Visuelle Analytik biologischer Daten

Kay Nieselt · Michael Kaufmann
Andreas Gerasch · Hans-Peter Lenhof
Marcel Spehr · Stefan Hesse
Stefan Gumhold

In memoriam Dirk Bartz

Einleitung

Biologische Daten sind heterogen, sehr komplex und oft sehr groß. Eine angemessene Visualisierung leistet hierbei einen entscheidenden Beitrag zum Verständnis der Daten. Die Visualisierung biologischer Daten spielt daher auch eine zentrale Rolle in der Bioinformatik. Hier reichen die Anwendungen von der Visualisierung einzelner Proteine oder ganzer Genome, von Familien von Genen, evolutionären Verwandtschaftsverhältnissen, makromolekularer Strukturen, mikroskopischer Bilddaten bis hin zur Darstellung von metabolischen oder regulatorischen Netzwerken und systembiologischer Daten (siehe Abb. 1 für einige Beispiele). Aufgrund der zunehmenden Komplexität und Verbundenheit biologischer Daten (man denke hier insbesondere an systembiologische Daten) ist eine integrative sowie standardisierte Visualisierung und die Entwicklung leistungsstarker und benutzerfreundlicher Werkzeuge von wachsender Bedeutung. Eine zunehmend große Rolle spielen dabei Werkzeuge, die zudem das Paradigma der visuellen Analytik (engl. „visual analytics“) verfolgen. Grundsätzlich integriert die visuelle Analytik Visualisierung, Datenanalyse und Interaktion durch den Menschen. Bei biologischen Daten hat der Einsatz der visuellen Analytik darüber hinaus das Ziel, die komplexen experimentellen Daten in Wissen umzusetzen. Forschern soll ein System zur Verfügung gestellt werden, das es erlaubt, Einsichten in biologische Prozesse in Zellen, Geweben und schließlich Organismen zu gewinnen sowie eine Modellierung biologischer Systeme vorzunehmen.

Die visuelle Analytik biologischer Daten hat erst vor kurzem auf den Information-Visualization- und Visual-Analytics-Konferenzen Beachtung gefunden. Einige auch in wissenschaftlichen Zeitschriften publizierte Artikel sind Themen wie Clustering von Expressionsdaten [17], der Genom-Assemblierung [18] oder der Bestimmung von Funktionen von Genen in neu sequenzierten Genomen [20] gewidmet. Die diesjährige IEEE VAST Challenge stand ganz im Zeichen einer biologischen Fragestellung.

Im folgenden Artikel möchten wir insbesondere auf drei der oben genannten Teilgebiete der Visualisierung biologischer Daten genauer eingehen: die visuelle Analytik von Genexpressionsdaten, die Visualisierung biologischer Netzwerke sowie die inhaltsbasierte Suche in zellbiologischen Bilddatenbanken.

DOI 10.1007/s00287-010-0482-y

Kay Nieselt
Zentrum für Bioinformatik Tübingen,
Universität Tübingen,
Sand 14, 72076 Tübingen
E-Mail: kay.nieselt@uni-tuebingen.de

Michael Kaufmann · Andreas Gerasch
Wilhelm-Schickard-Institut für Informatik,
Universität Tübingen,
Sand 13, 72076 Tübingen
E-Mail: {mk, gerasch}@informatik.uni-tuebingen.de

Hans-Peter Lenhof
Saarland University,
Im Stadtwald, 66123 Saarbrücken
E-Mail: lenhof@bioinf.uni-sb.de

Marcel Spehr · Stefan Hesse · Stefan Gumhold
Fakultät Informatik, TU Dresden,
Nöthnitzer Straße 46, 01187 Dresden
E-Mail: {marcel.spehr, stefan.hesse, stefan.gumhold}
@tu-dresden.de

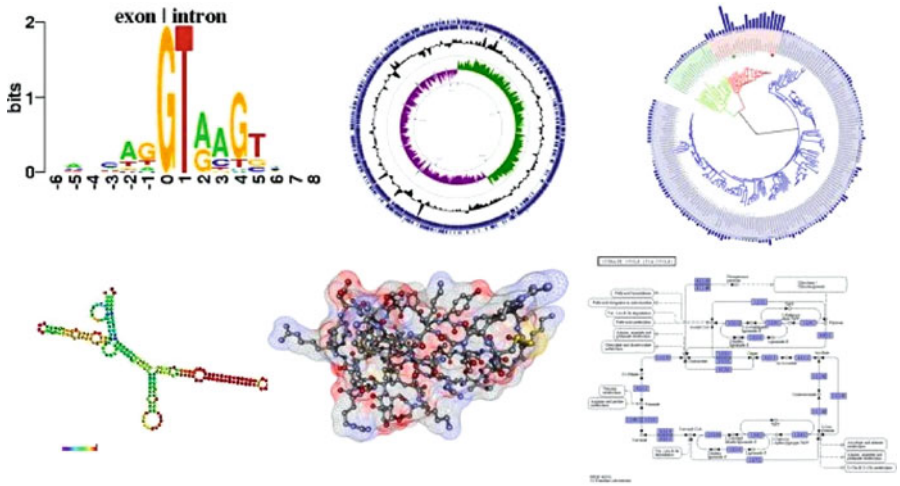

Abb. 1 Beispiele für typische Visualisierungen biologischer Daten. Von* oben, links *nach* rechts*: Profil-Alignments als SeqLogos (http://weblogo.berkeley.edu), zirkuläre Darstellung des Genoms und weiterer Annotationen des Bakteriums* Staphylococcus carnosus*, Phylogenie, RNA-Sekundärstruktur mit einer farblichen Kodierung der Basenpaarwahrscheinlichkeit, Proteinstrukturmodellierung (BallView), metabolisches Netzwerk des Zitratzyklus (KEGG)

Visuelle Analytik von Genexpressionsdaten

Kay Nieselt

DNA-Microarrays sind Werkzeuge, die eine parallele Analyse der Genexpression erlauben. Unter Genexpression versteht man den Prozess des Umsetzens der im Gen enthaltenen Information in das entsprechende Genprodukt (z. B. Protein). Aufgrund der Verfügbarkeit vollständiger Genome sowie der Miniaturisierung der Microarrays erlauben moderne Microarrays bis zu 1,5 Mio. Sonden pro Array. Mithilfe von Microarrays lassen sich unter anderem Momentaufnahmen der Gesamtmenge der in einem Organismus zu einem bestimmten Zeitpunkt vorhandenen Transkripte von Genen parallel erheben. Dieses nennt man auch Expressionsprofilierung: Wann, wo und in welcher Höhe sind alle Gene in der Zelle exprimiert. Dabei werden üblicherweise mehrere Zustände in einem vergleichenden Experiment untersucht. Eines der Ziele dabei ist es, Gruppen von Genen zu identifizieren, die unter bestimmten Zuständen ein ähnliches Expressionsverhalten zeigen.

Daten von Microarray-Experimenten werden typischerweise in einer sogenannten Expressionsmatrix erfasst, wobei in den Zeilen die Expressionswerte von Genen unter den verschiedenen Bedingungen, die den Spalten entsprechen, erfasst werden. Diese Matrix hat die Größe $n \times p$, wobei n die Anzahl Gene und p die Anzahl der Experimente ist. Zumeist ist n sehr viel größer als p.

Eine typische analytische Methode, Synexpressionsgruppen zu identifizieren, ist Clustering. Clusteralgorithmen produzieren entweder eine hierarchische Zerlegung oder Partitionierung in nichtüberlappende Teilmengen der Originaldaten. Das Ergebnis von Clustering bewirkt eine Ordnung bzw. Sortierung der Daten, die sich dann geeigneter Weise zumeist in sogenannten Heatmaps visualisiert lassen. Grundsätzlich präsentieren Heatmaps einen tabularen Blick auf eine Datenmatrix, d. h. das Experimentergebnis, bei dem die Farbe die Änderung des gemessenen Genexpressionswerts zu einem Kontrollexperiment zeigt. Typischerweise wird dabei ein Grün-schwarz-rot-Farbgradient verwendet. Daher auch der Name „Heatmap“: die rote Farbe (warm) repräsentiert die erhöhte Aktivität des Gens, während grün (kalt) die verringerte Aktivität visualisiert. Heatmaps sind visuell am Eingängigsten, wenn die Zeilen und Spalten der Expressionsmatrix so geordnet sind, dass die Muster der Ähnlichkeit deutlich identifizierbar sind. Clustering und Diskriminantenanalyse werden oft verwendet, um die Gruppierungen zu berechnen. Dann werden die Zeilen bzw. Spalten der Expressionsmatrix so sortiert, dass Elemente (Gene oder Experimente) einer Gruppe benachbart sind. Die meisten Heatmap-Implementierungen beschrän-

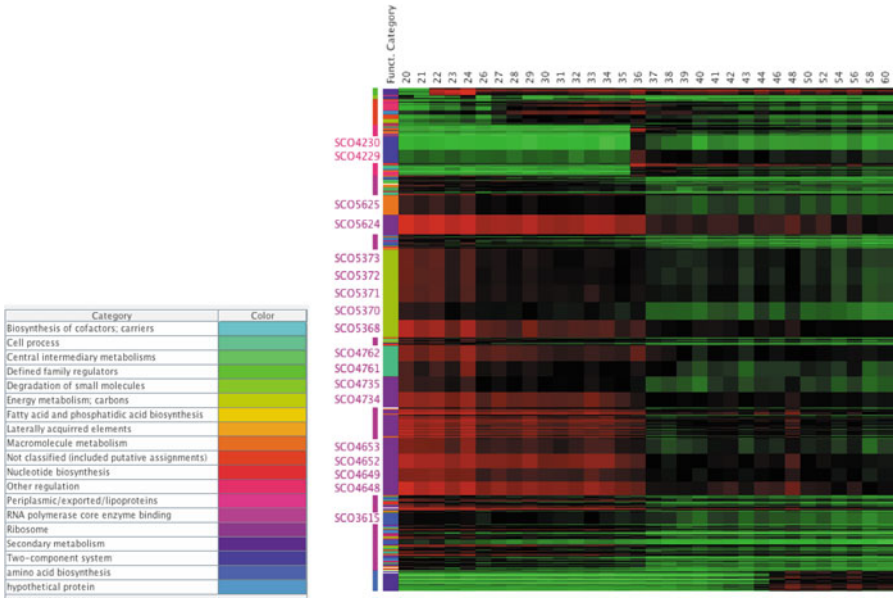

Abb. 2 Erweiterte Heatmap (enhanced Heatmap), hier einer Zeitserie von unter Phosphatmangel gewachsenen Bakterien [13]. Zusätzlich zur Darstellung der eigentlichen Expressionsdaten (mit einem Grün-schwarz-rot-Gradienten, der die Höhe der Expression widerspiegelt) aller varianten Gene, sind hier zwei weitere Spalten, die farblich kodiert sowohl die Cluster als auch die funktionale Kategorisierung (Farbzuordnung der Kategorien links) der Gene darstellen, visualisiert. Zudem sind alle Zeilen höhenskaliert bzgl. eines abgeleiteten p-Wertes für die statistische Signifikanz differenzieller Expression um den Zeitpunkt der Phosphaterschöpfung (35 h). Erkennbar ist, dass Gene mit ähnlichem Expressionsprofil oft in einer gemeinsamen funktionalen Gruppe liegen. Durch die Höhenskalierung werden die Gene visuell hervorgehoben, die eine differenzielle Regulation um den Zeitpunkt der Phosphaterschöpfung aufweisen

ken sich auf die Visualisierung der eigentlichen Expressionsdaten. Mit der sogenannten erweiterten Heatmap [5] (enhanced heatmap) haben wir in unserer Microarray-Software Mayday [3] die Möglichkeit geschaffen, durch einen zusätzlichen Farbgradienten Höhenskalierung der Zeilen sowie zusätzlichen Spalten zur Darstellung von z. B. abgeleiteten statistischen Parametern (z. B. p-Wert) oder Annotationsdaten die gemessenen Daten mit weiteren Daten in einer gemeinsamen Visualisierung darzustellen (Abb. 2).

Eine weitere wichtige visuelle Darstellung von Expressionsprofilen ist der Profilplot. Der Profilplot ist eine Adaptierung des Konzepts der parallelen Koordinaten, bei dem eine Koordinatenachse einem Experiment entspricht. Parallele Koordinatenplots (PKP) haben sich insbesondere bei dieser Art hochdimensionaler Daten als sehr geeignet erwiesen. In SpRay [4] ist das Konzept des PKPs implementiert und die Darstellung der experimentellen Daten um abgeleitete statistische Daten erweitert, die als zusätzliche Koordinaten modelliert werden (Abb. 3 links). Diese Kombination erlaubt es, Zusammenhänge zu detektieren, die allein mit visuellen Methoden möglicherweise unerkannt blieben. Damit verfolgen wir mit SpRay das Paradigma der visuellen Analytik zur Analyse von komplexen Expressionsdatensätzen umzusetzen. PKPs sind insbesondere in Verbindung mit Clusteralgorithmen sehr mächtig. Der PKP ist vornehmlich für die Visualisierung von Clustern geeignet, die mithilfe von Partitionierungsverfahren wie k-means oder Qt berechnet wurden. Jedes Cluster wird dann mit einem eigenen PKP visualisiert, womit das Problem der visuellen Überlappung verschiedener Expressionsmuster reduziert wird. Die Kombination von Clustering und Visualisierung mit PKPs haben wir unter anderem in Mayday implementiert (s. Abb. 3 rechts).

Eine wichtige Erweiterung ist die visuelle Darstellung differenziell exprimierter Gene im Zusammenhang mit ihrer genomischen Lokalisierung (sogenannte Chromosomenplots). Hier ist eine besondere Visualisierung erforderlich, die einerseits der inhärenten Linearität (Eindimensionalität) des Chromosoms als auch der beobachteten Expressionsveränderung genüge tut. Typische

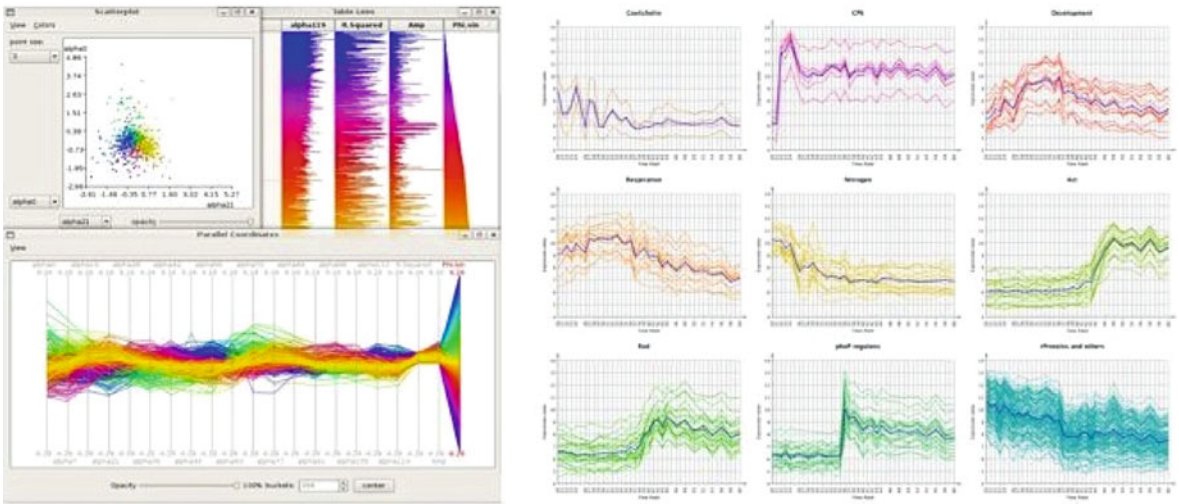

Abb. 3 **Links***: Screenshot der visuellen Analytik-Software SpRay, dabei ist* **unten** *der parallele Koordinatenplot mit erweiterten Dimensionen, die abgeleitete statistische Daten darstellen und somit den explorativen Ansatz der Datenanalyse unterstützen.* **Rechts***: Mit Mayday erzeugter Multi-Parallel-Koordinatenplot nach Clustering*

Visualisierungen verwenden daher einen „Track"-basierten Ansatz, der auch in den mithilfe moderner Sequenzierungsmethoden erzeugten Daten eingesetzt wird. Der UCSC-GenomeBrowser [15] gehört dabei zu den am meisten eingesetzten Webtools dieser Art.

Auch Expressionsdaten werden heutzutage mithilfe moderner Sequenzierungsmethoden erzeugt. Dies bezeichnet man als RNA-Seq. Die Größe der Daten (Rohdaten erzeugen zwischen 0.45–50 Gb pro Experiment) sowie die Ungenauigkeit des Abbildungsvorgangs, bei dem die sehr kurzen Sequenzfragmente ihrer zugehörigen genomischen Position im Genom zugeordnet werden, erfordert eine differenzierte und skalierbare Software mit einem neuen Ansatz für die Informationsvisualisierung. Bei der Visualisierung steht dabei derzeit die Darstellung der Anzahl der abgebildeten Sequenzfragmente bzgl. ihres zugehörigen genomischen Lokus (digitales Profil) im Vordergrund. Visuelle Analytik von RNA-Seqdaten spielt derzeit noch eine untergeordnete Rolle. In Zukunft werden jedoch Werkzeuge, die beispielsweise die Vorhersage und gleichzeitige visuelle Darstellung von Genarchitekturen (Stichwort alternatives splicing oder Transkriptgrenzen), Genfusionen oder die Detektion von bislang unentdeckten Genen erlauben, eine große Rolle spielen.

Grundsätzlich gilt es bei Expressionsdaten wie auch bei all den biologischen Daten, die mithilfe von Hochdurchsatzverfahren erzeugt werden, neue Methoden der visuellen Exploration und Interaktion zusammen mit fortgeschrittenen Statistikmethoden zu entwickeln, um die relevante Information aus den potenziell riesigen Datenmengen zu extrahieren. Dabei sind integrative und benutzerfreundliche Visualisierungswerkzeuge, die zudem die menschliche Kognitionsfähigkeit unterstützen, für diese hochkomplexen biologischen Daten in Zukunft von essenzieller Bedeutung.

Visualisierung biologischer Netzwerke

Michael Kaufmann, Andreas Gerasch, Hans-Peter Lenhof

Der Stoffwechsel (Metabolismus) von Zellen oder auch die Regulation von Genen kann als Netzwerk verstanden und modelliert werden, welches das Zusammenspiel der beteiligten Moleküle beschreibt. Unterschiedliche Anwendungen führten zur Entwicklung verschiedener Klassen von biologischen Netzwerken, wobei wir im Folgenden nur die drei wichtigsten Klassen, metabolische, regulatorische und Protein-Protein-Interaktion-Netzwerke (PPI-Netzwerke), diskutieren werden. In letzteren werden Proteine als Knoten und deren direkte physikalische Interaktion als Kanten dargestellt. Regulatorische Netzwerke hingegen, zu denen man auch Signalkaskaden zählt, beschreiben unter anderem die Hemmung oder Aktivierung der Expression eines

Gens durch Signalkaskaden von Interaktionen und Reaktionen zwischen anderen Genen bzw. Proteinen. Kanten in diesem Netzwerk beschreiben oft auch indirekte Interaktionen und können in einigen Fällen sogar Teile eines PPI-Netzwerks repräsentieren. Ein biologisches Ziel von Genregulation ist es, die Aktivität von Enzymen und somit den Stoffwechsel einer Zelle zu steuern, also den Umbau von biochemischen Verbindungen in die gerade benötigten Komponenten zu kontrollieren. Diese Stoffwechselprozesse werden durch metabolische Netzwerke modelliert.

Die Analyse und die Simulation biochemischer Prozesse, basierend auf biologischen Netzwerken, sind heutzutage wichtige Werkzeuge der Forschung in den Lebenswissenschaften. Dabei ist die anschauliche Visualisierung der Netzwerke sowie der Resultate der Analysen und Simulationen wesentlich für das Verständnis der komplexen biochemischen Prozesse, zumal durch die fortscheitende Entwicklung neuer Hochdurchsatztechnologien die Anzahl bekannter Regulationen und Interaktionen extrem zugenommen haben.

Die besonderen Anforderungen an die Visualisierung biologischer Netzwerke verdeutlicht Abb. 4, welche nur eine einzige Reaktion aus einem metabolischen Netzwerk mit tausenden Reaktionen zeigt. Die Abbildung zeigt die Umwandlung von Citrat zu Oxalacetat mit allen beteiligten Molekülen und deren Rollen. Diese Art der Darstellung wurde durch Gerhard Michals Karte von Stoffwechselpfaden [12] populär. Weltweit arbeiten viele Forschergruppen an Werkzeugen zur Visualisierung und Analyse biologischer Netzwerke mittels Verfahren aus dem Graphenzeichnen, wobei Cytoscape [19] das wohl bekannteste Tool ist. Einen guten Überblick findet man in dem Übersichtsartikel von Suderman und Hallet [21]. Bei fast allen Tools gibt es jedoch die Einschränkung, dass die biologischen Netzwerke

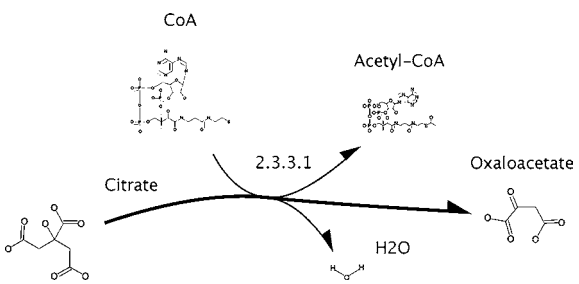

Abb. 4 Citratoxaloacetat-lyase

getrennt voneinander modelliert und visualisiert werden, obwohl sie stark ineinandergreifen. Ziel sollte es daher sein, die verschiedenen Netzwerke integriert zu visualisieren, um so das Erkennen von Zusammenhängen über die Grenzen hinweg zu erleichtern. Die einfache Anwendung von Graphenzeichenmethoden reicht hierfür jedoch nicht mehr aus, zu verschieden sind dafür die Modelle und Bedürfnisse an die Visualisierung.

Wir haben das Tool BiNA (Biological Network Analyzer) entwickelt, das dem Nutzer die Visualisierung von biologischen Netzwerken und Daten erlaubt und das auch eine Reihe von Werkzeugen für die Analyse der Netzwerke zur Verfügung stellt. Darüber hinaus zeichnet sich BiNA dadurch aus, dass es als Frontend für das umfangreiche Data-Warehouse System BN++ [8] genutzt werden kann, welches Zugriff auf eine Reihe von integrierten Netzwerkdatenbanken bietet, die aus ganz verschiedenen Netzwerktypen bestehen. Das nichttriviale Problem der Integration von Netzwerken aus verschiedenen Datenbanken wird durch ein einheitliches Datenmodell BioCore [8] unterstützt. Durch die Plugin-basierte Architektur ist es einfach möglich, BiNA um neue Algorithmen oder grafische Features zu erweitern, um so anwendungsspezifische Visualisierungen der Daten zu realisieren.

Ein Schwerpunkt von BiNA liegt auf der Visualisierung von unterschiedlichen Netzwerktypen, die sowohl separat als auch ineinander integriert dargestellt werden können. Durch die Verknüpfung klassischer Graphlayouts mit neuen, aus der biologischen Anwendung abgeleiteten Verfahren, ist es möglich, große Netzwerke übersichtlich, informativ und interaktiv darzustellen. Übersichtlichkeit wird unter anderem durch verschiedene Algorithmen zur Mehrfachdarstellung von Molekülen und der Möglichkeit, Teile des Netzwerks zusammenzufassen, erreicht. Ferner erlaubt es BiNA dem Nutzer, verschiedenste Daten auf die Netzwerke abbilden zu können, um so die Auswertung von Experimenten zu unterstützen. Im einfachsten Fall sind dies skalare Werte (z. B. Genexpressionsdaten), die auf die Knotenfarbe abgebildet werden, aber fast jede andere denkbare Art der Veränderung der Darstellung ist ebenso möglich, siehe Abb. 5. Abbildung 6 zeigt eine aus der biologischen Anwendung abgeleitete Darstellung, wobei die unterschiedlich gefärbten Schichten verschiedene Zellkompartimente repräsentieren. Hierbei werden die Knoten den Schichten

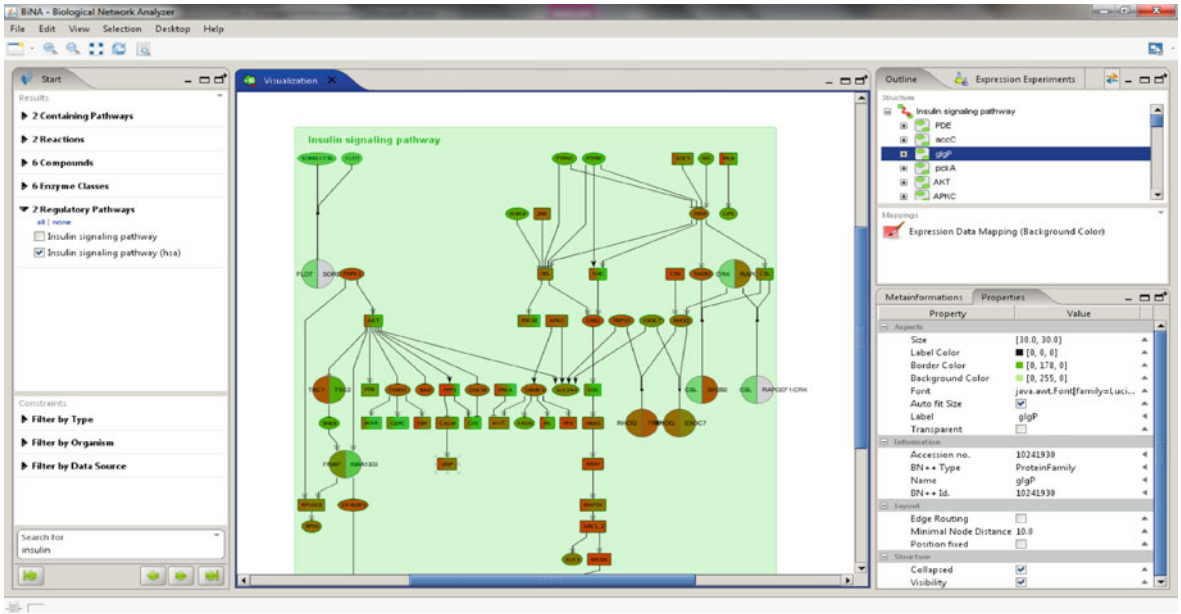

Abb. 5 Screenshot von BiNA, bei dem Genexpressionen in einem kleinen regulatorischen Netzwerk dargestellt werden. Die Formen der Knoten repräsentieren verschiedene Molekülklassen bzw. Proteinkomplexe, die Farben kodieren die Stärke der Expression der beteiligten Gene bzw. Proteine (grün = schwach, rot = stark)

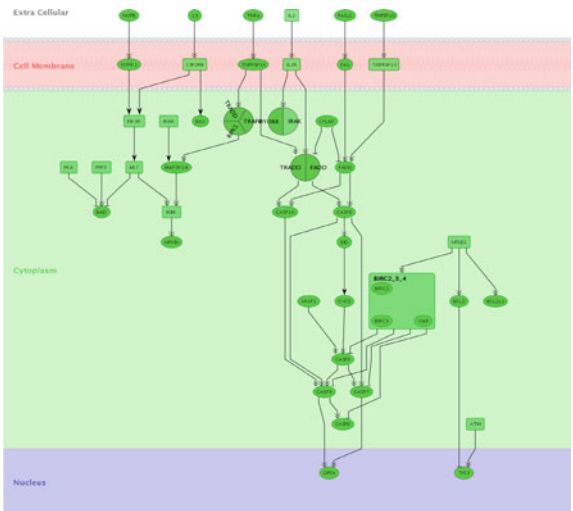

Abb. 6 Darstellung der Apoptose mittels eines vereinfachen Zellschemas

zugeordnet, in denen die entsprechenden Proteine lokalisiert und aktiv sind. Diese Art der Darstellung veranschaulicht den Informationsfluss durch molekulare Signalkaskaden in der Zelle.

BiNA wird zum Beispiel für die Visualisierung von Netzwerken vom Webservice GeneTrail [2] eingesetzt. GeneTrail ist ein leistungsfähiges Tool, das es dem Nutzer ermöglicht, die An- oder Abreicherung von Gen- oder Proteinmengen in einer Vielzahl von biologischen Kategorien und Netzwerken zu untersuchen. Nach einer GeneTrail-Analyse können die resultierenden signifikanten Netzwerke mittels BiNA visualisiert, weiter bearbeitet und analysiert werden. BiNA wird auch zur Visualisierung deregulierter Teilnetzwerke eingesetzt [7]. Diese liefert Ansatzpunkte für die Suche nach Pfaden in Genregulationsnetzwerken oder Signalkaskaden, die zu pathogenen Änderungen im entsprechenden Stoffwechsel führen, welche durch eine Krankheit, z. B. Krebs, ausgelöst wurden.

Inhaltsbasierte Suche in zellbiologischen Bilddatenbanken

Marcel Spehr, Stefan Hesse, Stefan Gumhold

Speicherplatz ist preiswert! Dies resultiert schnell in einem bekannten Problem – der Ansammlung einer Unmenge unsortierter, schwer zuordenbarer Fotos auf der Festplatte. Die gleiche Problematik zeigt sich, wenn diese Bilddaten nicht aus Urlaubserinnerungen bestehen, sondern Resultate aktueller zellbiologischer Experimente und Simulationen sind. Hierbei fallen immense Mengen von Aufnahmen an, welche katalogisiert und mit

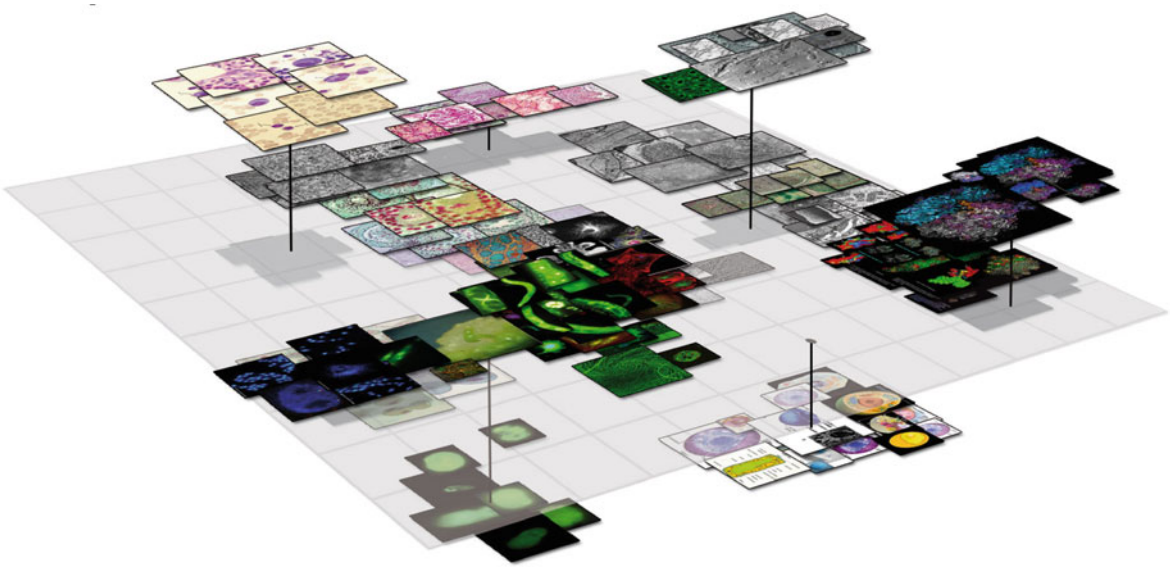

Abb. 7 Merkmalsbasierte Einbettung typischer Vertreter zellbiologischer Bilddaten in 3D. Strukturierung nach Bilddomänen wie Protein-Renderings, Fluoreszenzaufnahmen, grafischen Darstellungen und Elektronenmikroskopieaufnahmen

semantischen Informationen versehen werden müssen. Ebenso wie im privaten Bereich geschieht dies wegen des hohen zeitlichen und personellen Aufwands jedoch oft nicht. Weitere Probleme ergeben sich nach der Veröffentlichung von Resultaten wissenschaftlicher Untersuchungen in Form von Bildern. Unabhängig davon, ob diese im Internet oder in Fachzeitschriften herausgegeben wurden, ist die Zuordnung der Bilder zu ihrem Kontext in vielen Fällen nicht mehr automatisch nachvollziehbar. Wenn Bildunterschriften und andere assoziierte Texte nicht vorliegen, ist die Suche nach einem Bild bestimmten Inhaltes sehr schwer durchzuführen.

Inhaltsbasierte Bildsuche

Die Arbeit mit zellbiologischen Bilddaten ist ein Spezialfall des etablierten Forschungsbereichs *Content Based Image Retrieval* (CBIR). Dieser Bereich versucht über automatisch detektierbare Bildeigenschaften und bildbegleitende Texte inhaltliche Gemeinsamkeiten zwischen verschiedenen Bildern zu erkennen. Diese Erkennung ermöglicht anschließend das automatische Abrufen gewünschter Bildinhalte (Abb. 7). Besonders im medizinischen Kontext ist dies weit verbreitet und es existieren bereits viele Prototypen und Anwendungen [9]. Im Gegensatz dazu konnten sich im grundlagenorientierten, zellbiologischen Bereich bisher relativ wenige Systeme etablieren. Die Herausforderung besteht an dieser Stelle in der Entwicklung und Bereitstellung geeigneter Schnittstellen zu den Daten sowie von Unterstützungsmechanismen zur Vereinfachung der Suche. Diese Suche kann als Prozess mithilfe des *Mantras der Visuellen Analytik* [6, S. 16] beschrieben werden, wie im Folgenden erläutert wird:

> *„I. Analysiere zuerst [Datenanalyse] –*
> *II. Zeige das Wichtige –*
> *III. Vergrößere, filtere und analysiere weiter –*
> *IV. Details auf Nachfrage".*

Der Ansatz der Visuellen Analytik kann zum Beispiel für die semantische Annotation von Bilddatenbanken genutzt werden [22].

Merkmalsdefinition

Grundlegend für die Datenanalyse ist die Spezifizierung geeigneter, automatisch berechenbarer, visueller Bildmerkmale, welche sich in sogenannten Bilddeskriptoren zusammenfassen lassen. Einträge dieser Deskriptoren können kategorischer, diskreter oder textueller Art sein und überführen Bilder in einen abstrakten Merkmalsraum, in dem Ähnlichkeiten quantitativ erfasst werden können. Die Art der Bildinhalte bestimmt hierfür die verschiedenartigen Anforderungen. Allgemeine

Merkmale umfassen zum einen statistische Informationen über einzelne Pixel, etwa parametrisierte Farbverteilungen sowie strukturelle und räumliche Charakteristiken als Abhängigkeiten zwischen mehreren Pixeln, wie beispielsweise Frequenzdeskriptoren [14] oder Gradientenhistogramme. Zum anderen können die Merkmale Informationen über die räumliche Anordnung, Kombination und Beschreibung einzelner hervorstehender Bildpunkte in sogenannten *bags-of-features* [10] umfassen. Neben diesen *low level* (bildnahen) Eigenschaften liefern weitere Verarbeitungsschritte, wie Segmentierungen oder Objekterkennung, spezifischere Möglichkeiten zur Erweiterung des Bilddeskriptors. An diesem Punkt werden zunehmend Modellannahmen und Vorwissen zur Definition geeigneter Bildmaße und Distanzmetriken relevant. Eine Zusammenfassung häufig genutzter Bildmerkmale findet sich im *MPEG 7*-Standard [11]. Da die ursprüngliche Dimension der Bilddeskriptoren für gewöhnlich sehr groß ist, wird oft eine Dimensionsreduktionstechnik, wie die *Hauptachsentransformation*, auf den Merkmalsraum angewendet. Die Durchführung einer *multidimensionalen Skalierung* oder einer *Kern PCA* bietet äquivalente Möglichkeiten, falls ausschließlich paarweise Distanzen zwischen Bildern bekannt sind.

Welche Merkmale eignen sich nun besonders gut für biologische Bilder? Diese Frage ist bisher, ebenso wie im gesamten *CBIR* Bereich, nicht abschließend zu beantworten und hängt wesentlich von der vorliegenden Auswahl an Bilddaten ab. Falls der Deskriptor bspw. kreisförmige Strukturen (Zellen oder andere Organellen) unterschiedlicher Größe beschreiben soll, hat sich die Transformation des Bildes in den *Hough-Raum* als nützlich erwiesen.

Merkmalsanalyse

Nachdem eine problemspezifische Menge an Merkmalen definiert wurde, können diese für die weitere Suche verwendet werden. Abhängig davon, ob für einzelne Bilder Informationen über Zugehörigkeiten zu Kategorien vorhanden sind (Kategorien bzw. Klassen in diesem Kontext sind etwa spezielle Zellarten, Zellteile oder Aufnahmemodi), können die verwendeten Algorithmen in zwei Arten unterteilt werden. Wenn Bilder unterschiedlicher und bekannter Klassenzugehörigkeit sich im Merkmalsraum unterscheiden lassen, werden Klassifikationsalgorithmen (bspw. eine *Support Vector Machine*) eingesetzt und bezüglich Genauigkeit und Trefferquote analysiert. Diese Klassifikatoren können anschließend für die Einordnung von Bildern mit unbekannten Klassenzuordnungen verwendet werden. Abbildung 8 illustriert den Prozess der Merkmalsextraktion und Merkmalsanalyse sowie die Suche mithilfe von Bilddeskriptoren.

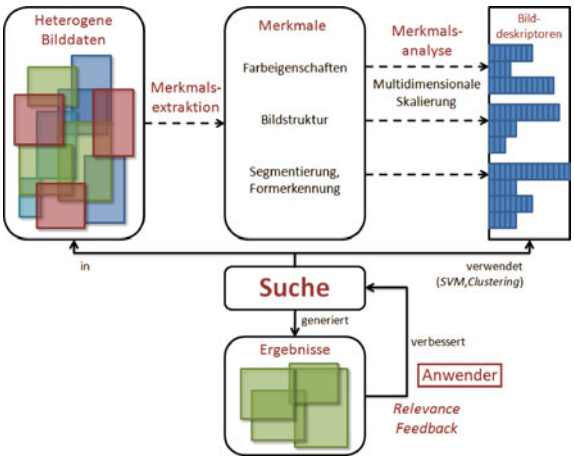

Abb. 8 Elemente eines Bildsuchsystems mit Anwenderfeedback (Relevance Feedback)

Die explorativen Analysestrategien des *Data Minings* finden Verwendung, falls keine Zugehörigkeiten bekannt sind. Dieses „Schürfen" nach Informationen zielt auf das Erkennen verborgener Muster in den Datenbeständen ab, die oft durch Häufungen im Merkmalsraum erzeugt werden. Algorithmen, die diese Häufungspunkte finden und beschreiben, sind bspw. *Mean Shift* oder *K-Means*.

Im Rahmen der *Visuellen Analytik* lassen sich beide Strategien (während der Schritte II und III des *Mantras der Visuellen Analytik*) um Nutzersteuerung und Navigationsmetaphern in der Visualisierungskomponente der Analyse erweitern. Bilder sind an sich schon wahrnehmbare Objekte. Die Visualisierung schafft dann einen Mehrwert, wenn dem Sucher über gleichzeitiges Zeigen einer Auswahl von Bildern Einblicke in Zusammenhänge und verborgene Muster des Datensatzes interaktiv ermöglicht werden. Hierfür wird dem Benutzer unter anderem das Werkzeug des *Relevance Feedbacks* [16] zur Verfügung gestellt, um die Suche im Merkmalsraum zu steuern, diesen dem Problem anzupassen und dem Anwender die für ihn relevanten Daten bereitzustellen. Die Relevanz der einzelnen Merkmale kann so automatisch an die Suchintention des Nutzers angepasst werden. Zusätzlich werden Distanzmaße abgeleitet, die Ähnlichkeiten zwischen

Bildern messbar machen. Abbildung 8 skizziert die wichtigsten Komponenten bei der nutzeradaptiven Suche mit *Relevance Feedback*. Die Präsentation vorhandener Metainformationen der Bilder beschließt Schritt IV des *Mantras*.

Bildarten

Es existieren vielfältige Aufnahmemethoden für zellbiologische Bilddaten. Durchlicht-, Fluoreszenz-, Rasterelektronen- und Transmissionselektronenmikroskopie sind einige ihrer Vertreter. Somit ist die Menge der betrachteten Bilddomänen strukturiert, aber in ihren einzelnen Ausprägungen sind die Bilder sehr verschieden. Abbildung 7 zeigt repräsentative Beispiele.

Auch die Herkunft der Bilddaten beeinflusst ihre Heterogenität. So findet man in einzelnen Forschungseinrichtungen sehr große Datenmengen mit speziellen Domäneninformationen (bspw. die verwendeten Aufnahmemethoden) und bekannten Zugehörigkeiten der Bilder zu inhaltlichen Kategorien. Hier bieten sich lokale Suchlösungen an, die sehr viele Zusatzinformationen des Datenmaterials verwenden können. Im allgemeineren Fall, einer heterogenen Bildherkunft (bspw. aus Publikationssammlungen), müssen diese Metainformationen zuerst extrahiert werden. Grundlegend ist hier das Ableiten von Bildkategorien. Hierbei ist domänenspezifisches Wissen gefragt, um sinnvolle, klar abgrenzbare, aber trotzdem allgemein gültige Objektklassen zu finden. Sehr hilfreich ist hierfür ein standardisiertes Vokabular. Ein Beispiel für ein derartiges Vokabular stellt die *Gene Ontology* [1] dar, die für diesen Zweck in einigen aktuellen Implementierungen eingesetzt wird.

In vielen Fällen liegen Bilder jedoch nicht mehr in ihrer Reinform vor. Falls sie etwa für eine Publikation in Übersichtsgrafiken integriert sind, müssen diese erst unter Zuhilfenahme von Schrifterkennungssystemen und weiteren Heuristiken in ihre Einzelteile zerlegt werden.

Anwendungsfälle

Wer wird eine inhaltsbasierte Bildersuche für Zellbilder verwenden? Und zu welchem Zweck? Die Antworten auf diese Fragen haben entscheidenden Einfluss darauf, welche Methoden der *Visuellen Analytik* für den Nutzer hilfreich sind. Jede Nutzergruppe hat eigene Anforderungen an und Verwendungsszenarien für eine Bildersuche. Ein Student möchte beispielsweise ein gelerntes Konzept vertiefen. Er ist somit nicht an der Auffindung eines einzelnen Bildes interessiert, sondern an einem explorativen Überblick über eine bestimmte Bildkategorie. Hier sind Visualisierungstechniken gefragt, die einen variablen Bereich des Merkmalsraumes abbilden können und gleichzeitig Ähnlichkeiten hervorheben. Das Auffinden eines ganz speziellen Bildes hingegen erfordert die Unterstützung der Navigation hin zu einem einzelnen festen Punkt in diesem Raum. Dem Nutzer muss ein Weg präsentiert werden, der möglichst kurz ist und intuitiv erscheint. Große visuelle Sprünge machen es dem Anwender schwer, Zusammenhänge zu erkennen und sind deshalb, wenn möglich, zu vermeiden. Morphing-Techniken, bekannt aus der Computeranimation, bieten eine Lösung, diese Sprünge zu vermeiden. Dabei werden merkmalsbasiert korrespondierende Bildteile identifiziert und während der Überblendung gezielt aufeinander abgebildet, um das Verständnis von Bildinhalten und Gemeinsamkeiten zu erhöhen.

Die Interaktionsdaten, welche während der Konfiguration der Suchmaschine durch Benutzerfeedback angefallen sind, können im Weiteren auch für die Trennung von untrainierten Bildern in Bildklassen verwendet werden.

20 Jahre intensive Forschung im Bereich der inhaltsbasierten Bildsuche haben eine immense Menge an Publikationen, Prototypen und kommerziellen Anwendungen produziert. Die in dieser Zeit gesammelten Erfahrungen werden von uns zur Realisierung eines 3D-Browsers verwendet, welcher eine kontinuierliche Navigation zwischen zellbiologischen Bilddatensätzen erlauben soll. Dadurch wird es möglich sein, durch die Markierung einzelner Zellstrukturen in einem Datensatz, Informationen zu korrespondierenden Teilen anderer Datensätze zu erhalten. Der einer abstrakten Zelle nachempfundene, modellierte 3D-Kontext ermöglicht dabei die intuitive Verknüpfung zellbiologischer Konzepte mit den dargestellten Bilddaten.

Literatur

1. Ashburner M, Ball CA, Blake JA et al. (2000) Gene ontology: tool for the unification of biology. Nat Genet 25(1):25–29
2. Backes C, Keller A, Küntzer J, Kneissl B, Comtesse N, Elnakady YA, Müller R, Meese E, Lenhof H-P (2007) GeneTrail – advanced gene set enrichment analysis. Nucleic Acids Res 35:WI86
3. Battke F, Symons S, Nieselt K (2010) Mayday – integrative analytics for expression data. BMC Bioinform 11:121

4. Dietzsch J, Heinrich J, Nieselt K, Bartz D (2009) SpRay: a visual analytics approach for gene expression data. IEEE Symposium on Visual Analytics Science and Technology (VAST), pp 179–186
5. Gehlenborg N, Dietzsch J, Nieselt K (2005) A framework for visualization of microarray data and integrated meta information. Inf Vis 4(3):164–175
6. Keim DA, Mansmann F, Schneidewind J, Ziegler H (2006) Challenges in visual data analysis. Tenth International Conference on Information Visualization, pp 9–16
7. Keller A, Backes C, Gerasch A, Kaufmann M, Kohlbacher O, Meese E, Lenhof H-P (2009) A novel algorithm for detecting differentially regulated paths based on gene set enrichment analysis. Bioinformatics 25:2787–2794
8. Küntzer J, Blum T, Gerasch A, Backes C, Hildebrandt A, Kaufmann M, Kohlbacher O, Lenhof HP (2006) BN++ – a biological information system. J Integr Bioinform 3
9. Lehmann TM, Gold MO, Thies C et al. (2004) Content-based image retrieval in medical applications. Methods Inf Medicine 43(4):354–361
10. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comp Vis 60(2):91–110
11. Manjunath BS, Salembrer P, Sikora T (2002) Introduction to MPEG-7: multimedia content description interface. Wiley
12. Michal G (1993) Biochemical Pathways (Poster). Boehringer, Mannheim
13. Nieselt K, Battke F, Herbig A et al. (2010) The dynamic architecture of the metabolic switch in *Streptomyces* coelicolor. BMC Genomics 11:10
14. Oliva A, Torralba A (2006) Building the gist of a scene: the role of global image features in recognition. Prog Brain Res 155:23–36
15. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita P, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ (2010) The UCSC genome browser database: update 2010. Nucleic Acids Res 38:D613-9
16. Rui Y, Huang TS, Ortega M, Mehrotra S (1998) Relevance feedback: a power tool for interactive content-based image retrieval. IEEE Transactions on circuits and systems for video technology 8(5):644–655
17. Santamaría R, Therón R, Quintales L (2008) A visual analytics approach for understanding biclustering results from microarray data. BMC Bioinform 9:247–247
18. Schatz MC, Phillippy AM, Sheiderman B, Salzberg BL (2007) Hawkeye, an interactive visual analytics tool for genome assembly. Genome Biol 8:R34
19. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504
20. Shaw CD, Dasch GA, Eremeeva ME (2007) IMAS: the interactive multigenomic analysis system. IEEE Symposium on Visual Analytics Science and Technology, pp 59–66
21. Suderman M, Hallett M (2007) Tools for visually exploring biological networks. Bioinformatics 23:2651–2659
22. Yang J, Fan J, Hubball D, Gao Y, Luo H, Ribarsky W, Ward M (2007) Semantic image browser: bridging information visualization with automated intelligent image analysis. IEEE Symposium on Visual Analytics Science and Technology (VAST), pp 59–66