



Integration von DNA-Array Daten mittels Bioinformatik – der Schlüssel zur Analyse komplexer Expressionsmuster

Georg Casari und Claus Kremoser

LION Bioscience AG, Heidelberg

Tab 1 Beispiele für Fragestellungen, bei denen Genregulation (Genexpression) eine entscheidende Rolle spielt

Ex vivo (Zellkulturen)

Änderung der Genexpression durch Wechsel des Nährmediums oder der Kultur- bzw. Fermentationsbedingungen bei einzelligen Mikroorganismen (aerob <-> anaerob, Nährsubstrate, Einfluss des pH-Wertes)

Änderung der Genexpression in Zellkulturen nach Induktion durch Hormone oder Wachstumsfaktoren

Änderung der Genexpression in Zellkulturen nach Transformation durch DNA-Vektoren oder nach Transduktion durch Viren

In vivo (Tiermodelle bzw. Biopsieproben)

Analyse der Genexpressionsmuster von Tumoren in verschiedenen Stadien zur Identifizierung von Tumormarkern, Angiogenese- oder Metastasierungsfaktoren

Analyse der Genexpressionsmuster von Geweben nach Pharmakagabe (z.B. Induktion von Phase I / Phase II Enzymen, Pharmakaresistenz u.a.)

Zeitabhängige Analyse der Genexpressionsmuster bei Entwicklungs-, Entzündungs- oder Regenerationsvorgängen (z.B. Wundheilung)

Mediziner wie Molekularbiologen versprechen sich durch den Einsatz von hochintegrierten DNA-Chips eine Fülle von Anwendungen, die von der Aufklärung neuer funktioneller Wechselwirkungen auf zellbiologischer Ebene bis zum Routineeinsatz in der medizinischen DNA-Diagnostik reichen. Auf dem Weg zur Erschließung dieser vielfältigen Möglichkeiten sind aber noch einige Probleme zu lösen. Eine der wichtigsten Fragen besteht darin, wie die riesigen Datenmengen, die DNA-Arrays produzieren, als Antworten auf biologisch-medizinische Fragen interpretiert werden können. Die Autoren dieses Artikels sind der Überzeugung, daß dies nur durch den intelligenten Einsatz von Bioinformatik gelöst werden kann.

Die Molekularbiologen der 80er und frühen 90er Jahre untersuchten einzelne mRNAs oder Proteine, die in einem engen funktionellem Zusammenhang standen. Funktionell zusammengehörige Proteine sind z.B. die Enzyme eines bestimmten Stoffwechselweges oder einer definierten Signaltransduktionskaskade. Komplexe Zusammenhänge wie die gegenseitige Beeinflussung von verschiedenen Systemen der hormonellen Signalübermittlung in die Zelle konnten in einem einzelnen Experiment nicht ermittelt werden. Biologische Zusammenhänge wurden indirekt über die Einbeziehung von Literaturdaten oder zeitaufwendige Serien von Experimenten erschlossen.

Der Paradigmenwechsel von der seriellen zur parallelen Betrachtung der

Genregulation von einzelnen Zellen und Geweben wurde durch Hochdurchsatz- und Automatisierungstechniken in der Sequenzanalyse von ESTs („expressed sequence tags“) und cDNAs und jüngst durch den Einsatz von DNA-Arrays („Chips“) erreicht. Im Gegensatz zu klassischen Hybridisierungsverfahren wie Northern Blots bieten DNA-Arrays die Möglichkeit, den Expressionsgrad tausender Gene parallel bestimmen zu können. Durch diese Parallelisierung kann theoretisch die komplette Biologie einer Zelle oder eines Gewebes auf der Ebene des sogenannten „Transkriptom“, der Gesamtheit aller aktiven Gene, zu einem gegebenen Zeitpunkt aufgelöst werden. Komplexe Phänomene der Genregulation, wie sie im Rahmen vieler Fragestellungen auftreten, können damit studiert werden (siehe Tab 1).

Bedingt durch diese Flut an Datenpunkten und Meßergebnissen ergibt sich aber folgende Schwierigkeit: Welcher einzelne Forscher ist in der Lage, aus der komplexen Fülle an Informationen diejenigen zu extrahieren, die sein spezifisches Problem betreffen? Gesetzt den Fall, daß der DNA-Chip mit definierten Sonden (= DNA-Proben) bekannter Sequenz beladen ist: Welcher Wissenschaftler oder welche Arbeitsgruppe kann die biologische Funktion tausender von Gene gleichzeitig überblicken?

Bioinformatiker und Genexpressionsexperten der LION Bioscience AG beschäftigen sich seit Jahren mit der Aufbereitung biologischer Daten auf ge-

gebene Fragestellungen hin. LION's Bioinformatik-Plattform bioSCOUT™ ist darauf konzipiert, zu einer gegebenen DNA-Sequenz alle verfügbaren Informationen zu Struktur, Funktion, Gewebsverteilung und verwandten Genen aus einem weltweiten Netzwerk an Datenbanken zu extrahieren und für den Anwender hin zu aufzubereiten. Deswegen lag es nahe, eine Auswertungssoftware für DNA-Array Experimente zu entwickeln, die dem Benutzer eine intelligente Auswahl der interessierenden Daten und eine direkte Verbindung zu den bioinformatischen Interpretationshilfen ermöglicht.

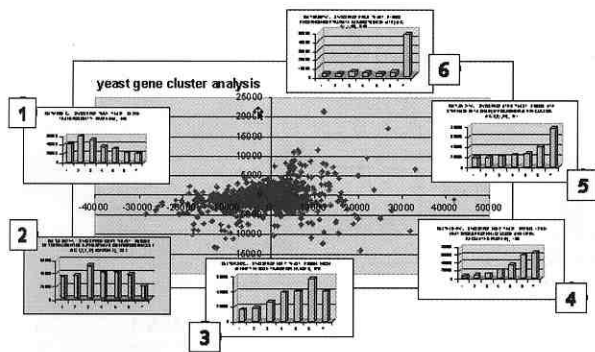
Die Funktionsweise dieses „Expression-Profiling Software Tools“ soll an einem experimentell gut untersuchten Beispiel erläutert werden, der Veränderung der Genexpression in der Bäckerhefe während der Fermentation von glucosehaltigem Medium (J.L. Derisi, V.R. Iyer, P.O. Brown; Science 278, Oct 1997).

Mit Hilfe eines DNA-Chips, der mit repräsentativen Sonden zur Abdeckung aller bekannten Hefegene (ca. 6000) beladen war, wurde das Gesamtexpressionsmuster mit sieben Meßproben von Beginn bis zum Ende der Fermentation ermittelt.

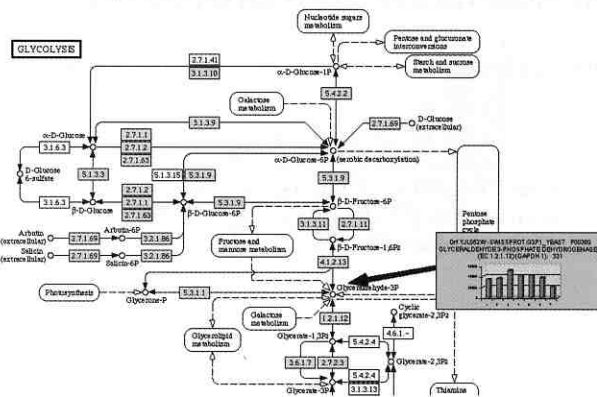
Die Auswertungssoftware geht zur Interpretation der Daten in zwei Schritten vor:

1. Identifizierung der Gene, die über die sieben Probenentnahmen während des Fermentationsverlaufs ein ähnliches Expressionsmuster auf-

Ident
 1
 2
 a) Im ch wa (4) At
 Abb 1
 w
 pe
 bl
 gi
 2. V
 va
 al
 b
 e
 s
 b
 F
 s
 fi
 F
 s
 C
 v
 Mit
 bio
 we
 wä
 gut
 Die
 ein
 pri
 Ko
 wie
 lau
 de
 Gl
 str
 ho
 ph
 frü
 wi
 hc
 St
 sy
 ne

Identifizierung der Gencluster


a) Im ersten Schritt werden die Gengruppen, die einem bestimmten zeitlichen Expressionsmuster zuzuordnen sind, identifiziert [(1) Zellwachstumskomponenten, (2) Glykolyseenzyme, (3) Zuckertransporter, (4) Streßproteine, (5) Proteine der oxidativen Phosphorylierung / Atmungskette, (6) Enzyme der Gluconeogenese]

Zuordnung zur Stoffwechselkarte


b) Im zweiten Schritt wird die Information über die identifizierten Gene / Proteine aus Datenbanken extrahiert und das Expressionsprofil des jeweiligen Gens in die dazugehörige elektronische Stoffwechselkarte eingeblendet

Abb1 Interpretation des Metabolismus von Bäckerhefe während der Fermentation von Glucose auf Basis von DNA-Chip Genexpressionsdaten

wiesen (Clusteranalyse oder Gruppen-Mustererkennung) und Einblendung der identifizierten Gengruppen;

2. Vollständige Extraktion aller relevanten Informationen, die über die als „interessant“ eingestuft Gen bekannt sind, aus Datenbanken, um eine zielgerichtete Interpretation dieser Ergebnisse zu unterstützen (Anbindung an bioSCOUT™, „Feature Report“ enthält eine zusammenfassende Charakterisierung des identifizierten Gens bzw. des codierten Proteins); Einblendung der Expressionsprofile der charakterisierten Gene in eine elektronische Stoffwechselkarte (siehe Abb 1).

Mit diesem Zwei-Stufen Modell der bioinformatikgestützten Versuchsauswertung läßt sich die Genexpression während des Fermentationsverlaufs gut erklären:

Die erste Gruppe von Genen, die bei einem Überangebot von Glucose exprimiert ist, besteht aus allgemeinen Komponenten des Zellwachstums (1), wie z.B. Ribosomen. Im weiteren Verlauf verschiebt sich der Schwerpunkt der Genexpression auf Enzyme der Glykolyse (2), um einen hohen Substratdurchsatz zu gewährleisten. Alkohol wird produziert, die oxidative Phosphorylierung ist nur begrenzt aktiv. Im frühen Stadium des Glucosemangels wird ein hochaffiner Zuckertransporter hochreguliert (3). Zusätzlich werden Stressproteine (Heat Shock Proteine) synthetisiert (4), da die Hefezelle zunehmend unter Glucosemangel leidet.

Unter massiver Glucoseverarmung zum Ende der Fermentation hin, wird der Alkohol, der zuvor fermentiert, wurde wieder dehydriert und die Reduktionsäquivalente in der Atmungskette zur ATP-Synthese eingesetzt (5). Enzyme der Gluconeogenese (6) werden hochreguliert, da Glucose als C-Quelle nicht mehr verfügbar ist.

Jeder Mediziner oder Biologe kann mit Hilfe dieses integrierten Systems aus Expressionsanalysenmodul und der bioSCOUT™-Bioinformatik-Plattform und Grundkenntnissen in Biochemie diese Interpretation vornehmen. Hinter dieser Bioinformatik steht das Prinzip, hochentwickelte Algorithmen und Datenbankabfragen mit intelligenter und zielorientierter Visualisierung so zu kombinieren, daß der Anwender sich auf die biologische Fragestellung konzentrieren kann. Dadurch können auch die in Tabelle 1 geschilderten, wesentlich komplexeren Fragestellungen angegangen werden. Nur die Integration aus robuster und leistungsfähiger DNA-Chip Hardware, sorgfältiger Entnahme und Aufbereitung biologischer Proben und Interpretation mittels intelligenter Bioinformatik-Software ermöglicht der biomedizinischen Forschung des beginnenden 21. Jahrhunderts das neue Paradigma der Parallelbetrachtung biologischer Prozesse mit Inhalt zu füllen. Das Ziel ist es, das weltweit verfügbare biologisch-medizinische Wissen optimal einzusetzen, um diese neue Ebene der Komplexität zu beherrschen und dadurch neue Ansätze für Diagnostik, Therapie und Präventivmedizin zu entwickeln.

Korrespondenzadresse

Dr. Georg Casari
 Dr. Claus Kremoser
 LION Bioscience AG
 Waldhofer Str. 98
 D-69123 Heidelberg
 Tel. 06221 / 4038-0
 Fax: 06221 / 4038-101
 contact@lion-ag.de