Parahistogramme innerhalb eines dreidimensionalen Interaktionsraumes

Albert Pritzkau, Dirk Bartz

Universität Leipzig, ICCAS/VCM
albert.pritzkau@medizin.uni-leipzig.de

Kurzfassung. Im Kontext der Genexpressionsanalyse haben sich parallele Koordinaten (PK) als ein hilfreiches Werkzeug erwiesen. Ein hochdimensionaler Datensatz kann auf diese Weise in einem zweidimensionalen Datenraum dargestellt werden. Dabei wird die Anzahl der darstellbaren Dimensionen lediglich durch den horizontal verfügbaren Platz begrenzt und ermöglicht ein relativ leichtes Erkennen inhärenter Zusammenhänge innerhalb eines Kontextes. Neben den offensichtlichen Vorteilen sind jedoch auch die damit verbundenen Beschränkungen zu berücksichtigen. Durch Überlagerung einzelner Linien kann diese Darstellung relativ schnell überzeichnet werden und verliert in diesen Bereichen an Informationsgehalt. Darüber hinaus stellt sich die Detektion von Korrelationen über mehrere Achsen als eine relativ schwierige Aufgabe dar. Um die Frequenz von Kantenzügen entlang der Koordinatenachsen zu erhalten, werden zur graphischen Darstellung Histogramme herangezogen, welche jeweils an den Koordinatenachsen ansetzen. Ohne den Kontext der PK-Darstellung zu verlassen, werden quantitative Aussagen über die Verdeckung in einem bestimmten Werteintervall ermöglicht.

1 Einleitung

Mit Einführung verschiedener Genom-und Proteom-Technologien wird die gleichzeitige Messung von vielen Hunderten oder Tausenden von biologischen Messgrößen ermöglicht. Diese ergeben sich aus unterschiedlichen medizinischen und biologischen Problemstellungen. Die Etablierung geeigneter Algorithmen zur Analyse dieser meist hochdimensionalen Daten stellt ein entscheidendes Kriterium dar. Die Herausforderung der Genexpressionsanalyse besteht beispielsweise darin, verlässliche Aussagen über die Aktivität von Genen unter unterschiedlichen Umständen zu treffen. Microarrays oder DNA-Chips dienen der Bestimmung relativer Änderungen der Genexpression. Das Ergebnis solcher Experimente setzt sich jeweils aus einer Liste der ermittelten Gensequenzen mit den dazugehörigen Expressionswerten dar. Mehrere Experimente ergänzen sich zu einer Datenmatrix dessen Spalten jeweils einem Experiment zuzuordnen sind. Zur Gewinnung aussagekräftiger statistischer und biologischer Daten greift man zunehmend auch auf grafische sowie andere Methoden zur Visualisierung solcher komplexer Datensätze zurück [1]. Bei der visuellen Analyse müssen dabei die hochdimensionale Daten auf niedrigdimensionale Sichten projiziert werden.

Bezüglich der visuellen Datenexploration von hochdimensionalen Daten hat sich die Darstellung der parallelen Koordinaten [2] (PK) als ein sehr hilfreiches Werkzeug erwiesen. Diese Form der Visualisierung veranschaulicht einen k-dimensionalen Datensatz anhand von zwei Darstellungsdimension. Sie besteht aus k parallelen und typischerweise äquidistanten Achsen. Diese repräsentieren jeweils den Wertebereich einer Dimension. In dem oben beschriebenen Ausgangsdatensatz ist eine Dimension jeweils einem Experiment zuzuordnen. Der dazugehörigen Wertebereich umfasst die minimalen und maximalen Expressionswerte. Generell werden benachbarte Achsen gleich orientiert. Über eine Nulllinie können die einzelnen Achsen zusätzlich miteinander verbunden werden. Zur Repräsentation werden die Datenwerte jeweils einer Zeile auf den entsprechenden Achsen abgetragen und ergeben einen Kantenzug. Das Ergebnis kann als Grundlage anschließender Analysen genutzt werden.

2 Material und Methoden

Die Darstellung der PK eignet sich ausgezeichnet zur Identifikation von Korrelationen zwischen Attributen benachbarter Achsen. Darüber hinaus sind auch Werteverteilungen oder Ausreißer gut erkennbar. Jedoch stellt sich in Bereichen hoher Linienkonzentrationen, die sich vor allem durch ähnliche oder mehrfach auftretende Attributwerte ergeben, die Interpretation der Darstellung als relativ schwierig dar. Dieses Problem wurde bereits durch eine Reihe unterschiedlicher Ansätze adressiert. Zur Darstellung einer Häufigkeitsverteilung entlang des Wertebereichs einer Koordinatenachse wurden beispielsweise Histogramme zur Integration in die PK-Darstellung vorgeschlagen [3]. Diese Kombination folgt generell dem weithin bekannten Konzept der Linked-Views, und findet auch in aktuelleren Beiträgen wie [4, 5] besondere Beachtung.

In der vorliegenden Arbeit wird die Einbettung der traditionell zweidimensionalen PK in einen dreidimensionalen Interaktionsraum vorgeschlagen. Es existieren bereits eine Reihe unterschiedlicher Ansätze, die hinzu gewonnene Dimension vorteilhaft zu nutzen. Beispielsweise erweitern [6, 7] die generelle Struktur der PK, um die Darstellung auf diese Weise zu entzerren. Im Gegensatz dazu wird in unserem Fall die traditionelle PK-Darstellung innerhalb des gegebenen Interaktionsraumes koplanar zur xy-Ebene angeordnet. Die räumliche Dimension wird dazu genutzt, eine Histogrammdarstellung zu integrieren. Sie ermöglicht anhand der angereicherten Häufigkeitsverteilung der Kantenzüge entlang der Koordinatenachsen eine differenzierte Interpretation kritischer Bereiche. Wie in Abb. 1 dargestellt, verlaufen die Histogrammsäulen, welche an den Koordinatenachsen ansetzen, orthogonal zur PK-Darstellung in z-Richtung des Interaktionsraumes. Die Höhe einer Säule repräsentiert dabei jeweils Frequenz von Kantenzügen in dem zugehörigen Intervall. Ist nun der Interaktionsraum entlang der Blickrichtung des Beobachters ausgerichtet, werden die Histogrammsäulen zunächst auf die jeweiligen Koordinatenachsen projiziert und sind für den Benutzer nicht sichtbar. Durch einfache Navigationsinteraktionen kann jedoch der Interaktionsraum in eine gewünschte Position rotiert werden und eröffnet damit den Blick

auf die Histogrammdarstellung ohne den visuellen Kontakt zu den PK zu unterbrechen.

3 Ergebnisse

Die Einbettung der PK-Darstellung in einen drei dimensionalen Interaktionsraum ermöglicht die Integration von weiteren Datenattributen. Durch Rotationen des Interaktionsraumes wird die Sichtbarkeit von diesen Zusatzinformationen - hier die Histogrammdarstellung - beeinflusst. Die Ausrichtung der Darstellung in räumliche Dimensionen ermöglicht eine visuelle Gewichtung einzelner Bestandteile des Interaktionsraumes, die je nach Bedarf von der vollständigen Sichtbarkeit bis hin zur Verdeckung reicht.

Der visuelle Vergleich einzelner Histogramme kann zudem als Ansatzpunkt weiterer Analysen dienen. So lassen sich beispielsweise Korrelationen zwischen Achsen mit ähnlicher Häufigkeitsverteilung vermuten. Zur differenzierten Untersuchung können zusätzlich die einzelnen Histogrammbalken markiert werden. Alle Linienzüge, die durch das entsprechende Intervall verlaufen, werden farblich hervorgehoben (Abb. 2). Anhand dieser Markierung ist es möglich Korrelationen über mehrere Dimensionen hinweg zu erörtern. Zur automatischen Selektion der einzelner Histogrammbalken wird eine Ebene eingesetzt werden, die parallel zur Ebene der parallelen Koordinaten angeordnet ist. Über einen Schieberegler kann die Höhe der Selektionsebene gesteuert, welche eine bestimmte Auftrittshäufigkeit repräsentiert. Alle Histogrammsäulen, die diese Ebene schneiden und damit den gegebenen Häufigkeitswert überschreiten, werden markiert. Diese Selektion berücksichtigt generell die Histogramme der gesamten PK-Darstellung. Jedoch kann die Auswahl auch auf eine Dimension beschränkt werden.

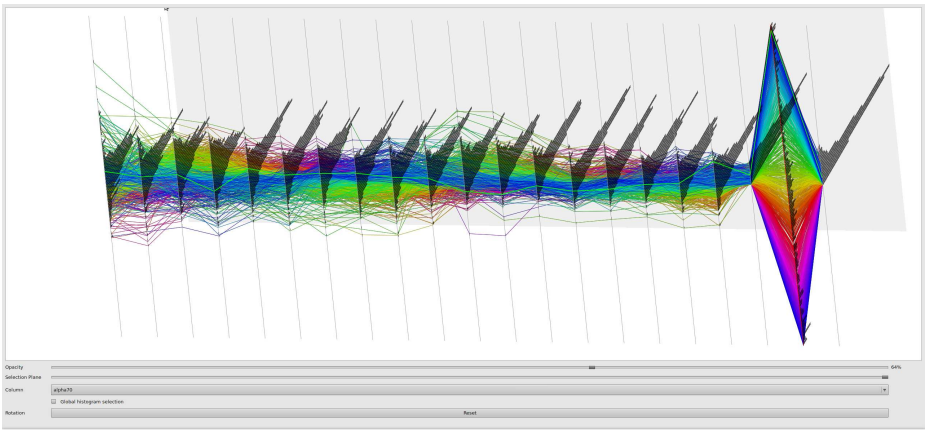

Abb. 1. Angereicherte Darstellung der Zeitreihe eines Zellzyklus-Experiment an Bäckerhefe (Saccharomyces cerevisiae [8, 1]) erweitert durch Histogramme.

4 Diskussion

Im Gegensatz zu den bekannten Verfahren, wird die Visualisierung der PK grundsätzlich nicht durch die Darstellung zusätzlicher Attribute überlagert. Jedoch ist der Benutzer in der Lage durch entsprechende Interaktionen den Fokus individuell so zu manipulieren, dass die erforderlichen Informationen für sichtbar werden. Anstelle der Histogrammdarstellung sind auch weitere Darstellungsformen denkbar, die die räumliche Dimension ausnutzen. Beispielsweise verfolgen Parallel Sets [9] einen ähnlichen Ansatz zur Analyse. Auch hier wird der Wertebereich einer Achse in eine Folge von Intervallen unterteilt und die Häufigkeit der betroffenen Linienzüge aufsummiert. Im Gegensatz zu Histogrammen rückt hier jedoch die Verteilung der Linienzüge zwischen den benachbarten Achsen in den Fokus der Betrachtung. Diese werden durch Bänder repräsentiert, die jeweils zwei Intervalle benachbarter Koordinatenachsen miteinander verbindet. In Zukunft soll die Integration dieser Methode in den vorhandenen Interaktionsraum evaluiert werden, um so vielleicht noch besser Korrelationen über mehrere Achsen hinweg verfolgen zu können.

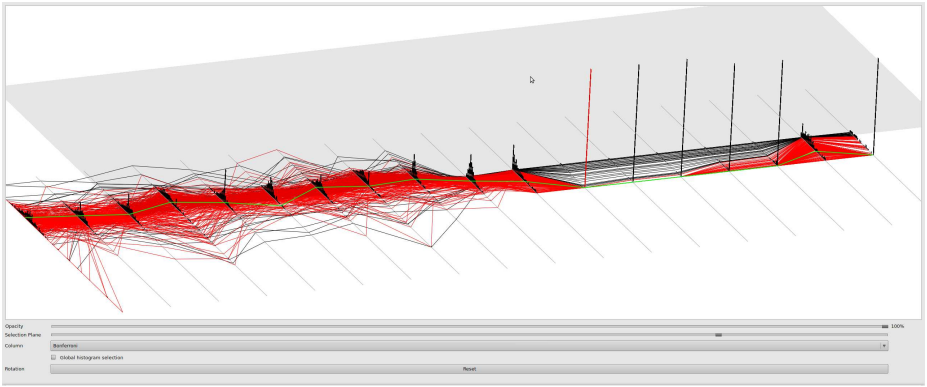

Abb. 2. Angereicherte Darstellung einer Expressionsstudie zur Auswirkung von sportlicher Anstrengung auf das Immunsystem [10, 1] erweitert durch eine Selektionsebene. Histogrammsäulen einer Koordinatenschase, welche die Selektionsebene schneiden, werden selektiert. Die Selektionsebene ist hier halbtransparent dargestellt.

Literaturverzeichnis

1. Dietzsch J, Heinrich J, Nieselt K, et al. SpRay: A visual analytics approach for gene expression data. In: Proc IEEE Symp Vis Anal Sci Technol; 2009.
2. Inselberg A, Dimsdale B. Parallel coordinates: a tool for visualizing multidimensional geometry. In: Proc IEEE Vis; 1990. p. 361–78.
3. Ong HL, Lee HY. WINVIZ: a visual data analysis tool. Comput Graph. 1996;20(1):83–4.

4. Hauser H, Ledermann F, Doleisch H. Angular brushing of extended parallel coordinates. In: Proc IEEE Symp Inform Vis; 2002. p. 127–30.
5. McDonnell KT, Mueller K. Illustrative parallel coordinates. Comput Graph Forum. 2008;27(3):1031–8.
6. Johansson J, Cooper M, Jern M. 3-Dimensional display for clustered multi-relational parallel coordinates. In: Proc IEEE Int Conf Inform Vis; 2005. p. 188–93.
7. Fanea E, Carpendale MST, Isenberg T. An interactive 3D integration of parallel coordinates and star glyphs. In: Proc IEEE Symp Inform Vis; 2005. p. 149–56.
8. Spellman PT, Sherlock G, Zhang MQ, et al. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell. 1998;9(12):3273–97.
9. Kosara R, Bendix F, Hauser H. Parallel sets: Interactive exploration and visual analysis of categorical data. IEEE Trans Vis Comput Graph. 2006;12(4):558–68.
10. Zieker D, Fehrenbach E, Dietzsch J, et al. cDNA microarray analysis reveals novel candidate genes expressed in human peripheral blood following exhaustive exercise. Physiol Genomics. 2005;23(3):287–94.