Towards zoomable multidimensional maps of the cell

Zhenjun Hu, Joe Mellor, Jie Wu, Minoru Kanehisa, Joshua M. Stuart and Charles DeLisi

Hauptseminar **„Visualisierung in der Bioinformatik“**

SS 2007

Ausarbeitung zum Seminarvortrag

von

Bianca Selzam

Inhaltsverzeichnis

1. Einleitung und Motivation 3
2. Typen des Zoomens 3
2.1. Geometrisches Zoomen 4
2.2. Semantisches Zoomen 4
3. Typen von Graphen 4
3.1. Einfacher Graph 4
3.2. Multigraph 5
3.3. Compound Graph 5
3.4. Metagraph 6
3.5. Hypergraph 6
4. Biologische Netzwerke und bekannte Darstellungen 7
4.1. KEGG Pathway Diagramm 7
4.2. Einfacher Graph 8
4.3. Compound Graph 9
4.4. Anforderungen an eine verbesserte Visualisierung 10
5. Metagraphen von biologischen Netzwerken 11
6. Anwendungen 12
7. Zusammenfassung und Fazit 14
8. Literaturverzeichnis 15

1. Einleitung und Motivation

Durch den technischen Fortschritt steigt die Größe der gesammelten Daten zu Genen und Proteinen seit Jahren kontinuierlich an. Um eine effiziente Arbeit innerhalb dieser Datenbestände zu ermöglichen, sind geeignete Visualisierungstechniken unverzichtbar, die komplexe Zusammenhänge und Abläufe möglichst intuitiv und transparent darstellen. Jedoch sind graphische Repräsentationen mit zunehmender Größe schnell höchst komplex und unübersichtlich; vor allem bei der Visualisierung von biologischen Bausteinen und Interaktionen, die häufig selbst extrem komplex sein können, führt dies sehr schnell zu nicht mehr überschaubaren graphischen Darstellungen. Gerade im biologischen Kontext sollten geeignete Visualisierungen den Kontext (zeitliche oder modulare Aktivitäten), die Hierarchie der Agenten (Proteine – Komplexe – Pathways usw.) sowie die Interaktionen zwischen den Agenten berücksichtigen und möglichst übersichtlich darstellen.

Die Autoren dieses Papers schlagen vor, die Darstellung der Zelle als Graph zu realisieren, um bereits existierende effiziente Algorithmen für Graphen auf den Visualisierungen anwenden zu können. Von den verschiedenen Graphtypen (s. Kap. 3) wird der Metagraph (s. Kap. 5) bevorzugt, da dieser sich für die Anwendung des Semantischen Zoomens (s. Kap. 2) am besten eignet. Die von den Autoren vorgestellte graphische Repräsentation stellt diverse biologische Zusammenhänge und Interaktionen einfach und effizient dar (s. Kap. 6) und verspricht daher, einen großen Beitrag zum vollständigen Verständnis der Zelle und ihrer Komplexität zu leisten.

2. Typen des Zoomens

Im allgemeinen kann zwischen zwei Typen des Zoomens unterschieden werden – dem geometrischen und dem semantischen Zoomen.

2.1. Geometrisches Zoomen

Beim geometrischen Zoomen wird der ausgewählte Bereich um einen bestimmten Prozentsatz vergrößert – die einzelnen Elemente werden hierbei lediglich in ihrer Größe verändert. Beschriftungen, Symbole und andere Teile der Graphik bleiben erhalten.

2.2. Semantisches Zoomen

Beim semantischen Ein- und Auszoomen werden Objekte der Graphik je nach Zoomstufe ein- oder ausgeblendet. Abhängig von der aktuellen Stufe und vom Kontext werden dynamisch alle Informationen angezeigt, die für das Verständnis der Graphik auf der aktuellen Zoomstufe nötig sind. Überflüssige Informationen, die nur die Übersichtlichkeit schmälern würden und auf einer anderen Ebene sinnvoller sind, werden hierbei verborgen.

Die Stärken des Semantischen Zoomens liegen in der Ausnutzung der hierarchischen Struktur der zu visualisierenden Daten sowie in der automatischen Anpassung der eingeblendeten Informationen zur jeweiligen Zoomstufe. Die Übertragung dieses Konzepts auf biologische Netzwerke würde die größtmögliche Übersichtlichkeit der Visualisierungen gewährleisten und dem oftmals hierarchischen Aufbau von biologischen Agenten Rechnung tragen.

3. Typen von Graphen

Um ein biologisches Netzwerk als Graph darzustellen, ist eine Reflektion über die verschiedenen Graphtypen nötig, um den am besten geeigneten Typ bestimmen zu können.

3.1. Einfacher Graph

Ein Einfacher Graph enthält keine reflexiven Kanten (Kanten zwischen ein und demselben

Knoten) oder mehrfache Kanten zwischen zwei Knoten.

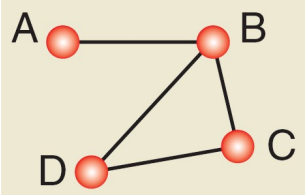

Abb. 1: Einfacher Graph [2]

3.2. Multigraph

Ein Multigraph basiert auf dem Einfachen Graphen, erlaubt jedoch auch mehrfache Kanten zwischen zwei Knoten.

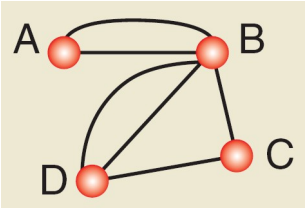

Abb. 2: Multigraph [2]

3.3. Compound Graph

Ein Compound Graph erweitert den Multigraph durch die Einführung von hierarchischen Knoten. Jeder Knoten darf hierbei beliebig viele Unterknoten enthalten, jedoch darf ein Unterknoten nur zu jeweils einem Oberknoten gehören.

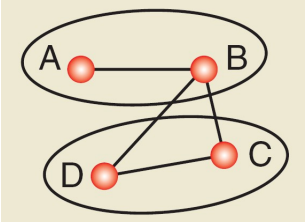

Abb. 3: Compound Graph [2]

3.4. Metagraph

Der Metagraph basiert auf dem Compound Graph und erlaubt auch Überlappungen zwischen den Unterknoten zweier Knoten. Somit kann ein Knoten Unterknoten von mehreren Oberknoten sein.

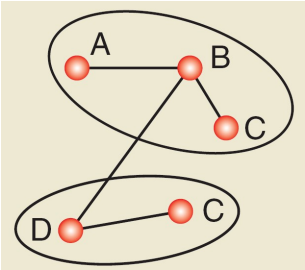

Abb. 4: Metagraph [2]

3.5. Hypergraph

Ein Hypergraph erlaubt die Verbindung von beliebig vielen Knoten durch eine Kante.

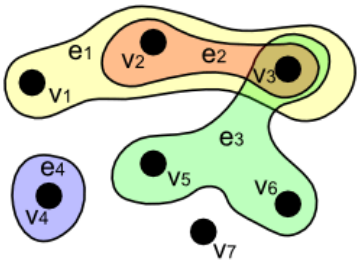

Abb. 5: Hypergraph [7]

Während Einfacher Graph, Multigraph und Compound Graph die hohe Komplexität von biologischen Interaktionen und Agenten nicht ausreichend abbilden können, ist der Hypergraph wegen seiner schnell auftretenden Unübersichtlichkeit ebenfalls nicht für die Visualisierung der biologischen Zelle geeignet. Der Metagraph dagegen scheint sich wegen seiner Fähigkeit, hierarchische und möglicherweise auch sich überlappende Strukturen abbilden zu können, für die Abbildung der Zelle als graphisches Netzwerk am besten zu eignen.

4. Biologische Netzwerke und bekannte Darstellungen

Um eine neue verbesserte Visualisierung für biologische Netzwerke zu entwickeln, müssen zunächst bereits existierende Darstellungen betrachtet und ihre Vor- und Nachteile eruiert werden.

4.1. KEGG Pathway Diagramm

Das KEGG Pathway Diagramm[3] stellt biologische Netzwerke in Graphenform dar; Knoten repräsentieren Proteine oder Proteinkomplexe, während die Kanten zwischen den Knoten Reaktionen oder Übergänge in andere Zustände darstellen. Knoten, die Proteinkomplexe vertreten, lassen sich durch ihre internen Unterteilungen in ihre einzelnen Proteinbestandteile erkennen (vgl. Abb. 6).

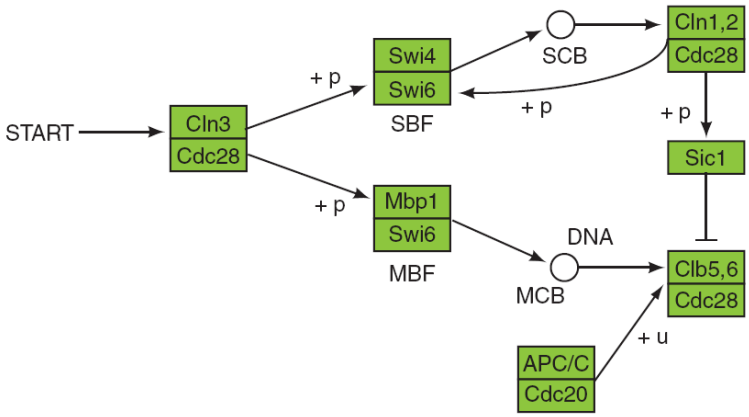

Abb. 6: KEGG Pathway Diagramm der G1-Phase von Bäckerhefe [2]

KEGG Pathway Diagramme nutzen zwar die hierarchische Struktur der biologischen Agenten aus, indem sie die einzelnen Bestandteile von Proteinkomplexen aufzeigen; falls jedoch ein Proteinkomplex zum Teil aus weiteren Proteinkomplexen besteht (z. B. setzt sich der Komplex Cln1,2 aus den Proteinen Cln1 und Cln2 zusammen), ist eine weitere Zerlegung hier nicht vorgesehen.

Das mehrfache Auftreten der Proteine Swi6 und Cdc28 wird im KEGG Pathway Diagramm

durch das simple Duplizieren ihrer Knoten realisiert – daher tritt z. B. in den Komplexen SBF und MBF jeweils ein Swi6-Protein auf. Jedoch wird durch diese Duplikate abgesehen von der identischen Benennung nicht sofort deutlich, dass es sich in beiden Fällen um dasselbe Protein handelt, von welchem zwei Instanzen an verschiedenen Stellen der G1-Phase mit anderen Proteinen interagieren und verschiedene Komplexe bilden.

4.2. Einfacher Graph

Das in Kap. 4.1 in einem KEGG Pathway Diagramm genannte Beispiel lässt sich auch als Einfacher Graph (vgl. Kap. 3.1) darstellen.

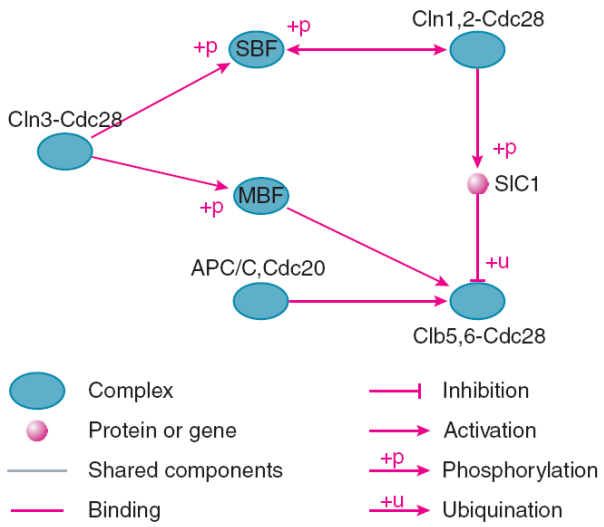

Abb. 7: Einfacher Graph der G1-Phase von Bäckerhefe [2]

Da die Knoten von Einfachen Graphen keine Unterknoten enthalten dürfen, können die Knoten in dieser Darstellung lediglich die auftretenden Proteinkomplexe repräsentieren, jedoch keine Informationen über ihren hierarchischen Aufbau liefern. Lediglich durch die Benennung der Knoten (z. B. Cln1,2-Cdc28) wird klar, wie sich diese Komplexe aus einzelnen Proteinen bzw. anderen Proteinkomplexen zusammensetzen.

Um die Hierarchie dieser Knoten zu visualisieren, müssen getrennte Graphen in Baumform erzeugt werden (vgl. Abb. 8). Diese Zweiteilung der Visualisierung ist sehr unintuitiv und für das Verständnis der komplexen Vorgänge, welche durch die Visualisierungen

abgebildet werden, nicht besonders förderlich.

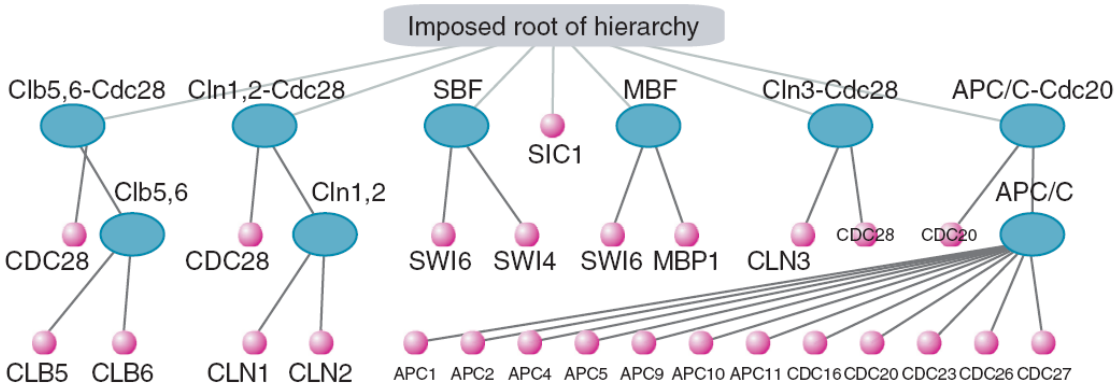

Abb. 8: Hierarchische Struktur der Knoten aus Abb. 7 [2]

4.3. Compound Graph

Um die Vereinigung der hierarchischen Struktur (vgl. Abb. 8) und dem Graphen der Interaktionen (vgl. Abb. 7) in einer einzigen Darstellung zu erreichen, eignet sich der Compound Graph (vgl. Kap. 3.3), da dieser Graphtyp hierarchisch aufgebaute Knoten erlaubt.

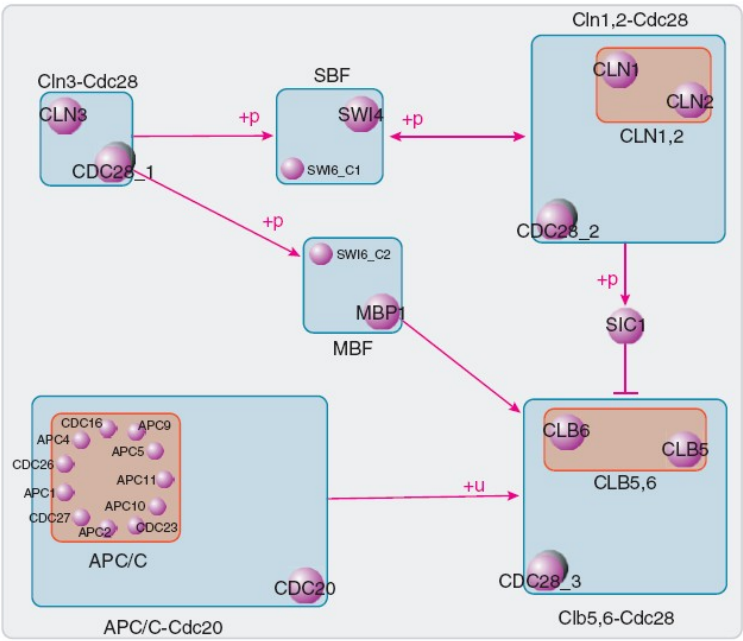

Abb. 9: Compound Graph der G1-Phase von Bäckerhefe [2]

Die Knoten des Compound Graph bilden den hierarchischen Aufbau der Proteinkomplexe (vgl. Abb. 8) ab, während die Kanten die Interaktionen und Zustandsübergänge der Proteinkomplexe verdeutlichen (vgl. Abb. 7). Somit ist die Zweiteilung von Hierarchie und Interaktion bei diesem Graphtyp nicht mehr nötig. Durch die Möglichkeit, mehr als zwei Hierarchieebenen in den Knoten abzubilden, lässt sich hierauf das Konzept des Semantischen Zoomens (vgl. Kap. 2.2) anwenden.

Die Einschränkung des Compound Graph, keine Überlappungen zwischen Knoten zu erlauben, stellt jedoch einen Nachteil für die Visualisierung von biologischen Netzwerken dar. Das mehrfache Auftreten von Proteinen in verschiedenen Stellen im Netzwerk wird, wie beim KEGG Pathway Diagramm (vgl. Kap. 4.1), im Compound Graph durch das Einführen von Duplikaten realisiert, wobei in diesem Fall die Duplikate durch verschiedene Benennungen (z. B. Swi6_C1 und Swi6_C2) unterschieden werden. Dieses Verfahren ist sehr mühselig und unpraktikabel, da durch diese Benennungskonvention nicht sofort deutlich wird, dass diese Elemente Instanzen ein- und desselben Proteins sind.

4.4. Anforderungen an eine verbesserte Visualisierung

Durch die in Kap. 4.1 - 4.3 beschriebenen Beispiele wird deutlich, welche Schwachstellen die bereits existierenden Visualisierungsmöglichkeiten aufweisen. Der Compound Graph (vgl. Kap. 4.3) scheint bislang die erfolgversprechendste Repräsentationsform für biologische Netzwerke zu sein, da er hierarchisch aufgebaute Knoten zulässt, was wiederum Semantisches Zoomen ermöglicht; jedoch erlaubt der Compound Graph keine Instantiierung von Knoten, was in Kap. 5 mit dem Einführen des Metagraphen als bevorzugten Graphtypen gelöst werden soll.

5. Metagraphen von biologischen Netzwerken

Der Metagraph (vgl. Kap. 3.4) erweitert den Compound Graph dahingehend, dass auch zwischen Subgraphen von Metaknoten Überlappungen existieren dürfen. Dies ist vor allem für die Fälle hilfreich, in denen Proteine in einem Pathway mehrfach auftreten, evtl. in verschiedenen Zuständen und in Verbindung mit verschiedenen anderen Proteinen oder Proteinkomplexen.

Um die Bildung von Duplikaten und somit Redundanzen zu vermeiden, sollen die Knoten von mehrfach auftretenden Proteinen als Instanzen verstanden werden, welche auf ein- und denselben Knoten (d. h. das Protein bzw. den Komplex) verweisen. Durch dieses Instanzenkonzept wird auf der Datenstrukturebene das Speichern von redundanten Informationen vermieden, da die bausteinspezifischen Daten nur einmal in der Datenstruktur des jeweiligen Bausteins gespeichert werden und die Instanzen sich lediglich ihre spezifischen Informationen wie Position oder Zustand merken.

Ein Knoten, der mindestens einen Unterknoten enthält, wird als Metaknoten bezeichnet. Die interne Struktur eines Metaknoten kann wieder als Graph verstanden werden; somit bildet ein Metaknoten einen Subgraphen des eigentlichen Graphen. Ein Metaknoten kann sich in zwei Zuständen befinden:

- **expandiert:** Der interne Subgraph des Metaknoten wird angezeigt, d. h. in der Visualisierung wird der Metaknoten als Graph dargestellt.
- **kontrahiert:** Der interne Subgraph des Metaknoten wird verborgen, d. h. in der Visualisierung wird der Metaknoten als einfacher Knoten dargestellt.

Die Verwendung von Metagraphen als bevorzugte Visualisierungsform wird u. a. durch die Entwicklung von Markup-Sprachen wie SBML[5] oder KGML[4] unterstützt. Das Instanzenkonzept sowie die Ausnutzung der hierarchischen Struktur der biologischen Elemente kommen auch in diesen Auszeichnungssprachen zum Tragen.

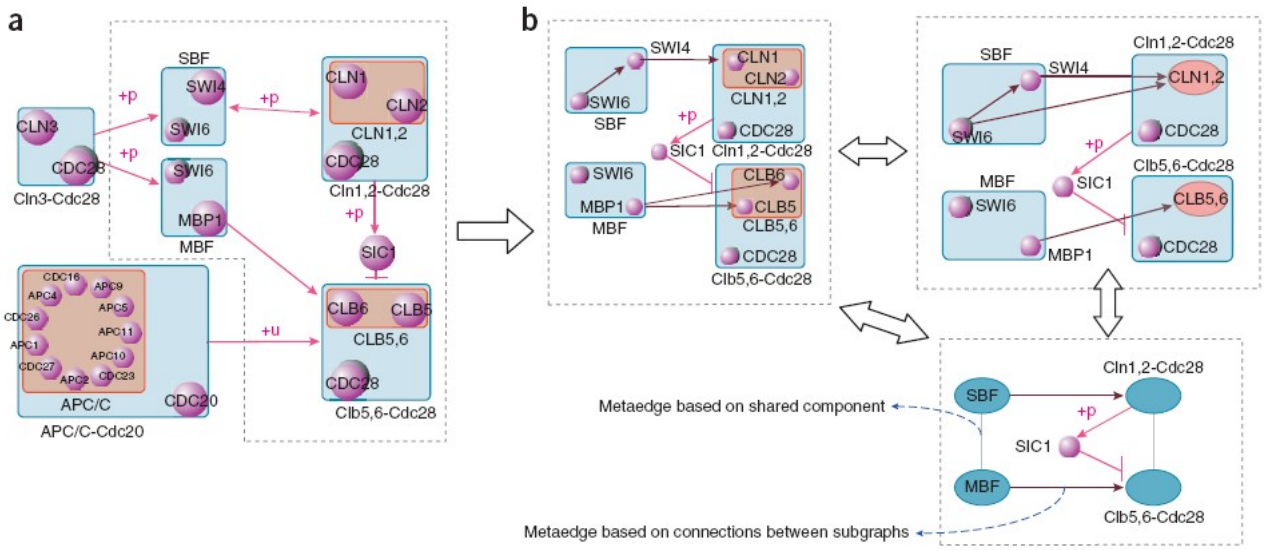

Abb. 10: Metagraphen der G1-Phase von Bäckerhefe in verschiedenen expandierten und kontrahierten Zuständen [2]

Abb. 10 zeigt eine Auswahl von teilweise expandierten und kontrahierten Zuständen des Beispiels der G1-Phase der Bäckerhefe. Graphik a stellt den vollständig expandierten Zustand des Metagraphen dar, da alle Subgraphen der Metaknoten angezeigt werden; Graphik b dagegen zeigt nur teilweise expandierte Zustände bis hin zum vollständig kontrahierten Zustand. Das Konzept des Semantischen Zoomens lässt sich hervorragend auf diese Darstellung übertragen, da die verschiedenen Expandierungszustände als einzelne Zoomstufen verstanden werden können. Der vollständig expandierte Zustand stellt den Graphen auf der höchsten Zoomebene dar, während der vollständig kontrahierte Zustand die niedrigste Zoomstufe repräsentiert. Die anderen Zwischenzustände bilden die Zoomlevel zwischen der höchsten und der niedrigsten Stufe. Die Instanzen von mehrfach auftretenden Proteinen (z. B. Swi6 oder Cdc28) werden in den Metagraphen als schattierte Knoten dargestellt.

6. Anwendungen

Das viel versprechende Konzept des Metagraphen als Visualisierungsform für biologische Netzwerke lässt sich leicht auf die unterschiedlichsten Anwendungsbereiche übertragen. Um bereits existierende Datenbestände in einem Metagraph zu visualisieren, können

hierarchisch annotierte Daten (z. B. in der Gene Ontology[1]) auf allen Ebenen in den Metagraphen übernommen werden. Für die unterste Ebene wird pro Element je ein Knoten definiert, für alle anderen Ebenen dagegen Metaknoten, die jeweils eine Menge von Unterknoten aus der nächst niedrigeren Ebene enthalten. Zwischen einem Metaknoten und den darin enthaltenen Unterknoten existieren demnach Inklusionsrelationen wie „is a“ oder „part of“, was sich mit dem Konzept des Semantischen Zoomens deckt.

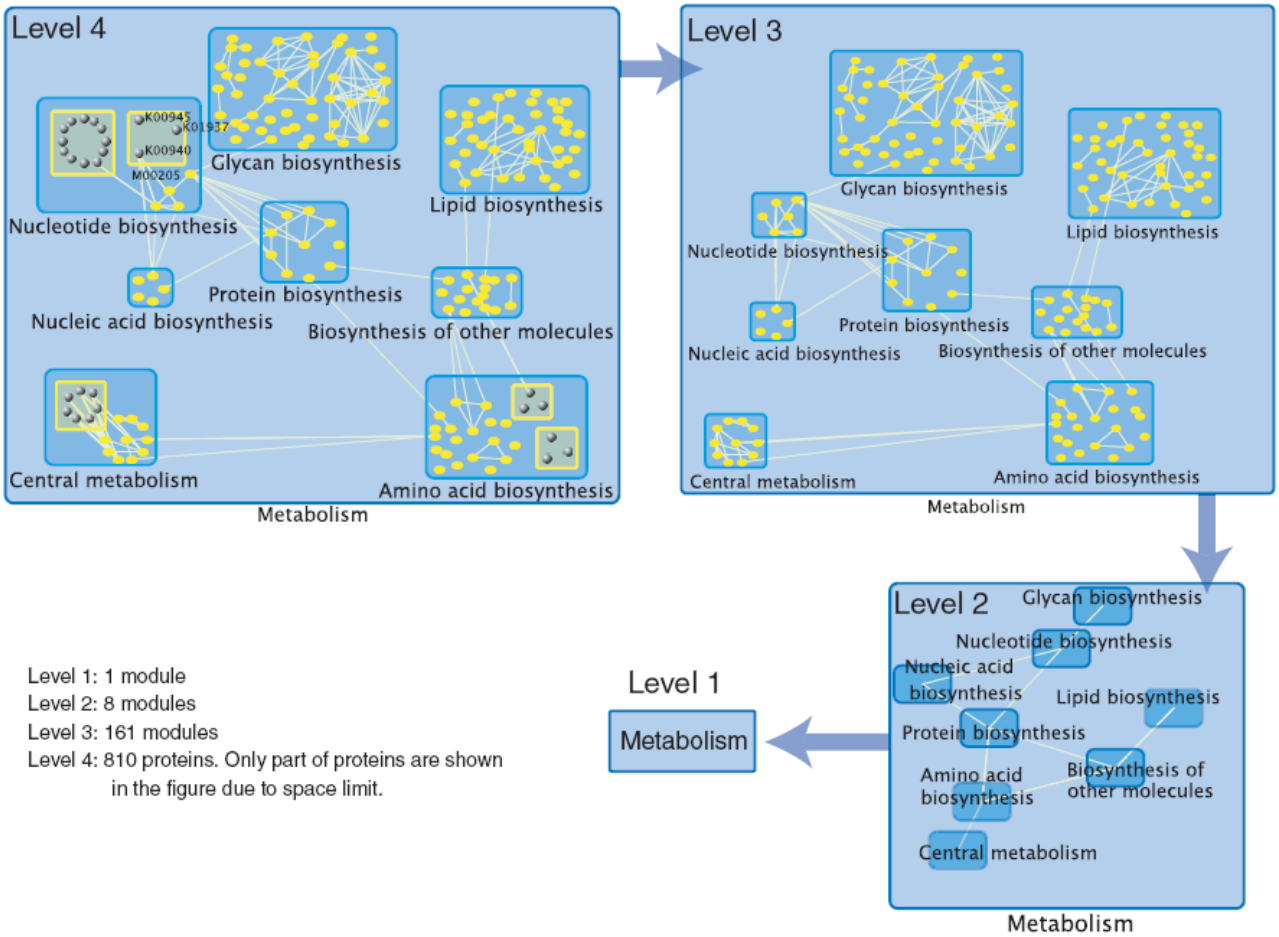

Abb. 11: Zoombare Darstellung des Metabolismus als Metagraph [2]

Abb. 11 zeigt vier verschiedene Zoomlevel eines Metagraphen am Beispiel des Metabolismus. Diese Graphik wurde mit dem Tool VisANT[6] aus den Daten der KEGG Datenbank[3] erzeugt und zeigt, wie sehr auch bei großen Datenmengen und großer Anzahl von Knoten die Übersichtlichkeit auf jeder Zoomstufe erhalten bleibt. Abb. 12 zeigt einen Screenshot des Tools, das frei heruntergeladen werden kann und in verschiedenen Java-

Ausführungen verfügbar ist.

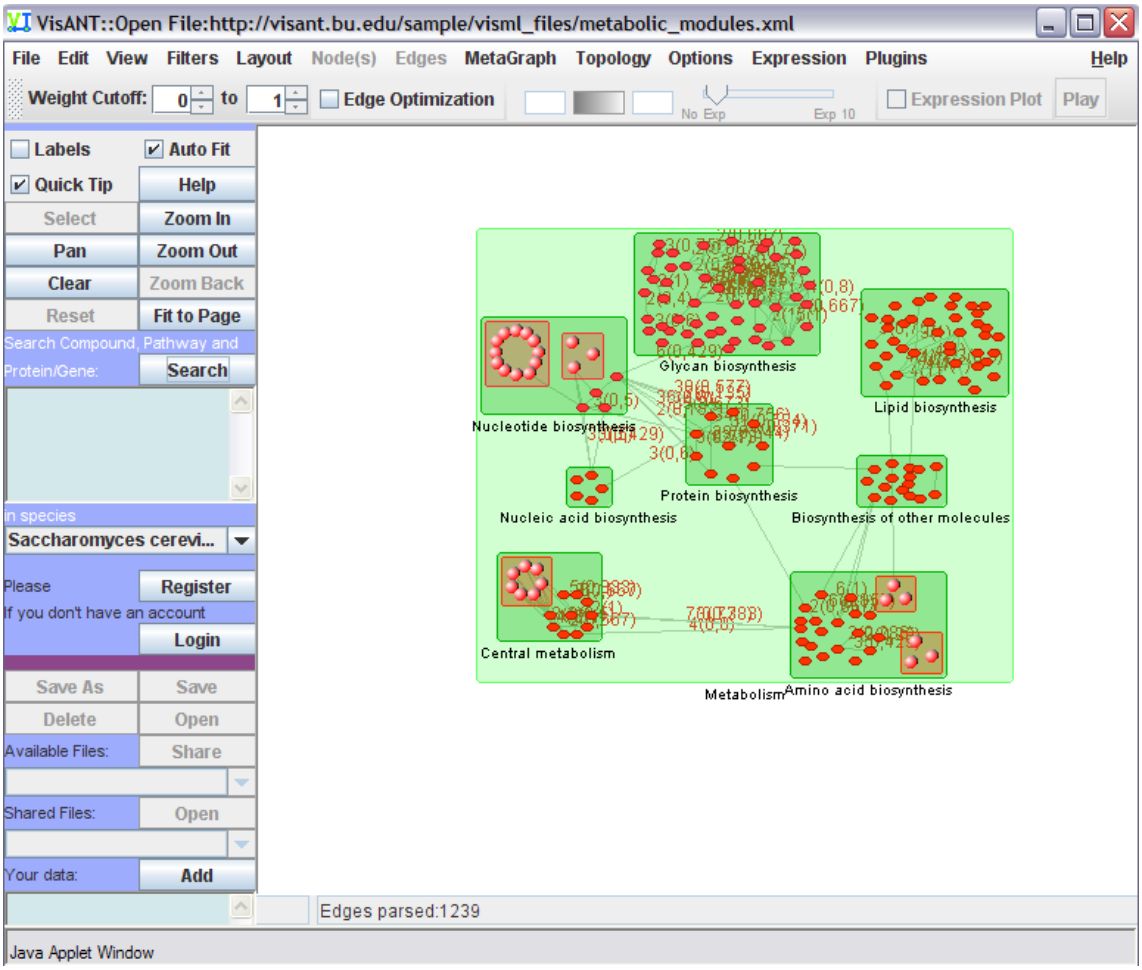

Abb. 12: Screenshot des Tools VisANT zur Erzeugung von Metagraphen[6]

7. Zusammenfassung und Fazit

Die Übertragung von Metagraphen auf biologische Netzwerke scheint ein viel versprechender Schritt zu sein auf dem Weg, die Anpassungsfähigkeit und Komplexität einer Zelle in einer Visualisierungsform abzubilden. Durch die Fähigkeit, hierarchisch aufgebaute Knoten darzustellen, kann das Semantische Zoomen auf Metagraphen angewendet werden, was die Anzeige von allen pro Zoomstufe relevanten Informationen bei gleichzeitiger höchstmöglicher Übersichtlichkeit gewährleistet. Durch die Instantiierung von Knoten werden redundante Informationen vermieden und die Fähigkeit der Proteine,

in unterschiedlichen Zuständen mit verschiedenen anderen Bausteinen zu interagieren, widergespiegelt.

Metagraphen können in der Zukunft eine wichtige Rolle spielen, wenn die Entwicklung von dynamischen Visualisierungen, die je nach Zustand und Position eines Bausteins mögliche Interaktionen oder Reaktionspartner anzeigen, voranschreitet – dies würde einen weiteren Schritt auf dem Weg zum umfassenden Verständnis der Komplexität und Adaptivität der biologischen Zellen darstellen.

8. Literaturverzeichnis

[1] Gene Ontology: http://www.geneontology.org/ (Stand 25.7.2007)

[2] Hu, Zhenjun et al.: Towards zoomable multidimensional maps of the cell. In: Nature Biotechnology 25, pp. 547 – 554, 2007.

[3] KEGG: http://www.genome.jp/kegg/ (Stand 25.7.2007)

[4] KGML: http://www.genome.jp/kegg/xml/ (Stand 25.7.2007)

[5] SBML: http://sbml.org/index.psp (Stand 25.7.2007)

[6] VisANT: http://visant.bu.edu/sample/metagraph_nature.htm (Stand 4.7.2007)

[7] Wikipedia: http://en.wikipedia.org/wiki/Hypergraph (Stand 4.7.2007)