Qualitätsstandards für DNA Chip-Analysen

Benedikt Brors

Theoretische Bioinformatik
Deutsches Krebsforschungszentrum
Heidelberg

Zusammenfassung
Als Methode zur Messung von Genexpressionsstärken werden zunehmend DNA Chips eingesetzt. Um die so gewonnenen Daten öffentlich zugänglich zu machen, wird von mehreren Institutionen die Einrichtung von Datenbanken geplant. Die Definition von Standards und Mindestanforderungen ist dazu unerlässlich. Auf dem 2nd International Meeting on Microarray Data Standards, Annotations, Ontologies and Databases (MGED2) vom 25.-27.5.2000 in Heidelberg wurden technische Standards für den Datenaustausch und Mindestanforderungen an die Beschreibung von DNA Chip-Experimenten vorgestellt. Daneben wurde auch ausführlich über Qualitätsstandards und notwendige Kontrollen diskutiert. Die Identität der auf einem DNA Chip enthaltenen Elemente (Spots) soll sichergestellt werden. Um die Vergleichbarkeit von Daten innerhalb der Datenbank zu ermöglichen, wird angestrebt, identische Kontrollelemente in DNA Chips mit einzuschließen und Normalisierungsverfahren zu vereinheitlichen.

Schlüsselwörter
Biochip Qualität; Genexpression; Array Standards;

Summary
The use of DNA chips to measure gene expression levels is increasing. To provide public access to the data of such experiments, several institutions are planning to establish databases. For this, standards need to be defined. During the 2nd International Meeting on Microarray Data Standards, Annotations, Ontologies and Databases (MGED2), May 25-27, 2000 in Heidelberg, technical standards for data exchange and minimal description guidelines have been proposed. Furthermore, quality standards and necessary controls have been discussed. The identity of elements (spots) in an microarray should be assured. The presence of identical control elements on DNA chips and acceptance of a common normalization procedure has been encouraged in order to make experiments in those databases comparable.

Key words
biochip quality; gene expression; array standards;

In den letzten Jahren haben DNA Chips zunehmend an Bedeutung für die Untersuchung der Expressionsstärke von Genen gewonnen (Duggan et al., 1999; Brown und Botstein, 1999; Khan et al., 1999; Berns, 2000). Die Möglichkeit, Tausende von Genen gleichzeitig untersuchen und die Änderung ihrer Expression verfolgen zu können, macht sie als Methode für den Forscher interessant. Wegen des ungünstigen Signal/Rausch-Verhältnisses und der hohen Variabilität der Messwerte ist die Validierung der Ergebnisse von entscheidender, wenn auch noch nicht allgemein beachteter Bedeutung.

Einrichtung von Datenbanken für DNA-Chip-Experiment-Daten

Zur Zeit sind mehrere Projekte in Vorbereitung, die auf die Einrichtung großer, öffentlich zugänglicher Datenbanken für Daten aus DNA Chip-Experimenten zielen, wie z.B. das GeneX Projekt des US National Center for Genome Ressources (http://www.ncgr.org/research/genex/), der Gene Expression Omnibus des US National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/geo/) sowie das ArrayExpress Projekt des European Bioinformatics Institute (EMBL) (http://www.ebi.ac.uk/arrayexpress/). Einen Überblick über aktuelle Projekte findet man unter http://www.ncgr.org/research/genex/other_tools.html. Vorbild hierfür sind Sequenzdatenbanken wie GenBank, EMBL und SwissProt. Datenbanken wie die hier geplanten werden auch zunehmend von den Verlagen der internationalen wissenschaftlichen Zeitungen gefor-

Minimale Informationen über ein publiziertes Mikroarray-Experiment

Entwurf von Empfehlungen basierend auf der MGED2 Tagung
Stand:14.6. 2000. Auszug (Volltext unter: http://www.ebi.ac.uk/microarray/MGED/Annotations-wg/annotations-wg.html)

Die minimale Beschreibung eines publizierten Genexpressionsexperiments, das mit Mikroarrays durchgeführt wurde, besteht aus 6 Teilen:

1. Beschreibung aller Hybridisierungsexperimente
2. Beschreibung der benutzten Arrays und jedes einzelnen Spots in dem Array
3. Beschreibung des Untersuchungsmaterials, der mRNA Präparationsmethode und des Markierungsprotokolls
4. Beschreibung der Hybridisierungsmethode
5. Beschreibung der Messung der Expressionsstärke
6. Kontrollen

Details:

ad 1) Beschreibung der Hybridisierungsexperimente (...)

- **für multiple Hybridisierungen:**
  - geordnet/ungeordnet
  - seriell (ja/nein)
  - Typ (z.B. Zeitreihe)
  - Gruppierung (ja/nein)
  - Typ (z.B. Dose/Response)
  - Liste der Untersuchungsmaterialien und Arrays, die in dem Experiment benutzt wurden, sowie Beschreibung der Beziehungen zwischen ihnen (jeder Probe und jedem Array sollte eine unverwechselbare ID zugewiesen werden, alle Beziehungen zwischen ihnen sollten mit den entsprechenden Kommentaren aufgelistet werden)
- **Qualitätsindikatoren**
  - gibt es eine Publikation in einer Peer-Review Fachzeitschrift?
  - Anzahl der Replikate (Wiederholungen) und Beschreibung derselben

(...)

ad 2)

- **jedes Element (Spot) in dem Array**
  - Kloninformation, obligatorisch für cDNA Elemente: Klon ID, erhältlich von, Datum, Verfügbarkeit
  - Sequenzinformation, obligatorisch für nicht-cDNA Elemente: Sequenz-Zugriffsnukleotide Nr. in DDBJ, EMBL od. GenBank, od. die Sequenzinfo. selbst
  - Anzahl der Oligos und Referenzsequenz (oder Zugriffsnukleotid Nr.) für Chips vom Affymetrix-Typ, sowie die Oligosequenzen, wenn erhältlich
  - Genname und Link zu den entsprechenden Datenbanken (z.B. SwissProt, oder organismusspezifische Datenbanken), sofern bekannt und relevant
  - PCR Status
  - Qualitätskontrolle der Template-DNA (keine, resequenziert, Qualitätsprüfung durch Gelelektrophorese, Menge der DNA)
  - Position in dem Array

ad 3)

- **Quelle des Untersuchungsmaterials und Behandlung:**
  - Organismus (NCBI Taxonomie)
  - Zelltyp und Quelle desselben (wenn aus Organismus isoliert)
  - Beschreibungen:
    Geschlecht
    Alter
    Entwicklungsstadium
    Gewebetyp / Organ
    Tier-, Pflanzenstamm oder -linie
    genetische Variationen (knock-out, transgen für ...)
    Patient
    individueller Genotyp (Krankheits-Allele, Polymorphismen etc.)
    Status (krank oder normal)
    zu untersuchender Zelltyp
    Trennungstechnik (keine, trimming, Mikrodissektion, FACS, ...)
    Zellinie und Quelle derselben
    in vivo Behandlungen
    in vitro Behandlungen (Zellkulturbedingungen)
    Behandlungsart (Substanz, Hitze-, Kälteschock, ...)

(...)

ad 6) Kontrollen

- **Art der Kontrollen** (bereits markiert und der Hybridisierung zugesetzt [Kalibrierung der Scan-Intensität gegen die Konzentration der Kontrolle]; der mRNA Präparation zugesetzt [Bestimmung der Markierungsrate]; der Amplifizierungsreaktion zugesetzt [in-vitro Transkriptions- oder PCR-Protokoll])
- **ID der Kontrollen**
- **zugehörige Elemente des Arrays**, die zur Normalisierung verwendet werden sollen.

dert, da die Rohdaten zu einer einzigen Veröffentlichung mehr als 1000 GB betragen können, und sie daher die Veröffentlichung dieser Daten nicht selbst leisten können.

Einführung von einheitlichen Standards

Im Zuge der Einrichtung solcher Datenbanken wird derzeit lebhaft über die Einführung von Standards für die zu erwartenden Daten diskutiert. Zum einen geht es um technische Standards, damit Daten zwischen den verschiedenen Datenbanken und Analysesystemen ausgetauscht werden können. Hier hat man sich bereits auf eine Variante von XML (Extended Markup Language) geeinigt, und eine Definition des Datenformats kann in Kürze erwartet werden (s. http://beamish.lbl.gov/). Zum andern wird diskutiert, wie Experimente mit DNA Chips beschrieben werden können. Da der Einfluss experimenteller Parameter auf die Genexpression nicht vollständig verstanden, aber allgemein für groß gehalten wird, sollen möglichst viele dieser Einflussgrößen festgehalten werden, auch solche, die dem Experimentator vielleicht unbedeutend erscheinen. Ausserdem soll die Beschreibung des Experiments nur mit festgelegten Schlüsselwörtern möglich sein. Das soll verhindern, dass identische Objekte innerhalb der Datenbank mit unterschiedlichen Namen beschrieben sein können, ein großes Problem der aktuellen Sequenzdatenbanken.

Im Rahmen dieser Diskussion fand vom 25. bis 27. Mai 2000 in Heidelberg das 2nd International Meeting on Microarray Data Standards, Annotations, Ontologies and Databases statt. Organisiert wurde es gemeinsam vom Deutschen Krebsforschungszentrum und vom European Molecular Biology Laboratory (EMBL). Das Tagungsprogramm sowie einige der Präsentationen sind über das Internet unter http://www.ebi.ac.uk/microarray/MGED/MGED25052000/mged25052000-agenda.html zugänglich. Eine erste auf der Tagung diskutierte Empfehlung kann unter http://www.ebi.ac.uk/microarray/MGED/Annotations-wg/ annotations-wg.html eingesehen werden (siehe auch Auflistung links). Die Ta-

gung ist Teil einer Serie, die letztes Jahr im November in Hinxton, UK, begann und nächstes Jahr in Stanford fortgesetzt werden soll. An der Konferenz haben sich die meisten akademischen Arbeitsgruppen aus Europa, den USA und Japan beteiligt, die eigene Datenbanken für Mikroarray-Daten planen, sowie Vertreter der Firmen Affymetrix, Incyte, GeneLogic, NetGenics, GlaxoWellcome und SmithKline Beecham. Die Diskussion über Standards für DNA Chips wird auch über mehrere email-Diskussionsforen weitergeführt. Eine Zusammenfassung der Ergebnisse der ersten Tagung vom November 1999 liegt ebenfalls vor (Brazma et al., 2000).

Schutz vor schlechter Qualität

Ein wichtiger Diskussionspunkt auf der Tagung war die Einführung von Qualitätsstandards. Sogenannte cDNA Chips werden z.B. heute überwiegend aus kurzen Stücken exprimierter Gene hergestellt, die öffentlich verfügbar sind. Deren Qualität ist aber so schlecht, dass Resequenzierungsprojekte gestartet wurden, um sicherzustellen, dass ein Sequenzabschnitt auf einem DNA Chip tatsächlich das vorgesehene Gen repräsentiert. Weiterhin wurden Kontroll-Spots auf dem DNA Chip gefordert. Mehrere Positiv- und Negativ-Kontrollen sollten durchgeführt werden, um z.B. eine Abschätzung vornehmen zu können, welcher Anteil eines Signals auf einem DNA Chip auf Kreuzhybridisierung oder unspezifische DNA/DNA-Wechselwirkung zurückzuführen ist.

Auf der Suche nach Normalisierungsparametern

Ein weiterer wichtiger Punkt ist die Normalisierung der Daten. Um Hybridisierungen vergleichen zu können, muss sichergestellt sein, dass die durchschnittliche Signalintensität in beiden Hybridisierungen übereinstimmt. Unterschiedliche Markierungseffizienzen für die verwendeten Fluoreszenzfarbstoffe, unterschiedliche mRNA Konzentrationen und andere, kaum zu kontrollierende Parameter führen dazu, dass eine Normalisierung nachträglich rechnerisch durchgeführt werden muss. Im Moment ist jedoch kein allgemein akzeptiertes Normalisierungsverfahren verfügbar. Daher sollten stets auch die Rohdaten erhältlich sein. Wenn die Bedingung, dass die große Mehrheit der auf einem DNA Chip repräsentierten Gene keine Änderung der Expressionsstärke zeigt, nicht mehr erfüllt ist, müssen auf dem Chip zusätzliche Standards in Form von heterologer DNA vorhanden sein. Die komplementären mRNAs werden dann durch in-vitro Transkription gewonnen und in definierter Konzentration der zu untersuchenden mRNA Präparation vor der Markierungsreaktion zugesetzt. Bei Chips, die aus einem kleinen Satz ausgesuchter Gene bestehen, ist eine Normalisierung anders nicht durchführbar. Auch muss überdacht werden, gegen welche Standardbedingungen Vergleiche der Genexpressionsstärken durchgeführt werden sollen. Ergebnisse, die auf Vergleich mit nicht identischen Standardbedingungen beruhen, sind im allgemeinen nicht direkt miteinander in Beziehung zu setzen. Die Möglichkeit, für einen bestimmten Organismus (Mensch, Maus, Bäckerhefe, Arabidopsis thaliana etc.) eine von allen akzeptierte und benutzte Referenzprobe zu definieren, wird allgemein skeptisch beurteilt.

Vergleichbarkeit von Experimenten aus unterschiedlichen Herstellungsserien

Ein weiteres Problem ist die Vergleichbarkeit von Experimenten, die mit ansonsten gleichen Chips unterschiedlicher Herstellungsserien gemacht wurden. Nur bei Oligonukleotidarrays kann gegenwärtig die Menge an DNA, die in einem Spot gebunden ist, konstant gehalten werden. Für Chips, bei denen EST Klone durch PCR amplifiziert werden, bevor sie auf das Array aufgebracht werden, variieren die DNA Mengen für einen bestimmten Spot erheblich zwischen verschiedenen Herstellungsserien. Dies macht Vergleiche zwischen Serien unmöglich.

Die Reproduzierbarkeit ist nicht immer gewährleistet

Vom statistischen Standpunkt aus ist es weiterhin wünschenswert, dass DNA Chip Untersuchungen in mehreren Wiederholungen durchgeführt werden. Eine erste Möglichkeit wäre, Repräsentanten für jedes Gen doppelt auf den DNA Chip aufzubringen. Nach unserer Erfahrung reicht das jedoch nicht aus, um eine ausreichende Reproduzierbarkeit der Ergebnisse zu gewährleisten (Beißbarth et al., 2000). Die mindestens einmalige Wiederholung des Hybridisierungsexperiments einschließlich der mRNA Präparation aus dem Untersuchungsmaterial sollte eigentlich ein Mindeststandard sein.

Insgesamt bleibt festzuhalten, dass Fragen der Validierung und Qualitätssicherung von DNA Chips langsam ins Interesse der beteiligten Forscher geraten. Sicherlich wird in Zukunft bei der Publikation von Ergebnissen in Fachzeitschriften und auch bei der Aufnahme von Experimenten in Datenbanken kritischer als bisher nachgefragt werden, ob und wie die erhaltenen Ergebnisse durch Kontrollen und Qualitätsprüfungen abgesichert sind.

Literatur

Beißbarth T, Fellenberg K, Brors B, Arribas-Prat R, Boer JM, Hauser NC, Scheideler M, Hoheisel J, Schütz G, Poustka A, Vingron M (2000) Processing and quality control of DNA hybridization data. Bioinformatics, im Druck.

Berns A (2000) Gene expression in diagnosis. Nature 403:491-492

Brazma A, Robinson A, Cameron G, Ashburner M (2000) One-stop shop for microarray data. Nature 403:699-700

Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. Nat Genet Suppl (Jan 99) 21:33-37

Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM (1999) Expression profiling using cDNA microarrays. Nat Genet Suppl (Jan 99) 21:10-14

Khan J, Bittner M, Chen Y, Meltzer PS, Trent JM (1999) DNA microarray technology: the anticipated impact on the study of human disease. Biochim Biophys Acta 1423:M17-M28

Korrespondenzadresse
Dr. Benedikt Brors
Theoretische Bioinformatik, Deutsches Krebsforschungszentrum, Heidelberg
Im Neuenheimer Feld 280
69120 Heidelberg
Tel. 0049-6221-42 2718
Fax. 0049-6221-42 2849
b.brors@dkfz-heidelberg.de