Computergestützte Auswertung von Protein-Interaktions-Screens

Jorge Silva[1], Rainer Stotzka[1], Nicole V. Ruiter[1] und Peter Uetz[2]

[1]Institut für Prozessdatenverarbeitung und Elektronik,
[2]Institut für Toxikologie und Genetik,
Forschungszentrum Karlsruhe, 76344 Eggenstein
Email: stotzka@ipe.fzk.de

Zusammenfassung. Das „Two-Hybrid"-System ist eine genetische Methode für die Detektion von Protein-Protein-Interaktionen. Dabei werden z. B. Matrizen mit 384 Feldern manuell auf wachsende Hefekolonien untersucht und die Ergebnisse in eine Datenbank eintragen. Die Auswertung ist subjektiv und zeitaufwändig. Die vorliegende Arbeit präsentiert ein erstes computergestütztes System, um digitale Bilder von solchen „Two-Hybrid"-Systemen automatisch auszuwerten. Das System zeigte im Test eine 97-prozentige Effektivität.

1 Protein-Protein-Interaktionen

Proteine sind die aktiven Bestandteile aller lebenden Zellen. Sie erfüllen sehr vielfältige Aufgaben, z. B. katalysieren sie als Enzyme fast alle Reaktionen eines Organismus und als Strukturproteine prägen sie dessen Gestalt von der Zelle bis zum kompletten Lebewesen. Dazu kommen regulatorische Proteine, welche die Dynamik lebender Systeme steuern und einige andere [1]. Traditionell werden Proteine einzeln untersucht. Man hat aber schon lange festgestellt, dass die meisten Proteine ihre Aufgabe in Zusammenarbeit mit anderen Proteinen erfüllen und nicht alleine. Daher ist eine Beschreibung der Protein-Interaktionen in einer Zelle für das Verständnis der Zellstruktur und der dynamischen Prozesse in der Zelle notwendig [2].

In der heutigen Forschung ist die Hefe der Modellorganismus schlechthin, weil sie der erste höhere Organismus ist, dessen Genom sequenziert und bei dem systematische Studien an allen Proteinen durchgeführt wurden. Nachdem das Hefegenom vollständig sequenziert war, kannte man zwar die darin kodierten 6000 Proteine, aber nicht deren Funktion und Anordnung in der Zelle. Aus diesem Grund wurde gleich nach Abschluss der Sequenzierung des Hefegenoms im Jahre 1996 begonnen, Protein-Interaktionen in der Hefe systematisch zu untersuchen [1].

Ein „Two-Hybrid"-System ist eine genetische Methode, Protein-Protein-Interaktionen zu detektieren. Es basiert auf einem genetischen Trick, bei dem die Zelle veranlasst wird, nur dann zu wachsen, wenn zwei bestimmte Proteine interagieren. Damit kann man eine Protein-Interaktion an der Entstehung

einer simplen Hefekolonie ablesen. Da alle Proteine in der Hefe bekannt sind, können alle Proteine systematisch paarweise auf solche Interaktionen getestet werden. Bei 6000 verschieden Proteinen der Hefe existieren 36 Millionen Kombinationsmöglichkeiten. Auf Grund der Komplexität von biologischen Prozessen ist es notwendig, die Gesamtzahl dieser Kombinationen zu untersuchen, um eine vollständige Beschreibung der Aktivitäten in diesem Organismus zu erzielen.

Zur Automatisierung der Versuche werden die Hefekolonien in Array-Form unter Verwendung eines Roboters angeordnet. In jeder Kolonie wird ein bestimmtes Paar Fusions-Proteine exprimiert. Diese Arrays ermöglichen die systematische Untersuchung aller möglichen Protein-Paare auf Interaktion. Zellen mit positiver Interaktion wachsen zu Kolonien heran, die als weiße Spots auf den Two-Hybrid-Screens zu erkennen sind. Durch das Array-Format werden die Einzelversuche reproduzierbar und vergleichbar und die Feststellung von einzelnen Fehldetektionen [2] wird vereinfacht. Zur Zeit erfolgt die Auswertung der Two-Hybrid-Screens manuell durch einen Experten, der die Ergebnisse sichtet und in eine Datenbank einträgt.

Der Nachteil solcher manuellen Auswertungen liegen in den subjektiven Interpretationen, die nicht immer zu reproduzierbaren Ergebnissen führen, dem Zeitaufwand der Experten und den damit verbundenen Kosten. Außerdem erlaubt diese Methode keine quantitative Aussage über die Größe der Kolonien. In der vorliegenden Arbeit wird ein erstes computergestütztes System für die automatische Auswertung von Two-Hybrid-Screens vorgestellt, die die Auswertezeit von mehreren Minuten auf wenige Sekunden pro Screen reduziert, besonders bei Screens mit vielen Signalen.

2 Computergestützte Auswertung

Für die automatische Auswertung der Two-Hybrid-Screens wurde eine System-Architektur [3] entwickelt, die mehrere Clients gleichzeitig bedienen kann (siehe Abb. 1). Die Datenaufnahme erfolgt durch eine handelsübliche Digitalkamera. Auf einem Client werden lokal die aufgenommenen Bilder gespeichert. Über ein Web-Interface werden die Bilder zu einem Auswerteserver gesendet. Auf dem Server wird die Bildauswertung durchgeführt. Die Ergebnisse werden in einer XML-Seite zusammengefasst und zurück zum Client geschickt. Dort überprüft ein Experte die Ergebnisse und kopiert sie in eine Datenbank.

Die Bildauswertung erfolgt auf dem Auswerteserver in fünf Schritten (siehe Abb. 1): Vorverarbeitung, Registrierung, Segmentierung, Quantifizierung und Klassifikation.

Durch die Bildaufnahme mittels einer Digitalkamera kann die Bildqualität aufgrund der Beleuchtung und Rauschen leicht variieren. Bei der **Vorverarbeitung** handelt es sich um Kontrastverbesserung und Filterung von Störungen. Zuerst wird das Farbbild, bei dem ein Weißabgleich von der Kamera automatisch durchgeführt wurde, in ein 8-bit-Grauwertbild umgewandelt. Durch eine Grauwertspreizung auf den gesamten Grauwertbereich (0 bis 255) wird der Kontrast

verstärkt. Anschließend wird durch mehrfache Anwendung eines Median-Filters mit einer 3x3-Maske das Pixelrauschen reduziert.

Das Hefewachstum an einer bestimmten Stelle soll einem der 384 Felder im Two-Hybrid-Screen zugeordnet werden. Die Lage der Hefekolonien kann auf den Bildern verschoben, rotiert und in leicht unterschiedlicher Skalierung vorliegen. Durch eine rigide **Registrierung** werden die Bilder mit einem Gitter-Template in Übereinstimmung gebracht. Die Quadrate an den Ecken des Screen's werden als Marker für die Ausrichtung verwendet. Nach der Anwendung eines geeigneten Schwellwertes auf das zu registrierende Bild wird nach den Ecken der Quadrate gesucht, die nach der Schwellwertoperation weiß erscheinen. Aus deren Position und der Position der Quadrate im Gitter-Template werden die Matrizen für die affine Transformation des Bildes berechnet und das Bild wird transformiert. Auf diese Art und Weise kann jedes Pixel des auszuwertenden Bildes eindeutig einer Gitterposition zugeordnet werden.

Für die **Segmentierung** der Spots wird das Grauwertbild mit einer Schwellwertoperation in ein Binärbild umgewandelt. Der Schwellwert wird so gewählt, dass weiße Punkte innerhalb der Gitter den Spots mit Hefewachstum entsprechen.

In den Arbeitsschritten **Quantifizierung** und **Klassifikation** wird die Größe der Spots gemessen und einer Klasse zugeordnet. Mit der Anzahl der weißen Pixel in einem Gitterfeld wird das Hefewachstum in eine der Klassen „none“ (kein Wachstum), „weak“ (schwaches Wachstum) und „strong“ (starkes Wachstum) eingeordnet. Die Ergebnisse werden in XML formatiert und zum Client gesendet.

Die Server-Software wurde als Web-Service in JAVA unter Verwendung von „ITK“ [4] und „KHOROS“ [5] auf einem Linux-Rechner implementiert und getestet.

3 Ergebnisse

Der Erfolg bei der Erkennung von Kolonien mit positiver Interaktion wurde mittels 58 Bilder aus sechs verschiedenen Proben evaluiert. Die sechs Proben wurden a-priori von einem Experten ausgewertet, wobei jedes Feld im Screen in eine der drei Klassen („none“, „weak“ oder „strong“) eingeordnet wurde. Diese qualitativen Aussagen wurden quantifiziert, um diese mit den quantitativen Ergebnissen der Mustererkennung, die angeben, wie viel Prozent von den Pixeln in einem Feld vom weißen Spot belegt sind, vergleichen zu können. Es wurde folgende Einteilung verwendet:

„none“ Weniger als 10 %,
„weak“ zwischen 10% und 24 % und
„strong“ größer als 24 %.

Von jeder Probe wurden bei unterschiedlicher Position, Skalierung, Rotation und Beleuchtung jeweils zehn Bilder aufgenommen, um mögliche Variationen bei der Bildaufnahme abzudecken. Dabei wurden auch extreme Situationen geschaffen, die in der Routine eigentlich nicht auftreten sollten.

Die computergestützte Auswertung wurde mit der menschlichen Auswertung verglichen. In 97 Prozent der Fälle bei 22272 untersuchten Feldern stimmen die Ergebnisse der Software mit der Auswertung des Experten überein.

4 Diskussion

Das vorgestellte System ermöglicht die automatische Erkennung von weißen Spots in den „Two-Hybrid-Screens“ und deren Lokalisation. Auf diese Weise können die Protein-Interaktionen in einem Screen innerhalb von Sekunden festgestellt werden. Dass die Auswertesoftware eine Effektivität von 97 Prozent besitzt, wird von unterschiedlichen Faktoren verursacht: Zum ersten existiert ein Interpolationsfehler bei der geometrischen Transformation des Bildes, die während der Registrierung stattfindet. Auf Grund des quadratischen Rasters des diskreten Bild ergeben sich Fehler in der Position von 0,5 Pixel. In der realen Welt übersetzt sich das in einen Abstand von 0,17 mm, etwa 3,33 Prozent der Feldgröße. Das heißt, dass 3,33 Prozent der Pixel werden einem falschen Feld in Gitter zugewiesen. Das Problem kann in Zukunft durch eine höhere Auflösung verringert werden, ist aber mit mehr Rechenzeit verbunden. Zum zweiten sind Spots vorhanden, die mehrere Felder überlappen. Bei der einfachen eingesetzten Segmentierungsmethode werden nicht alle Pixel eines solchen Spot's als zusammenhängend identifiziert. Manche davon werden sogar einem Nachbarfeld zugewiesen. Folglich werden die Ergebnisse für zwei Positionen gleichzeitig verfälscht. An dieser Stelle kann eine andere Segmentierung verwendet werden.

Zusätzlich sind andere wichtige Funktionen implementiert worden. Das Eintragen der Daten in eine Datenbank wird durch das XML-Format erleichtert, um die Experimente einfach dokumentieren zu können. Die Softwarearchitektur durch ihren modularen Aufbau erlaubt eine einfache Handhabbarkeit und Erweiterbarkeit des Systems. Infolgedessen kann die Bildauswertung sukzessiv an neue Anforderungen angepasst werden. Seit Januar 2004 ist eine erste Version im Betrieb und wird von den Biologen getestet.

Literaturverzeichnis

1. Uetz P: Protein-Protein-Interaktionen im Modell Hefe. BIOforum 25(1-2):2–4, 2002.
2. Uetz P: Two-Hybrid Arrays. Curr Opin Chem Biol 6:57–62, 2001.
3. Stotzka R, Silva J, Uetz P, et al.: Computer-aided analysis of protein-protein-interactions In German Conference of Bioinformatics GCB'03:14–15, 2003.
4. Ibanez L, Schroeder W: The ITK software guide. Kitware, 2003.
5. Argiro D, Farrar K, Kubica S: Cantata In Visualization, Imaging and Image Processing Conference Proceedings, 2001.

Abb. 1. Software-Struktur. Die Datenaufnahme erfolgt durch eine Digitalkamera. Im Originalbild ist ein Two-Hybrid-Screen zu sehen. Die weißen Spots sind Hefekolonien mit interagierenden Proteinen (und einige potentiell falsch-positive Spots). Im Hintergrund sieht man das Template mit einem Raster und den entsprechenden Koordinaten (Spalten mit Zahlen und Zeilen mit Buchstaben durchnummeriert) zur Lokalisierung der Spots. In den Ecken des Bildes sind vier Marker-Quadrate für die Registrierung sichtbar. Über ein Web-Interface kommt das Bild zum Server. Dort wird der Kontrast verbessert und das Pixelrauschen reduziert. Mit Hilfe der vier Marker im Referenzbild wird das Bild so registriert, dass alle Gitter-Koordinaten bekannt sind. Nach der Segmentierung entsteht ein Binärbild, auf dem die Spots weiß erscheinen und sich vom Hintergrund abheben. Die weißen Pixel in jeder Gitter-Zelle werden quantifiziert und in eine Klasse eingeordnet. Auf diese Weise entsteht eine Tabelle, in der jede Position im Gitter klassifiziert wird. Anschließend wird das Ergebnis in XML formatiert und zurück zum Client gesendet. Ein Experte kontrolliert das Ergebnis und überträgt es in eine Datenbank.