Vom Suchen und Finden funktioneller Module in biologischen Netzwerken: Ein neuer Ansatz zur integrierten Netzwerkanalyse in der Systembiologie

Marcus Dittrich

Lehrstuhl für Bioinformatik
Julius-Maximilians-Universität Würzburg
marcus.dittrich@biozentrum.uni-wuerzburg.de

Abstract: Die funktionelle Analyse großer Interaktionsnetzwerke hat sich in den vergangenen Jahren zu einem wichtigen Kerngebiet der Bioinformatik und Systembiologie entwickelt. Hier wird ein neuer, verbesserter Ansatz zur Identifizierung funktioneller Module in großen biologischen Netzwerken präsentiert. Der vorgestellte Ansatz wurde anhand eines gut untersuchten Satzes von Microarray- und Survival-Daten von Lymphoma-Patienten im Kontext von Protein-Protein-Netzwerken entwickelt und getestet. Zur Kombination der funktionellen Daten mit dem Netzwerk wurde eine flexible Gewichtungsfunktion hergeleitet und damit jedes Protein im Netzwerk gewichtet. Dies ermöglicht es nun, die Suche nach funktionellen Modulen im graphentheoretischen Sinne als Suche nach dem Subgraphen mit maximalem Gewicht zu formulieren. Durch die explizite Modellierung des Signal- und Rauschanteils liefert die Methode zugleich eine quantitative Abschätzung des Informationsgehaltes und erlaubt es, die erwartete Anzahl falsch-positiver Knoten im resultierenden Subnetzwerk zu kontrollieren. Die Anwendung dieses Algorithmus auf das Netzwerk zeigt, dass sich damit bekannte medizinisch relevante Module wiederfinden und ergänzen lassen. Intensive Simulationsexperimente belegen, dass dieser exakte Ansatz deutlich bessere Ergebnisse liefert als bereits beschriebene heuristische Methoden.

1 Netzwerke in der Systembiologie

In der biomedizinischen Forschung haben systematische Untersuchungen zellulärer Systeme in den vergangenen Jahren zunehmend an Bedeutung gewonnen. Die großen Fortschritte der experimentellen Technologien, insbesondere Weiter- und Neuentwicklungen im Bereich der Hochdurchsatzverfahren (z.B. Microarray, Massenspektrometrie, Sequenzierungstechnologien der neuen Generation), ermöglichen heutzutage die Analyse des gesamten zellulären mRNA-Pools (Transkriptom) sowie der Gesamtheit aller zellulären Proteine (Proteom) und aller Protein-Protein-Interaktionen (Interaktom). In diesem Zusammenhang erhält die Bioinformatik eine immer größere Relevanz bei der Verarbeitung und Analyse experimenteller Daten [DBM+08, DBM+05]. Um aus diesen Daten jedoch eine verwertbare Information zu extrahieren und damit letzlich auch neue Einsichten in das komplexe Zusammenspiel verschiedener zellulären Prozesse auf der Ebene des Gesamtsystems zu gewinnen, sind geeignete Analyseverfahren von herausragender Bedeutung. Eine

besondere Herausforderung ist die Integration der oft heterogenen Daten, die aus verschiedenen Studien mit unterschiedlichen Technologien gewonnen wurden und verschiedene, oftmals komplementäre Aspekte eines zellulären Systems beschreiben. Dies bildet heute eine wichtige Domäne der Systembiologie und angewandten Bioinformatik. Im Gegensatz zum klassischen reduktionistischen Ansatz, bei dem einzelne Teile des Systems (Gene, Proteine) getrennt untersucht werden, versucht die Systembiologie zelluläre Systeme in ihrer Ganzheit zu betrachten und zu modellieren. Dies folgt dem Paradigma, dass sich bei der Untersuchung komplexer Systeme die Funktion nicht durch die isolierte Betrachtung einzelner Komponenten, sondern letztlich nur durch das Verständnis des Zusammenspiels der beteiligten Komponenten und Subsysteme erschließt.

2 Neue Methoden zur Integrierten Netzwerkanalyse

Klassische Verfahren der Netzwerkanalyse fokussieren vor allem auf eine Untersuchung der Struktur, also der Topologie des zu Grunde liegenden Graphen. Obwohl viele dieser Methoden auch Anwendungen bei der Analyse biologischer Netzwerke gefunden haben, z.B. zur Detektion von potentiellen Proteinkomplexen, reicht zur Beantwortung zahlreicher biologischer Fragestellungen die in der Struktur des Netzwerks enthaltene Information allein nicht aus. Hierzu muss man sich den Modellcharakter des Graphen noch einmal vor Augen führen: Im hier betrachteten Kontext von Protein-Protein-Interaktionsnetzwerken repräsentieren die Kanten potentielle Interaktionen zwischen den Proteinen (Knoten). Diese enthalten jedoch keinerlei Informationen über zustandsabhängige Aktivierungen und Deaktivierungen unterschiedlicher Teilnetzwerke beziehungsweise Module. Um indes die biologisch bedeutsamen Fragestellungen nach zustandsabhängigen aktivierten Subnetzwerken zu beantworten, ist es notwendig das reduzierte Modell eines statischen Netzwerkes als einfachen Graphen zu erweitern und die Information über Systemzustände in das Modell zu integrieren. Dazu ist es nötig (i) entsprechende Daten zu erheben, (ii) eine adäquate Abbildung dieser Information im Modell zu formulieren und (iii) eine klare Beschreibung der biologischen Fragestellung zu finden um diese algorithmisch lösen zu können. Der hier präsentierte Ansatz zur integrierten Netzwerkanalyse formuliert genau dieses Problem neu und erlaubt so durch ein neues statistischen Modell zur Integration externer Daten in den Netzwerkkontext eine algorithmische Formulierung zur optimalen Lösung des Problems. Als konkretes Beispiel werden hier mRNA-Expressionsdaten (Microarray) und Survivaldaten von Lymphoma-Patienten mit zwei unterschiedlichen Subtypen (ABC und GCB) [RWC+02] im Kontext von Protein-Protein-Interaktionsnetzwerken analysiert. Die folgende Darstellung fokussiert vor allem auf die statistische Modellierung sowie auf die Evaluation des vorgestellten exakten Ansatzes im Vergleich zu bislang beschriebenen heuristischen Verfahren. Eine weiterführende Darstellung der biologischen Daten und insbesondere des algorithmischen Vorgehens ist in [DKR+08] zu finden.

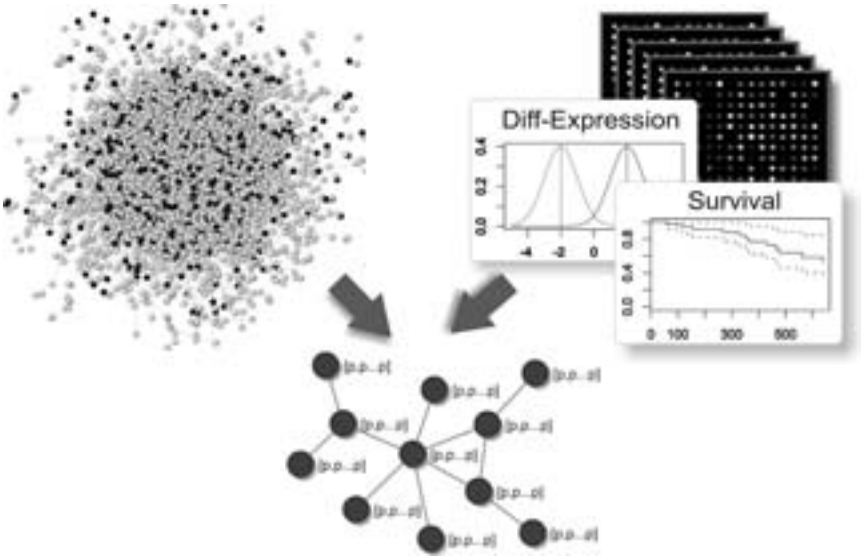

Abbildung 1: Erstellung des Integrierten Netzwerkes: Die Knoten des Gesamtnetzwerkes (humanes Interaktom mit 9.392 Knoten und 36.504 Kanten) wurden mit dem Vektor aus p-Werten der Microarray- und Survivaldaten annotiert (gefärbte Knoten; weiße Knoten bezeichnen Gene, für die keine Expressionswerte gemessen wurden). Die größte Zusammenhangskomponente (2.034 Knoten und 8.399 Kanten) ist dann der Ausgangsgraph für die Suche der Module.

2.1 Integration der funktionellen Daten in das Netzwerk

Die Problem der Identifizierung funktioneller Module lässt sich in dem hier gegebenen Kontext grob in zwei Subprobleme zerlegen. Das erste umfasst die Definition einer geeigneten Gewichtungsfunktion (Score), die die externe Information auf den Knoten des Netzwerks abbildet und so die zu suchenden Netzwerkregionen gewichtet. Das zweite Subproblem ist dann die Formulierung eines Algorithmus zur Suche nach maximal gewichteten (also optimal scorenden) Subgraphen. In unserem Kontext verwenden wir die Signifikanzwerte (p-Werte) aus dem T-Test der Expressionsanalyse als Maß für die differentielle Expression zwischen den beiden Subtypen, sowie die p-Werte des Regressionskoeffizienten aus der Cox-Regression als Maß für die Risikoassoziation der einzelnen Gene. Diese Details werden hier nicht weiter erläutert, für das weitere Verständnis genügt es, sich klar zu machen, dass durch die Signifikanzwerte die funktionelle Information (differentielle Expression und Risikoassoziation) quantitativ erfasst wird. Unser Problem besteht nun darin, aus den Signifikanzwerten der Netzwerkknoten (in diesem Falle zwei) eine Gewichtungsfunktion für den Suchalgorithmus zu formulieren (Abbildung 1).

Nachdem jeder Knoten im Netzwerk mit den dazugehörigen Signifikanzwerten annotiert wurde, muss jeder dieser Vektoren aus p-Werten in ein geeignetes Knotengewicht transformiert werden. Der hier vorgestellte Ansatz geht dafür im Wesentlichen in drei Schritten vor: Als erstes werden die p-Werte durch eine Ordnungsstatistik in einen einzelnen p-Wert

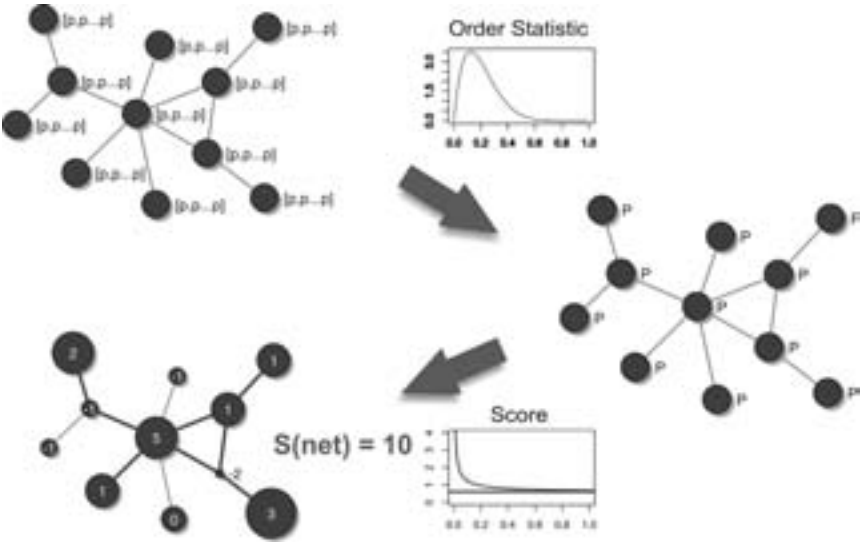

Abbildung 2: Definition der Gewichtungsfunktion für die Knoten. Als erstes werden die Signifikanzwerte der einzelnen Analysen über eine Ordnungsstatistik zusammengezogen. Anschließend wird die Verteilung der aggregierten p-Werte als Mischungsverteilung einer Signal- und Rauschkomponente modelliert und darauf aufbauend eine Knotengewichtung definiert. Da der Score additiv, ist ergibt sich hier der Score des optimalen Subnetzwerkes (dicke Kanten) als Summe der Knotenscores (10).

zusammengezogen. In einem zweiten Schritt wird die Verteilung der aggregierten p-Werte als Mischungsverteilung einer Rausch- und Signalkomponente modelliert. Im letzten Schritt wird dann die Gewichtungsfunktion basierend auf dem Logarithmus des Likelihood des Verhältnisses aus Signal- und Rauschkomponente errechnet (Abbildung 2).

Um für jeden Knoten einen p-Wert zu erhalten, berechnen wir die Ordnungsstatistik über den Vektor der Signifikanzwerte jedes einzelnen Knotens. Seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen aus der gleichen Verteilung (iid), dann bezeichnet $X_{(i)}$ die i kleinste Beobachtung. So ist also $X_{(1)} = \min_i X_i$ und $X_{(n)} = \max_i X_i$. Die Dichte der i kleinsten Beobachtung ist nun wie folgt gegeben

$$f_{X_{(i)}}(x) = \frac{n!}{(n-i)!(i-1)!} f(x) F(x)^{i-1} (1 - F(x))^{n-i} \tag{1}$$

für $i \in 1, \ldots, n$. Für eine detaillierte Herleitung siehe [DKR$^+$08]. Hieraus lässt sich eine zusammenfassende Statistik über die p-Werte aus der $i$ten Ordnungsstatistik ableiten. Da die p-Werte unter der Nullhypothese gleichverteilt sind, erhält man mit Gleichung (1) für $f_X(x) = 1$ und $F_X(x) = x$ die Dichte der i kleinsten Beobachtung (p-Wert) wie folgt:

$$X_{(i)} \sim \frac{n!}{(n-i)!(i-1)!} \quad 1 \quad x^{i-1}(1-x)^{n-i} \qquad 0 \leq x \leq 1 \tag{2}$$

und kann damit den Vektor von p-Werten in einem einzelnen Signifikanzwert für jeden Knoten im Netzwerk zusammenfassen.

Diese Verteilung der aggregierten Signifikanzwerte kann man nun als eine Mischungsverteilung betrachten, die aus einer Signal- und einer Rauschkomponente zusammengesetzt ist. Ausgehend von der beobachteten Verteilung der Signifikanzwerte (Abbildung 3) lässt sich die Signalkomponente als Beta(a,1) modellieren, während die Rauschkomponente durch die Gleichverteilung beschrieben werden kann, die auch der Beta(1,1) Verteilung entspricht. Dabei ist die Beta(a,b) gegeben durch

$$f(x) \quad = \quad \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, x^{a-1}(1-x)^{b-1},$$

wobei $\Gamma(\cdot)$ die Gammafunktion bezeichnet. Damit reduziert sich die Verteilung der p-Werte auf

$$f(x \mid a, \lambda) = \lambda + (1-\lambda)ax^{a-1} \quad \text{für} \quad 0 < x \leq 1; 0 < a < 1$$

mit dem Mischungsparameter λ und dem Formparameter a der Beta-Verteilung. Für die Beobachtungen $x = x_1 \ldots x_n$ ist dann die Log-Likelihood definiert als

$$\log \mathcal{L}(\lambda, a; x) \quad = \quad \sum_{i=1}^{n} \log(\lambda + (1-\lambda)ax_i^{a-1}) \ .$$

Daraus lässt sich nun ein Maximum-Likelihood-Schätzer für die beiden unbekannten Parameter angeben.

$$[\hat{\lambda}, \hat{a}] = \underset{\lambda,a}{\operatorname{argmax}}\, \mathcal{L}(\lambda, a; x).$$

Beide Parameter können durch numerische Optimierung bestimmt werden. Für den Datensatz der Lymphoma-Patienten erhielten wir 0.536 für den Mischungsparameter λ und 0.276 für den Parameter a der Beta-Verteilung.

Eine geeignete Gewichtungsfunktion sollte additiv sein, so dass sich der Gesamtscore des Netzwerkes als Summe der einzelnen Knotenscores ergibt. Inbesondere sollte das Signal eine positive Gewichtung erhalten, wohingegen das Rauschen negativ gewichtet werden sollte. In Anlehnung an den Likelihood-Quotienten-Test kann man mit Hilfe der im Mischungsmodell definierten Verteilungen eine Gewichtungsfunktion mit den gewünschten Eigenschaften definieren. Hierbei gibt Beta(1,a) die Likelihood der Beobachtungen unter dem Signalmodell und Beta(1,1) die Likelihood der Beobachtungen unter dem Rauschmodell an. Als Log-Likelihood der Signal- und Rauschkomponente erhält man so

$$S(x) = \quad \log\left(\tfrac{\mathrm{Beta}(a,1)(x)}{\mathrm{Beta}(1,1)(x)}\right) \quad = \log(a) + (a-1)\log(x) \ .$$

Aus diesem Modell kann man nun ähnlich wie bei den klassischen statistischen Testverfahren ein Signifikanzniveau ableiten. Wie in [PM03] beschrieben, erlaubt das Modell der Mischungsverteilung eine Abschätzung des Anteils falsch-positiver Klassifikationen (FDR: False Discovery Rate). Daraus kann man nun einen Schwellenwert $\tau(\mathrm{FDR})$ für die p-Werte berechnen und damit eine flexibel anpassbare Gewichtungsfunktion definieren:

$$\begin{aligned} S^{\mathrm{FDR}}(x) \quad &= \quad \log\left(\frac{ax^{a-1}}{a\tau^{a-1}}\right) \\ &= \quad (a-1)\left(\log(x) - \log(\tau(\mathrm{FDR}))\right) \ . \end{aligned} \tag{3}$$

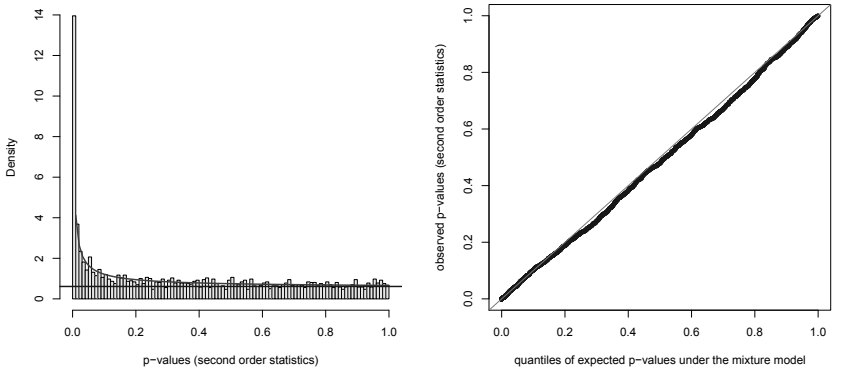

Abbildung 3: Linke Seite: Das verwendete Modell der Mischungsverteilung (Beta-Uniform Mixture Model) deckt sich hervorragend mit der empirischen Verteilung. Die Parameter des gefitteten Modells sind $a = 0.276$ and $\lambda = 0.563$. Das Histogramm der beobachteten p-Werte ist konsistent mit der modellierten Dichtefunktion (rote Linie). Die blaue Linie beschreibt den Anteil der unter dem Nullmodell (Rauschen) erwarteten p-Werte. Rechte Seite: Die diagonale Gerade im Quantile-Quantile-Diagramm der modellierten und beobachteten Verteilung verdeutlicht die hohe Präzision der Übereinstimmung.

Diese Gewichtungsfunktion unterscheidet sich letztlich nur durch die additive Konstante τ von dem Log-Likelihood-Quotienten. Damit werden p-Werte unterhalb dieses Schwellenwertes (τ) der Signalkomponente zugordnet und positiv gewichtet, während p-Werte oberhalb davon der Rauschkomponente zugeordnet und negativ gewichtet werden. Da die einzelnen Beobachtungen x_i unter dem Nullmodell unabhängig sind, lässt sich der Score für ein gegebenes (Sub-)Netzwerk als Summe der einzelnen Knotenscores berechnen. Damit skaliert der Erwartungswert des Netzwerkscores $S_{\text{net}}^{\text{FDR}}$ linear mit der Anzahl der Knoten im Netzwerk. Und so wird auch die wichtige Rolle der FDR offenbar: Die geeignete Wahl dieses Parameters adjustiert den Nullpunkt des Scores und stellt damit die Lokalität der Lösungen sicher.

2.2 Anwendung und Evaluation des Ansatzes

Mit der oben beschriebenen statistischen Modellierung haben wir ausgehend von den Signifikanzwerten heterogener Analysen eine integrierte Gewichtungsfunktion hergeleitet, die es jetzt erlaubt nach Regionen von hoher Signifikanz, also mit starkem Signal, zu suchen. Insbesondere aufgrund der Additiviät des Scores kann damit das Problem der Identifizierung funktioneller Module als Suche nach zusammenhängenden maximal gewichteten Subgraphen in einem Knoten-gewichteten Graphen formuliert werden:

Problem 1 (Maximum-Weight Connected Subgraph Problem, MWCS)**.** *Für einen gegebenen ungerichteten Knoten-gewichteten Graphen $\mathcal{G} = (V, E, w)$ mit den Gewichten $w : V \to \mathbb{R}$, finde einen zusammenhängenden Subgraphen $\mathcal{H} = (V_\mathcal{H}, E_\mathcal{H})$ in $\mathcal{G}$, $V_\mathcal{H} \in V$,*

$E_{\mathcal{H}} \in E$ der die folgende Zielfunktion maximiert

$$score(\mathcal{H}) = \sum_{v \in V_{\mathcal{H}}} w(v) \ .$$

Es konnte gezeigt werden, dass dieses Problem *NP* vollständig ist [IOSS02, DKR$^+$08]. Ein auf mathematischer Programmierung basierender Algorithmus zur Suche der optimalen Lösungen wird in [DKR$^+$08] ausführlich beschrieben. Eine Lösung für das hier untersuchte Lymphoma-Netzwerk ist in Abbildung 4 dargestellt. Basierend auf Signifikanzwerten aus dem einseitigen T-Test werden durch die Gewichtungsfunktion Gene, die im GCB-Subtyp relativ überexprimiert sind, hoch gewichtet. Interessanterweise werden in diesem Modul die Gene der medizinisch bekannten "Bystander Signatur" (FN1, SPARC, MMP9, CTSK, ITGA5, und ITGB5) miteinander verknüpft.

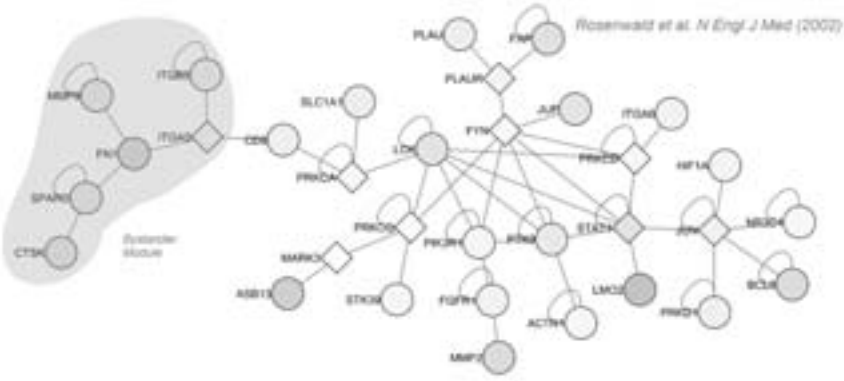

Abbildung 4: Graphische Darstellung einer optimalen Subnetzwerk-Lösung. Dieses Modul repräsentiert vernetzte Gene, die im GCB-Subtyp gemeinsam hochreguliert sind (p-Werte aus einseitigem T-Test). Die unterschiedlichen Farbschattierungen visualisieren die Stärke der relativen Überexpression im GCB-Subtyp.

In der Literatur sind heuristische Verfahren zur Suche von funktionellen Modulen in Protein-Protein-Interaktionsnetzwerken beschrieben. Der wahrscheinlich bekannteste ist der erste von [IOSS02] beschriebene Simulated Annealing Algorithmus; er wurde hier zum Vergleich herangezogen[1]. Zur Evaluation wurden simulierte Signalmodule von biologisch relevanter Größe (46 und 143 Knoten) im Netzwerk versteckt. Die p-Werte der Signalkomponente wurden gleichverteilt zwischen 0 und 10^{-3} gezogen, während die restlichen Knoten (Rauschkomponente) gleichverteilt zwischen 0 und 1 gezogen wurden. Für den exakten Ansatz wurden Gewichtungen über einen weiten FDR-Bereich ausgewertet. Da das heuristische Verfahren keinen vergleichbaren Parameter zur Adjustierung bietet, wurde hier der Algorithmus drei Mal rekursiv auf die jeweiligen Lösungen angewandt. Die *Recall-Precision* Diagramme (Abbildung 5) zeigen deutlich die Überlegenheit des exakten

[1] Viele der nachfolgenden Heuristiken basieren stark auf diesem Ansatz, insbesondere auf der Gewichtungsfunktion, die sich von der hier entwickelten grundlegend unterscheidet. Dort werden anstelle einzelner Knoten ganze Subnetzwerke in jedem iterativen Schritt neu gewichtet

Ansatzes: Während er sogar perfekte Lösungen liefert (*Precision* und *Recall* von 1), erreicht keine der heuristischen Lösungen die obere rechte Ecke des Diagramms.

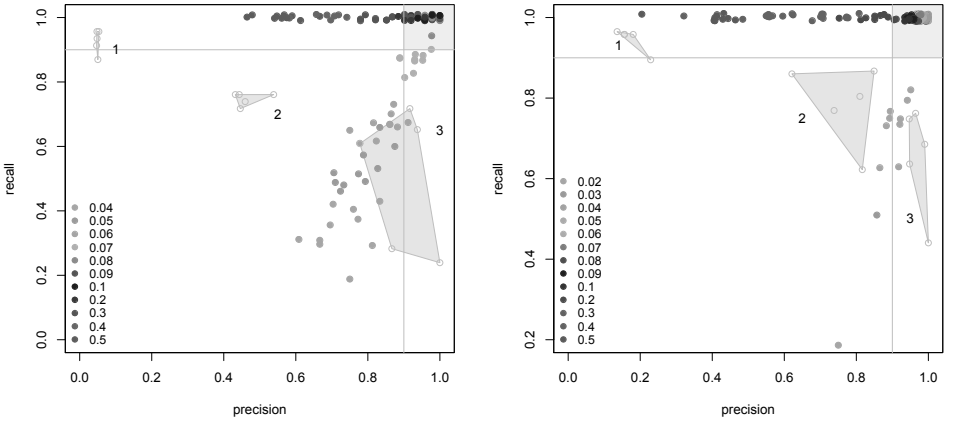

Abbildung 5: Darstellung des *Recall* (Sensitivität) gegen die *Precision* (positiver Vorhersagewert) für mehrere Lösungen über einen weiten FDR-Bereich. Bei den heuristischen Lösungen [IOSS02] ist jeweils die konvexe Hülle der Lösungen gegeben. Hier wurde der Algorithmus jeweils drei Mal rekursiv für fünf verschiedene Simulationen angewandt. Es wurden zwei verschiedene Signalgrößen evaluiert (46, links und 143 rechts). Es zeigt sich deutlich, dass der hier vorgestellte exakte Algorithmus der heuristischen Methode überlegen ist: Im Gegensatz zu dem exakten Ansatz erreicht keine der heuristischen Lösungen den Bereich hoher *Precision* und *Recall* oben rechts im Graphen.

3 Ausblick

Die Anwendungsmöglichkeiten des vorgestellten Verfahrens reichen deutlich über das hier gezeigte Beispiel hinaus. Im Prinzip lassen sich jegliche Analyseresultate, die sich als Signifikanzwerte ausdrücken lassen, in unser Modell integrieren. Neue und erweiterte Anwendungsmöglichkeiten werden sich natürlicherweise aus der stetig steigenden Menge und zunehmenden Qualität quantitativer Proteom- und Phosphoproteom-Daten ergeben. Anzumerken ist allerdings, dass eine große Herausforderung bei der Analyse dieser komplexen Modelle aus heterogenen Daten und Stukturen in der Interpretation der gewonnenenen Ergebnisse selbst liegt. Im Gegensatz zu homogenen Daten, die sich mit klassischen Verfahren der linearen Algebra oder Graphentheorie analysieren lassen, erfordert diese Form der integrativen Analyse sowohl ein konzeptionelles Verständnis der zugrundeliegenden mathematischen und algorithmischen Verfahren, als auch notwendigerweise Einsichten in die biologische Semantik der Daten und Modelle. Man darf jedoch erwarten, dass in Zukunft solche integrativen Verfahren eine zunehmend wichtige Rolle spielen werden, da nur dadurch die große Menge an heterogenen Daten sinnvoll zu kombinieren ist und verschiedene funktionelle Aspekte eines biologischen Systems gleichzeitig in einem Modell abgebildet werden können.

Literatur

[DBM+05] Marcus Dittrich, Ingvild Birschmann, Silke Mietner, Albert Sickmann, Ulrich Walter und Thomas Dandekar. Understanding platelets. Lessons from proteomics, genomics and promises from network analysis. *Thromb Haemost*, 94(5):916–25, 2005. 0340-6245 (Print) Journal Article Review.

[DBM+08] Marcus Dittrich, Ingvild Birschmann, Silke Mietner, Albert Sickmann, Ulrich Walter und Thomas Dandekar. Platelet protein interactions: map, signaling components, and phosphorylation groundstate. *Arterioscler Thromb Vasc Biol*, 28(7):1326–1331, Jul 2008.

[DKR+08] Marcus Dittrich, Gunnar W Klau, Andreas Rosenwald, Thomas Dandekar und Tobias Müller. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics*, 24(13):i223–i231, Jul 2008.

[IOSS02] Trey Ideker, Owen Ozier, Benno Schwikowski und Andrew F Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 Suppl 1:S233–S240, 2002.

[PM03] Stan Pounds und Stephan W Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):1236–1242, Jul 2003.

[RWC+02] Andreas Rosenwald, George Wright, Wing C Chan, Joseph M Connors, Elias Campo, Richard I Fisher, Randy D Gascoyne und H. Konrad Muller-Hermelink et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med*, 346(25):1937–1947, Jun 2002.

Marcus Dittrich hat von 1995 bis 1999 in Tübingen Humanmedizin studiert und dann an der Case Western Reserve University in Cleveland (Ohio) in der experimentellen Immunologie geforscht. Das Praktische Jahr seiner klinischen Ausbildung leistete er an der Universitätsklinik Zürich. Im Jahr 2003 wechselte er nach Würzburg, wo er die ärztliche Approbation erwarb. Nach einem Aufbaustudium Biologie (MD/PhD Stipendium) promovierte er sowohl in der Bioinformatik, als auch in der Medizin. Als Habilitant am Lehrstuhl für Bioinformatik der Universität Würzburg beschäftigt er sich heute vor allem mit der Entwicklung neuer Methoden zur integrierten Datenanalyse in der Systembiologie. Mit seinen Co-Autoren wurde er im Jahr 2008 für eine neuartige Methode zur Netzwerkanalyse mit dem ”Outstanding Paper Award” auf der ISMB 2008 ausgezeichnet.