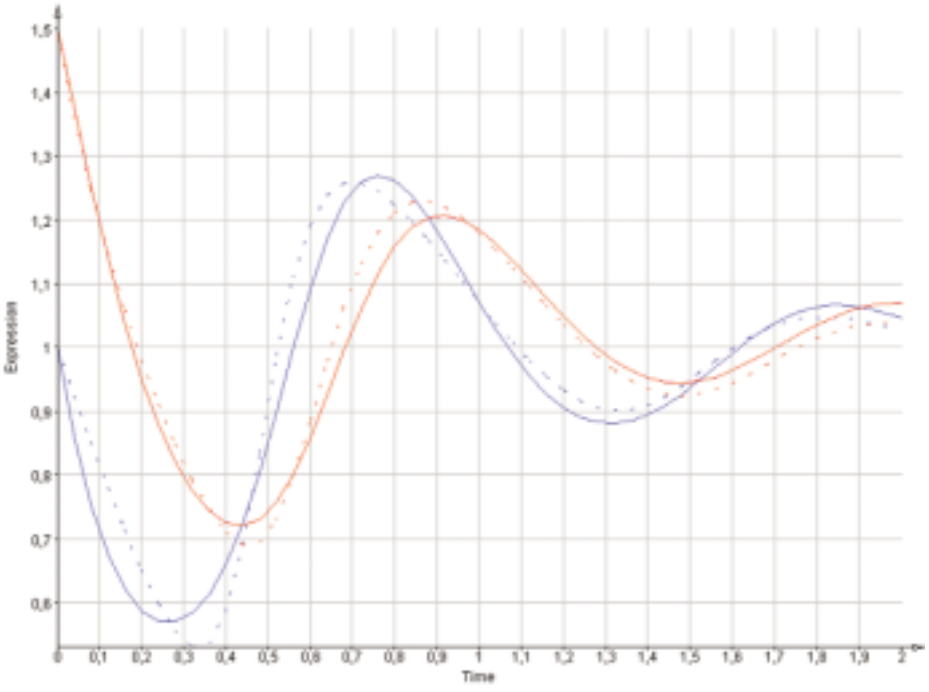

Abbildung 1: Dargestellt sind die gemessenen Experimentdaten (gestrichelte Linie) zusammen mit den Daten (durchgezogene Linie), die aus der Simulation des besten gefundenen mathematischen Modells stammen.

auftretenden Metaboliten innerhalb der Zelle:

$$x_i(t+1) = h(\vec{x}(t)), \quad \vec{x}(t) = (x_1 \mathrm{K}\ x_n)$$

wobei h eine Funktion der Änderung der einzelnen RNA-Level, abhängig von allen oder von einzelnen RNA-Konzentrationen, zum vorherigen Zeitpunkt darstellt.

Ziel der Inferenz ist es dann, Parameter der mathematischen Modelle zu finden, sodass diese Modelle simuliert über die Zeit eine ähnliche Dynamik zeigen, wie die Experimentdaten. Im Inferenzprozess werden die Parameter des Modells derart adaptiert, dass die Abweichungen zwischen Experimentdaten und simulierten Daten gegen Null gehen (siehe Abbildung 1). Dies kann durch direkte Heuristiken oder durch kombinatorische Optimierungsverfahren (Evolutionäre Algorithmen) erfolgen.

v) Validierung Ein wichtiger Teil des Inferenzprozesses ist die Verifikation der gefundenen Modelle. Bei Verwendung von künstlichen Datensätzen kann dieser Validierungsschritt in silico durchgeführt werden. Im Falle von echten biologischen Experimentdaten müssen zu diesem Zweck Biologen hinzugezogen werden, um die Qualität der errechneten Modelle bewerten zu können.

Werkzeug für Analysen Ziel des Forschungsprojektes ist die Entwicklung bzw. Weiterentwicklung eines Softwarewerkzeugs (JCell) für Analysen im Bereich der Systembiologie, welches metabolische und genregulatorische Netzwerke simulieren und ihre Parameter bestimmen kann. Damit sollen Wissenschaftler in die Lage versetzt werden, ihre experimentellen Genexpressions- oder Metabolismus-Daten auf einem wesentlich höheren Level zu untersuchen, als es bisher mit den verwendeten Standardverfahren möglich ist. So ermöglicht die Bestimmung der Topologie von genregulatorischen Netzwerken, Abhängigkeiten innerhalb des Genoms zu erkennen. Die Ergebnisse dieser neuen Art der Analyse ermöglichen es, strukturelle und quantitative Hypothesen über Teilnetze oder komplette regulatorische Systeme aufzustellen. Durch quantitative Bestimmung der kinetischen Modellparameter könnten Simulationen von genregulatorischen Prozessen und deren pathogen veränderten Varianten durchgeführt werden und dadurch möglicherweise Ansätze zur Therapie der zugrunde liegenden Krankheiten gefunden werden.

JCell beinhaltet bereits die meisten mathematischen Modelle, die zur Simulation von regulatorischen Systemen verwendet werden können, wie zum Beispiel lineare Gewichtsmatrizen, pseudolineare Modelle und S-Systeme. Daneben wurden bereits spezielle Inferenzstrategien für die Rekonstruktion von regulatorischen Netzwerken entwickelt, die zusätzlich zu den experimentellen Daten auch biologisches Wissen über Pathways aus öffentlichen Datenbanken wie KEGG zur Parameterbestimmung nutzen (siehe Abbildung 2).

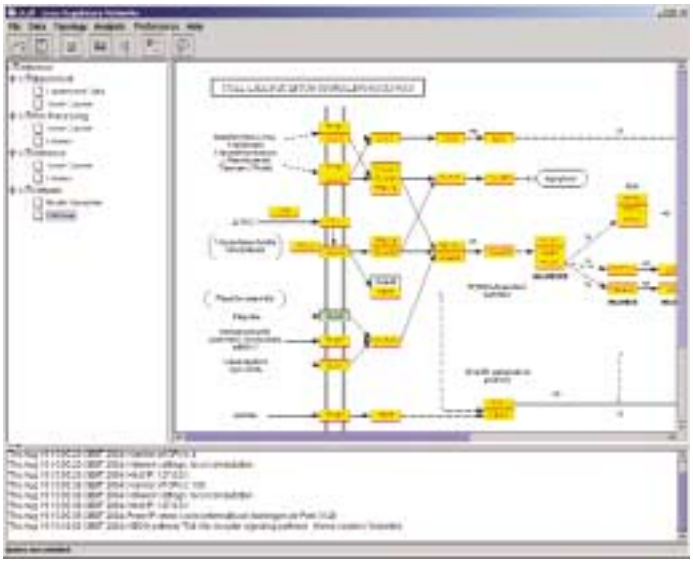

Abbildung 2: Hauptfenster von JCell mit einer schematischen Abbildung eines Signalwegs, die von KEGG importiert wurde.

Kontakt

Christian Spieth
Zentrum für Bioinformatik Tübingen
Universität Tübingen
Sand 1 · 72076 Tübingen
www-ra.informatik.uni-tuebingen.de/software/JCell

Glossar

Algorithmus ist ein prozeduraler Arbeitsvorgang, bei dem durch Wiederholung einfacher (Rechen)vorgänge auch komplexere Probleme lösbar werden.

Heuristik dient der methodischen Gewinnung neuer Erkenntnisse, mit Hilfe der Erfahrung. Sie beruht in der künstlichen Intelligenz meist auf Faustregeln bzw. Algorithmen. Heuristische Verfahren nützen häufig die sehr spezielle Struktur von Problemen aus, damit sie zu effizienten Verfahren werden und somit im Gegensatz zu exakten Verfahren schnell zulässige Lösungen finden.

Inferenz bedeutet hier die Rekonstitution von dynamischen Systemen, ausgehend von gemessenen Daten (z.B: Expressionslevel) bzw. die Bestimmung der Abhängigkeiten zwischen den Systemkomponenten (z.B. Gene) eines mathematischen Modells des betrachteten Netzwerks.

Gene Ontology ermöglicht es, Genprodukte methodisch und eindeutig einer oder mehreren Funktionsklassen zuzuordnen. Drei strukturierte und kontrollierte Vokabulare stellen Beschreibungen bezüglich

- der biologischen Prozesse, an denen ein Protein beteiligt sein kann (biological processes),
- des Zellkompartementes, in dem sich ein Protein befindet (cellular components), sowie
- der Funktion des Proteins auf molekularer Ebene (molecular functions) zur Verfügung, und zwar unabhängig vom Organismus.

Hierarchisch gliedern sich diese drei Ontologien in immer detailliertere Beschreibungen.