Automatische Auswertung von Mikroarraybildern

Mathias Katzer, Franz Kummert und Gerhard Sagerer

Technische Fakultät,AG Angewandte Informatik
Graduiertenkolleg Bioinformatik
Universität Bielefeld, 33501 Bielefeld
Email: {mkatzer|franz|sagerer}@techfak.uni-bielefeld.de

Zusammenfassung. Wir beschreiben ein Verfahren, das die automatische Segmentierung von Mikroarraybildern in unkalibrierten Umgebungen ermöglicht. Insbesondere behandeln wir die als Adressierung oder in der englischsprachigen Literatur als „Gridding" bezeichnete Segmentierung der Messpunktgitter. Wir verwenden klassische Regionensegmentierung, Häufigkeitsverteilungen der Abstände zum nächsten Nachbarn, Achsenprojektionen und optimierte Segmentierung durch dynamische Programmierung. Das Verfahren ist für die automatisierte Segmentierung von Bildserien aus umfangreicheren Experimenten besonders geeignet.

1 Einleitung

Unsere Arbeit befasst sich mit der automatischen Auswertung von Mikroarraybildern, die in differentiellen Genexpressionsexperimenten gewonnen werden.

Genexpressionsanalyse durch Mikroarray-Hybridisierung ist eine Methode, die in der molekulargenetischen Grundlagenforschung für Experimente zur Genregulation entwickelt worden ist. Inzwischen gibt es auch viele Beispiele von Anwendungen in der Medizin, etwa das Cancer-Chip Projekt am DKFZ oder die Arbeit von Alizadeh et al. [1]. Neben Anwendungen in der Grundlagenforschung könnte die Mikroarraytechnologie in Zukunft für die Charakterisierung von Geweben zur Diagnose und allgemein zu Zwecken der Genotypisierung eingesetzt werden.

Mikroarrays sind Hybridisierungssubstrate (Glas) auf denen mit Hilfe eines Roboters DNA-Bibliotheken punktweise aufgedruckt werden. Gegen diese Bibliothek werden dann im eigentlichen Experiment mit zwei verschiedenen Fluoreszenzfarbstoffen markierte DNA-Proben konkurrierend hybridisiert. Eine der beiden Proben dient bei der Auswertung als Referenz, so dass die Verhältnisse der Fluoreszenzintensitäten der Farbstoffe als Veränderung einer DNA-Konzentration gegenüber der Referenzprobe interpretiert werden können (siehe [4]).

Zur Auswertung werden Fluoreszenzbilder aufgenommen, in denen zunächst die Messpunkte segmentiert werden müssen. Die Messpunkte werden per Roboter gedruckt und sind in der Regel in mehreren einfachen, wenn auch nicht

unbedingt untereinander ausgerichteten Rechteckgittern angeordnet. Die Segmentierung ist wegen häufig auftretender Verunreinigungen, unspezifischer Hybridisierung (Hintergrund) und großer Signaldynamik eine sehr anspruchsvolle Aufgabe , wenn auf ein kalibriertes System verzichtet werden soll.

Deshalb werden heute in aller Regel halbautomatische, interaktive Programme zur Auswertung verwendet. Die weitergehende Automatisierung scheint jedoch erforderlich, weil die Methode häufig als Hochdurchsatzverfahren angewandt wird, weshalb durch automatische Bildauswertung viel monotone und fehlerträchtige Arbeit gespart werden kann. Auch sollte optimalerweise jeder einzelne Messpunkt auf Verunreinigungen kontrolliert werden, was bei Hunderttausenden von Punkten im Gesamtexperiment ohne Automatisierung kaum zu leisten ist. Die einzelnen Messpunkte zeigen sehr variable Formen ('Möndchen' und 'Krater' sind möglich), die durch die Eigenschaften der gedruckten DNA-Lösungen und mechanische Unzulänglichkeiten der auf Miniaturisierung optimierten Geräte verursacht sind. Es sind deshalb automatische Verfahren erforderlich, die gegen variable Bedingungen bei der Arrayherstellung und Bildaufnahme (farbstoffabhängige Scannerempfindlichkeit u.ä.), Verschmutzungen und die Deformation der Messpunkte robust sind.

2 Methoden

Zur Segmentierung der Gitter von Messpunkten haben wir ein mehrstufiges Verfahren entwickelt, das sich wegen der gegebenen Einschränkungen vorwiegend auf die lokal periodische Anordnung der Punkte stützt. Zunächst werden mit einem lokalen Schwellwertverfahren Regionen segmentiert, die vermuteten Messpunkten entsprechen. Die Schwellwertbestimmung kann mit gewichteten Histogrammen erfolgen, da die meisten Verunreinigungen nur einen oder wenige Pixel groß sind. Aus Häufigkeitsverteilungen der Abstände zur nächsten Nachbarregion können Spalten- und Zeilenabstände der Messpunktgitter geschätzt werden. Damit kann die periodische Gitteranordnung der Meßpunkte stückweise rekonstruiert werden, wodurch sich die oft gegen die Bildkoordinaten gedrehten Gitterachsen bestimmen lassen. Mit Hilfe von Achsenprojektionen und den vorgegebenen Spalten- und Zeilenanzahlen der Gitter werden Segmentierungshypothesen erzeugt, aus denen mit dynamischer Optimierung eine optimale Gesamtsegmentierung, d.h eine überlappungsfreie, möglichst gut die Regionen überdeckende Anordnung von Gittern im Bild, ausgewählt wird.

Insbesondere wenn Randbereiche von Gittern von Verunreinigungen überdeckt sind oder kein Hybridisierungssignal tragen, kann die Segmentierung kein eindeutiges Ergebnis liefern. Abhilfe kann hier die Benutzung von Segmentierungsergebnissen anderer Arraybilder aus der gleichen Herstellungsserie bringen, da die relativen Abstände der Messpunktgitter innerhalb einer Serie näherungsweise konstant sind.

Zur Berechnung der Intensitätsverhältnisse je Messpunkt benutzen wir das allgemein angenommene Modell der konkurrierenden Hybridisierung, nach dem die Farbstofffluoreszenzintensitäten der Bildpunkte in den beiden Kanälen ab-

züglich eines konstanten Hintergrundanteils, den es zu schätzen gilt, linear abhängig sind. Pixelweise Intensitätsverhältnisse lassen sich hier jedoch nicht immer zuverlässig berechnen, weil die direkte Korrespondenz der Kanäle durch lokal variable, von der Bildaufnahme herrührende Verschiebungen gestört ist. Aus dem Hybridisierungsmodell folgt, dass es ausreicht, nur die helleren Bereiche der Messpunkte mit gutem Signal-Rauschverhältnis für die Berechnung zu benutzen [5]. Deshalb wenden wir für die Bestimmung signaltragender Regionen die von Y. Chen beschriebene Mann-Whitney-Segmentierung an [3]. Mit den Ergebnissen der Gittersegmentierung läßt sich die Mann-Whitney-Methode automatisch initialisieren. Zusätzlich korrigieren wir die lokale Verschiebung entsprechend der Schwerpunktsverschiebungen zusammengehörender Regionen in beiden Kanälen, um die vom Modell geforderte Korrespondenz möglichst gut zu erhalten.

3 Ergebnisse

Wir haben die Gittersegmentierung mit einer Sammlung von Bildern aus 100 Mikroarrayexperimenten aus vier verschiedenen gemeinsam gedruckten Serien von Mikroarrays getestet (4 bis 32 Gitter pro Bild). Die Segmentierung der einzelnen Gitter verlief auf den Einzelbildern in 60 Prozent der Fälle erfolgreich. Bei 22 Bildern wurden alle Gitter korrekt segmentiert. Die häufigsten Fehlerursachen waren dunkle Randzeilen oder großflächige, Messpunktgitter überdeckende Kontaminationen. Bei Arrays aus einer Serie mit nur 24 Punkten je Gitter versagte mehrmals die Schätzung der Spaltenabstände wegen der zu kleinen Stichproben von Nächste-Nachbar-Abständen. Benutzt man die Zusammengehörigkeit der Serien, werden die Ergebnisse drastisch besser: Drei der vier Serien werden fehlerfrei verarbeitet, und 55 (statt vorher 2) der 62 Bilder mit 24 Punkten je Gitter werden richtig segmentiert, insgesamt also über 90 Prozent der Bilder.

Außerdem haben wir die Qualität der berechneten Einzelspotintensitätsverhältnisse durch Vergleich mit Ergebnissen aus manuellen Auswertungen untersucht, wobei wir die von Buhler et al. vorgeschlagene Methode verwendet haben [2]. Danach liefert unser automatisches Verfahren ähnlich gute bis geringfügig bessere Werte [5] als die manuelle Auswertung mit dem weitverbreiteten Programm Scanalyze (M. Eisen, http://www.microarrays.org/software.html), dem ersten und längere Zeit einzigen Programm seiner Art.

Die Verarbeitung dauert auf einer Compaq Alpha XP1000 Workstation (500 MHz, SPECint95 26.5, SPECfp95 52.2) etwa 15-25 Sekunden bei kleinen Bildern (288 Messpunkte) und 220-260 Sekunden bei den größten Bildern (24192 Punkte). Der Speicherbedarf liegt etwa bei der sechsfachen Größe der Fluoreszenzbilder, und damit bei unseren Testbeispielen zwischen 60 und 230 MByte.

4 Schlussfolgerungen und Ausblick

Die Ergebnisse zeigen, dass in dem praktisch relevanten Szenario der Auswertung von Arrayserien die Gittersegmentierung sehr gut automatisierbar ist. Insbeson-

dere bei größeren Experimenten mit vielen Messpunkten lassen sich hervorragende Ergebnisse erzielen. Die beschriebenen Probleme mit zu kleinen Gittern lassen sich in der Experimentvorbereitung vermeiden, indem Messpunkte ausreichend oft mehrfach gedruckt werden, wovon auch die weitere Datenauswertung profitiert.

Durch unser Verfahren können mit immens verringertem Zeit- und Arbeitsaufwand Ergebnisse erreicht werden, deren Genauigkeit den interaktiv hergestellten Auswertungen mindestens äquivalent ist.

Mit der Rechenleistung von Standardhardware ist die Bildauswertung schneller als die Bildaufnahme, die etwa 5-10 Minuten dauert. Damit können große Mengen von Bildern ohne Zeitverzögerung verarbeitet werden.Weitere Information und Beispiele sind im World Wide Web unter der Adresse

http://www.techfak.uni-bielefeld.de/ags/ai/projects/microarray

zu finden. Unsere Implementation wird Interessierten auf Anfrage für wissenschaftliche Zwecke zur Verfügung gestellt. Einzelne Bilder können auf dem Bielefelder Bioinformatikserver ausgewertet werden (Zugang z.B. über o.g. Adresse).

Danksagung: Diese Arbeit ist in Teilen durch das Graduiertenkolleg Bioinformatik (GK 635) der Deutschen Forschungsgemeinschaft gefördert. Wir danken unseren Kooperationspartnern für die freundliche Bereitstellung der Bilddaten: Terry Gaasterland (The Rockefeller University, New York), Alfred Sporman und Mike Cherry (Stanford University), Shixia Huang und Agnes Viale (Memorial Sloan Kettering Cancer Center, New York), Anke Becker (Zentrum für Genomforschung, Universität Bielefeld)

References

1. A. Alizadeh et al.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 403:503–511, 2000.
2. Buhler J, Ideker T, Haynor D: Dapple: Improved Techiques for Finding Spots on DNA Microarrays. Technischer Bericht der University of Washington , UWTR 2000-08-05, 2000.
3. Chen Y, Dougherty E, Bittner M: Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images. J Biomed Opt, 2:364–374, 1997.
4. Eisen M, Brown P: DNA arrays for analysis of gene expression. Methods Enzymol, 303:179–205, 1999.
5. Katzer M, Kummert F, Sagerer G: Robust microarray image analysis. Proceedings of the International Conference on Bioinformatics, Bangkok, 2002 (accepted).