Rechnergestützte Suche nach Korrelationen in komplexen Datensätzen der Biowissenschaften

Stephan Heymann, Katja Tham, Peter Rieger and Johann-Christoph Freytag

Lehrstuhl für Datenbanken und Informationssysteme,
Institut für Informatik, Humboldt Universität zu Berlin,
Unter den Linden 6, D-10099 Berlin.
E-mail: {heymann, tham, rieger, freytag}@dbis.informatik.hu-berlin.de

Abstract: Umfangreiche Datensammlungen, wie sie im Ergebnis molekularbiologischer Hochdurchsatzexperimente entstehen, sind zu einer wichtigen Triebfeder der funktionalen Genomforschung geworden. Eine wesentliche Anforderung an die Informationstechnologie besteht darin, Werkzeuge bereitzustellen, die es ermöglichen, implizite Zusammenhänge in Datensätzen unterschiedlicher methodischer Herkunft aufzudecken. Diese Arbeit beschreibt eine Vorgehensweise, die von der Aufbereitung der Daten bis zur Visualisierung der Ergebnisse einen integrierten Lösungsansatz bereitstellt. Im ersten Schritt werden die Experimentaldaten so genannten Datenkategorien zugeordnet. Entsprechend dieser Systematik erfolgt die Integration in ein relationales Datenbank-Managementsystem. In einer Aufbereitungsphase werden die Daten nach den Anforderungen des Anwenders in ein komplexes Analysemodell überführt. Dieses dient dem Auffinden unbekannter Zusammenhänge, als Eingangsformat für die grafische Visualisierung sowie der interaktiven Navigation. Im Rahmen dieser Vorgehensweise wurde ein Verfahren entwickelt, das es ermöglicht, qualitative Aussagen über Korrelationen in Daten unterschiedlicher Kategorien abzuleiten.

1 Einleitung

Die Erkennung von Erklärungsmustern für komplexe Lebensvorgänge aus Genomdaten ist eines der zentralen Anliegen der Bioinformatik. Es existiert jedoch keine einheitliche "Genomtheorie", die erklärte, *wie* das Erbgut eine Spezies und einzelne Individuen befähigt zu leben, zu agieren, zu reagieren, sich anzupassen, zu entwickeln oder auszusterben.

Historisch gewachsenes Domänenwissen
Die Facetten des komplexen funktionellen Wechselspiels der Gene und ihre wechselseitigen Abhängigkeiten werden im Ergebnis eines ganzen Arsenals adäquater Experimentaltechniken und computergestützter Untersuchungen mosaikartig zusammengetragen. Die Erkenntnisse bleiben jedoch in den Grenzen der Aussagefähigkeit der jeweiligen Methode gefangen. Demzufolge hängt der Erkenntniszuwachs in hohem Maße von Technologien ab, die einen Forscher beim Aufdecken impliziter Korrelationen zwischen Datensätzen unterschiedlichen methodischen Ursprungs unterstützen. Einen besonderen Schwerpunkt bildet dabei die Auswertung von Messreihen aus Hochdurchsatzverfahren, deren Querbezüge dann offenbar werden, wenn ein experimenteller Datensatz zu einem zweiten – oder zu mehreren – in Bezug gesetzt und ausgewertet wird.

Der gemeinsame Kern heterogener Experimentaldatensätze
Abstrakt betrachtet baut die Analyse von Experimentaldatensätzen auf folgenden Überlegungen auf: Gegeben sei ein Sachverhalt, der n Objekte O einer Klasse überspannt. Die Menge $M = \{O_i\}$, i=1, ...,n werde k methodisch unterschiedlichen Untersuchungen unterzogen, wodurch jeweils Aussagen zu Untermengen $M^{(k)} \subseteq M$ gewonnen werden. Untermengen deswegen, weil im Regelfalle nicht alle Elemente aus M in die Untersuchungen einbezogen sind oder aber die Untersuchungsmethode nicht zu allen untersuchten Elementen Ergebnisse liefert. Der Charakter der Untersuchungen wiederum bedingt, dass die Resultate in drei Datenkategorien fallen:

(i) Eigenschaften und Merkmale, die *einzelnen* Objekten der Teilmenge $M^{(k)}$ als Deskriptoren zugeordnet werden können. Ein Objekt hat die betreffende(n) Eigenschaft(en) oder es hat sie nicht.

(ii) Beziehungen einer bestimmten Qualität zwischen je zwei Objekten einer Teilmenge $M^{(k)}$. Die Vereinigungsmenge aller *paarweisen* Beziehungen einer Qualität lässt sich am besten durch ein Netzwerk darstellen. Es ist unerheblich, ob die paarweisen Beziehungen in ihrer Gesamtheit ein zusammenhängendes Netzwerk bilden oder nicht.

(iii) Beziehungen und/oder Attribute, die einer Gruppe von Objekten – nicht einzelnen Gruppenmitgliedern – innerhalb einer Teilmenge $M^{(k)}$ eigen sind und *mengenwertige* Unterscheidungs- und Zuordnungskriterien darstellen.

Auf dem Gebiet der funktionellen Genomanalyse liegt es nahe, die Gene/Genprodukte als Objekte zu behandeln und M als die Vereinigungsmenge der Gene einer Spezies anzusetzen. Bei der Bäckerhefe *Saccharomyces cerevisiae* beispielsweise beinhalten ~6.300 Gene die Synthesevorschrift für die entsprechenden Proteine. Vom Standpunkt moderner Datenhaltung aus gesehen ist das ein leicht zu handhabender Basisdatensatz geringen Umfangs, der sich seit dem Abschluss des Hefe-Sequenzierungsprojekts 1997 (siehe z.B. [Me97]) nur noch marginal veränderte. Demgegenüber führt die weltweite Erforschung der Funktionalität dieser Objekte auch weiterhin zu erheblichem Datenaufkommen in allen drei oben genannte Kategorien, die zum Verständnis der Lebensvorgänge im Modell- und Produktionsorganismus Hefe beitragen.

Hefegenomik – Sachstand, Trends und Erfordernisse aus Informatiksicht
Mit dem Aufkommen experimenteller Hochdurchsatztechniken und im Ergebnis systematischer Studien lässt sich beobachten, dass sich der Erkenntniszuwachs von Kategorie (i) mehr und mehr in die Kategorie (iii) verschiebt. Folglich erscheint es von besonderer Wichtigkeit, solche Herangehensweisen zu forcieren, die einen ganzheitlichen Blick auf die ständig wachsenden Daten erlauben, indem die o. g. Kategorien kombiniert und zueinander in Beziehung gesetzt werden.

Die im Rahmen der Experimente gesammelten Daten, werden heute in der Regel in Form von Dokumenten im Internet bereitgestellt. Um Zusammenhänge in den Daten aufzudecken, ist es jedoch erforderlich, dass spezifische Sichten auf diese heterogene Datenbestände definiert werden können. Diese Notwendigkeit erfordert eine Integration, die über die Inhalte einzelner Datenquellen hinausgeht. Jedoch ist im Bereich der Le-

benswissenschaften die Integration häufig auf die Bereitstellung von Dokument-Dokument-Querbezügen mit Hilfe von Hypertext-Markups beschränkt. Das ist jedoch für die Anforderungen einer integrierten Analyse unzureichend, da in der Regel weder die exakte Semantik eines Querbezugs definiert ist[1], noch können dokumentübergreifende Anfragen oder mengenwertige Ergebnisse verarbeitet werden.

Für die Hefedaten (siehe Abschnitt 3) sind Dienste verfügbar, die den Zugriff auf mehrere Datenquellen unter einer gemeinsamen Nutzerschnittstelle ermöglichen. Es werden dabei jedoch nur solche Zusammenhänge berücksichtigt, die bereits bekannt und beschrieben sind. Die *function junction* and *expression connection* Schnittstellen von SGD [Ce02] sind gute Beispiele für diese Art von Diensten. Nur wenige Dienste, wie EPCLUST/ EP:PPI [Ke02B][BV00], ermöglichen es dem Nutzer selbst, die Ergebnisse verschiedener experimentelle Studien miteinander zu kombinieren. Im Ergebnis erhält er eine Liste von Gennamen, welche die Anforderungen seiner Anfrage erfüllen. Die Weiterverarbeitung dieser Daten im Sinne einer Interpretationsunterstützung und Navigation wird nicht angeboten. Daher beklagen die Nutzer in den Lebenswissenschaften generell, dass sie zu wenig Unterstützung in Form integrierter Analyseumgebungen und Softwarewerkzeuge erhalten.

2. Vorgehensweise

Der in diesem Artikel beschriebene Ansatz, der sich während der letzten zwei Jahre herauskristallisiert hat, unterstützt den Wissensgewinn aus Biodaten. Auf der Grundlage von Experimentaldaten, die aus allen oben genannten Datenkategorien herangezogen werden, kann eine effiziente, interessengeleitete Datendurchsicht nach impliziten Zusammenhängen vorgenommen werden. Sowohl technische als auch theoretische Aspekte dieser Vorgehensweise werden in diesem Abschnitt beschrieben und diskutiert.

2.1 Datenmodellierung

Das Konzept der Datenverwaltung basiert auf einem relationalen Datenbanksystem. Dessen Entity-Relationship Modellierung orientiert sich sowohl an den vorhandenen Datensätzen sowie deren Zugehörigkeit zu einer der drei Datenkategorien (siehe Abb. 1 für einen schematischen Überblick). An zentraler Position des Datenmodells steht ein Basismodell, welches als Bindeglied zwischen den oben genannten Datenkategorien fungiert. Es orientiert sich am zentralen Dogma der Biologie und beschreibt u.a. Gene, Genvarianten, Proteine (*gene*, *gene_variant*, *protein*) und deren Beziehungen untereinander in ihrer zellulären Umgebung. Die gewählte Modellierung erlaubt es, jeden Datensatz eindeutig seinem biologischen Kontext zuzuordnen und gewährleistet damit eine sinnvolle Verknüpfung der einzelnen Submodelle (siehe Abb. 1, Submodelle gruppiert um ein zentrales Basismodell).

[1] D.h. sogenannte Hyperlinks werden gelegt, wenn *irgendein* Bezug zwischen zwei Dokumenten konstruiert werden kann.

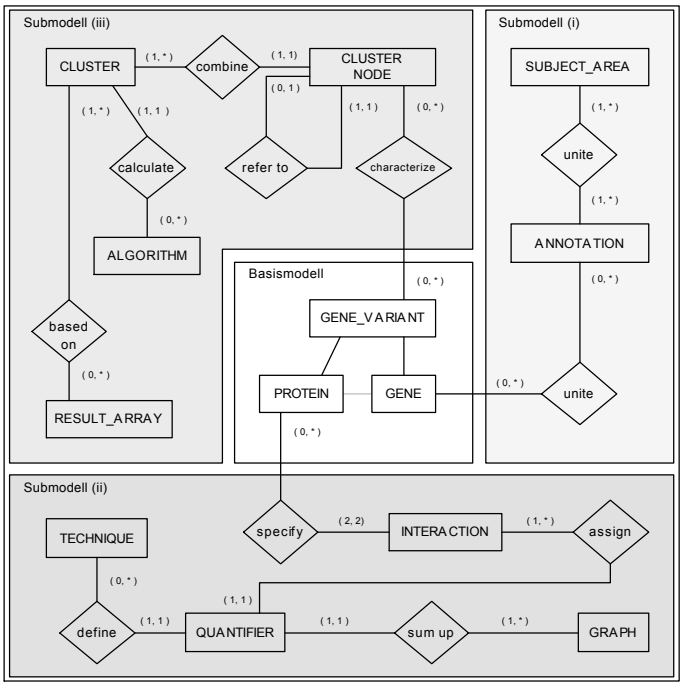

Abb. 1: Schematische Darstellung des gewählten Datenmodells

Zum besseren Verständnis wird in der folgenden Erläuterung eine konkrete Ausprägung (Genomforschung) des Datenbankmodells herangezogen. Ein oft wiederkehrendes und zentrales Problem der Datenintegration auf dem Gebiet der funktionellen Genomforschung besteht in den divergierenden Nomenklaturkonventionen[2]. Bei Vernachlässigung dieses Faktums bieten sich jedoch folgende Submodelle (orientiert an den erwähnten Datenkategorien) für die Datenintegration an. Der Bezug zum Datenmodell wird über die Angabe der Entitäten (kursiv) gewährleistet und wird für jede Datenkategorie einzeln dargestellt:

(i) Die Zuordnung von Eigenschaften zu Objekten (im Beispiel Gene/Genprodukte) erfolgt über ein festgelegtes Vokabular (*annotation*), welches in einzelne Themenkreise (*subject_areas*) gegliedert ist. Dieses Vorgehen sichert aus Sicht der Inhaltsverwaltung neben der redundanzfreien Verwaltung auch eine klare Unterscheidung zwischen den Objekten und Konzepten ab, denen die Eigenschaften zugeordnet sind (Abb. 1, Submodell (i)).

(ii) Eine Beziehung (*interaction*) zwischen zwei Objekten wird mit einem Quantitätsmaß (*quantifier*) und einem Methodenindikator (*technique*) spezifiziert. Der

[2] Die korrekte biologische Zuordnung für ein gegebenes Attribut (SWISSPROT z.B. unterscheidet in der Regel nicht nach Spleißformen bei der Angabe der Gewebe, in welchen das Protein als exprimiert gefunden wurde.) muß zumeist manuell vorgenommen werden, da bis zum heutigen Zeitpunkt keine generische, computergestützte Lösung abzusehen ist.

Methodenindikator bezeichnet dabei die Methode oder Technologie, mit der die Beziehung erkannt bzw. vorhergesagt wurde. Die Interpretation des Quantitätsmaßes, soweit angegeben, ist nur im Kontext der verwendeten Methode gültig. Dessenungeachtet stellt sie einen eindeutigen Bezugspunkt zu jeder der beteiligten ternären Beziehungsentitäten *interaction*, *technique*, und *graph* dar. Für die Verwaltung innerhalb des Datenmodells ist es jedoch unwesentlich, inwieweit der aus den paarweisen Beziehungen entstehende Graph (*graph*) gerichtet oder ungerichtet ist. Es bleibt den Algorithmen zum Abgriff und zur Weiterverwendung überlassen, die korrekte Interpretation sicherzustellen (Abb. 1, Submodell (ii)).

(iii) Eigenschaftszuordnung können sich gleichermaßen auf eine Menge von Objekten beziehen, die ihrerseits eine biologische Konzeptebene bzw. ein Partitionierungsprinzip in Form von Clustern[3] begründen können. Die Modellierung der Clusterbäume (*cluster*) erfolgt auf der Grundlage von Experimentaldaten (*result_array*). Parallel wird die hierarchische Struktur der einzelnen Cluster (*cluster_node*) in Verbindung zu den beteiligten Objekten (*gene_variant*) verwaltet. Mit dem Einsatz von Algorithmen oder Beschreibungsprozeduren (*algorithm*) kann zudem das rechnerische Vorgehen belegt werden. Die gewählte Informationsstruktur des Datenmodells lässt jedoch keine Aussagen über die Auswahl oder Abfolge von Objekten innerhalb eines Clusters zu. Mathematisch gesehen sind Cluster demnach schlicht Mengen von Objekten (Abb. 1, Submodell (iii)).

2.2 Datenaufbereitung

Nachdem beschrieben wurde, wie die Datensammlungen in der Datenbank verwaltet werden, wird sich im nächsten Schritt der Fragestellung zugewandt, wie ein Nutzer Datensätze für seine spezifischen Anfragen kombinieren kann. Die Datenaufbereitung orientiert sich einerseits an vordefinierten Formatvorlagen, die eine manuelle Zusammenstellung der Datensätze gestatten und wird anderseits von der Anwendung *DataReader* unterstützt. Letzterer verbindet sich via ODBC mit der Datenbank und ermöglicht es so, eine interaktive Auswahl aus allen vorhandenen Datensätze zu treffen.

Zur Datenvisualisierung und -analyse wird ein induzierter Graph verwendet, welcher auf der Grundlage der ausgewählten Datensätze gebildet wird. Mit der Zusammenstellung des Graphen wird demnach eine erste Einschränkung der Datensätze auf die zu betrachtenden Fragestellung vorgenommen.

[3] In dieser Arbeitbesiehen sich alle Textpassagen auf hierarchische Clusterbäume, ungeachtet dessen, inwieweit es sich dabei um gewichtete Cluster handelt oder nicht.

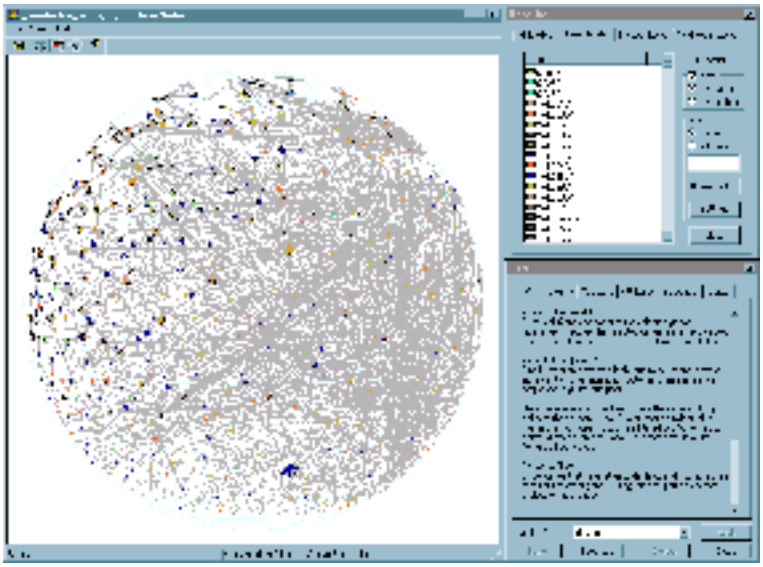

Abb. 2: Screenshot der GeneViator Anwendung

Um alle zum heutigen Zeitpunkt existenten und in Zukunft als realistisch anzusehenden Größenordnungen von Datensätzen zu unterstützen, bedient sich das Visualisierungsprinzip des GeneViator (siehe Abb. 2) der hyperbolisch verzerrten Projektion von Graphenstrukturen in einer Kugel, wie sie von Tamara Munzner als H3-API [Mu00][Ke02A][Hy02] entwickelt bzw. erweitert wurde. Dieser Ansatz erlaubt es, einen theoretisch unendlichen Satz von Knoten auf einem typischen Computerbildschirm darzustellen. Jeder Knoten des Graphen spiegelt ein Objekt des Gegenstandsbereiches wieder, welche über eine/mehrere Beziehung(en) miteinander verbunden sein können. Nachfolgend wird erläutert, wie die einzelnen Datenkategorien zur Erzeugung des Graphen genutzt werden[4].

(i) Die Datensätze der Kategorie (i) werden den Knoten als Attributliste zugeordnet. Jedem Attribut steht bei der späteren Visualisierung genau ein Farbwert gegenüber. Die Kombination der Attribute eines Knoten spiegelt sich demnach in der entsprechenden Färbung jedes einzelnen Knotens wieder. Vorläufig geschieht die Farbzuweisung automatisch und kann vom Nutzer nicht beeinflusst werden.

(ii) Die Kanten des Graphen werden aus den gewählten Datensätzen der Kategorie (ii) erzeugt, d.h. jede vorhandene paarweise Beziehung wird als eine Kante zwischen den beteiligten Knoten wiedergegeben. Sollten mehrere Kanten zwischen zwei Knoten auftreten, wird aus Gründen der Übersichtlichkeit nur eine Kante dargestellt. Sind im Anschluss daran nicht alle Knoten in einem zusammenhängenden Graphen verbunden, wird ein virtueller Wurzelknoten eingefügt, der alle entstandenen Subgraphen über Kanten verbindet. Das ist dem Fakt geschul-

[4] Ein illustrierendes Beispiel wird im Abschnitt Anwendungen dargestellt.

det, dass die H3-API wenigstens einen Wurzelknoten für die Verarbeitung von Wäldern[5] benötigt.

(iii) Die Datensätze der Kategorie (iii) besitzen keinen Einfluss auf die Struktur des Graphen, können jedoch genutzt werden, um mittels Filterfunktionen Teile des Graphen auszublenden. Zusammenfassend spiegeln die gewählten Datensätze in ihrer Kombination eine, von der Datenbank losgelöste statische Datensammlung wieder, die zur Visualisierung des Graphen im GeneViator [He02] verwendet wird.

2.3 Datenverarbeitung

Auf Grundlage der in Abschnitt 2.2 erläuterten Datenzusammenstellung können Experimentaldaten aller drei Datenkategorien im GeneViator untersucht werden. Diese ermöglichen nicht nur eine willkürliche Auswahl der Datensätze unabhängig ihrer Kategoriezuordnung, sondern erlauben zudem die interaktive Navigation innerhalb der Datenrepräsentation und bieten Strategien zur Datenanalyse an. Die implementierten Methoden der Datenanalyse zielen darauf ab, eine breite Spanne von Nutzerstrategien zu unterstützen, die für das Aufdecken von wissenschaftlich relevanten Korrelationen innerhalb des untersuchten Systems adäquat erscheinen.

Für alle drei Datenkategorien wird die Filterung in zwei Richtungen unterstützt (siehe Abb. 3), zum einen *top-down* (Ausgehend von der Eigenschaften) zum anderen *bottom-up* (Ausgehend von den Objekten). Dieses Vorgehen verfolgt die Zielstellung den aufgabenspezifischen Anforderungen bestmöglich zu genügen. Einschränkungen bezüglich der Analysekomplexität müssen lediglich im Bereich der filterübergreifenden 'oder' bzw. 'und/oder nicht ' Verknüpfungen der gewählten Kriterien hingenommen werden. Für die nachstehende Erläuterungen der Analysesichtweise wird das beschriebene graphentheoretische Vokabular beibehalten.

	Datenkategorie (i)	Datenkategorie (ii)	Datenkategorie (iii)
top-down Vorgehen	Attributfilter	Strukturfilter Selektivitätsfilter	Clusterfilter
bottom-up Vorgehen	-	Nachbarschaftsfilter	Clusterfilter

Abb. 3: Tabelle der Filterstrategien des GeneViator

[5] Die H3-API basiert auf Bäumen, d.h. azyklischen Graphen. Graphen, die aus nicht verbundenen Bäumen bestehen, heißen Wälder. Kanten, die Zyklen induzieren, werden intern anders gehandhabt, das ist jedoch nicht Thema der Darlegungen in dieser Arbeit.

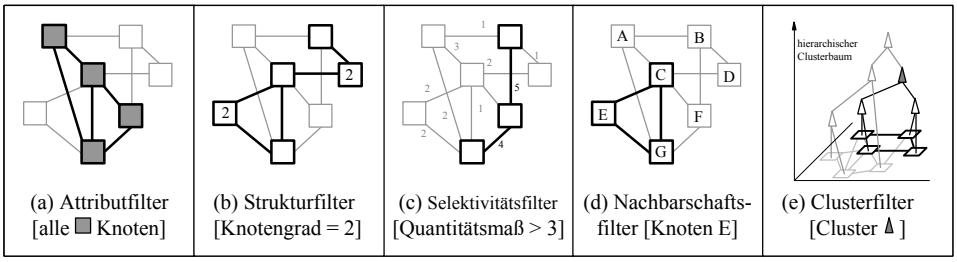

Abb. 4: Skizzierte Darstellung der Filterstrategien des GeneViator am einfachen Beispiel

Das *bottom-up* Vorgehen unterstützt den knotenorientierten (im Beispiel Gen getriebenen) Prozess der Suche nach Korrelationen. Durch Festlegung einer Knotenauswahl als Anfrage kann der Nutzer untersuchen, welche biologischen Konzepte (auch Kombinationen solcher) seine spezifische Auswahl abdecken. Die biologisch-konzeptuell getriebene Suche nach Korrelationen hingegen wird durch das *top-down* Vorgehen unterstützt. Der Nutzer wählt biologische Konzepte, die für ihn von Interesse sind, und bestimmt so Teilmengen von Knoten (im Beispiel Gene), die seinen vorgegebenen Kriterien entsprechen.
Im folgenden wird beschrieben, wie diese Vorgehensweisen auf die beschriebenen Datenkategorien angewandt und kombiniert werden können. Außerdem wird aufgezeigt, wie die entstehende Komplexität mit dem Visualisierungs- und Navigationswerkzeug GeneViator beherrscht werden kann:

Datenkategorie (i)
Für die Auswahl von Knoten anhand ihren Eigenschaften ist die *top-down* Vorgehensweise als der spezifische Attributfilter (siehe Abb. 4a) implementiert. Es kann eine frei zu wählende Kombination von Attributen festlegen werden. Die graphische Repräsentation des Gesamtgraphen reduziert sich daraufhin auf diejenigen Knoten (im Beispiel Gene/Genprodukte), die wahlweise mindestens einem der Attribute oder allen ausgewählten Attributen genügen.

Datenkategorie (ii)
Wie eingangs erwähnt werden paarweise Beziehungen als Kanten zwischen den betreffenden Knoten dargestellt. Diese graphenorientierte Sichtweise lässt beide oben genannten Vorgehensweisen zur Datenanalyse sinnvoll erscheinen. Das *top-down* Vorgehen wird durch zwei Filtertypen unterstützt. Der Strukturfilter (siehe Abb. 4b) ermöglicht es, den induzierten Graphen auf Subgraphen mit einem definierten minimalem und/oder maximalem Knotengrad[6] zu reduzieren. Wenn die untersuchten Kanten zudem mit einem quantitativen Maß (siehe Abschnitt 2.1) versehen sind, kann der Graph auch auf Kanten oberhalb einer einstellbaren minimalen Stärke weiter reduziert werden. Diese Stellgröße wird über den implementierten Selektivitätsfilter (siehe Abb. 4c) geregelt. Unterstützung für das *bottom-up* Vorgehen bietet der Nachbarschaftsfilter (siehe Abb. 4d). Der Nutzer kann eine beliebige Gruppe von Anfrageknoten (im Beispiel Gene) auswählen und sich dadurch selektiv diejenigen Knoten einblenden, die zu einem belie-

[6] Anzahl der Nachbarn eines Knotens

bigen der Anfrageknoten benachbart sind. Wiederholte Anwendung dieses Filters führt zur Aufdeckung weiterer Nachbarschaftskreise, die entsprechend über 1 bis n-1 Kanten (wobei n für die maximale Anzahl von Knoten in einer Abfolge steht) erreichbar sind.

Datenkategorie (iii)
Unterstützung beim Navigieren in und Filtern von mengenwertigen Experimentaldaten (z.B. hierarchisch strukturierte Clusterbäume) bietet die jüngste Funktionalitätserweiterung der Clusterfilter. Dieser ermöglicht es einen/mehrere Clusterbäum(e) dem Graphen orthogonal gegenüberzustellen (siehe Abb. 4e). Jeder Knoten des Graphen kann somit gleichzeitig einen Blattknoten innerhalb eines oder mehrerer Clusterbäume darstellen. Der *top-down* orientierte Zugriff verfolgt eine geradlinige Strategie der Clusterauswahl, d.h. der Nutzer wählt zunächst den/die für ihn relevanten Cluster auf der entsprechenden Hierarchieebene des Clusterbaums aus. Der GeneViator reduziert den Graphen anschließend auf diejenigen Knoten, die als Blattknoten unterhalb dem/den ausgewählten Cluster(n) liegen. Die *bottom-up* Vorgehensweise stellt sich dagegen komplexer dar[7], hierbei trifft der Nutzer im ersten Schritt eine Auswahl der Knoten des Graphen, die im weiteren als Anfragemenge (Abb. 5, "A, B, G, H") über einem oder mehreren Clusterbäumen genutzt werden können. Um die folgende Identifikation von (hoffentlich) bedeutsamen Clustern auf möglichst tiefer Hierarchieebene des Clusterbaums zu unterstützen, wurden zwei Suchparameter eingeführt: *Trennschärfe* und *Unterstützung*. Mit Hilfe der *Trennschärfe* kann die Suche auf eine gegebene minimale und maximale Mächtigkeit, d.h. die zulässige Anzahl der Blattknoten unterhalb eines Clusters einschränkt werden. Für das in Abb. 5 gewählte Beispiel (Trennschärfe = [3, 5]) bedeutet dies, dass mindestens 3 und maximal 5 Blattknoten in einem Cluster enthalten sein müssen, damit dieser ausgewählt wird. Die Wahl der *Trennschärfe* ermöglicht es damit, indirekt die Hierachieebenen von Interesse[8] einzugrenzen. Der Parameter *Unterstützung* spezifiziert, wieviele der ausgewählten Anfrageknoten mindestens in einem Cluster enthalten sein müssen, damit dieser als "Treffer" angezeigt wird. Das in Abb. 5 gewählte Beispiel (Unterstützung = 2) impliziert, dass mindestens zwei Knoten der Anfrageliste im Cluster enthalten sein müssen, damit dieser ausgewählt wird. Der verwendete Algorithmus wurde dahingehend optimiert, dass ein nachträgliches Erweitern bzw. Ausdünnen der Anfrageknoten lediglich eine Berechnung der veränderten Komponenten nach sich zieht. Diese Strategie macht Sinn, wenn bedacht wird, dass speziell im Forschungsbereich der Autoren (Genom-, Transkriptom- und Proteomforschung) eine iterative Annäherung an die endgültige Anfrageformulierung realistisch ist. Es ist demnach ausdrücklich erwünscht, dass ein Nutzer interaktiv die Einstellungen der getroffenen Knotenanfrage variiert, bis er auf Cluster stößt, die sein Interesse auf sich ziehen. Das wird in den Fällen gelingen, wenn eine intuitiv postulierte (vermutete) Korrelation auch tatsächlich hinter den Daten versteckt ist.

[7] Eine naive Implementation für Hierarchien hätte immer ein triviales Ergebnis, nämlich, daß die oberste Hierachieebene (Wurzelknoten) alle in die Untersuchung einbezogenen Blattknoten umfaßt.
[8] Die Mächtigkeit wurde gegenüber der Ebene bevorzugt, weil die Auswertung nicht auf balancierte Hierarchien einschränkt werden sollen. Besonders für tief gestaffelte binäre Hierarchien erlaubt dies eine intuitivere Kennung.

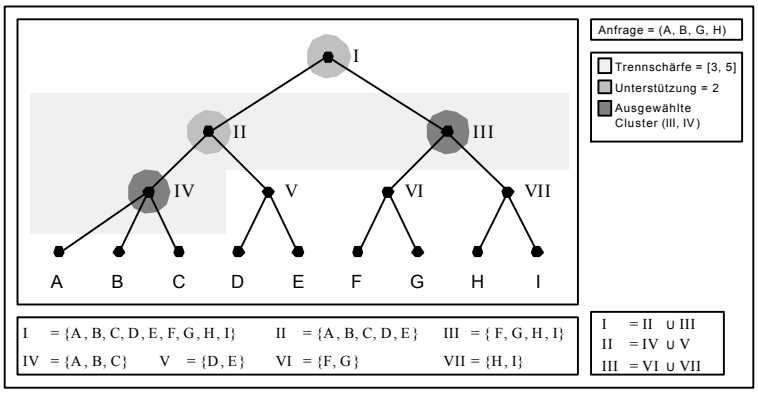

Abb. 5: Einschränkungen einer hierarchischen Suchabfrage
mit den Parametern *Trennschärfe* und *Unterstützung*

Wie oben ausgeführt, können alle beschriebenen Filtervorgänge in beliebiger Reihenfolge und Kombination ausgeführt werden. Dadurch ist eine variable Untersuchung eines interessierenden Systems, gerade in Bezug auf die Überlagerung unterschiedlicher Informationsdimensionen gewährleistet. Angesichts der Komplexität des Forschungsprozesses und bei gegebener Datenlage kann jedoch nicht von vornherein garantiert werden, ob die Daten überhaupt Korrelationen beinhalten und erkennen lassen. Um dem Nutzer eine verlässliche Führungshilfe durch multidimensionale Datensammlungen zu geben, haben die Autoren eine Art "Lackmustest" für eine gegebene Datenkombination entwickelt, der im Abschnitt Anwendung dargelegt wird.

3 Anwendung

Nachdem die vollständige Bestimmung aller Gensequenzen einer Spezies zur wissenschaftlichen Alltäglichkeit geworden ist, besteht nunmehr die Herausforderung, diesen Genen ihre Funktionen in den Lebensprozessen zuzuordnen. Hochdurchsatzexperimenten, in denen jeweils ein molekularbiologischer Aspekt für das gesamte Genom erhoben wird, sollen die dafür notwendigen Daten liefern. Als besonders vielversprechend für die Funktionsaufklärung gelten derzeit Untersuchungen zur sogenannten Genexpression und zur Proteinwechselwirkung. Die Genexpression beschreibt, welche Gene in einem bestimmten Zusammenhang gemeinsam aktiviert sind. Proteinwechselwirkungsstudien versuchen, systematisch zu erheben, welche Genprodukte in ihrem unmittelbaren Zusammenspiel Lebensfunktionen übernehmen können. Die Daten, die im Rahmen dieser Verfahren bestimmt werden, sind besonders geeignet, den Nutzen des hier beschriebenen Ansatzes zu demonstrieren. Zum einen ist es erforderlich, dass zwei Gene, deren Produkte gemeinsam eine Funktion erfüllen, auch zur gleichen Zeit aktiv sind. Zum anderen hat sich gezeigt, dass klassische, statistische Verfahren nicht vermögen, die Korrelationen innerhalb einer einzigen Verfahrensweise aufzudecken [KS02], [Ku02].

Die Bäckerhefe ist die Spezies, für die derzeit die umfangreichsten und vollständigsten Datensammlungen zu den beschriebenen Fragestellungen verfügbar sind. In der hier beschriebenen Anwendung wird gezeigt, wie ausgehend von den Genomdaten der Hefe, Korrelationen in systematisch zusammengetragenen Experimentaldaten aufgedeckt werden können.

3.1 Datenintegration

Entsprechend der in Abschnitt 2 dargelegten Vorgehensweise sind die notwendigen Daten modelliert, zusammengestellt und in eine Datenbankinstanz[9] integriert worden:

Das Bezugsystem der Analyse bildet das Hefegenom, so wie es in der *Saccharomyces Genome Database* (SGD) [Ce02] bereitgestellt wird. Dort sind alle bekannten Hefegene mit ihren jeweiligen Sequenzen, Namen und eindeutigen Bezeichnungen zusammengestellt. In der Hefeforschung haben sich die sogenannten ORF-Namen als eindeutige Genbezeichner durchgesetzt. Ein ORF (*open reading frame*) bezeichnet einen Abschnitt des Hefegenoms, der die Bildungsvorschrift für das dem Gen zugeordnete Genprodukt enthält. Für die einzelnen Datenkategorien wurden folgende Datenquellen verwendet:

(i) Zur Charakterisierung von Genen wird die *Gene Ontology* (GO) verwendet [As00]. GO stellt Begriffshierarchien für die Beschreibung von *biologischen Prozessen, Zellkomponenten* und *molekularen Funktionen* bereit. Die Charakterisierung eines Gens erfolgt, indem einem ORF-Namen der jeweils spezifischste Begriff aus den drei Hierarchien zugeordnet wird. Damit ergibt sich im Idealfall ein Begriffstripel, das einem ORF-Namen als Attribut zugeordnet wird. Auch wenn diese Zuordnungen noch im Wesentlichen als *work in progress* bewertet werden müssen[10], kann für die Hefe in Anspruch genommen werden, dass aufgrund der intensiven internationalen Forschungsbemühungen die Fehlerquote im Vergleich zu anderen Spezies als gering einzuschätzen ist.

(ii) Binäre Beziehungen zwischen Genen werden aus der 'Physical Interaction Table' abgeleitet, die von der *Comprehensive Yeast Genome Database* (CYGD) [Me02] verwaltet wird. In dieser Tabelle werden weltweit publizierte Daten zu experimentell nachgewiesenen Proteinwechselwirkungen zusammengestellt[11]. Ein Eintrag in dieser Tabelle besteht aus zwei ORF-Namen und der Methode, mit der eine Wechselwirkung festgestellt wurde.

(iii) Mengenmäßige Attributzuschreibungen wurden aus einer Studie mit dem Namen *Yeast Cell Cycle Analysis Project* aufbereitet, die von Spellman et al. 1998 publiziert worden ist [Sp98][Ce02]. In dieser Studie wurde die Genexpression der Hefe im Zusammenhang mit dem Zellzyklus bestimmt, den eine Zelle zwischen zwei Teilungen durchläuft.

[9] DB2 UDB V7.2 auf einem Linux Server

[10] Häufig ist einem ORF-Namen auch zugeordnet, dass über Prozess, Funktion und Lokalisierung keine Aussage gemacht werden kann.

[11] Die Tabelle enthält zudem Wechselwirkungen zwischen Proteinen und funktionalen RNA Molekülen. Diese wurden in unserer Anwendung nicht berücksichtigt.

Abbildung 1 zeigte das Datenmodell, in das die Daten integriert wurden. Dabei konnten 1.892 der ca. 6.300[12] Hefegene, Daten aller beschriebenen Kategorien zugeordnet werden. Diese Gene bilden die Basis für die hier dargestellten Analysen. Die hierarchische Clusteranalyse der Expressionsdaten wurde mit dem Werkzeug J-Express [Mo02] durchgeführt.

3.2 Suche nach Korrelationen

Das Explizieren von Korrelationen, die in den experimentellen Daten vorhanden sind, ist eine wichtige methodische Voraussetzung für das Verständnis des Zusammenspiels der Gene in lebenden Organismen. Wie bereits erwähnt, sind die bekannten statistischen Verfahren zur Korrelationsanalyse jedoch häufig nicht geeignet, die wesentlichen Zusammenhänge aufzudecken. Dies ist insbesondere in dem Umstand begründet, dass kein linearer Zusammenhang zwischen den wahren Werten und den Beobachtungswerten einer Größe, die experimentell erfasst werden soll, unterstellt werden kann. Insbesondere das Protokoll zur Erfassung von Expressionsdaten umfasst mehrere Schritte, in denen jeweils nicht lineare Transformationen der Beobachtungsgröße erfolgen[13]. Bei der Erfassung von Proteinwechselwirkungsdaten muss erheblich in das zu beobachtende System eingegriffen werden. Daher muss regelmäßig angenommen werden, dass sich nicht die „natürlichen Verhältnisse“ einstellen können. Dennoch wird allgemein angenommen, dass die verfügbaren Beobachtungsdaten zumindest qualitativ ein zutreffendes Bild der untersuchten Lebensprozesse liefern. Derzeit gibt es weltweit Anstrengungen, eine Vielzahl von Data-Mining-Methoden für die Analyse von Genomdaten nutzbar zu machen. Stellvertretend seien hier das *frequent subgraph mining* [KK01] und der Aufbau von *Bayesian networks* [Pe01] genannt, die erste, vielversprechende Ergebnisse geliefert haben. Das *frequent subgraph mining* ist aus Effizienzgründen auf Probleme moderater Größe beschränkt. Das Ableiten von *Bayesian networks* aus Expressionsdaten erfordert für jeden Anwendungsfall das Aufstellen komplexer wahrscheinlichkeitstheoretischer Modelle, die einem standardmäßigen Einsatz der Methode im Wege stehen. Wir schlagen daher im Folgenden eine Vorgehensweise vor, die zumindest qualitativ eine Aussage über die Korrelation von Genen ermöglicht.

Gesucht werden also Korrelationen zwischen Objekten, in diesem Falle Gene, für die Beobachtungsdaten unterschiedlicher Art vorliegen. Wir werden diese Fragestellung im Folgenden dahingehend interpretieren, dass für eine Teilmenge der Gene entschieden werden soll, ob sie hinsichtlich mehrerer Kriterien in ein gemeinsames Klassifikationsraster fallen. Die Vorgehensweise lehnt sich an die *bottom-up*-Navigation in Mengenattributen an, so wie sie in Abschnitt 2 beschrieben wurde. Dabei werden folgende Beobachtungen für die Entscheidung herangezogen:

[12] Die exakte Anzahl von Hefegenen ist immer noch nicht bekannt. Jedoch werden ca. 6.300 ORF-Namen von mehreren Datenbanken seit Jahren als gesicherte Gene verwaltet.
[13] Würden diese Transformationen für alle gleichzeitig beobachteten Gene gleichförmig sein, wäre das kein prinzipielles Problem. Jedoch wirken sich die Prozesse in Abhängigkeit von der Länge und Nukleotidsequenz einzelner Genen (nicht vorhersagbar) unterschiedlich auf die Messwerte aus.

1. Beobachtung:
Werden Gene ausgewählt, die hinsichtlich der Genexpression ähnliche Merkmalsausprägungen aufweisen, finden sich viele Cluster, die auf niedriger Stufe in der Hierarchie bereits mehrere Gene aus der Auswahl enthalten. Diese Cluster, die bei einem niedrigen Wert des Filterparameters *Unterstützung* gefunden werden, enthalten darüber hinaus nur wenige zusätzliche Gene, die nicht in der Auswahlmenge vorhanden sind. Das führt insgesamt dazu, dass bei niedrigem Wert für *Unterstützung* der Anteil der Gene aus der Auswahl größer ist als der Anteil von Genen die insgesamt angezeigt werden. Erhöht man ausgehend von dieser Situation die geforderte *Unterstützung* stufenweise, so gelangt man zu einem Punkt, an dem nur noch in einer vergleichsweise hohen Hierarchiestufe die Auswahlanforderung erfüllt wird und zwar nur noch für ein einzelnes Cluster. Daher gibt es an dieser Stelle einen Knick im Kurvenverlauf. Ab dieser Stelle finden sich dann alle gewählten Gene im gleichen Teilbaum der Hierarchie. Der Anteil der Gene aus der Auswahl, die sich in dem jeweiligen Cluster finden, ist stets größer als der Anteil der Gene, die insgesamt angezeigt werden.

2. Beobachtung:
Wählt man Gene aus, deren Expressionsverhalten keine Ähnlichkeit aufweist, so findet man zwar für niedrige *Unterstützung* eine Reihe von Clustern, die diese Gene enthalten, aber bereits hier werden viele Gene mit angezeigt, die nicht in der Auswahl vorhanden sind. Erhöht man die geforderte Unterstützung, dann „springen“ die gefundenen Cluster in der Hierarchie und die Anzahl der gefundenen Cluster kann sich bei jeder Erhöhung der Anforderung ändern. Daher kommt es zu einem oszillierenden Kurvenverlauf. Ab einem gewissen Punkt findet sich auch hier nur noch ein Teilbaum in der Hierarchie, in dem durch stetiges „Hochklettern“ die Anforderungen erfüllt werden. Jedoch ist in diesem Fall der Anteil der insgesamt angezeigten Gene stets größer als der Anteil der Gene aus der Auswahl, die im jeweiligen Cluster vorhanden sind.

In Abbildung 6 werden diese Beobachtungen anhand schematischer Kurvenverläufe dargestellt. Die gestrichelte Linie (aufgetragen gegen die rechte Y-Koordinate) gibt an, wie viele Gene aus der Auswahl in das Klassifizierungsraster fallen. Die durchgezogene Linie (aufgetragen gegen die linke Y-Koordinate) gibt an, wie viele Gene der Grundgesamtheit bei der jeweiligen Filtersetzung insgesamt zur Anzeige gelangen. Da beide Y-Koordinaten prozentual skaliert sind, treffen sich also beide Kurven in dem Punkt der vollständigen Abdeckung der jeweiligen Gesamtheit.

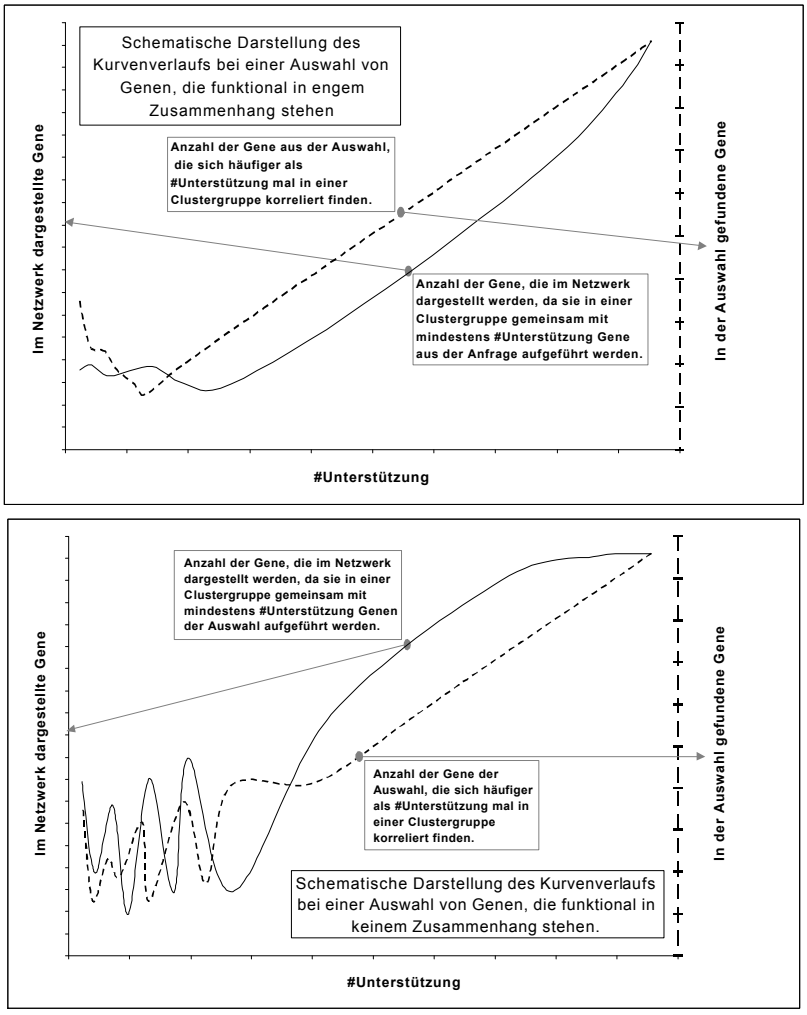

Abb. 6: Schematische Wechselbeziehung bei
a) korrelierter und b) nicht korrelierter Genauswahl

Die beschriebenen Beobachtungen lassen sich nun auf die allgemeine Auswahl von Genen übertragen: wählt man Gene basierend auf Kriterien anderer Kategorien, so kann man aus den analog abgeleiteten Kurvenverläufen schließen, ob diese Gene auch im Hinblick auf ihre Expression eine Beziehung aufweisen. Um zu prüfen, ob diese Schlussfolgerung einer Überprüfung anhand der integrierten Beobachtungswerte standhält, wurden Szenarios festgelegt, in denen bereits durch die Fragestellung die Entscheidung für oder wider Korrelation induziert wird.

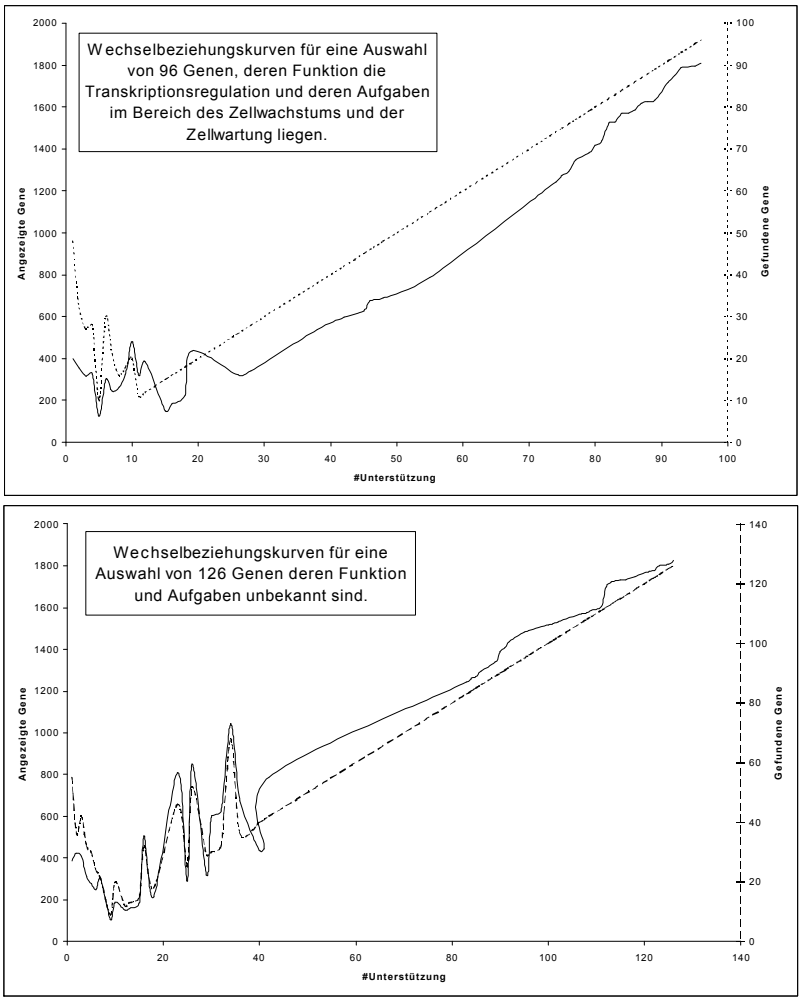

Abb. 7: Wechselbeziehung für
a) funktionell verwandte Gene und b) Gene unbekannter Funktion

Im **ersten Szenario** wurden alle Gene ausgewählt, die eine bekannte Rolle im *Zellwachstum und/oder bei der Zellwartung* spielen und außerdem die Funktion eines *Transkriptionsregulators*[14] ausüben. Diese Auswahl (auf Basis der Datenkategorie i) umfasst 96 Gene. Von diesen kann mit hoher Sicherheit angenommen werden, dass sie über ähnliche Expressionsmuster verfügen, da in den zu Grunde liegenden Messungen insbesondere die Genaktivität in Abhängigkeit vom Zellzyklus erfasst wurde.

Im **zweiten Szenario** wurden anhand der GO-Charakterisierung Gene ausgewählt, für die sowohl die *molekulare Funktion* als auch der *biologische Prozess unbekannt* sind.

[14] D.h. die Genprodukte dieser Gene regulieren das Aktivitätsniveau anderer Gene

Die Auswahl enthält 126 Gene, von denen mit hoher Sicherheit anzunehmen ist, das sie kein übereinstimmendes Expressionsmuster aufweisen.

Die für Szenario eins und zwei erstellten Kurvenverläufe sind in Abbildung 7 dargestellt und weisen im wesentlichen die Charakteristika auf, die in Beobachtung eins bzw. zwei dargelegt wurden. Insbesondere der Umstand, dass der Anteil der insgesamt angezeigten Gene unterhalb bzw. oberhalb des Anteils der *unterstützten* ausgewählten Gene liegt, zeichnet sich als geeignetes Entscheidungskriterium für oder wider Korrelation ab.

Diese Beobachtung konnte auch in weiteren Szenarios untermauert werden, in denen z.B. Gruppen von Genen ausgewählt wurden, die über Daten der Kategorie ii vernetzt waren (Proteinwechselwirkung) bzw. für die keinerlei Hinweis auf eine physische Interaktion vorlagen.

Abbildung 8 zeigt den Ausschnitt des komplexen Netzwerkes, der sich bei Verwendung der in Abschnitt 2 beschriebenen Filter auf eine konkrete molekularbiologische Fragestellung ergibt.

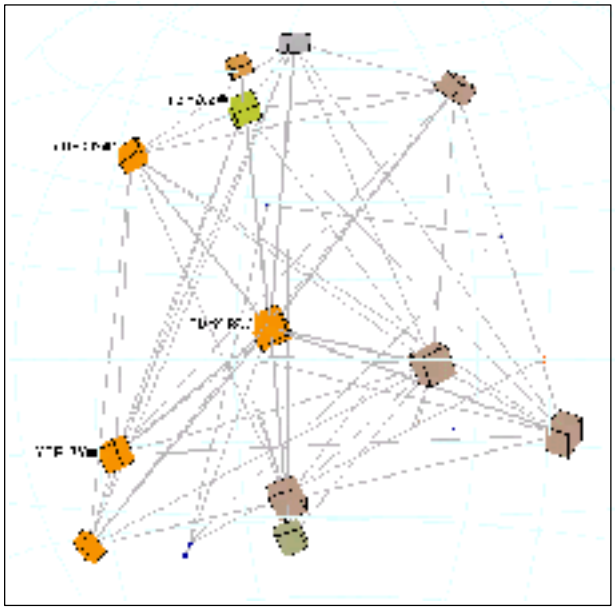

Abb. 8: GeneViator Screenshot einer bottom-up-Auswahl. Die Anfrage bildeten die physischen Wechselwirkungspartner der primären Stufe (19 Gene) und sekundären Stufe (51 Gene) des Hefegens YDR448W. Ausgehend von diesen wurden 27 Gene als korreguliert mit YDR448W identifiziert, übereinstimmend zu Spellman's Untersuchung [Sp98].

4 Zusammenfassung und Ausblick

Mit der vorliegenden Arbeit wird ein weitgehend werkzeuggestütztes Verfahren zur Suche nach Korrelationen in heterogenen Datensammlungen unterschiedlicher konzeptueller Kategorien vorgestellt. Das Vorgehen verfolgt den Zweck, versteckte und potentiell bedeutsame Korrelationen qualitativ festzustellen. Bei der Anwendung auf eine typische Zusammenstellung verschiedener Datenquellen, konnten die Machbarkeit und der Nutzen des Ansatzes gezeigt werden.

Derzeit erlaubt der Ansatz jedoch nur eine Überprüfung auf Korrelation in einer vom Anwender selbst erstellten Auswahlmenge. Es wäre wünschenswert, dass das System selbständig Gengruppen identifiziert, die vor dem Hintergrund der vorliegenden Beobachtungsdaten korreliert sind. Da jedoch prinzipiell eine in der Anzahl der Gene exponentielle Anzahl von Teilmengen existiert, müssen hier erst leistungsfähige Heuristiken für den Aufbau und die Auswahl potentiell interessanter Gengruppen entwickelt werden.

Weiterhin sind bereits neue Hochdurchsatzverfahren im Einsatz, die Ergebnisse liefern, die sich nicht ohne weiteres in eine der drei beschriebenen Datenkategorien einordnen lassen[15]. Wir streben an, auch diese Datensätze in unsere integrierte Verwaltungs- und Analyseumgebung einzubeziehen.

Glossar

Aminosäure	Grundbaustein der Proteine. Aminosäuren haben den gleichen chemischen Grundaufbau und unterscheiden sich voneinander durch die Beschaffenheit ihrer Seitenketten.
DNA	Desoxyribonukleinsäure, der materielle Träger der biologischen Erbinformation. DNA ist ein lineares Polymer aus vier *Nukleotiden* mit den Symbolen A, T, G, C in Doppelhelixform
Expressionshöhe	hier: Anzahl der Transkripte von einem Gen.
Gen	ein Abschnitt auf der *DNA* mit einer bestimmten funktionellen Bedeutung
Genetischer Code	Zuordnung von je drei aufeinanderfolgenden Nukleotiden zu einer *Aminosäure*. Die Gesamtheit der 4^3=64 möglichen Dreierkombinationen (Nukleotid-Triplets) und der zugeordneten Aminosäuren wird auch als Translationstabelle bezeichnet. In der Tabelle sind auch die Tripletts enthalten, die das Signal für Beginn bzw. Abbruch der Proteinsynthese vermitteln.
Gensequenz	die Abfolge der Nukleotide entlang eines Gens, in der Informatik als Zeichenkette aus den Nukleotid-Symbolen dargestellt

[15] Die experimentelle Feststellung von Proteinkomplexen mit Hilfe der Massenspektrometrie ist eines dieser Verfahren.

Genexpression	der Prozess der Realisierung des Informationsgehalts eines Gens. Wird ein Gen aktiviert, kommt es zunächst zur Anfertigung von n Kopien seiner Sequenz. Dieser Teilprozess heißt Transkription. Die Transkripte werden zurechtgeschnitten und als *mRNA* zum Ort der Proteinsynthese transportiert. Dort wird im Teilprozess der Translation die von der mRNA überbrachte Synthesevorschrift für ein *Protein* realisiert, indem je drei Bausteine nach dem *genetischen Code* festlegen, welche *Aminosäure* in das Protein eingebaut wird
Genom	die Vereinigungsmenge der Erbanlagen eines Organismus oder einer Spezies
mRNA	Messenger RNA, ein lineares Boten-Molekül, das die kodierende Information zum Ort der Proteinsynthese überträgt
Nukleotid	hier: Bausteine der linearen Polymere, die Erbinformation speichern bzw. transportieren
Offenes Leseraster	Abschnitt einer kodierenden Gensequenz zwischen Start- und Stoppkodon (s.a. Genetischer Code). Da die Einbaureihenfolge der Aminosäuren in ein Protein durch Tripletts festgelegt ist, ist die Länge Offener Leseraster ein Vielfaches von 3.
Protein	Lineares Polymer aus 20 *Aminosäuren*. Proteine bilden den Hauptteil der eigentlich operativen Bestandteile der Zelle. Sie bewirken Strukturbildung, Stoffwechsel, Energiehaushalt, Reproduktion und alle anderen wichtigen Lebensvorgänge in einer Zelle oder sind zumindest daran beteiligt.
Zentrales Dogma	Die *Genexpression* ist ein Prozess des horizontalen Informationsflusses (durchgezogene Pfeile) zwischen DNA, RNA und Proteinen im zentralen Dogma (oder Zentralen Lehrsatz) der Biologie, das sich folgendermaßen darstellen lässt:

Literaturverzeichnis

[As00] Ashburner, M. et. al.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25, 2000; S. 25-29. http://www.geneontology.org/cgi-bin/GO/downloadGOGA.pl/gene/association.sgd

[BV00] Brazma A., Vilo J.: Gene Expression Data Analysis. FEBS Letters Volume 480, Issue 1, 2000; S. 17-24. http://eptest.ebi.ac.uk/EP/EPCLUST/

[Ce02] Cherry, J. M.et. al.: Saccharomyces Genome Database, 2002. http://genome-www.stanford.edu/Saccharomyces/ ftp://genome-ftp.stanford.edu/pub/yeast/

[He02] Heymann S. et. al.: Viator - A Tool Family for Graphical Networking and Data View Creation. Very Large Data Bases, HongKong SAR, China, 2002. http://www.dbis.informatik.hu-berlin.de/

[Hy02] Hyun Y.: Walrus - Graph Visualization Tool. 2002. http://www.caida.org/tools/visualization/walrus/

[Ke02A] Keim D. A.: Datenvisualisierung und Data Mining. Datenbank-Spektrum 2, 2002; S. 30-39.

[Ke02B] Kemmeren P. et. al.: Protein Interaction Verification and Functional Annotation by Integrated Analysis of Genome-Scale Data. Molecular Cell 9(5), 2002; S. 1133-1143. http://eptest.ebi.ac.uk/EP/PPI/

[KK01] Kuramochi M., Karypis G.: Frequent Subgraph Discovery. Department of Computer Science/Army HPC Research Center University of Minnesota nr. 02-026, 2001.

[KS02] Kumar A., Snyder M.: Protein complexes take the bait. Nature Vol. 415, 2002; S 123-124.

[Ku02] Kuo W. P. et. al.: Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics vol. 18, 2002; S. 405-412. http://bioinformatics.oupjournals.org/cgi/content/abstract/18/3/405

[Me97] Mewes, H. et. al.: Overview of the yeast genome. Nature 387, 1997; S. 7-8.

[Me02] Mewes H. W. et. al.: MIPS: a database for genomes and protein sequences. Nucleic Acids Research 30(1), 2002; S. 31-4. http://mips.gsf.de/proj/yeast

[Mo02] MolMineAS, Norway, 2002; J-Express v. 2.1. http://www.molmine.com/

[Mu00] Munzner T.: Interactive Visualization of Large Graphs and Networks, Ph.D. Dissertation, Stanford University, 2000. http://graphics.stanford.edu/papers/munzner/thesis/

[Pe01] Peèr D. et. al.: Inferring Subnetworks from Perturbed Expression Profiles. ISMB, 2001.

[Sp98] Spellman P. T. et al.: Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. Molecular Biology of the Cell 9, 1998; S. 3273-3297.