Microarrays, Genexpressionsanalyse und Bioinformatik

Daniel Schober

Max Delbrück Zentrum für Molekulare Medizin Berlin-Buch, AG Bioinformatik

In den letzten Jahren wurden große Fortschritte bei der Entwicklung von Ultrahochdurchsatz-Techniken zur Datenerfassung in den Biowissenschaften erzielt. Schnelle Sequenzierungstechniken ermöglichten die nahezu vollständige Entschlüsselung des menschlichen Genoms und einiger weiterer Modellorganismen. Der nächste konsequente Schritt in Richtung auf ein holistisches Verständnis des synergetischen Zusammenspiels der molekularen Bestandteile des Organismus im Rahmen der funktionalen Genomik ist die Entschlüsselung des Transkriptoms und des Proteoms. Eine hierfür geeignete Technik, besonders zur parallelen Massendatenerfassung der RNA-Mengen sehr vieler Gen-Transkripte, ist die in den letzten Jahren weiter perfektionierte Microarray-Technik. Die Auswertung der hierbei anfallenden sehr großen Datenmengen stellt die Wissenschaft vor Probleme, denen durch eine weitergehende Automatisierung, auch des Auswertungsprozesses, begegnet wird.

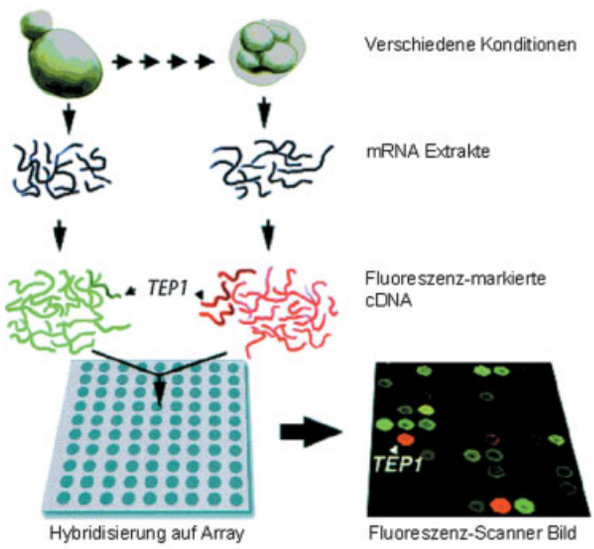

Abb. 1: Der schematische Ablauf eines Mikroarray-Experimentes. Genaue Beschreibung siehe Text. Bild modifiziert nach [6].

Das Funktionsprinzip

Die Microarray-Technik basiert auf der Hybridisierung von Nukleinsäuren. Komplementäre Nukleinsäure-Einzelstränge lagern sich dabei spezifisch über Wasserstoffbrückenbindungen zwischen ihren Purin- und Pyrimidin-Basen aneinander. Auf einer Immobilisierungsmatrix wie beschichteten Glas-Objektträgern, Siliziumchips oder Nitrocellulose-Membranen immobilisiert man an definierten Positionen Nukleinsäuren zu untersuchender Gene bekannter Sequenz, die sogenannten Proben-Nukleinsäuren. Diese hybridisieren dann mit unterschiedlich fluoreszenzmarkierten Target-Nukleinsäuren aus verschiedenen zu untersuchenden Geweben oder Gewebekonditionen. Die im Gewebe vorhandenen Target-Nukleinsäuren werden nach Abwaschen unspezifisch gebundener Target-Nukleinsäuren über einen Fluoreszenz-Scanner durch den Ort der Hybridisierung auf dem Chip identifiziert und über das Verhältnis der Fluoreszenzintensitäten bei den für die Markierungen der zu vergleichenden Target-Samples charakteristischen Wellenlängen quantifiziert.

Microarrays mit besonders kleiner meist siliziumbasierter Immobilisierungsmatrix, auf der die Proben-Nukleinsäuren besonders dicht aufgetragen sind, nennt man Microarray-Chips. Wenn ein vollständiges Genom auf einem Chip untersucht werden kann spricht man vom Genom-Chip. Dieser Begriff schließt die zunehmend wichtiger werdende Protein-Chip Technik mit ein. Eine allgemeine Bezeichnung für alle Microarray-Typen ist Bio-Chip. Nach der immobilisierten Proben-DNA unterscheidet man auch Format I Microarrays, bei denen 500 bis 5,000 Basen lange cDNAs als Proben-DNA immobilisiert wird und Format II Microarrays, bei denen 20 bis 25 Basen lange Oligonukleotide als Proben-DNA dienen. Diese können in situ direkt auf dem Chip synthetisiert oder erst synthetisiert und dann auf dem Chip immobilisiert werden. Sind die Nukleinsäure-Spots kleiner als 250 µm, spricht man auch von hochauflösenden Microarrays.

Herstellung der Microarrays

Der wichtigste kommerzielle Anbieter von Microarrays, die Firma Affimetrix, benutzt eine photolithographische Festphasen-Synthese oder Very Large Scale Immobilized Polymer Synthesis (VLSIPS) genannte Technik der Herstellung[3]. Photolabile Schutzgruppen auf Glassubstrat werden durch Licht, das selektiv durch eine photolithografische-Maske strahlt, ortsgebunden für die On-the-spot-Oligosynthese aktiviert. Das Glassubstrat wird dann mit einer photolabilen DNA-Base geflutet, die an die definierten vorher beleuchteten Arraystellen binden. Für die nächsten Positionen in den Sequenzen werden dann entsprechend andere photolithographische Masken benutzt und der Vorgang wiederholt. Für Jede Base im Proben-Oligo (pro Position) werden also 4 Masken benötigt. Der Vorteil ist die direkte Herstellung auf dem Chip, wodurch das mechanische Spotten entfällt. Proben-Sequenzen müssen nicht extrahiert, sondern können direkt aus bekannten Sequenzdatenbanken abgeleitet werden. Die normierte Herstellungsweise der Affymetrix-Microarrays trägt zu einer Standardisierung der Expressionsanalyse und der Ergebnisse bei. Ein anderes Verfahren, das mechanische Microspotting wird in vielen Laboren selbst durchgeführt. Dabei werden die präparierten DNA-Proben über Kapillarkräfte durch einen Spotting-Roboter direkt auf die Trägermatrix aufgebracht. Bei einem sehr schnellen Ink Jetting genannten Verfahren wird die DNA-Probe über eine Art Tintenstrahldrucker-Düse piezoelektrisch auf das Substrat aufgeschossen.

Verschiedene Analysen mit Microarrays

Microarrays können in verschiedenen Analyse-Ansätzen verwendet werden. Bei DNA-

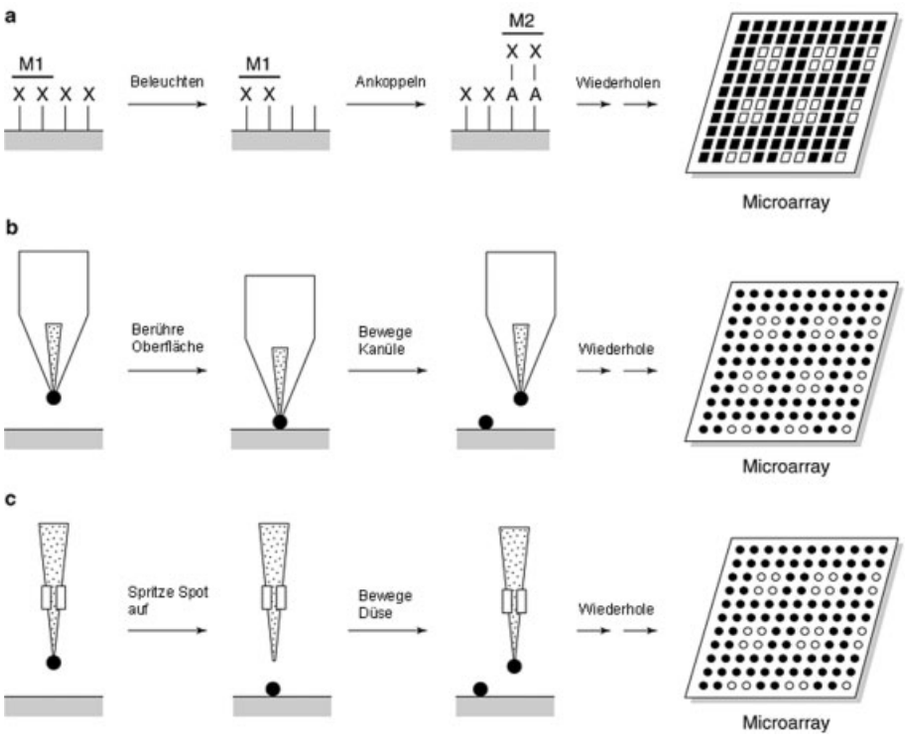

Abb. 2: Übersicht verschiedener Herstellungsmethoden für Microarrays: a Photolitografische Festphasen-Synthese, b Mechanisches Microspotting, c Ink Jetting. Beschreibung siehe Text. Bild modifiziert nach [7].

Analysen werden auf einem Microarray PCR-Amplifikate von DNA-Klonen der interessierenden Gene mit DNA Targets zweier verschiedener Genotypen oder einem Gewebe aus verschiedenen Experimenten bzw. Konditionen und der Kontrolle hybridisiert[1]. Die DNA-Targets sind aus genomischer DNA extrahierte und über Nick-Translation oder Random Primer markierte Gen-Sequenzen. DNA-Arrays benutzt man, wenn schnell sehr viele Gene parallel untersucht werden sollen (Highthroughput Genemapping), zum Auffinden von Polymorphismen wie SNPs oder Mutationen, kleineren Deletionen und Insertionen (Genotyping). So können beispielsweise Korrelationen zwischen der Reaktion eines Patienten-Phenotyps auf Arzneimittel bzw. toxischen Substanzen und dem genetischen Profil des Patienten untersucht werden (Pharmako-/Toxikogenomik). Das ermöglicht Vorhersagen von Krankheitsrisiken und an das genetische Profil des Patienten angepasste Therapiewege im Rahmen der individualisierten Pharmazeutik. Weiter benutzt man DNA-Chips zum Auffinden resistenter Viren- und Bakterienstämme und für die Abschätzung der Variation in einem Genkomplex in der Bevölkerung oder zwischen verschiedenen Völkern.

Genfunktionen sind nicht ausschließlich durch die sie kodierenden Sequenzen determiniert, sondern auch durch die räumliche und zeitliche Expressionskontrolle der die Gene unterliegen. Microrrays bieten die Möglichkeit, die Transkriptmengen aller Gene eines Organismus, das heißt das gesamte Transkriptom parallel und damit die Dynamik der Genexpression zu analysieren[2]. Für die Hochdurchsatz-Expressionsanalyse mit Microarrays werden cDNA-Proben oder aus Datenbanksequenzen abgeleitete Oligonukleotid-Proben auf dem Array immobilisiert. Diese werden mit markierten cDNA-Targets aus den zu untersuchenden Geweben hybridisiert. Zur Herstellung der cDNA-Targets wird die mRNA extrahiert, aufgereinigt und über reverse Transkriptase mit fluoreszenz-markierten Nukleotiden in markierte cDNA-Targets umgeschrieben. Expressionsanalysen erlauben Vergleiche von Transkriptmengen in gesundem und krankem Gewebe oder die Überwachung der zeitlichen Veränderung der Genexpression nach Gabe eines Stimulus wie z.B. einer Arznei. Bei letzteren, als Zeitserien-Analysen gestalteten Expressionsanalysen erhält man Expressions-Profile der Gene über die Zeit. Neuerdings werden Microarrays auch zur Massenanalyse von Proteinen eingesetzt. Die benutzten Protein-Microarrays bestehen aus mit Robotern auf Polyacrylamid-beschichtete Glas-Substrate gespotteten Proteinen, z.B. Antikörpern. Die immobilisierten Proteine behalten dabei bedingt ihre Fähigkeit, mit anderen Proteinen oder niedermolekularen Substanzen wie Arzneimitteln zu interagieren. Dadurch sind sie für die schnelle parallele Ultrahochdurchsatz-Proteinanalyse in kleinen Volumina geeignet. Mit Protein-Microarrays kann man Protein-Protein Interaktionen detektieren und so zum Beispiel neue Enzymsubstrate identifizieren oder Proteintargets für Arzneimittel finden. Die größte Schwierigkeit besteht darin, die Proteine so auf dem Array zu immobilisieren, dass sie nicht denaturieren und möglichst viele ihrer Eigenschaften beibehalten.

Automatisierung von Microarray Experimenten

Die Schlüsseltechnologien zur Effizienzsteigerung, Miniaturisierung, Parallelisierung und Automatisierung, finden vermehrt auch in der Microarray-Technologie Anwendung. Die Miniaturisierung und Parallelisierung zeigt sich in der zunehmend höheren Dichte in der die Gene parallel als Spots auf den Microarrays repräsentiert sind. Zur Automatisierung auf Hardwarebasis gehört zum einen der Einsatz von schnellen Spotting-Robotern, welche die Proben-Nukleinsäuren auf die Trägermatrix aufbringen und zum anderen das Auslesen der Datenpunkte durch den Fluoreszenz-Scanner. Weiter kann der Gesamtprozess des Hybridisierungsexperiments als Prozess-Pipeline durch Roboter automatisiert werden, was sich jedoch nur bei sehr großen Projekten lohnt. Zur Automatisierung auf Softwarebasis gehört die Regelung des Prozess- und Informationsflusses bei der Hardwaresteuerung, das gesamte Datenmanagement in modernen Datenbanktechnologien und die bioinformatische Auswertung der Microarray Ergebnisse.

Bioinformatische Auswertung von Microarray-Daten

Microarrays liefern ihrem parallelen Ultrahochdurchsatz-Anspruch gemäß extrem große Datenmengen. Allein eine Hybridisierung ergibt oft mehr als zehn Werte pro Spot. Neben der relativen Fluoreszensintensitäten sind oft viele statistische Werte zur Beurteilung der Datengüte vorhanden. Eine Versuchserie mit einem Chip von 40.000 humanen Genen kann mehrere Millionen Datenpunkte generieren. Diese Datenmengen können trotz modernster Datenbank-Management-Systeme nicht vollständig von Menschen ausgewertet werden. Hier ist die Bioinformatik als Lieferant von Automatisierungswerkzeugen besonders gefordert. Schon bei der Bildverarbeitung kommt eine automatische Auswertung zum Einsatz. Ein hybridisiertes Microarray mit Spots ver-

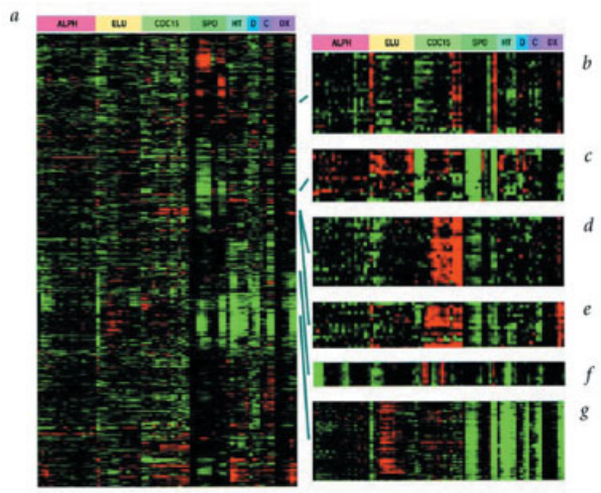

Abb. 3: a. Über Hierarchisches Clustering nach Ähnlichkeiten im zeitlichen Expressionsmuster gruppierte Hefe-Gene. Die Zeilen repräsentieren 2473 Hefe-Gene, die Spalten repräsentieren zeitliche Expressionsprofile bei acht verschiedenen Stoffwechsel-Konditionen. Rot bedeutet im Vergleich zu einem Referenz-Sample stärkere Expression, grün schwächere. b.-g. Einige in ihren zellulären Funktionen kohärente Expressionsprofil-Cluster. Bild aus [6].

schiedener Intensitäten muss nach dem sogenannten Grid-Alignment an die Spots über einen A/D-Wandler in eine numerische Tabelle mit Intensitätswerten umgewandelt werden. Hierbei muss beispielsweise eine Bestimmung der auswertbaren Größe der Spots auf dem Array unter Berücksichtigung von Fluoreszenz-Inhomogenitäten innerhalb der Spotfläche erfolgen. Durch die Bildverarbeitung werden zahlreiche Daten zur Feature/Artefakt-Differrenzierung und statistische Qualitätsparameter generiert. Dazu gehört auch die Normierung der Genexpressions-Werte etwa anhand eines Vergleiches mit stets vorhandenen Haushaltsgenen.

Ein weiteres wichtiges Gebiet ist das Datamining genannte semi- bis vollautomatische Aufdecken aussagekräftiger Strukturen in den großen Datenmengen. Hier ermöglichen bioinformatische Methoden eine Datenreduktion auf für den Anwender besonders relevante Daten. Zur Reduzierung der Dimensionalität bzw. Komplexität der Ergebnisse werden Gruppierungsverfahren eingesetzt, die Gene in überschaubare Gruppen, sogenannte Cluster mit ähnlichen Expressionsmustern aufteilen[6]. Solche Expressionsprofil-Alignments fassen co-exprimierte Gene zusammen, die ähnliche zelluläre Funktionen haben können, oder durch ähnliche Mechanismen reguliert werden. Auf Basis ihrer Expressionsprofile fallen viele Gene unbekannter Funktion mit Genen bekannter Funktion in dieselben Cluster, was Rückschlüsse über deren Funktion ermöglicht. Wichtige Clustering-Verfahren sind Hierarchical Clustering, Self Organising Maps und K-Means Clustering.

Expressions-Datenbanken

Datenbanken und Datenbank-Management Systeme ermöglichen die Verwaltung, Zentralisierung, Visualisierung und Verbreitung der zahlreichen Microarray-Daten. Die Stanford Microarray Datenbank (SMD) zum Beispiel speichert Rohdaten und normalisierte Daten aus ca. 10.000 Microarray Experimenten verschiedener Modellorganismen in online-zugänglicher Form und setzt die Daten über Hyperlinks in semantischen Kontext zu Daten aus anderen biologischen Datenbanken[4]. Das ermöglicht eine interdisziplinäre Resourcennutzung und Synergieeffekte. SMD kann nach Kriterien wie Namen des Experimentators, Experimentalparameter und Organismus abgefragt werden und macht dem Benutzer außerdem verschiedene Auswertungswerkzeuge zugänglich. Weitere online zugängliche Expressions-Datenbanken sind der vom NCBI entwickelte Gene Expression Omnibus GEO und ArrayExpress vom European Bioinformatics Institute.

Datenstandards und Ontologien für Microarray-Daten

Es gibt viele unterschiedliche Microarray-Technologien mit jeweils wiederum einer Vielzahl von Parametern. Dies führt dazu, dass die durch Microarrays gewonnenen Ergebnisse selten quantitativ vergleichbar und oft sogar bei ein und demselben Experiment schwer reproduzierbar sind. Um die Microarray-Ergebnisse miteinander vergleichen zu können, gibt es Ansätze diese Daten zu standardisieren. Standardisierte Microarray-Ergebnisse in online zugänglichen Datenbanken sind z. B. wichtig für Kreuz-Vergleiche zwischen Daten verschiedener Organismen, Experimente oder Techniken, sowie für die Erstellung von Eichstandards und Fehlerraten. Das MGED-Konsortium (Microarray Gene Expression Database) entwickelt solche Standards. Ein erster Schritt in diese Richtung ist MIAME (Minimum information about microarray experiments). MIAME gibt Richtlinien welche Informationen zur eindeutigen Beschreibung von Microarray-Daten benötigt werden und soll möglichst die Kerndatentypen, die allen Microarray-Experimenten gemein sind, erfassen[8]. Ein weiterer Standard der die Verbreitung und automatische Weiterverarbeitung vereinfachen wird, ist die XML-basierte Web-Sprache Gene Expression Markup Language (GEML), die zum Austausch von Microarray-Daten zwischen verschiedenen Menschen, Datenbanken oder Computeranwendungen benutzt werden soll.

Eine komplexere, aus der Künstlichen Intelligenz (KI) kommende Methode, die vielen semantisch und syntaktisch heterogenen Microarray-Ergebnisse über einen einheitlichen Beschreibungsstandard zu integrieren, bieten sogenannte Ontologien[5]. Das sind Wissensrepräsentationen, die zentrale konzeptionelle Elemente von Wissensgebieten und die semantischen Beziehungen zwischen ihnen definieren. Als gemeinsamer Standard für die Beschreibung und den Austausch von Microarray-Daten definieren Ontologien die wesentlichen Begriffe für Konzepte, Eigenschaften und Relationen, die der Beschreibung von Microarray-Daten dienen und ermöglichen so eine explizite, gemeinsame Sichtweise auf diese. Dadurch erleichtern sie die Kommunikation zwischen verschiedenen Arbeitsgruppen und Datenbanken und ermöglichen den direkten Vergleich dieser Daten.

Ein weiterer Schritt in der Informations-Pipeline ist die automatische Ableitung von Regeln und die Herleitung neuen Wissens aus den Microarray-Daten. Diese Aufgabe kommt normalerweise dem Anwender zu, kann jedoch neuerdings auch bedingt automatisiert werden. Vorraussetzung für diesen Schritt vom Datamining zum Knowledgemining ist jedoch eine formale, standardisierte und für den Computer „verständliche" Repräsentation der Daten als Wissen, z.B. als Ontologie. Auch arbeitet man an Agenten genannten mobilen autonomen Software-Entitäten, die selbstständig verschiedenste Datenbanken nach vom Benutzer definierten Dateneinträgen absuchen, zusammenfassen und diese in aufschlussreicher Form an den Benutzer zurückgeben. Auch diese Software-Agenten kommunizieren über Ontologien. Dieser Bereich des KI-Datamining steht jedoch noch ganz am Anfang der Entwicklung. Abschließend ist zu sagen, dass in den letzten Jahren vermehrt Techniken aus der KI Eingang in die Biowissenschaften gefunden haben.

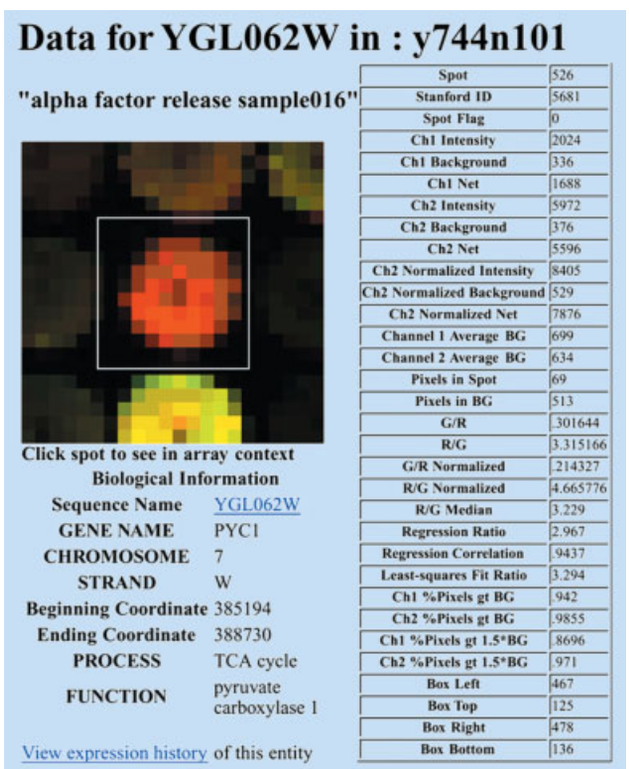

Abb. 4: Dieses Bild zeigt die visuelle Repräsentation der Expressions-Daten eines Gens in der Standford Microarray Database SMD und gibt einen Endruck von der Menge an Daten, die allein zur Evaluierung eines Spots angegeben wird. Bild aus [4].

Literatur

[1] ***Hacia J.,*** *Nature Genetics, 1999, Vol 21 (Suppl.), S 42–47,* Resequencing and mutation analysis using oligonucleotide microarrays

[2] ***DeRisi J., Iyer V., Brown P.,*** *Science, 1997, Vol 278, S 680–686,* Exploring the metabolic and genetic control of gene expression on a genomic scale

[3] ***Lipshutz R., Fodor S., Gingeras T., Lockhart D.,*** *Nature Genetics, 1999, Vol 21 (Suppl.), S 20,* High density synthetic oligonucleotide arrays

[4] ***Eisen M., Sherlock G., Cherry M.,*** *Nucleic Acid Research, 2001, Vol29, No1, S 152–155,* The Stanford Microarray Database

[5] ***Köhler J., et al..*** *in Bioinformatics and Biomedical Engineering. 2000. Arlington, Virginia, USA.* Logical and Semantic Database Integration

[6] ***Brown O., Botstein D.,*** *Nature Genetics, 1999, Vol 21 (Suppl.), S 33,* Exploring the new world of the genome with DNA microarrays

[7] ***Viraphong L.,*** *Faculty of Medicine, Khon Kaen University, Thailand,* DNA Microarrays and Expression Bioinformatics

[8] ***Brazma A., et al.,*** *Nature Genetics, 2001, Vol 29, S 365–371,* MIAME-toward standards for microarray data

Kontaktadresse

Dipl. Biol. Daniel Schober
Abt. Bioinformatik
Walter Friedrich Haus, Raum 6
Max Delbrück Zentrum für molekulare Medizin, Berlin-Buch
Robert Rössele Straße 10
D-13125 Berlin
Telefon: +49-30-9406-3378
Fax: +49-30-949-2294
E-mail: schober@mdc-berlin.de
Homepage: www.heuristic.de.vu