Bedeutung der Bioinformatik für die Auswertung von Microarray-Experimenten am Beispiel der Molekularen Onkologie

Significance of Bioinformatics for the Evaluation of Microarray Experiments taking Molecular Oncology as an Example

H. A. G. Müller[1], G. E. Hoffmann[2]

Zusammenfassung: Die Microarray-Technik erlaubt eine umfassende Analyse der Genexpression auf mRNA- und Proteinebene und ermöglicht so insbesondere in der molekularen Onkologie Einblicke in das Krankheitsgeschehen, die bis vor wenigen Jahren noch unvorstellbar gewesen wären: Da mit einer einzigen Untersuchung mehrere 1 000 Gene gleichzeitig analysiert werden können, erhält man Momentaufnahmen komplexer Expressionsmuster, die – sofern sie reproduzierbar sind – für die Klassifikation von Tumorentitäten sowie zur Früherkennung, Prognostik und Wahl einer individualisierten Therapie geeignet sein können.
Ehe man allerdings zu relevanten Aussagen gelangt, ist eine intensive Vorverarbeitung der Rohdaten und statistische Analyse notwendig. Die drei Hauptziele dieser zur Bioinformatik zählenden Verfahren sind
- kalibrierte und auf einen Referenzwert normierte Ausgangswerte zu erzeugen;
- Gene mit ähnlicher Expression in funktionelle Gruppen einzuteilen;
- Zellen mit ähnlichen Expressionsmustern klinischen Phänotypen zuzuordnen.

Die aktuelle Herausforderung an die Labordiagnostik sehen wir darin, die derzeit nur für die klinische Forschung zugelassenen Microarrays in die Diagnostik einzuführen.
Wir haben an Hand von simulierten und realen Datensätzen die in der Literatur beschriebenen mathematischen Verfahren geprüft und eine Auswahl mit Blick auf ihre Nützlichkeit für die Labordiagnostik getroffen. Ein Überblick über aktuelle onkologische Studien beweist den klinischen Wert derartiger Untersuchungen und macht gleichzeitig die Notwendigkeit einer intensiven Auseinandersetzung mit der Bioinformatik für diagnostische Belange deutlich.

Schlüsselwörter: Bioinformatik; Microarrays; Biochips; Onkologie.

Summary: The microarray technique enables a comprehensive gene expression analysis to be performed both on the mRNA and protein level. It has created more insight into pathological processes than expected just few years ago: Since a single test analyses several 1000 genes in parallel, snapshots of complex expression patterns are generated, which – if they are reproducible – may be suitable for classification of tumor entities as well as for early detection, prognosis, and the choice of individual therapies.
However, before relevant statements can be made, data need to be intensively pre-processed and evaluated with statistical methods. The three major goals of bioinformatic procedures are
- calibration and normalization of raw signals with respect to reference values,
- classification of genes with similar expression into functional groups,
- assignment of clinical phenotypes to cells with similar expression patterns.

Since microarrays have been approved only for scientific purposes so far, the major challenge for clinical pathology is in our view their introduction into diagnostics.
We have used simulated and real data sets to test published mathematical procedures and to select some of them with respect to their diagnostic utility. An overview of current studies in oncology proves the clinical value of such examinations and at the same time illustrates the necessity to deal intensively with bioinformatics for diagnostic purposes.

Keywords: bioinformatics; microarrays; biochips; oncology.

Das Hauptmerkmal maligner Erkrankungen ist eine Anhäufung von zumeist erworbenen Mutationen, die in ihrer Summe zu einem Ungleichgewicht zwischen Zellvermehrung und Zelltod und damit zu invasivem und metastasierendem Wachstum führen [1, 2]. Die Microarray-Technik erlaubt eine umfassende Analyse

[1]Institut für Laboratoriumsmedizin, Klinik am Eichert, Göppingen, Deutschland
[2]Trillium GmbH, Grafrath bei München, Deutschland
Korrespondenz: Prof. Dr. med. Georg Hoffmann, Trillium GmbH, Hauptstraße 12b, 82284 Grafrath, Deutschland
Fax: +49-81 44-9 81 69
E-mail: ghoffmann@trillium.de
Web: www.trillium.de

der Genexpression vor allem auf der Ebene der Transkription (m-RNA-Quantifizierung mit DNA-Arrays) [3], zunehmend auch der Translation (Protein-Arrays) [4]. Gemeinsam mit anderen Technologien der massiv parallelen Analyse von genomischer Information haben gerade diese als „Biochips" bezeichneten Reagenzträger zur Etablierung allgemein akzeptierter „Stoffwechselwege des Krebses" (Cancer Pathways) [5] geführt. Es würde den Rahmen dieser auf die Bioinformatik ausgerichteten Arbeit sprengen, zu den verschiedenen nasschemischen Techniken detailliert Stellung zu nehmen. Der interessierte Leser sei vor allem auf die Serie „Methods in Molecular Medicine" des Humana Press Verlags, New Jersey hingewiesen [1–4], die derzeit wohl umfangreichste Sammlung methodisch orientierter Monografien. Zwei kurze und verständliche Übersichten „Biochips – Hoffnungsschimmer im Kampf gegen den Krebs" und „Bändigung einer ständig steigenden Datenflut" haben die Autoren kürzlich in *LaborManagement Aktuell* 11/2002 publiziert.

Aus labordiagnostischer Sicht besteht das Hauptproblem in der Fülle von Messwerten bei meist limitierter Anzahl von Experimenten. Typischerweise erhält man bei einer einzigen Messung 10^2 bis 10^4 Signale. Führt man solche Messungen mehrfach, z. B. an verschiedenen Zelltypen oder zu verschiedenen Zeitpunkten, durch, so erhält man bei nur 25 Experimenten mit 4 000 Genen bereits 100 000 Messpunkte. Gerade in der molekularen Onkologie kommt es ferner darauf an, die räumlichen Verhältnisse innerhalb einer Gewebeprobe sowie den zeitlichen Verlauf über die verschiedenen Tumorstadien hinweg korrekt abzubilden. Dadurch erweitert sich die im Beispiel genannte zweidimensionale Matrix von 25 mal 4 000 Punkten für ein Einzelexperiment rasch in eine dritte und vierte Dimension, was für die so genannte „New Biology" eine enorme datentechnische Herausforderung darstellt [6].

Da nun aber der Preis je Biochip mehrere Hundert oder Tausend Euro betragen kann, sind allein aus Kostengründen Experimente mit einer statistisch ausreichenden Fallzahl eher die Ausnahme. Es gilt deshalb vor allem, voreilige Schlüsse bei mangelhafter Signifikanz, falsch positive Resultate durch ungezielte Profile und ähnlich grundlegende Fehler zu vermeiden. Das Problem liegt also weniger in der Herstellung oder Anwendung von Microarrays als vielmehr in der richtigen Planung und Auswertung der Experimente: Wie eliminiert man das Rauschen unter Tausenden von Messwerten, wann liegt eine signifikante Differenz bei der Expression eines Gens vor, wie erkennt man Zusammenhänge zwischen ähnlich regulierten Genen? Die Autoren haben sich die Mühe gemacht, Literatur und Internet nach geeigneten Verfahren zu durchsuchen, diese an simulierten und realen Datensätzen praktisch zu testen und in verständlicher Form darzustellen. Die Tabellen 1 und 2 listen Werkzeuge und Quellen auf, die

Tabelle 1 Ausgewählte Programme und Anleitungen im Internet

Internet-Adresse	Beschreibung/Produktname	Bemerkungen
ep.ebi.ac.uk/EP/	Expression Profiler	Diverse Programme zur online-Analyse von Daten
www.genome.wi.mit.edu/cancer/software/	GeneCluster I, GeneCluster II	Programme kostenlos; Registrierung erforderlich
rana.lbl.gov/	ScanAlyze® Cluster® TreeView® von M. Eisen, Stanford University	Programme kostenlos; Registrierung erforderlich
www.trillium.de	Datenfilter auf Basis neuronaler Netze	Excelprogramm, kostenlos, keine Registrierung
www.xlstat.com/	Allgemeines Statistikprogramm auf Excel-Basis mit Clusteralgorithmen	Kommerzielles Statistikprogramm

Tabelle 2 Allgemeine Informationen im Internet

Internet-Adresse	Beschreibung	Bemerkungen
www.mged.org/index.html	Microarray Gene Expression Data Society – MGED Society	Informationen für Autoren zum Design von Microarray-Experimenten
research.nhgri.nih.gov/microarray/analysis.html	National Human Genome Research Institut (USA)	Übersicht über Genanalyse und Datenaufbereitung
www.nature.com/nrc/journal/v2/n5/weinberg_poster	Internetseite zu Literaturzitat Weinberg *et al.* („Metro-Plan")	Erläuterung onkologisch relevanter Gene
genome-www5.stanford.edu	Standard-Datensätze der University of Stanford	Für eigene Experimente geeignet, gute Erläuterungen

im Internet zugänglich sind. Im Literaturverzeichnis finden sich vor allem Lehrbuch- und Review-Zitate sowie einige ausgewählte Grundlagenarbeiten. Die meisten Verfahren der Bioinformatik basieren ohnehin auf Rechenvorschriften (Algorithmen), die seit Jahrzehnten (z. B. multivariate Analyse) oder Jahrhunderten (Gauß, Euklid) bekannt sind. Hierzu wird auf für Labordiagnostiker besonders geeignete Lehrbücher der Biostatistik [7, 8] verwiesen.

Einsatz der Bioinformatik in der Labordiagnostik

Die geschilderte Problematik ist keineswegs neu. Seit den frühen 70er Jahren sieht sich die mechanisierte Labordiagnostik mit der Situation konfrontiert, mehr Daten zu produzieren, als der diagnostizierende Arzt verarbeiten kann. Entsprechend umfangreich ist die Literatur zur Computer-assistierten Analyse von Analyseprofilen [9], auch wenn der Begriff *Bioinformatik* damals noch nicht explizit verwendet wurde. Stattdessen sprach man von *Medicometrics* [10] oder *Chemometrics* [11] und blieb wegen der geringen Leistungsfähigkeit damaliger Computer mehr in theoretischen Diskursen gefangen.

Es ist sicher ein Verdienst der jüngeren Forschung, die aus der SMAC-Ära [12] wohlbekannten Probleme im Licht der Microarrays neu zu betrachten, doch erstaunt es, wie wenig von den damals gewonnenen Erkenntnissen in der aktuellen Literatur auftaucht. Ab den 50er Jahren wurde insbesondere die Vereinheitlichung von Referenzbereichen durch Normierung von Laborresultaten auf einen Referenzwert intensiv, aber ohne nennenswerte Erfolge erforscht (Übersicht bei Goldschmidt [9]). Angesichts der Datenmengen, die die Labordiagnostik auch ohne Microarrays produziert, wären Fortschritte in dieser Richtung durchaus willkommen.

Im Rahmen der aktuellen Kosten- und insbesondere der künftigen DRG-Diskussion dürfte das Thema einerseits überflüssiger und andererseits preisgünstiger Profiluntersuchungen neue Brisanz gewinnen. Auch die zunehmende Annäherung zwischen Proteomics und klassischer Proteindiagnostik gibt der Bioinformatik womöglich zusätzliches Gewicht.

Der Begriff *Bioinformatik* entstand erst am Ende des 20. Jahrhunderts parallel zum *Human Genome Projekt* und ist bis heute noch nicht einheitlich definiert [13]. Man fasst darin meist alle Datenverarbeitungs-Techniken zusammen, die dazu dienen, die Ergebnisse der molekularen Biologie und Medizin zu strukturieren, zu analysieren und nutzbar zu machen. Es geht dabei vor allem darum, „die Struktur und Funktion von Genen und Proteinen aufzuklären“ [14], eine überaus komplexe Aufgabe, wenn man bedenkt, dass allein am Krebsgeschehen etwa 5 000 Gene beteiligt sein sollen, die in mindestens 200 typischen Wertemustern unterschiedlichste diagnostische und therapeutische Konsequenzen haben [15].

Probenvorbereitung und Hybridisierung

Typischerweise wird ein mRNA- oder Proteingemisch aus Zellen oder Gewebeproben extrahiert. Hierzu werden die Proben homogenisiert und im Fall der empfindlichen mRNA gegen enzymatischen Abbau stabilisiert. Die Frage lautet, ob das Genexpressionsmuster der untersuchten Proben Unterschiede aufweist oder auch, ob sich das Muster eines bestimmten Zelltyps im Verlauf einer Entwicklung ändert. Die Wahl des Ausgangsmaterials hängt somit von der Fragestellung ab. Im Folgenden gehen wir davon aus, dass das mRNA-Muster einer maligne entarteten Probe P mit dem einer gesunden Referenzprobe R verglichen werden soll. Beide Proben können aus unterschiedlichen Quellen stammen. Bessere Vergleichbarkeit wird aber erreicht, wenn sie aus entarteten bzw. unverdächtigen Anteilen derselben Probe mittels Laser-Mikrodissektion [16] gewonnen wurden.

Sobald die zu untersuchende mRNA beider Proben in stabiler, wässriger Lösung vorliegt, folgt ihre differentielle Markierung mit Signalmolekülen, typischerweise mit den Fluoreszenzfarbstoffen Cy3 (grün) und Cy5 (rot). Probe und Referenz werden nach Bestimmung der Gesamt-RNA-Gehalte im Verhältnis 1 : 1 gemischt und auf einen DNA-Chip aufgetragen. Sie binden dort an komplementäre Einzelstränge der punktförmig aufgebrachten Fängermoleküle, sog. Sonden.

Als mRNA-Sonden kann man Oligonukleotide mit definierter Sequenz synthetisieren oder cDNA aus natürlicher mRNA mit Hilfe der reversen Transkriptase erzeugen. Im Fall von Protein-Microarrays tritt an die Stelle der genomischen Hybridisierung meist eine Protein-Protein-Interaktion, z. B. in Form einer Antigen-Antikörper-Reaktion. In jedem Fall muss die Menge der gebundenen Analyte im Vergleich zur Gesamtmenge vernachlässigbar sein, so dass die Signale der gebundenen Marker im Sinne der Ambient Analyte Theorie von Ekins [17] vom aufgebrachten Probenvolumen unabhängig und proportional zur Konzentration in der Probe sind. Dies ist wichtig, denn andernfalls müsste man ähnlich wie bei Immuno-Assays für jede Sonde eine Standardkurve erstellen.

Standardisierung und Normalisierung der Rohdaten

Um die Microarray-Ergebnisse verschiedener internationaler Arbeitsgruppen untereinander vergleichbar und transferierbar zu machen, wurden vor allem in Nordamerika, Europa und Japan standardisierte, öffentlich zugängliche Datenbanken geschaffen [18]. Sie stehen allerdings in Konkurrenz mit proprietären Datenbanken von Microarray- und Geräteherstellern wie z. B. Affymetrix sowie mit Bioinformatikprojekten großer Computerfirmen wie IBM und sind auch untereinander nur bedingt vergleichbar. Insofern ist die Hoffnung auf völlige Vergleichbarkeit von Biochipdaten vorläufig utopisch. Interessanterweise hat sich die amerikanische

Food and Drug Administration FDA erst kürzlich gegen eine „Knebelung" des technischen Fortschritts durch verfrühte Vorschriften zur Standardisierung ausgesprochen [19]. Für den Fall einer diagnostischen Nutzung stellt die Behörde allerdings strenge Qualitätsauflagen in Aussicht und warnt vor allem vor einer Flut falsch positiver Ergebnisse: Mit dem heute üblichen 95 %-Referenz-Bereich ergäbe ein einziges Experiment mit 4 000 Sonden für eine gesunde Person bereits 200 pathologische Werte. Allerdings sind für solch umfassende Microarrays ohnehin Referenzbereiche einzelner Sonden ohne Kenntnis der abzugrenzenden Krankheiten kaum bestimmbar.

Der wichtigste Schritt in Richtung Vergleichbarkeit ist die Umwandlung absoluter in relative Messwerte: Um eine Über- bzw. Unterexpression feststellen zu können, muss man die Signale von Probe und Referenz zueinander in Bezug setzen. Dies kann nach getrennter Messung rein rechnerisch durch Quotientenbildung (Probe/Referenz) oder durch Vermischung von Probe (z. B. rot) und Referenz (z. B. grün) in einem Ansatz geschehen. Im letzteren Fall kommt es zu kompetitiver Bindung beider Anteile unter exakt identischen Analysebedingungen. Ist das Verhältnis genau 1 : 1, so kann man – bei gleichen Molekulargewichten – eine identische Anzahl von mRNA-Molekülen annehmen. Aus dem Rot-Grün-Verhältnis jedes Messpunkts errechnet man dann den Grad der Unter- bzw. Überexpression in Vielfachen bzw. Bruchteilen von 1. In der Praxis ist diese Bedingung allerdings meist nicht vollständig erfüllt. Deshalb muss man bei jedem Experiment prüfen, ob der auf die Referenz normierte mittlere Quotient sogenannter Housekeeping-Gene dem Erwartungswert von 1 entspricht. Ist dies nicht der Fall, so sind Korrekturen fällig. Eine sorgfältige Prüfung der Gene, die als Housekeeping-Gene Verwendung finden, ist von großer Bedeutung für die Qualität der Information.

Jedes Signal zeigt eine Reihe von Abhängigkeiten, die sich wie folgt definieren lassen [20]: Die Expressions-Intensität eines Genes ist eine Funktion der Ausbeute bei der mRNA Reinigung, der Markierung, Hybridisierung und Bildgebung, im Fall einer vorherigen Amplifikation auch von deren Ausbeute. Die einzelnen Glieder dieser Ereigniskette werden von diversen Faktoren beeinflusst: So sind z. B. die Farbstoffe Cy3 und Cy5 unterschiedlich stabil und zeigen unterschiedliche Quenching-Charakteristik, und die Träger des Arrays beeinflussen Richtigkeit und Präzision der Messung in Abhängigkeit von Material, Eigenfluoreszenz, Größe der aufgebrachten Sonden usw. Alle diese Schritte tragen zur Gesamtvariabilität der Ergebnisse bei. Der Verarbeitungsweg von den Rohsignalen bis zu den schlussendlich normalisierten Relativsignalen besteht aus folgenden Schritten:

Segmentierung der Microarrays: Die Lage der einzelnen Elemente auf einem Microarray wird mittels Bildverarbeitung ermittelt und in einer Datenbasis, der sog. GIPO-Datei (gene in plate order) verwaltet.

Elimination des Hintergrunds: Eine vorhandene Hintergrundsfluoreszenz wird gemessen und subtrahiert oder über eine Regressionsrechnung eliminiert. Problematisch ist vor allem eine Überkorrektur des Hintergrundsignals, da sie zu negativen Messwerten mit nicht definiertem Logarithmus führt.

Bestimmung der Intensität der Fluoreszenzsignale: Das eigentliche Fluoreszenzsignal ist nicht homogen, sondern die Summe einzelner Bildpixel. Diese können je nach Messtechnik aufaddiert, integriert oder als Mittel- bzw. Medianwerte erfasst werden. Wird eine Mischung aus Probe und Referenzprobe analysiert, müssen die roten und grünen Signale getrennt gemessen und dann zueinander in Beziehung gesetzt werden. Typischerweise ist das Resultat der Messung ein Grauwert im 16-Bit-Format, also eine Zahl zwischen 1 und 65 536.

Transformation des roten und grünen Signals: Man bildet zunächst einen Quotienten aus Probe P und Referenz R ($Q = P/R$). Er ist bei gleicher Signalintensität von Probe und Referenz 1. Überwiegt der Probenanteil, so steigt der Quotient ohne Limitierung an, überwiegt das Referenzsignal, so geht er gegen Null. Um diese Limitierung nach unten zu vermeiden, bildet man den Logarithmus zur Basis 2 und erhält ein symmetrisches Spektrum von negativen und positiven Werten: $\log_2(1) = 0$, $\log_2(2) = 1$, $\log_2(4) = 2$, $\log_2(0{,}5) = -1$, $\log_2(0{,}25) = -2$ usw.

Normalisierung des Expressionssignals: Die Messwerte innerhalb eines Experiments und zwischen verschiedenen Experimenten müssen vergleichbar sein, damit sinnvolle Aussagen möglich werden. Dies erfordert, dass die Signalintensitäten, die durch die oben beschriebenen Unterschiede bedingt sind, ausgeglichen werden. Die Grundannahme für das Verfahren ist, dass in einem Microarray mit zufälliger Anordnung beliebiger Gene ein ausgeglichenes Verhältnis von Referenz und Probesignal vorliegt. Das einfachste Korrekturverfahren besteht darin, alle P/R-Quotienten durch das Verhältnis N_{ges} der Signalsummen zu dividieren:

$Q_i = 1/N_{ges} * P_i/R_i$.

In der logarithmierten Form entspricht dies der Subtraktion einer Konstanten vom Logarithmus des P/R-Quotienten:

$\log_2(Q'_i) = \log_2(Q_i) - \log_2(N_{ges.})$.

Grafisch dargestellt ergibt sich eine Parallelverschiebung aller Messpunkte nach der Seite des unterrepräsentierten Fluoreszenzfarbstoffs hin (Abb. 1).

Bei einem Verfahren mit der Bezeichnung lowess (locally weighted linear regression) erkennt man systematische Abweichungen durch Auftragen von log (R * G) gegen log (R/G). Der mathematische Hintergrund dieser Darstellung ist der Vergleich des mittleren Signals (R + G)/2 gegen die Differenz (R–G)/2. Auf der logarithmischen Ebene wird aus der Summe ein Produkt, aus der Differenz ein Quotient. Mit der lokalen gewichteten linearen Regression werden für einzelne Bereiche der Bezugskurve Korrekturen ermittelt [21].

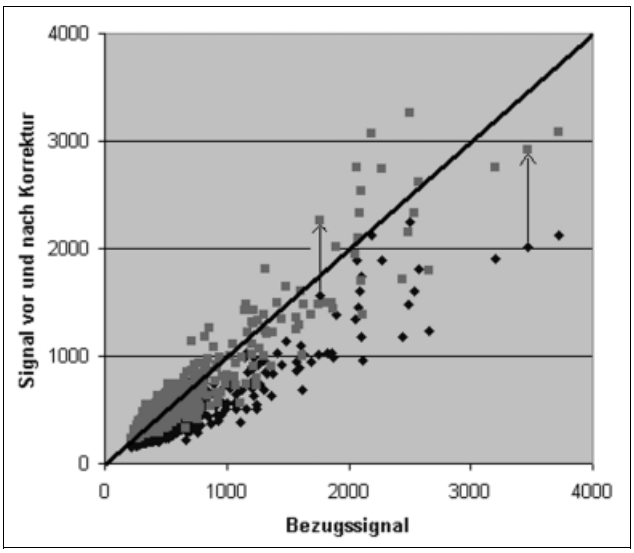

Abbildung 1 Expressionssignale von 1109 Oligonukleotid-Sonden vor und nach Kalibration gegen die Summe aller Signale in einer unabhängigen Referenzprobe. Dieses Verfahren ist nur dann anwendbar, wenn die Zahl der fehlregulierten Gene im Vergleich zur Gesamtmenge gering ist. Andernfalls käme es zu systematischen Fehlkorrekturen. Dunkle Rauten = unkorrigierte Signale; Helle Quadrate = korrigierte Signale; Die eingetragene Linie entspricht der 45°-Geraden y = x.

Ist das Verhältnis rot : grün eines Messpunkts nach Durchführung aller Korrekturen deutlich kleiner als 1, so nimmt man eine Unterexpression an, ist es deutlich höher als 1, ist eine Überexpression wahrscheinlich. Da Richtigkeit und Streuung der Messungen im Fall von mehreren 1 000 Sonden oft nicht im Detail analysierbar sind, werden typischerweise nur Abweichungen um mehr als Faktor 2 vom Referenzwert als signifikant angesehen. Diese sog. Two-Fold-Rule ist allerdings eine Notlösung, die nach Möglichkeit zugunsten einer sauberen Analyse der einzelnen Varianzen aufgegeben werden sollte [22], da andernfalls die Signifikanz von Abweichungen nicht angegeben werden kann. Wir haben in einem Test mit etwa 5 000 Oligonukleotid-Sonden [23] an gesunden Referenzpersonen beispielsweise Intraassay-Varianzen dadurch ermittelt, dass jedes Gen auf demselben Microarray durch fünf Sonden repräsentiert war (Abb. 2). Für die Interassay-Varianz kann man Mittelwerte oder Mediane solcher Mehrfachmessungen einsetzen, um aussagekräftige Resultate zu erhalten.

Identifizierung von unterschiedlich exprimierten Genen

Es ist von Interesse festzustellen, welche Gene sich in ihrer Expression wie stark von der Referenz unterscheiden. Hierzu bildet man meist Rangstufen, denen man in der grafischen Computeraufbereitung Farben gibt. Die typische Darstellungsform ist eine Tabelle, in der jede Zeile einem Gen und jede Spalte einem Microarray-Experiment (also z. B. einem Patienten) entspricht.

In einer grundlegenden Arbeit von 1998 haben Eisen *et al.* [24] für die „intuitive" Erfassung von Massendaten jeden Messwert im Referenzbereich (Log_2 von –1 bis +1) schwarz, erhöhte Log-Werte über 1 rot und verminderte Log-Werte unter –1 grün gefärbt. Das Computerprogramm kann unter Java direkt im Internet ausgeführt werden (s. Tab. 1), doch ist es auch sehr einfach, entsprechende Bilder mit MS Excel® herzustellen, wobei man dann die Farben nach eigenem Ermessen (z. B. gelb für Normalwerte) wählen kann (Abb. 3).

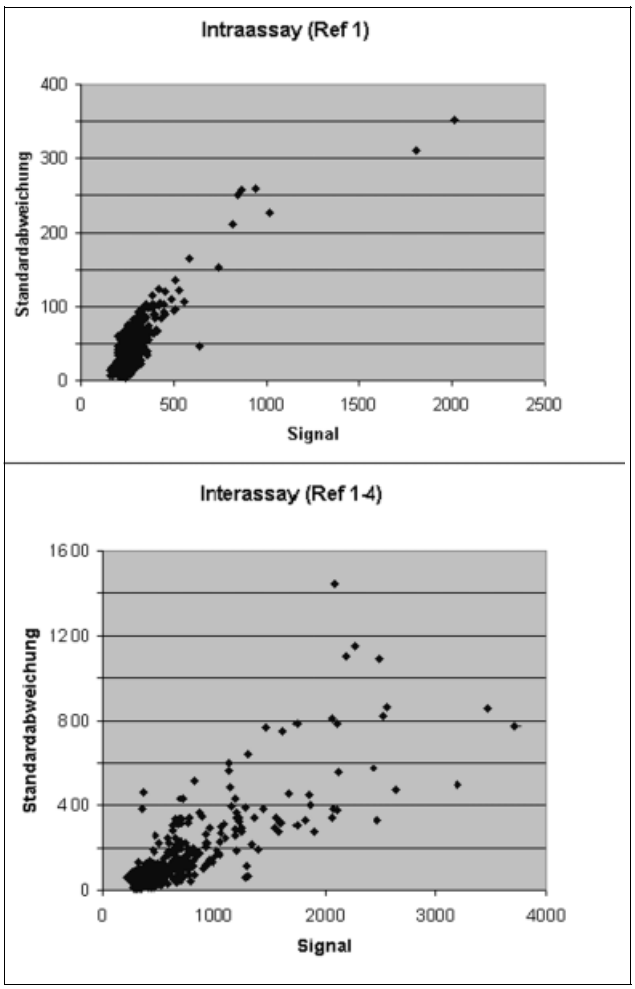

Abbildung 2 Typische Profile der Intra- und Interassay-Streuung von Microarray-Signalen, erhoben an vier gesunden Referenzpersonen mit jeweils fünf Sonden je Gen. Für die Interassay-Streuung wurden die jeweiligen Mittelwerte nach Kalibration (vergleiche Abbildung 1) verwendet. Die z. T. erhebliche Streuung der Genexpressionen zwischen den vier gesunden Referenzpersonen zeigt, dass der Begriff „Normalität" bei über 1 000 Einzelwerten kritisch hinterfragt werden muss. Bei jeder solchen Messung sind einzelne Gene über- oder unterexprimiert, je nachdem, welche Referenz man als Bezugswert wählt.

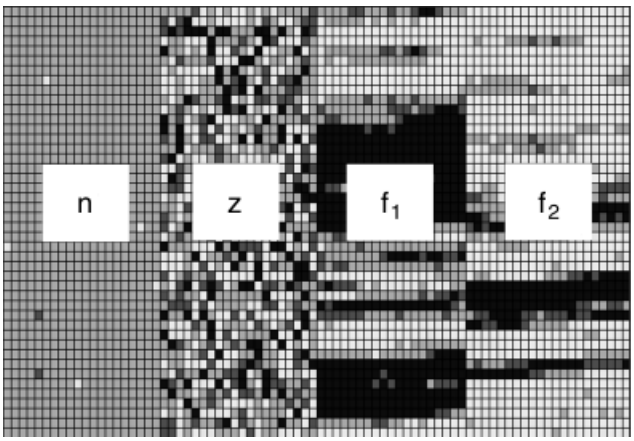

Abbildung 3 Erstellung experimenteller Daten zur Testung von Verfahren der Bioinformatik mit MS Excel® und Visual Basic for Applications (VBA®). Jede Zeile entspricht einem Gen, jede Spalte einem Experiment (Fall). Durch die Simulation lassen sich weitgehend natürliche Muster erzeugen. Im dargestellten Beispiel wurden je 20 Patienten mit Gauß'scher Verteilung (normale Expression mit wenigen Werten außerhalb des 2-SD-Bereichs), Gleichverteilung (Zufallsgenerator) und zwei regelhaften Mustern (stochastische Brown'sche Fraktale) erzeugt. Hellgrau: Unterexpression; Mittelgrau: Normalexpression; Dunkelgrau: Überexpression. n = normal verteilt, z = zufallsverteilt, f_1, f_2 = zwei fraktale Muster

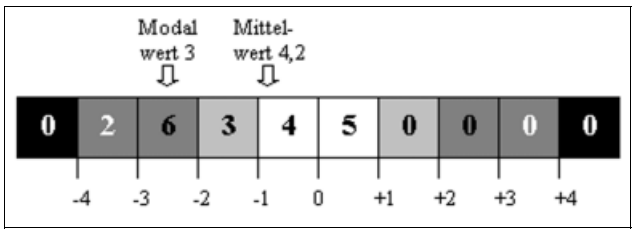

Abbildung 4 Eine aus labordiagnostischer Sicht interessante Verteilung der dualen Logarithmen von Genexpressionsdaten in zwei Kollektiven von je 10 Patienten: Die in 10 Ränge sortierten Messungen weisen zwei Gipfel auf, von denen einer zwischen 0,125 (2^{-3}) und 0,25 (2^{-2}), der andere zwischen 1 (2^0) und 2 (2^1) liegt. Dieses Gen ist ein guter Kandidat für die Trennung von Tumorpatienten (Unterexpression) und Referenzpersonen (normale Expression). Für die Erkennung solcher Konstellationen eignet sich z. B. die Rangnummer des höchsten Werts (sog. Modalwert = 3) im Vergleich zum Rangmittelwert (= 4,2).

Ein in der Bioinformatik derzeit intensiv bearbeitetes Feld ist die methodisch bedingte Instabilität der Varianz über den weiten dynamischen Bereich der Fluoreszenzsignale. Es ist leicht ersichtlich, dass die Varianz der Rohsignale im Bereich geringer Expression klein sein muss, während sie bei Genen mit starker Überexpression im Bereich der oberen Messbereichsgrenze extrem zunehmen kann (s. Abb. 2). Die oben beschriebene Logarithmierung kehrt die Verhältnisse um: Im oberen Messbereich werden die Varianzen verschwindend klein, während der Logarithmus bei Werten nahe Null gegen minus unendlich geht. Es gibt unterschiedliche Verfahren zur mathematischen Varianzstabilisierung, z. B. das auf dem arcsin hyperbolicus basierende Fehlermodell des DKFZ Heidelberg [25], doch ist es im Allgemeinen ausreichend, die Genexpression wie oben beschrieben in Kategorien (z. B. ––, –, normal, +, ++) einzuteilen und dann mit diskreten Rängen weiterzuarbeiten (s. Abb. 4).

Die Diskussion, wie und mit welchen Verfahren Signifikanzprüfungen durchgeführt werden, ist im Gange, ohne dass derzeit ein Verfahren als „Standard" angesehen werden kann. Wir haben uns entschieden, die im folgenden beschriebenen Aussagen mit simulierten und realen Datensätzen zu überprüfen, um Empfehlungen aussprechen zu können. Ohne an dieser Stelle auf methodische Details einzugehen, kamen zur Simulation vorwiegend sogenannte „Brown'sche Fraktale" [26] zum Einsatz. Dabei handelt es sich um selbstähnliche Datenreihen, die durch einstellbare stochastische Schwankungsbreiten (die sog. Fraktale Dimension) sowohl normal verteilte Referenzwerte als auch naturgetreue Abweichungen von der Norm mit statistisch auswertbaren Mustern liefern (s. Abb. 3). Der Vorteil eines solchen Verfahrens ist neben dem geringen Kosten der Umstand, dass man die Daten selbst modellieren kann, während bei nasschemischen Messungen in Tumorzellen unberechenbare Randkonditionen die Bewertung der Analysenergebnisse oft schwierig bis unmöglich machen.

Verfahren zur Datenverarbeitung

Nachdem die einzelnen Gen-Daten standardisiert und normalisiert sind, folgen die Verfahren der Datenanalyse. Hierbei geht es darum, charakteristische Gruppen von Genen zu finden, die die Zustände *normal* und *abnormal* eindeutig beschreiben, also etwa in dem Sinne, dass Gene des Apoptose-Systems beim Übergang zur Malignität inaktiviert werden, so dass sich Zellen mit schwerwiegenden genetischen Defekten beliebig unbegrenzt teilen können. Aus den Über- und Unterexpressionsmustern solcher Gengruppen leiten sich beispielsweise neue, biochemisch definierte Tumorklassifizierungen oder tumorspezifische Therapieansätze ab.

Aus labordiagnostischer Sicht besteht eine zweite wichtige Aufgabenstellung darin, aufgrund der beobachteten Expressionsmuster Vorschläge für ein möglichst effizientes Chipdesign zu machen. Die derzeit kommerziell erhältlichen Microarrays, beispielsweise GeneChip® von Affymetrix oder MatriXarray® von Roche, sind nur für wissenschaftliche Zwecke zugelassen und auch dafür optimiert. Sie dienen dem Verständnis des pathologischen Geschehens und weniger der zuverlässigen, kostengünstigen Diagnostik eines bestimmten Tumors. Das Ziel muss es nun sein, aus der Fülle der angebotenen Sonden die geeignetsten auszuwählen.

Datenreduktion: Ehe man sich der eigentlichen Musteranalyse widmen kann, muss man möglichst viele Signale, die mit der getesteten pathologischen Bedingung nichts zu tun haben, aus dem Rohdatensatz entfernen, um die gewünschten Muster nicht im Rauschen der Daten untergehen zu lassen. Im ersten Schritt bietet sich eine univariate Verteilungsanalyse an, ähnlich wie sie bei der Einführung neuer Tests auch in der klassischen Labordiagnostik üblich ist. Allerdings muss man die logarithmische Verteilung der Rohdaten berücksichtigen. Hat man beispielsweise 10 Kontrollpersonen und 10 Tumorpatienten mit je 500 oder 5 000 Sonden vermessen, so sollte man zunächst die Expressionsdaten aller Sonden eliminieren, deren Mittelwert und Gesamtstreuung für beide Kollektive im Referenzbereich liegen: Sie sind als Referenzgene („Housekeeping-Gene") wichtig für die Überprüfung und Normalisierung des Experiments (s. o.), tragen aber vermutlich nichts zur Differenzierung der beiden Klassen bei. Zur Prüfung bildet man beispielsweise 10 diskrete Kategorien oder Ränge (z. B. normal, hoch, sehr hoch, niedrig, sehr niedrig usw.) und zählt ab, wie viele Werte jeweils in eine Kategorie fallen.

In Abbildung 4 ist ein solches Beispiel konkret mit 10 Rängen dargestellt. Man kann solche diskreten Verteilungen entweder optisch mit Hilfe von Histogrammen oder rechnerisch mit den traditionellen Verfahren der Statistik (Mittelwert, Varianz, Median, Modalwert, Anpassung an Poisson- und Binomialverteilung etc.) prüfen (Abb. 5). Kleine Varianzen von weniger als 20 % des Mittelwerts, Modalwerte um 0 oder auch zufällige Gleichverteilungen über mehrere Kategorien hinweg sprechen gegen die Nützlichkeit der entsprechenden Sonde für die gestellte Frage, große Varianzen, schiefe und mehrgipfelige Verteilungen sind günstig.

Es ist grundsätzlich lohnend, die univariate Analyse nicht nur in den Zeilen, sondern auch in den Spalten der Tabelle durchzuführen. Anstelle der Verteilungskurven je Gen erhält man so die Werteverteilung je Fall und erkennt oft bereits mit freiem Auge typische Muster. So kann man z. B. erwarten, dass die Expressionsdaten von Normalpersonen eng um den Mittelwert von 1 streuen, während pathologische Proben zusätzliche Gipfel aufweisen sollten. Hierfür ist allerdings die obige Vorfilterung eine Grundvoraussetzung, da andernfalls die interessierenden Muster von der großen Anzahl normaler Genexpressionen völlig überdeckt würden.

Falls die Zuordnung der Fälle zu bestimmten Klassen durch das experimentelle Design bereits zweifelsfrei bekannt ist, eignet sich ein von uns entwickeltes einfaches Filterverfahren zur raschen Erkennung aussagekräftiger Gene [27]. Hierbei wird für jedes Gen eine Matrix mit x-Kategorien und y-Klassen angelegt, z. B. 3 Kategorien (*niedrig, normal, hoch*) und 3 Klassen (*Kontrolle, Tumor A, Tumor B*). Man zählt aus, wie viele Ergebnisse in jedes Feld dieser Tabelle fallen, bildet Spalten- und Zeilensummen und drückt jedes einzelne Zählergebnis in Prozent dieser zwei Summen aus. Der Mittelwert der beiden so erhaltenen „Diskriminatoren" ist ein Maß für die Nützlichkeit dieses Gens zur Mustererkennung. Weist ein Gen in einem der Felder z. B. einen 100 %-Wert auf, so heißt dies, dass es als Indikator für die jeweilige Klasse in Frage kommt. Interessierte finden unter www.trillium.de ein Excelprogramm, das die Berechnung mit simulierten Daten durchführt. Sowohl in der Simulation als auch an realen Daten von Leukämiepatienten [23, 27] erwiesen sich Gene, deren bester Diskriminator unter 80 % lag, für die weitere Analyse als entbehrlich bzw. eher störend, während Gene mit 100 %-Diskriminatoren verständlicherweise hervorragende Kandidaten für die Differenzierung zwischen normal und pathologisch waren.

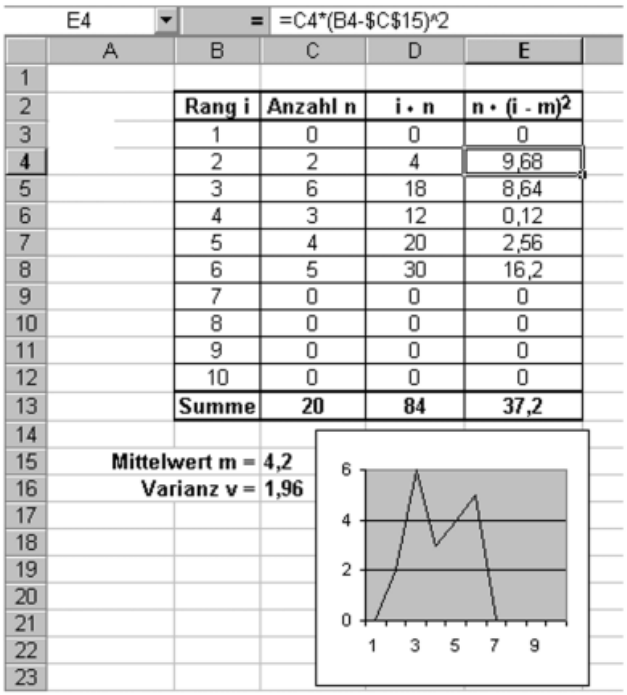

Abbildung 5 Darstellung und Berechnung von Schlüsseldaten mit MS Excel®. In Spalte B sind die Rangnummern 1 bis 10, in Spalte C die entsprechenden Zählergebnisse aufgetragen. Alle übrigen Werte sowie die zweigipfelige Histogramm-Darstellung sind mit Standardfunktionen erstellt. So werden beispielsweise für die Berechnung der Varianz die Differenzen zwischen Rangwert und Rangmittelwert quadriert und aufaddiert (Spalte E) und dann durch n–1 geteilt (37,2/19 = 1,96). Eine wertvolle Anleitung zum Einsatz von MS Excel® für Fragestellungen dieser Art findet sich in [35].

Clustering: Hat man den Datensatz auf potenziell aussagekräftige Daten reduziert, so gilt es nun, relevante Muster darin zu erkennen. Diese Bildung von Gruppen wird als „Clustering" bezeichnet. Dabei sollen die einzelnen Gruppen in sich möglichst homogen, die Unterschiede zwischen den Gruppen möglichst groß sein. Die Vorarbeit besteht in der Definition eines Proximitätsmaßes, wobei man Korrelationsmaße, Distanzmaße, Ähnlichkeitsmaße und probabilistische Proximitätsmaße unterscheidet. Am verbreitetsten sind Distanzmaße. Sie setzen eine metrische Skalierung, z. B. in Form der dua-

len Logarithmen von P/R-Quotienten oder eine Rangmetrik wie in Abbildung 4 beschrieben, voraus und stellen – bildlich gesprochen – Entfernungen in einem multidimensionalen geometrischen Raum dar.

Auf der zweidimensionalen Fläche sind uns solche Entfernungstabellen von Autokarten her vertraut. Zeilen und Spalten tragen dort Städtenamen, die Kreuzungsstellen enthalten die Entfernung, beispielsweise von München nach Berlin oder von Düsseldorf nach Köln. Zur Erläuterung sei dieses illustrative Beispiel im Detail besprochen.

Im ersten Schritt werden die Objekte mit dem geringsten Abstand (z. B. Düsseldorf und Köln) zu einem „Cluster" zusammengefasst, die beiden Originalnamen werden dafür aus der Tabelle entfernt und die Distanzen neu berechnet. Als Abstand des Clusters „Düsseldorf-Köln" zu den übrigen Städten setzt man wahlweise den Mittelwert (average linkage), das Minimum (single linkage) oder das Maximum (complete linkage) der ursprünglichen Entfernungen ein. So fährt man fort, bis aus n einzelnen Städtenamen ein einziger langer Name geworden ist, d. h. bis *n-Cluster mit je einem Element in einen Cluster mit n-Elementen* umgewandelt sind. Darin werden die Großstädte des Ruhrgebiets entsprechend ihrer geographischen Entfernung eng beisammen stehen, andere Städte dagegen verstreut angeordnet sein.

In Analogie zu den Städtedistanzen kann man auch die Distanzen von Genen berechnen, wobei man allerdings anstelle der Kilometer andere Maße entwickeln muss. In vielen Arbeiten wird die euklidische Vektordistanz der transformierten und normalisierten Fluoreszenzen verwendet. In der bereits erwähnten Arbeit von Eisen *et al.* [24] wird ein Ähnlichkeitsmaß statt eines Distanzmaßes bevorzugt. Der Korrelationskoeffizient dient dabei als Maß der Ähnlichkeit. In beiden Fällen vergleicht man die Messwerte von je zwei Sonden über alle Experimente (bivariate Analyse). Wären z. B. alle Werte identisch, so müsste als Distanz 0 resultieren. Je mehr Werte deutlich voneinander abweichen, desto größer soll die Maßzahl werden. Die euklidische Distanz berechnet sich analog zum Satz des Pythagoras aus der Wurzel der Distanzquadrate auf der x- und y-Achse ($x^2 + y^2 = \text{Distanz}^2$), die „Manhattan-Distanz" oder „City-Block-Distanz" stellt einfach die Summe beider Distanzen dar ($x + y = \text{Distanz}$). Zu bevorzugen ist die euklidische oder auch die quadrierte euklidische Distanz. Beide ergeben identische und relativ stabile Resultate. Ob die Distanz oder die Ähnlichkeit als Grundlage der Clusterbildung dient, hängt von der Fragestellung ab: Um beispielsweise Fälle mit unterschiedlicher Genexpression verschiedenen Clustern zuzuordnen, eignet sich eher die Distanz, während man ähnlich exprimierte Gene relativ gut mit Korrelationen clustern kann.

Wären z. B. die Werte der einen Sonde immer genau halb so hoch wie die einer anderen (z. B. da Gen 1 von Gen 2 gehemmt wird), so ergäbe die euklidische Distanz eine große Entfernung zwischen beiden Sonden, die den Zusammenhang verschleiert, während der Korrelationskoeffizient r die Ähnlichkeit klar darstellen würde.

Es sind zahlreiche weitere Verfahren mit oft wesentlich höherem Rechenaufwand im Einsatz, z. B. die von neuronalen Netzen abgeleiteten Self-Organizing Maps (SOMs) [28] oder das so genannte k-means [29] Verfahren. Viele Autoren haben sich intensiv mit den Unterschieden zwischen verschiedenen Cluster-Algorithmen befasst [30], und auch die Tagungen der GMDS [31] und anderer Fachgesellschaften sind voll von derartigen Berichten. Als Fazit kann nur festgehalten werden, dass jedes Verfahren Vor- und Nachteile besitzt und dass eine Einigung auf identische Verfahren innerhalb einer bestimmten Fachrichtung für den Datenaustausch und die gemeinsame Diskussion wichtiger ist als die Suche nach dem „besten" Verfahren. Dabei ist hervorzuheben, dass es sich bei allen genannten Verfahren grundsätzlich um rein mathematische Methoden handelt, die ohne biologische Begründung angewendet werden. Biologische Konzepte kann ohnehin nur der mit der biologischen Materie vertraute Experte entwickeln. Allerdings ist der Computer unabdingbare Voraussetzung dafür, dass der Experte überhaupt Anhaltspunkte für plausible Konzepte innerhalb des Datenflut findet.

Ausblick

Wie werden die Microarrays die Laboratoriumsmedizin in Zukunft beeinflussen? Die Zahl der Publikationen, die sich mit der Anwendung dieser Technik auf konkrete medizinische Fragen beschäftigen, hat in den letzten Jahren sprunghaft zugenommen. Hierzu zählen vor allem Studien zum Vorhersagewert der Genexpression, beispielsweise für den Verlauf des Brustkrebses [32]. Die Autoren zeigen einrucksvoll die Überlegenheit der Genexpressionsanalyse gegenüber klinischen und histologischen Verfahren. Untersucht wurden 295 Patientinnen, bei denen die Erstdiagnose zwischen 1984 und 1995 gestellt worden war. Unter etwa 5 000 Kandidatengenen wurden 70 identifiziert, die mit früher Fernmetastasierung assoziiert waren. Zu einem ähnlichen Ergebnis kam eine Studie aus der Universität Frankfurt, bei der mit Clusterverfahren und SOMs ein Satz von 41 Markergenen gefunden wurde, die prädiktiv für rasche Metastasierung waren (23 % bei Aufnahme in die Studie, 50 % nach zwei Jahren) [33].

Ähnlich Aussagen konnten für MLL (mixed lineage leukemia), eine besonders aggressive Leukämieform, gemacht werden [34]. Gegenüber einer ALL (acute lymphatic leukemia) fanden sich 1 000 Gene mit deutlicher Unterexpression, viele davon in der frühen Phase der B-Zell-Entwicklung involviert, daneben auch eine ganze Reihe von überexprimierten Genen. Hierzu ist anzumerken, dass die meisten Studien weit weniger fehlexprimierte Gene finden. Der Grund für solche Unterschiede ist in der Aggressivität bzw. im Fortschreiten einer malignen Erkrankung zu suchen, da Malignome im Verlauf immer mehr Mutationen anhäufen und damit „zu leben lernen".

Die Forderung der Labordiagnostik an die Hersteller von Microarrays wird in Richtung ausgewählter Sonden

für wichtige (weil besonders häufige oder besonders aggressive) Krankheitsbilder gehen. In der Regel dürften bis zu 100 Gensonden für eine spezifische Aussage genügen, jede möglichst in drei- bis fünffacher Ausfertigung, um die Intraassay-Präzision und Signifikanz von Abweichungen zu berechnen und Ausreißer zu erkennen. Dazu kommt ein Satz der oben erwähnten Housekeeping-Gene, deren Expression mit größter Wahrscheinlichkeit normal ist. Anhand dieser Gene kann man die Daten normalisieren und misslungene Experimente eliminieren. Alles in allem dürfte die Idealzahl von Sonden je Microarray für die Onkologie um 1 000, eher darunter liegen.

Weiterhin sind standardisierte, robuste Auswertealgorithmen der Bioinformatik für die Praxis wichtig. Die vorliegende Arbeit hat versucht, diejenigen vorzustellen, die bereits eine gewisse Verbreitung gefunden haben. Das Fach ist aber noch so jung, dass in den nächsten Jahren noch Weiterentwicklungen zu erwarten sind.

Inwieweit die hier erarbeiteten Verfahren zur Referenzwertnormalisierung und zur computerisierten Auswertung komplexer Muster auch positive Rückwirkung auf die klassische Labordiagnostik haben werden, bleibt abzuwarten. Gerade im Hinblick auf die aktuelle Diskussion um neue Normalwerte bei der Temperaturumstellung von Enzymen wäre es durchaus ein Segen, wenn zum Beispiel die Datennormalisierung bereits auf theoretisch wohlfundierten Füßen stände: Dadurch hätte viel Aufwand und Geld gespart werden können. Schließlich muss man die Weiterbildung der Labordiagnostiker in Bioinformatik verstärken und allgemein zugängliche, einfach zu bedienende Computerprogramme schaffen, damit unser Fach den künftigen Aufgaben gewachsen ist.

Literatur

1. Coleman W, Tsongalis G, editors. The Molecular Basis of Human Cancer. Totowa (New Jersey, USA): Humana Press, 2002.

2. Boultwood J, Fidler C, editors. Molecular Analysis of Cancer. Totowa (New Jersey, USA): Humana Press, 2002.

3. Rampal JB, editor. DNA Arrays – Methods and Protocols. Totowa (New Jersey, USA): Humana Press, 2001.

4. Liebler DC. Introduction to Proteomics – Tools for the New Biology. Totowa (New Jersey, USA): Humana Press, 2002.

5. Hahn W, Weinberg R. Modelling the Molecular Circuitry of Cancer. Nature Reviews Cancer 2002;2:331–41.

6. Duyk G. Sharper Tools and Simpler Methods. Nature supplement 2002;32:465–8.

7. Ehrenberg ASC. Statistik oder der Umgang mit Daten. Weinheim (Deutschland): VCH Verlagsgesellschaft mbH, 1986.

8. Keller H. Klinisch-chemische Labordiagnostik für die Praxis. Stuttgart (Deutschland): Thieme Verlag, 1986.

9. Goldschmidt HMJ. The Application of Multivariate Statistical Analysis in Clinical Chemistry and Haematology (doctoral thesis). Dordrecht (Niederlande): ICG Printing, 1987.

10. Goldschmidt HMJ, Leijten JF. Medicometrics: A New Promising Discipline. Trendelenburg C, editor. Proceedings of the 5th Int. Conf. on Computing in Clinical Laboratories. Kohl GmbH, 1985.

11. Kowalski BR. Chemometrics. Theory and Application. Washington DC (USA): ACS Symposium Series Vol 52, 1977

12. Schwartz MK, Bethune VG, Fleisher M *et al.* Chemical and Clinical Evaluation of the Continuous-Flow Analyzer SMAC. Clin Chem 1974;20:1062–70.

13. Rashidi HH, Bühler LK. Grundriss der Bioinformatik. Heidelberg Berlin (Deutschland): Spektrum-Verlag, 2001.

14. Dugas M, Schmidt K. Medizinische Informatik und Bioinformatik. Berlin Heidelberg (Deutschland): Springer-Verlag, 2003.

15. Bahls C, Fogarty M. Reining in a Killer Disease. The Scientist 2002;27:16–8.

16. Willingham E. Laser Microdissection Systems. The Scientist 2002;13:42–4.

17. Ekins R. Multi-Analyte Immunoassay. J Pharm Biomed Anal 1989;7:155–168.

18. Stoeckert C, Causton H, Ball C. Microarray Databases: Standards and Ontologies. Nature supplement 2002;32:469–73.

19. Petricoin EF, Hackett JL, Lesko LJ *et .al* Medical applications of microarray technologies: a regulatory science perspective. Nature supplement 2002;32:474–9.

20. Liao JC, Sabatti C. Microanalysis of DNA-Microarrays. ASM News 2002;68(9):432–7.

21. Quackenbush J. Microarray Data Normalization and Transformation. Nature supplement 2002;32:496–501.

22. Wolfinger R, Gibson G, Wolfinger E *et al.* Assessing Gene Significance from cDNA Microarray Expression Data via Mixed Models. J Comp Biol 2001;8:625–37.

23. Donner, H: Persönliche Mitteilung, MWG Biotech AG, Ebersberg, 2001.

24. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster Analysis and Display of Genome-Wide Expression Patterns. Proc Natl Acad Sci USA 1998;95:1463–8.

25. Huber W, von Heydebreck A, Sültmann H *et al*: Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression. Bioinformatics 2002;1:1–9.

26. Mandelbrot B. Die fraktale Geometrie der Natur (dt. Ausgabe). Basel (Schweiz): Birkhäuser Verlag, 1991.

27. Hoffmann G. A Simple Filter Algorithm for Gene Expression Profiling Data. JALA 2002;7:95–97.

28. Kohonen T. Automatic formation of topological maps of patterns in a self-organizing system. In: Oja E, Simula O, editors. Proceedings of 2nd Scand. Conference on Image Analysis. Helsinki (Finland) 1981;214–20.

29. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Le Cam LM, Neyman J, editors. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley (Califonia, USA): University of California Press, 1967;1:281–97.

30. Slonim DK. From Patterns to Pathways: Gene Expression Data Analysis Comes of Age. Nature supplement 2002;32:502–8.

31. Kruse E, Hüsing J, Hölter T *et al.* Einfluss verschiedener Klassifizierungsmethoden auf die Interpretation von Microarray-Daten. Abstract 148 der 47. Jahrestagung der GMDS, 2002.

32. Van de Vijver MJ, Yndong D, Van't Veer LJ *et al.* A gene-expression signature as a predictor of survival in breast cancer. NEJM 2002;347:1999–2009.

33. Ahr A, Karn T, Solbach C *et al.* Identification of high risk breast-cancer patients by gene expression profiling. Lancet 2002; 359:131–2.

34. Armstrong SA, Staunton JE, Silverman LB *et al.* MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature Genetics 2002;30:41–7.

35. Fleischhauer C. Excel in Naturwissenschaft und Technik. München (Deutschland): Addison-Wesley, Pearson Education Deutschland GmbH, 2000.