Kapitel 5

Bioinformatik und Mustererkennung

Die Mustererkennungsgruppe des ZPR hat sich im Laufe der Jahre mit den Themenbereichen Qualitätssicherung durch automatische Bildverarbeitung und maschinelle Beleglesung beschäftigt. Aktiv wird über die Anwendbarkeit von Methoden der statistischen Mustererkennung in der medizinischen Informatik geforscht.

5.1 Automatisierte Belegverarbeitung

Über einen Zeitraum von fünf Jahren hat die Mustererkennungsgruppe des ZPR auf dem Gebiet der maschinellen Beleglesung die im folgenden beschriebenen Problemstellungen wissenschaftlich bearbeitet und bis hin zur industriellen Serienreife entwickelt. Die erstellten Produkte werden inzwischen von kommerziellen Trägern gepflegt.

Schrifterkennung

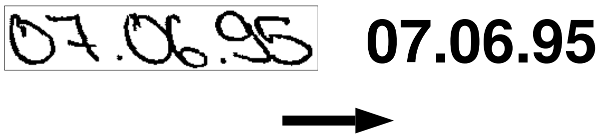

Abbildung 5.1: Schriftzug und Erkennungsergebnis

Der erste Schritt unserer Aktivitäten war die Entwicklung von „COGNITUS“, einer Erkennungssoftware, die auf leistungsfähigen *hierarchischen* Nächste-Nachbar-Klassifikationsverfahren basiert. Diese erlauben es, eine große Vielfalt von Schriftarten abzudecken und sehr niedrige Fehlerraten zu erzielen. In einfachen Fällen wird dabei eine Klassifikation durch Vergleich des zu lesenden Zeichens mit einigen wenigen Prototypen vorgenommen, während nur in Zweifelsfällen aufwendigere Operationen eingesetzt werden, um ein höchstes Maß an Sicherheit zu garantieren. Auf diese Weise können etwa 350 Ziffern pro Sekunde bei einer Fehlerrate von nur 0.5% erkannt werden. In enger Zusammenarbeit zwischen der Segmentierungs-, Klassifizierungs- und Kontextkorrektureinheit wird anschließend die global wahrscheinlichste Interpretation ermittelt, wodurch eine sehr hohe Zuverlässigkeit erreicht wird. Das Programm wurde zunächst zur Erfassung hand- und maschinengeschriebener Überweisungsformulare entwickelt, ist jedoch leicht vom Benutzer auf andere Anwendungen anpaßbar. Durch eine effiziente Implementierung wird eine hohe Verarbeitungsleistung von etwa fünf Formularen pro Sekunde auf einem Pentium Prozessor erreicht. Für noch höhere Leistungsanforderungen ist der Einsatz von Parallelrechnern problemlos möglich. Die Entwicklung von „COGNITUS“ ist weitgehend abgeschlossen.

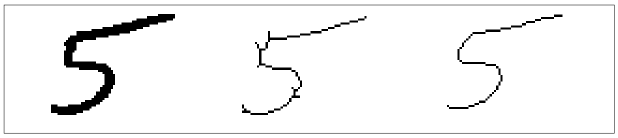

Abbildung 5.2: Ergebnis des Skelettierungsverfahrens

Im Rahmen dieser Entwicklung ist an unserem Lehrstuhl im Berichtszeitraum eine Promotionsarbeit angefertigt worden, die sich auf das Teilproblem der dynamischen Segmentierung konzentriert. Hier geht es um die Aufteilung des Dokumentes in einzelne Zeichen. Sie stellt ein grundlegendes Problem bei der Handschriftenerkennung dar. Reicht es bei Maschinenschriften noch aus, das Bild in Zusammenhangskomponenten aufzulösen und lediglich in einigen Fällen sich berührende Zeichen durch eine senkrechte Linie zu trennen, so ist die Situation bei Handschriften im Normalfall viel komplizierter. Es ist nötig, zuerst die morphologische Grundstruktur der Zeichen zu erfassen. Dazu wurde eine neues Skelettierungsverfahren entwickelt, das gegenüber Rauschen sehr robust ist und an die Auflösung der Zeichen angepasst werden kann. Skelettierungsverfahren reduzieren die Zeichen zunächst auf Linien von einem Pixel Breite. Danach zerlegt man die Zeichenfolgen in ihre topologischen Bestandteile (d. h. die einzelnen Linienstücke). Der darauf folgende Schritt des Schrifterkennungsverfahrens fügt nun diese elementaren Bestandteile zu Gruppen zusammen, die jeweils einem Buchstaben entsprechen. In vielen Fällen müssen dabei mehrere Möglichkeiten der Gruppierung generiert und anschließend mit dem Buchstabenerkenner ausgewertet werden, um die beste Zerlegung zu ermitteln. Dies stellt sicher, daß auch Zeichen, die sich überlappen, noch richtig erkannt werden, und damit ein optimales Erkennungsergebnis erreicht wird.

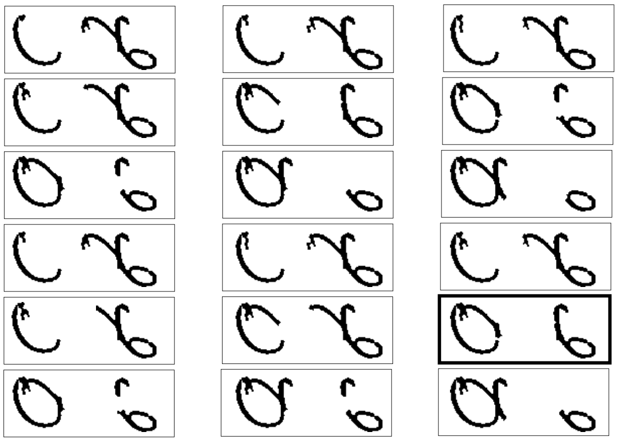

Abbildung 5.3: Rekombinationsmöglichkeiten des Monatsfeldes aus Bild (5.1)

Datenbank-Abgleich

Nachdem ein Schriftstück, z. B. ein Bankformular, durch den Schrifterkenner ausgewertet wurde, liegt meist immer noch keine eindeutige Interpretation der Zeichen vor. Es hat sich als unerläßlich heraus gestellt, die Ergebnisse von maschinellen Schriftenerkennungssystemen durch Einsatz von großen Erfahrungsdatenbanken auf Kontextzusammenhänge zu prüfen und gegebenenfalls zu korrigieren. Der nächste Schritt unserer Arbeit war daher die Entwicklung der assoziativen Datenbank „DACCORD", die diese Aufgabe erfüllt. Um einen möglichst hohen Datendurchsatz zu erzielen, werden die Daten komprimiert im Hauptspeicher abgelegt, wobei ein Kompromiß zwischen Kompressionsfaktor und schnellem Datenzugriff gefunden werden mußte. Der eigentliche Abgleich zwischen dem Ergebnis der Schrifterkennung und den in der Datenbank vorhanden Einträgen (z. B. den Adressen der Bankkunden) wird über die sogenannte Levenshteindistanz gesteuert. Diese ermittelt einen Wert für die „Ähnlichkeit" zweier Zeichenfolgen. Die Levenshteindistanz wird wortweise berechnet und wurde außerdem dahingehend modifiziert, daß für einzelne Zeichen gewisse Wahrscheinlichkeiten berücksichtigt werden können. Ein großer Satz von Parametern erlaubt es, das Verhalten der Datenbank auf die Wünsche der einzelnen Kreditinstitute individuell anzupassen. Auf unserem Teststapel konnte die Fehlerrate pro Feld von 15.3 % auf unter 0.1 % gesenkt werden wobei die Zurückweisungsrate allerdings auf 23.3 % ansteigt. Die Durchsatzrate von „DACCORD" selbst bei Datenbankgrößen von mehreren 100 MByte liegt bei über 50 Formularen pro Sekunde.

Lokale Kontextanalyse und Betragsfeld-Behandlung

Neben der reinen Datenbankfunktion stellt „DACCORD" auch eine „lokale Kontext-Korrektur" zu Verfügung, die auf sehr einfachen Regeln beruht. Beispielsweise kann in einem Wort jede „0" kommentarlos durch ein „O" ersetzt werden, da eine Erkennungsmaschine diese beiden Zeichen in der Praxis sowieso nicht unterscheiden kann. Bei gewissen Schriftarten und ganz besonders bei Handschriften kommen noch weitere Verwechslungsmöglichkeiten hinzu. Eine „5" in einem Wort ist höchstwahrscheinlich ein „S", eine „8" ein „B" etc. Diese Standardersetzungen lassen sich von außen kontrollieren, so daß eine Anpassung an die speziellen Stärken und Schwächen des Erkennungssystemes möglich ist.

Die Kontextdatenbank „DACCORD" wird derzeit insbesondere im Zusammenspiel mit einer Schrifterkennungssoftware eingesetzt, die von der Firma Parsytec GmbH entwickelt und vertrieben wird. Im täglichen Einsatz als unerwartet schwierig hat sich dabei die Behandlung des Betragsfeldes auf Bankformularen erwiesen. Die Lösung dieses Problems hat die neueste Erweiterung von „DACCORD" eingeleitet. Um zunächst einen Eindruck zu vermitteln, mit welchen Schwierigkeiten man zu kämpfen hat, folgen nun ein paar Beispiele aus der Praxis, die wir aus diesem Schrifterkennungssystem erhalten haben:

`3,701`

```
62<,1>6<86>
-10 ,90-
3 6 0,--
100,1
500-----
69.47.-
150.001111111
2.74 ,04 -
36<.@><-@>3
**4*1<0?><??><??><??><??><*?>*50,00****
```

Die in eckigen Winkelklammern eingeschlossenen Zeichen sind Alternativen, zwischen denen sich der Erkenner nicht definitiv entscheiden konnte. Eine Eigenheit des Erkennungssystemes, das die obigen Ergebnisse geliefert hat, ist es offenbar, aus Querstrichen Einsen zu machen. So lautete der tatsächliche Betrag auf dem Beleg im ersten Fall „3,70-“ und der Eintrag von „150.001111111“ war in Wirklichkeit „150.00——-“. Die sonstigen Schwierigkeiten lassen sich aus den gegebenen Beispielen nur erahnen. Das „Monstrum“ am Schluß ist übrigens aus „***********50,00****“ entstanden und der vorletzte Betrag war ursprünglich „36.–“!

Um den tatsächlichen Betrag zu rekonstruieren, wurde das Aussehen von in der Praxis vorkommenden Beträgen als regulärer Ausdruck formuliert, um dann eine „Levenshteindistanz“ zwischen diesem Ausdruck und dem Erkennungsergebnis zu berechnen. Mit Hilfe eines Backtracking-Verfahrens wird dann ermittelt, welche Zeichen aus dem Leseergebnis zu entfernen beziehungsweise zu modifizieren sind. Auch hier stehen wieder sehr viele Einstellungsparameter zu Verfügung, um die Eigenheiten der eingesetzten Erkennungssoftware optimal zu berücksichtigen. Die Bearbeitungsgeschwindigkeit liegt bei mehr als 1000 Beträgen pro Sekunde auf einem Pentium-166 Prozessor und ist damit für das Gesamtsystem nicht zeitkritisch.

Das Programm „DACCORD“ ist sehr schnell von den Banken akzeptiert worden. Mittlerweile ist es in über 40 Kreditinstituten im Einsatz und läuft absolut stabil.

5.2 Bioinformatik

Die Fortschritte in der experimentellen Technik haben den Biowissenschaften ein völlig neues Gesicht gegeben. Statt der Beschreibung phänomenologischer Beobachtungen ist nun, durch die Sequenzierung und Entschlüsselung des Genoms, die systematische Aufklärung molekularer Mechanismen komplexer Lebensvorgänge möglich. Die Menge an DNA- und Proteinsequenzen, die sich in öffentlich zugänglichen Datenbanken befindet, wächst exponentiell; die Sequenzierung findet heute schon nicht mehr in Laboren sondern in „Sequencing Factories“ statt.

Die Bioinformatik stellt hierbei ein neues, eigenständiges, interdisziplinäres Arbeitsfeld dar, in dem Methoden der Mathematik, Informatik und Statistik benutzt werden. Das Ziel ist es einerseits, im engen Dialog mit Biologen unter Einbeziehung von Wissen aus der Biologie, Biochemie, Chemie und Physik die experimentelle Technik voranzutreiben, und so die Kosten z.B. des Human Genome Projects zu reduzieren. Andererseits muß die Beantwortung der relevanten biologischen Fragen durch neue Ansätze in der Datenanalyse und Modellierung möglich gemacht werden.

5.2.1 Klassifizierung von NMR-Spektren

Die Kernmagnetresonanzspektroskopie (Nuclear Magnetic Resonance, NMR) ist ein sowohl quantitativ als auch qualitativ einsetzbares Verfahren, um Lösungen auf die in ihnen enthaltenen chemischen Bestandteile zu untersuchen. Durch einen im Vergleich mit konkurrierenden Labormethoden hohen Durchsatz, bietet sich der Einsatz der NMR-Spektroskopie für das routinemäßige Untersuchen kompletter Patientenkollektive an.

Wir haben ein Verfahren der statistischen Mustererkennung erfolgreich eingesetzt, um automatisch seltene Stoffwechselkrankheiten bei Neugeborenen zu erkennen. Hierbei werden Proben von Babyurin mit einem NMR-Spektroskop untersucht, und die so erhaltenen Spektren statistisch in Bezug auf die Zugehörigkeit zum Normalkollektiv untersucht. Das Normalkollektiv ist durch einen annotierten Trainingsdatensatz aus ca. 200 Spektren gegeben. Eine Unterscheidung zwischen verschiedenen Krankheiten ist aufgrund des geringen Umfangs des Datenmaterials noch nicht möglich. Unser Ansatz kann aber leicht dahingehend erweitert werden.

5.2.2 Effizientes Screenen von Klon-Bibliotheken

In einer wissenschaftlichen Zusammenarbeit mit der Arbeitsgruppe „Theoretical Biology and Biophysics“ (T-10) am Los Alamos National Laboratory (LANL) in Los Alamos, New Mexico, USA wurden, u.a. im Rahmen mehrerer

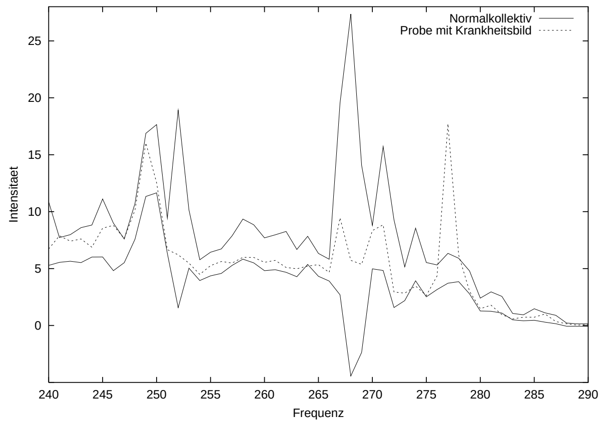

Abbildung 5.4: *Das Normalkollektiv ist durch den Bereich zwischen den druchgezogenen Linien gekennzeichnet. Das Spektrum der Patientenprobe weicht von diesem signifikant ab.*

Forschungsaufenthalte eines ZPR-Mitarbeiters in Los Alamos, Verfahren entwickelt, um die experimentellen Kosten des Erstellens von Physikalischen Kartierungen, einem wichtigen Schritt im Rahmen des Human Genome Projects, zu minimieren.

Gruppentests - Fortschritt dank Syphilis

Zur medizinischen Tauglichkeistuntersuchung für amerikanische Rekruten gehörte während des zweiten Weltkriegs unter anderem ein Test auf Syphilis. Das Blut jedes einzelnen Rekruten wurde im Labor untersucht, um gezielt die Syphilis-Erreger nachzuweisen. Da die Krankheit nur sehr selten auftrat, gab es zumeist negative Testergebnisse, – die hohen Ausgaben für den Einzeltest waren also nicht gerechtfertigt. Gruppentesten – oder Pooling – beruhte auf der simplen Idee, daß man das Blut vieler Rekruten „zusammenschütten“ (poolen) konnte, um dann den Pool auf Syphilis zu testen. Der Test lieferte ein positives Ergebnis, wenn zumindest einer der Rekruten, deren Blut im Pool war, an Syphilis erkrankt war. Bei geeigneter Poolgröße war auch bei den Pools meist mit einem negativen Ergebnis zu rechnen. Weitere Einzeltests waren damit nur bei Rekruten nötig, deren Blut in einem Pool mit positivem Befund enthalten war. Diese Vorgehensweise ergab massive Einsparungen.

Allgemein lassen sich Gruppentests immer dann erfolgreich einsetzen, wenn eine große Anzahl gleichartiger Objekte demselben Test unterzogen wird. Dabei muß der Test sensitiv genug sein, dann ein positives Ergebnis zu liefern, wenn zumindest ein Objekt der getesteten Gruppe positiv ist. Bei Gruppentests unterscheidet man kombinatorische und statistische Tests. Bei kombinatorischen Tests geht man davon aus, daß das Testergebnis immer korrekt ist, d.h. ein negatives Testergebnis besagt, daß alle Objekte negativ sind, und ein positives Testergebnis besagt, daß zumindest ein Objekt positiv ist. Statistische Tests erlauben Fehler und sind damit beim Einsatz in der experimentellen Praxis meist nicht vermeidbar. Es gibt hierbei Wahrscheinlichkeiten für das Auftreten falsch negativer Pooltests, also einem negativen Testergebnis trotz positiver Objekte im Pool, und ebenso Wahrscheinlichkeiten für falsch positiver Testergebnisse. Häufig hängen diese Wahrscheinlichkeiten von der Anzahl der positiven (negativen) Objekte, bei denen man ein falsch negatives (positives) Testergebnis erwarten kann, ab — je mehr positive Objekte existieren desto unwahrscheinlicher ist ein falsch negativer Pooltest.

Weiterhin unterscheidet man zwischen adaptiven und nicht-adaptiven Gruppentesten: Während beim nicht-adaptiven Gruppentesten die Zuordnung der Objekte zu Pools a priori festgelegt wird, wird beim adaptiven Vorgehen, die Zusammenstellung der Gruppen von den Ergebnissen der vorhergehenden Tests abhängig gemacht.

Kombinatorische, adaptive Gruppentests sind – wenn auch nicht unter diesem Namen – beliebte Knobelaufgaben: Gegeben seien 16 Münzen und eine Balkenwaage. Eine der Münzen ist gefälscht, was sich durch ein geringeres Gewicht bemerkbar macht. Wie häufig muß man wiegen, um die gefälschte Münze zu erkennen. Wie man hierbei sehen kann, hängt die Schwierigkeit eine gute Poolingstrategie zu finden stark davon ab, ob man die Anzahl der positiven Objekte kennt, und wie groß diese Anzahl ist.

Klon-Bibliotheken und Pooling

Ziel des Humangenomprojektes ist die Entschlüsselung des menschlichen Erbgutes. Ein wichtiger Zwischenschritt ist dabei die Erstellung physikalischer Kartierungen. Um die sehr langen chromosomalen DNA-Moleküle analysieren und sequenzieren – also die Abfolge der Basen A, C, G, T bestimmen – zu können, ist es nötig, eine ausreichende Menge einheitlichen Materials zu gewinnen. Dazu wird die DNA in kleinere Fragmente zerlegt. Chromosomen oder komplette Genome werden mit Restriktionsenzymen in mehrere zehn- oder hunderttausende, einander überlappende Bruchstücke unterteilt. Die Sammlungen solcher Bruchstücke werden als Bibliotheken bezeichnet.

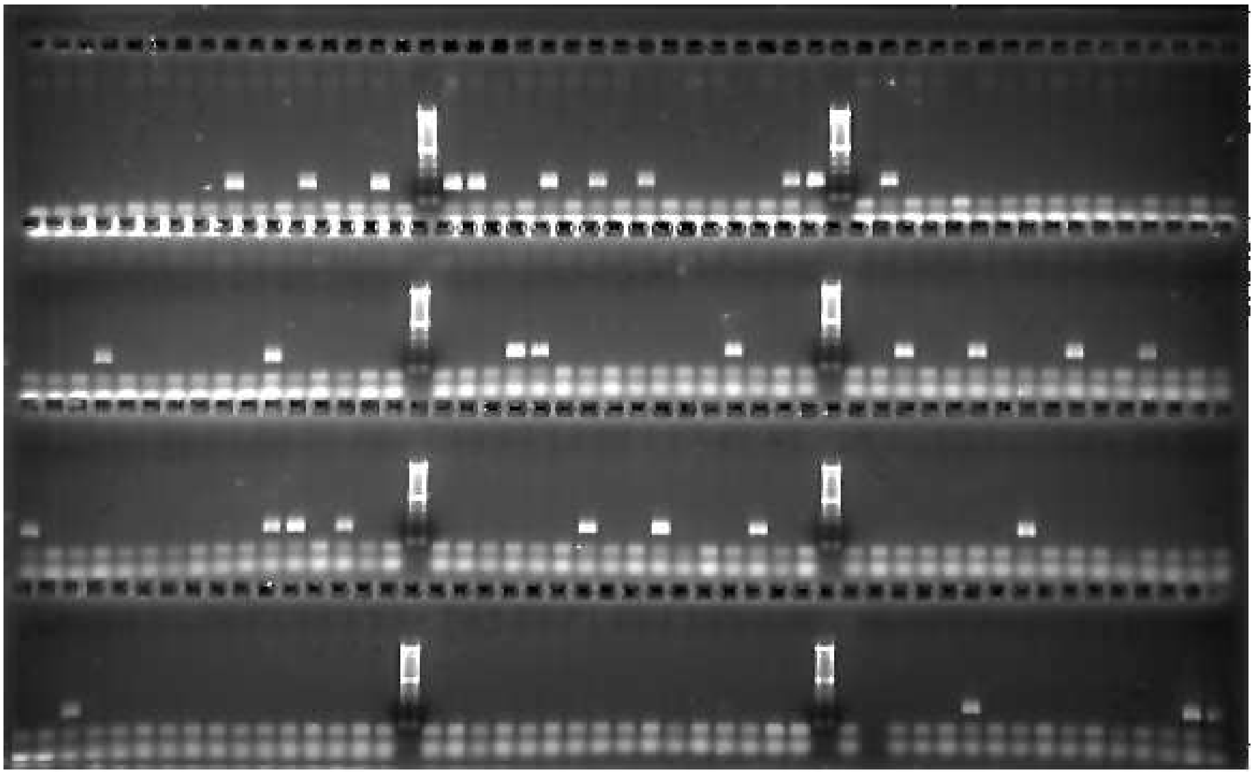

Abbildung 5.5: Das Ergebnis einer PCR-Reaktion wird durch Gel-Elektrophorese sichtbar gemacht. Die hellen Stellen sind genomische DNA.

Einzelne Fragmente werden jeweils z.B. in künstliche Hefechromosomen (YACs, yeast artificial chromosomes) eingebaut, damit man sie vermehren (klonen) kann. Das Zerteilen mit Restriktionsenzymen stellt sich bei unbekannter Sequenz als nicht deterministisch dar: Die Lage der Schnittstellen ist unbekannt und es ist nicht sicher, daß an allen möglichen Schnittstellen geschnitten wird. Daher muß, um sicherzustellen, daß die gesamte DNA in der Bibliothek enthalten ist, jede Position in mehreren Klonen enthalten sein. Dieses Überdeckungstiefe nennt man Coverage: typisch sind Werte zwischen 3 und 24.

Nach dem Zerschneiden hat man keine Information über die Lage der in einem Klon enthaltenen DNA in Bezug auf das Genom. Diese muß anhand der Überlappungen der Klones rekonstruiert werden. Hierzu werden sogenannte Sequence-Tagged-Sites (STS) Marker benutzt. Alle Klone, die den selben STS-Marker enthalten, müssen sich überlappen. Zur Überprüfung muß für jede STS und für jeden Klon eine geeignete Polymerase-Chain-Reaction (PCR) vorgenommen werden. Bei der großen Anzahl an Klonen und zu testenden STSs bieten sich hier Gruppentests an, um den experimentellen Aufwand in den Griff zu bekommen.

Um für alle STSs nur einmalig die Pools erstellen zu müssen und damit experimentelle Fehler zu minimieren, haben wir uns für eine nicht-adaptive Poolingstrategie entschieden. Die größte bisher gepoolte Klon-Bibliothek hat 442.368 Klone bei 94 Pools. Hierbei befindet sich jeder Klon in 8 Pools, und jeder der Pools enthält ca. 4.600 Klone.

Statistische Modellierung und Dekodierung

Im Laboratorium wird für jede STS jeweils für alle Pools eine PCR vorgenommen, deren Ergebnisse – enthält einer der Klones im Pool die STS – als „negativ“, „schwach positiv“ oder „stark positiv“ klassifiziert werden. Mit Hilfe eines Bayes’schen Ansatzes können wir ausgehend von den Ergebnissen aller Pools die Wahrscheinlichkeit ausrechnen, mit der ein bestimmter Klon positiv ist.

Der Ansatz führt allerdings zu einer Summierung über exponentiell viele Terme, was praktisch nicht möglich ist. Um diese Schwierigkeit zu umgehen, benutzen wir eine Markov-Chain-Monte-Carlo-Methode (MCMC). Hierbei bedeutet Monte-Carlo-Methode, daß man sich zufällige Stichproben erzeugt, deren Stichprobenmittelwert z.B. genau die Größe ist, an der man interessiert ist. Bei geeigneten qualitativen Voraussetzungen gibt es Aussagen über die Konvergenz dieses Verfahrens. Bei dem vorliegenden Problem ist allerdings schon das Erzeugen der zufälligen Stichproben äußerst schwierig. Markov-Ketten liefern in aufeinanderfolgenden Schritten zwar keine zufälligen Stichproben, aber es läßt sich unter bestimmten Bedingungen (z.B. mit Hilfe von Gibbs-Sampling) eine Kette erzeugen, deren Samples für das Mitteln geeignet sind. Dabei garantieren theoretische Resultate die Konvergenz der Kette und damit die Anwendbarkeit des gesamten Verfahrens.

Unser Problem erlaubt eine Berechnung durch die MCMC-Methode, die nicht nur die notwendigen qualitativen Anforderungen erfüllt sondern mit einer Laufzeit von knapp 30 Minuten (auf einer CPU einer Sun Enterprise 2000) für das Design mit 442.368 Klones auch ausreichend schnell ist.

Ausblick

Unsere Poolingstrategie und die Dekodersoftware laufen seit Anfang 1995 im Produktionsbetrieb im Human Genome Center des Los Alamos National Laboratory. Der Ansatz hat sich im täglichen Einsatz als robust, statistisch zuverlässig und ressourcenschonend erwiesen. Die Performance liegt selbst bei Fehlerraten, jeweils für falsch positive und falsch negative Fehler, von zehn Prozent oberhalb von 85 Prozent.

Bei dem ersten implementierten Poolingdesign war das Verfahren sensitiv genug, Korrelationen zwischen Klones anzuzeigen, die auf den Mikrotiterplatten benachbart waren. Dies wurde durch weitere Experimente als ein „Überschwappen“ auf den Mikrotiterplatten der Klonbibliothek zurückgeführt, ein Fehler, der ohne das Dekodierungsverfahren vermutlich nie gefunden worden wäre.

Dies ist eine gemeinsame Arbeit mit David C. Torney (Group T-10, LANL). Weiterhin beteiligt sind Manny Knill, Norman Dogget, Jon. A. Longmire, Judy Tesmer, David Bruce, Bill Bruno (LANL).

5.2.3 Analyse von Proteinsequenzen für die Strukturvorhersage von Proteinen

Proteine stehen im Mittelpunkt jedes biologischen Prozesses. Sie katalysieren als Enzyme einen komplexen Ablauf

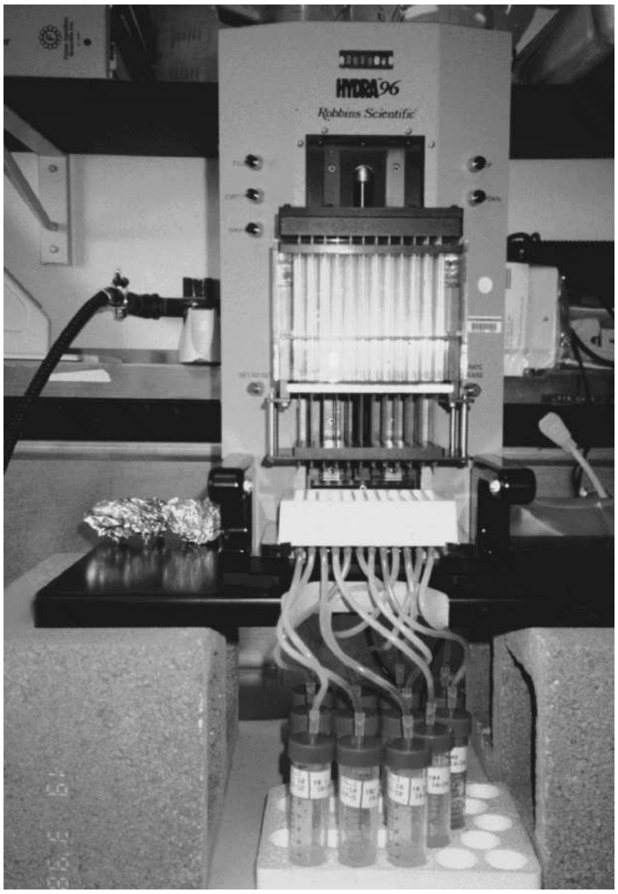

Abbildung 5.6: Mit der Pipetiermaschine *HYDRA* wurden große Teile der Pools erstellt.

biochemischer Reaktionen, die in ihrer Gesamtheit „ das Leben“ ausmachen. Um die molekularen Mechanismen der enzymkatalysierten Reaktionen zu verstehen und z.B. in der Medikamentenentwicklung (drug design) in sie eingreifen zu können, ist es notwendig die 3D-Struktur der beteiligten Proteine zu kennen.

Die experimentelle Bestimmung der Struktur ist allerdings wesentlich zeitaufwendiger und teurer als die Bestimmung von Proteinsequenzen: Dies wird sichtbar durch die exponentiell wachsende Menge an bekannten Sequenzen im Vergleich zu der wesentlich langsamer wachsenden Anzahl an aufgeklärten Strukturen. Daher ist die Vorhersage der räumlichen Struktur ausgehend von einer Sequenz eines der zentralen Probleme der Bioinformatik.

Die Strukturvorhersage funktioniert derzeit nur zufriedenstellend, wenn man die Struktur eines homologen Proteins, d.h. eines Proteins mit gleicher Abstammung, kennt. Dann kann man davon ausgehen, daß eine ähnliche 3D-Struktur vorliegt und die bekannte Struktur kann als Vorlage (Template) benutzt werden, um mittels Homology Modelling ein Modell für die 3D-Struktur des anderen Proteins zu erstellen.

Im Rahmen einer gemeinsam von der Arbeitsgruppe Schomburg, Institut für Biochemie, und der Arbeitsgruppe Bachem/Schrader, ZAIK, betreuten Diplomarbeit wird daher versucht, das Auffinden entfernt homologer Proteine zu verbessern. Das Prinzip des Ansatzes beruht auf der akzeptierten Annahme, daß zwei Sequenzen mit ausreichend hoher Sequenzähnlichkeit eine ähnliche Struktur besitzen. Für die Qualität der Sequenz–Struktur–Zuordnung ist die Sequenzidentität ein entscheidendes Maß. Aber auch bei geringer Sequenzähnlichkeit können zwei Proteine homolog sein. Eine Eigenschaft der Homologie ist, daß sie transitiv ist: wenn A und B sowie B und C sich aus dem gleichen Vorfahren ableiten, muß A auch einen Vorfahren mit C gemeinsam haben. Die Transitivität der Homologie wird in diesem Projekt genutzt, um auch entfernt homologe Proteine mit niedriger Sequenzidentität zu finden.

Hierzu werden auf einem großen Datensatz an Proteinsequenzen qualitativ hochwertige Alignments berechnet. Die Ergebnisse werden als Eingabe für ein Graphen–basiertes Cluster–Verfahren benutzt. Ziel der Arbeit ist es zu zeigen, ob sich mittels der resultierenden Cluster tatsächlich mehr entfernt homologe Sequenzen auffinden lassen als bisher. Dies wird durch den Vergleich mit bestehenden Strukturdatensätzen untersucht.