Genomics und Genprofil-Analyse

Vor acht Jahren wurde mit „differential display“ die erste einfache Technik zur Analyse von Genexpresssionsprofilen entwickelt. Das neuartige Prinzip war die Kombination von randomisierten (zufällig ausgewählten) PCR-Primern, was zur Amplifikation und Detektion einer Vielzahl von unterschiedlichen Genen führte.[1)] Damals waren lediglich einige Tausend Gene bekannt. Bis heute hat sich diese Zahl exponentiell vermehrt. Parallel hierzu entwickelten sich die wissenschaftlichen Analysenmethoden rasant weiter.

Die Datenflut wurde durch den Einsatz von DNA-Sequenzierungen im großem Stil initiiert. Das Humangenomprojekt beschleunigte den Prozeß der Gendatenerzeugung zusätzlich. Es enstanden schnell wachsende EST(expressed sequence tag)-Datenbanken. Bis heute wurden die Genome von über 30 Organismen vollständig entschlüsselt, weitere 100 Organismen befinden sich noch in der Phase ihrer Sequenzierung (www.tigr.org oder www.ncbi.nlm. nih.gov). Die Sequenzierung der beiden menschlichen Chromosomen 21 und 22 wurde vor kurzem erfolgreich abgeschlossen.[2,3)] Man geht davon aus, daß sehr bald der Großteil des menschlichen Erbgutes bekannt sein wird. Schätzungen gehen von ca. 100000 humanen Genen aus, von denen jedoch lediglich etwa 5000 eine Funktion zugeordnet werden kann (weitere Informationen dazu auch im Abschnitt „Das Humangenomprojekt“, S. 321).

Dies illustriert das gegenwärtige Problem: Es sind zwar Miliarden von Basenpaaren entziffert worden, die biologische Information und Funktion ist jedoch in den meisten Fällen unklar. Einzelne isolierte DNA-Sequenzen sagen nichts über die Funktion eines Gens, die biochemischen Abläufe in einer Zelle oder die Ausbildung einer Krankheit aus. Bildlich gesprochen liegt zwar die Liste der Wörter eines Buches vor, die Sätze, Abschnitte, und Kapitel sind jedoch unbekannt. Dies ist genau der Punkt, an dem eine neue Disziplin der Forschung, „functional genomics“, ansetzt. Dabei beschränkt man sich jedoch nicht darauf, eine Liste sämtlicher Gene und ihrer Funktion zu erstellen. Das ehrgeizige Ziel der Genomicsforschung ist es, die Wechselwirkungen zwischen den Genen, den Genprodukten (Proteinen) und letztendlich die damit zusammenhängenden Mechanismen zu ergründen. Hierzu werden die Ebenen der RNA-Transkripte, der Proteine (Proteomics) sowie der Proteinwechselwirkungen untersucht.

Dieser Trendbericht soll insbesondere die Möglichkeiten der RNA-Analytik beleuchten. Das experimentelle Methodenspektrum umfaßt hier so unterschiedliche Ansätze wie auf Hybridisierung basierende DNA-Arrays[4,5)] oder die statistische Sequenzierung Sage (serial analysis of gene expression).[6)] Gleichzeitig hat in den letzten Jahren eine Renaissance von auf PCR basierenden Analysenmethoden stattgefunden, die wichtige Alternativen darstellen und in mancher Hinsicht zu favorisieren sind. Die Analyse und Interpretation von Expressionsdaten und -profilen ist ein weiterer zentraler Punkt im Genomicsfeld. Mit steigender Datenmenge wird die Bedeutung der Bioinformatk und Statistik weiter anwachsen.

DNA-Arrays

Den Grundstein für die Entwicklung von DNA-Arrays legte eine Erfindung von Ed Southerns. Es gelang ihm nachzuweisen, daß man auf Membranen fixierte Nukleinsäuren durch die Hybridisierung mit einer geeigneten Sonde detektieren kann.[7)] Dadurch wurde es zum ersten Mal möglich, Informationen über die Anwesenheit bestimmter Nukleinsäuren in einem Gewebe oder einem Zelltyp zu erhalten.

Das Interesse an DNA-Arrays hat in den letzten Jahren stark zugenommen, und die Techik selbst hat sich rasant weiterentwickelt. Fluoreszierende Moleküle zur Markierung von DNA ersetzten radioaktive Nachweismethoden. Gleichzeitig wurden Techniken eingeführt, um Nukleinsäuren auf soliden Trägermaterialien wie Glas zu fixieren, ohne dabei ihre molekularen Eigenschaften zu verändern. Das ermöglichte eine Miniaturisierung der DNA-Arrays und somit einen erhöhten Durchsatz.[8,9)] Gleichzeitig erforderte dies die Entwicklung von geeigneten Robotern, die mehrere tausend verschiedene Arten von DNA-Molekülen auf eine Glasfläche von wenigen cm^2 aufbringen können.

In den letzten Jahren haben sich zwei verschiedene Arten von DNA-Arrays etabliert: Im ersten Fall befinden sich isolierte cDNA-Klone auf dem Chip. Hierzu müssen entsprechende cDNA-Klone per Roboter zuerst auf die Glasoberfläche aufgebracht und dann fixiert werden. Im zweiten Typ von Array werden statt cDNA-Klonen kurze DNA-Oligonukleotide verwendet, die man direkt auf dem Glas-Chip synthetisiert.[10)]

Die Erfassung von Genexpressionsprofilen ist in beiden Fällen ähnlich. Fluoreszenz-markierte cDNA, die aus dem zu untersuchenden Gewebe gewonnen wurde, wird gegen die cDNA-Klone oder die Oligonukleotide auf den Arrays hybridisiert. Anschließend detektiert man die Signale, die von der gebundenen cDNA ausgehen. Dadurch erhält man zum einen die Information, welche Gene in dem Gewebe exprimiert werden. Desweiteren ist eine Quantifizierung der Expressionsstärke dieser Gene möglich. Werden im Experiment cDNAs aus unterschiedlichen Geweben oder Zelltypen verwendet, kann man direkt ableiten, wie sich die untersuchten Gewebetypen in Bezug auf ihre Genexpression unterscheiden.

Mit Hilfe von DNA-Arrays lassen sich eine große Anzahl von Genen innerhalb eines Experiments untersuchen. Insofern ist die Erfassung eines Genprofils über viele verschiedene Zustände eines Gewebes oder Zelltyps im Prinzip

möglich. Die Hauptproblematik der DNA-Array-Technik liegt jedoch darin, daß nur cDNAs analysiert werden können, deren genetische Information bereits bekannt ist. Neue, unbekannte Gene kann man mit dieser Methode nicht untersuchen. Desweiteren ist durch die zugrundeliegende Technik der Hybridisierung die Sensitivität der Methode limitiert. Moleküle, die nur in geringer Anzahl in der Zelle vorliegen, können daher nicht detektiert werden. Aufgrund dieser Einschränkungen läßt sich mittels Arrays kein vollständiges und umfassendes Genprofil erstellen. Die Stärke des Verfahrens liegt hier eher im Bereich der Diagnostik, um mit möglichst großem Durchsatz eine begrenzte Anzahl Gene gleichzeitig zu analysieren.

Sage

1995 entwickelten V. E. Velculescu, B. Vogelstein und K. W. Kinzler die „Serial analysis of gene expression" (Sage).[11] Mit Sage läßt sich theoretisch die gesamte Genexpression in einem RNA-Gemisch untersuchen. Es handelt sich dabei um einen experimentellen Ansatz, der auf der mit hohem Durchsatz durchgeführten Sequenzierung von 10 bis 20 bp langen DNA-Fragmenten (Tags) beruht. Das Prinzip der Sage-Methode ist einfach und einleuchtend: Ein Tag, das z. B. in einem erkrankten Gewebe häufiger sequenziert und damit öfter aufgefunden wurde als in einem gesunden, ist dort überrepräsentiert. Das zum Tag korrespondierende Gen ist demnach im erkrankten Zustand induziert. Im umgekehrten Fall, wenn das entsprechende Tag seltener sequenziert wird, liegt eine Repression des entsprechenden Gens im kranken Gewebe vor.

Methodisch steht am Anfang des Prozesses die Konstruktion einer spezifischen cDNA-Bibliothek. Eine typische Sage-Bibliothek enthält ungefähr $2 \cdot 10^6$ Tags (oder Transkripte). Etwa 40 unterschiedliche Tags liegen in Klonen von ca. 500 bp hintereinander und bilden Konkatemere (miteinander verknüpfte DNA-Sequenzen).[12] Verkettet man die Tags, reduziert sich der notwendige Sequenzieraufwand zur Identifizierung der Transkripte deutlich. In einem typischen Experiment werden ca. 2000 individuelle Klone sequenziert, was ungefähr 50000 Sage-Tags entspricht. Der Sequenzieraufwand beläuft sich demnach auf $5 \cdot 10^5$ bis $1 \cdot 10^6$ bp. Die durch die Sequenzierung gewonnenen Rohdaten werden dann aufwendigen Computeranalysen unterzogen. Dabei ermittelt man aufgrund der in jedem Tag enthaltenen Sequenzinformation das zugrundeliegende Gen oder EST. Die Software zählt außerdem die Häufigkeit, mit der ein spezifisches Tag aufgefunden wurde, und bestimmt damit das Expressionsniveau des entsprechenden Gens. Desweiterten kann ein Vergleich sowohl zwischen gleichen oder verschiedenen Sage-Projekten als auch mit Sage-Referenzdatenbanken durchgeführt werden.

Die Stärken von Sage liegen vor allen Dingen in ihrer einfachen Handhabung und in der Vergleichbarkeit und Reproduzierbarkeit der Daten. Man muß jedoch auch festhalten, daß folgende grundlegende Effekte zu beachten sind: Seltene, also schwach exprimierte Gene, können nicht mehr mit ausreichender statistischer Genauigkeit nachgewiesen werden, da sie in einem Pool von 50000 Tags kaum oder überhaupt nicht mehr vorhanden sind. Dies gilt insbesondere für seltene RNA-Moleküle, wie sie beispielsweise im Gehirn anzutreffen sind.[13,14] RNA-Moleküle kommen in verschiedenen Geweben in unterschiedlicher Konzentration vor und können in vier Klassen eingeteilt werden: Sehr selten ($1:1^{05}$ bis $1:1^{06}$), selten (bis 1:20000), mittel (bis 1:1000) und häufig (bis 1:25). Häufig auftretende Proteine sind z. B. Komponenten des Cytoskeletts oder der Ribosomen, die in 25 zufällig ausgesuchten Transkripten immer vertreten sind. Vorraussetzung für die Erstellung eines umfassenden Genprofils des Gehirns (was auch die seltenenen und sehr seltenen Gene umfaßt) ist eine Sensitivität von mindestens 1: 500000 (Abbildung 1A). Im Gehirn sind über 80 % der einzelnen Transkripte, (z. B. Rezeptoren für Neurotransmitter oder Ionenkanäle), selten oder sogar sehr selten exprimiert. Da aber schon die Klasse der seltenen Gene (Häufigkeit bis 1:20000) mit nur 50000 Sage-Tags aus statistischen Gründen nicht mehr quantifiziert werden kann, ist die Erstellung eines vollständigen Genprofils des Gehirns mittels Sage nicht möglich (Abbildung 1B). Desweiteren liegt der Schwellenwert für den Nachweis eines Gens in einem typischen Sage-Experiment bei einer mindestens acht- bis zehnfachen Induktion oder Repression. Schwächer differentiell regulierte Gene können aus statistischen Gründen nicht mehr identifiziert werden. Beide Limitierungen können durch eine größere Anzahl von sequenzierten Tags kompensiert werden, was sich allerdings auf der Aufwand- und Kostenseite deutlich bemerkbar macht.

Depd, Reads, Toga und Gene calling

Parallel zu DNA-Arrays und Sage wurden in den letzten Jahren verschiedene auf PCR-basierende Techniken entwickelt. Hier seien als Beispiele Reads (restriction enzyme analysis of differentially expressed sequences),[15] Toga (Total gene expression analysis),[16] Depd (digital expression pattern display)[17] und Gene calling[18] genannt.

Allen Methoden gemeinsam ist, daß zuerst eine enzymatische Spaltung der cDNA durchgeführt wird. An die entstehenenden Fragmente werden DNA-Adaptoren ligiert und so Bindungsstellen für PCR-Primer geschaffen. Mit komplementären Primern lassen sich dann bei hohen Temperaturen PCR-Reaktionen durchführen und Gene so reproduzierbar und verläßlich nachweisen. Dies stellt einen entscheidenden Unterschied zum ursprünglichen Differential Display dar, bei dem die Verwendung von randomisierten Primern in der PCR nur niedrige Temperaturen erlaubt, was letzlich zu einer hohen Fehlerrate führt.

Exemplarisch soll hier die Depd vorgestellt werden: Als Ausgangsmaterial dienen Gewebeproben oder Zellen, aus denen im ersten Schritt mRNA isoliert wird. Daraufhin wird die mRNA in cDNA umgeschrieben und mit Restriktionsenzymen gespalten. Die eigentlichen Depd-Schritte bestehen aus der Ligation spezifischer Adaptoren an die DNA-Fragmente und der Durchführung von 3072 verschiedenen PCR-Reaktionen. Man erzielt dadurch eine Auftrennung der ursprünglichen cDNA (und der in ihr enthaltenen Gene) in 3072 Fraktionen. In jeder der 3072 PCR-Reaktionen können 50 bis 100 Fragmente analysiert werden, was einer Auflösung von bis zu 300000 Banden (bei einer Redundanz von 2,5 x) entspricht. Diese hohe Auflösung ist notwendig, um beispielsweise ein umfassendes Genprofil des Gehirns zu erstellen. Depd-PCRs werden in der Regel mit Fluoreszenz-markierten Primern durchgeführt und anschließend auf 96-Kapillar-Sequenzierungsgeräten aufgetrennt (Abbildung 2). Genau dies stellt auch die eigentliche Stärke der Methode dar. Wie schon angesprochen, geht man von etwa 100000 Genen im menschlichen Genom aus. Man benötigt daher eine möglichst hohe Auflösung, um alle Gene, auch die seltenen, detektieren zu können.

Letztendlich erhält man in einem Depd-Experiment Informationen über die differentielle Regulation von Genen, ESTs und noch unbekannnten, neuen Sequenzen. Im Gegensatz zu DNA-Arrays ist man also nicht auf schon publizierte und klonierte Sequenzen beschränkt. Ein weiterer wichtiger Punkt betrifft die Sensitivität. Depd ist in der Lage, ein Molekül unter 750000 zu detektieren. Desweiteren können schon zweifache Unterschiede im Expressionsniveau nachgewiesen werden (Abbildung 3). Der Grund für die im direkten Vergleich mit den Arrays sehr viel höhere Sensitivität beruht auf der Tatsache, daß die PCR in dieser Beziehung der Hybridisierung überlegen ist.

Der Nachteil von Depd ist, daß man aufgrund der Komplexität einen sehr hohen Aufwand betreiben muß. Es ist daher keine Methode, die für den laboreigenen Gebrauch konzipiert wurde (es handelt sich hierbei um eine Technologie von Biofrontera Pharmaceuticals). Auch für Reads (Gene logic), Toga (DGT), und Gene calling (Curagene) sind keine kommerziellen Produkte zur eigenen Anwendung erhältlich. Im Unterschied zur Depd werden jedoch bei den letztgenannten Methoden deutlich weniger PCR-Reaktionen generiert, was zu einer stark verringerten Auflösung und Sensitivität führt.

Ausblick

Jede vorgestellte Methode aus der RNA-Analyse in der Genomicsforschung weist ihre charakteristischen Stärken auf. Um dies auszunutzen, bietet sich eine Kombinantion der verschiedenen Methoden an. So ist es beispielsweise denkbar, im ersten Schritt eines Projekts etwa durch die Depd-Technologie eine Anzahl von relevanten, interessanten Genen zu identifizieren. Anhand dieser Informationen kann man Arrays mit den entsprechenden Sequenzen herstellen. Diese spezifischen DNA-Arrays können dann in der Routine wie beim Screening eingesetzt werden.

Unmittelbar auf die Erstellung von Expressionsprofilen mit den zuvor beschriebenen Methoden folgt die Analyse und Interpretation von Expressionsdaten und -profilen. Dieser wichtige Prozeß steht im Fokus der Genomicsforschung. Als eine schon häufig angewendete Technik ist hier die Clusterbildung (clustering) von Expressionsdaten zu nennen. Durch Clustering werden Gruppen von Genen identifiziert, deren Regulation zeitgleich und gemeinsam erfolgt und die funktionell zusammenwirken.[19,20)]

In den nächsten Jahren wird mit der steigenden Datenmenge die Bedeutung der Bioinformatk und Statistik für den Genomicsbereich weiter anwachsen. Die Interpretation der Massen von Genexpressionsdaten stellt daher die eigentliche Herausforderung für die Zukunft dar.

Stefan Zwilling, Ralf Hoffmann, Hermann Lübbert, Biofrontera Pharmaceuticals, Leverkusen

1) P. Liang, A. B. Pardee, Science *1992,* 257, *967–971.*
2) I. Dunham et al., Nature *1999,* 402, *489–495.*
3) M. Hattori et al., Nature *2000,* 405, *311–319.*
4) G. G. Lennon, H. Lerrach, Trends Genet. *1991,* 7, *314–317.*
5) N. Zhao et al., Gene *1995,* 156, *207–213.*
6) V. E. Velculescu et al., Science *1995,* 270, *484–487.*
7) E. M. Southern et al., Nucleic Acids Res. *1994* 22, *1368–1373.*
8) S. P. Fodor et al., Science *1991,* 251, *767–773.*
9) D. Shalon, S. Smith, P. O. Brown, Genome Res. *1996,* 6, *639–645.*
10) E. M. Southern, K. Mir, M. Shchepinov, Nature Genet. *1999,* 21, *5–9.*
11) V. E. Velculescu et al., Science *1995,* 270, *484–487.*
12) L. M. Madden et al., Drug Discov. Today *2000,* 5, *415–425.*
13) B. B. Kaplan et al., Biochemistry *1978,* 17, *5516–24.*
14) N. D. Hastie, J. O. Bishop, Cell *1976,* 9, *761–774.*
15) Y. Prashar, S. M. Weissman, Methods Enzymol. *1999,* 303, *258–272.*
16) J. G. Sutcliffe et al., Proc. Natl. Acad. Sci. USA *2000,* 97, *1976–1981.*
17) R. Hoffmann, S. Zwilling, H. Luebbert, Dt. Patent/*Nr. 19806431.4-41.*
18) R. A. Shimkets et al., Nature Biotech. *1999,* 17, *798–803.*
19) M. B. Eisen et al., Proc. Natl. Acad. Sci. USA *1998,* 95, *14863–14868.*

20) X. Wen et al., Proc. Natl. Acad. Sci. USA *1998*, 95, *334–339.*

Abb. 1. **A) RNA-Moleküle kommen in verschiedenen Geweben in unterschiedlicher Konzentration vor und können in vier Klassen eingeteilt werden. B) Über 80 % der einzelnen Transkripte, (z. B. Rezeptoren für Neurotransmitter oder Ionenkanäle), sind im Gehirn selten oder sogar sehr selten exprimiert.**

Abb. 2. **Die einzelnen Schritte eines Depd (Digital expression pattern display)-Experiments: Aus Gewebeproben oder Zellen isolierte mRNA wird in cDNA umgeschrieben und enzymatisch gespalten. Es folgt die Ligation spezifischer Adaptoren an die DNA-Fragmente und die Durchführung von 3072 verschiedenen PCR-Reaktionen. Depd-PCRs werden in der Regel mit Fluoreszenz-markierten Primern durchgeführt und anschließend auf 96 Kapillar-Sequenzierungsgeräten (MegaBACE1000) aufgetrennt.**

Abb. 3. **Depd-Experiment: Ratten wurde ein pharmazeutischer Wirkstoff appliziert. Zu sehen ist ein kurzer Abschnitt von 100 bis 300 bp aus einem Genprofil des Gehirns (im Vergleich zu Kontrolltieren). Insgesamt erscheinen elf Signale, die PCR-Fragmenten von Genen entsprechen. Deckungsgleich (gleich stark exprimiert) sind zehn Gene. Ein Fragment der Größe 279,1 bp ist im behandelten Tier im Vergleich zum Kontrolltier etwa zweifach reprimiert. Der Wirkstoff unterdrückt also die Expression des entsprechenden Gens um 50 %.**