Kapitel 3

Bioinformatik

Die moderne Molekularbiologie, ihre Umsetzung in der Forschung und Entwicklung, sowie die Bioinformatik als Motor des Fortschritts finden inzwischen selbst im Feuilleton von Tageszeitungen regelmäßig ihren Niederschlag. Die Biotechnologie ist inzwischen einer der wichtigen wirtschaftlichen Wachstumsfaktoren, sowie Inhalt zahlreicher Studiengänge an deutschen Universitäten.

Bemerkenswert bei dieser Entwicklung ist die Tatsache, dass z.B. die Erfolge von Celera Genomics beim viel publizierten Wettstreit um die Erstsequenzierung des menschlichen Genoms letztendlich Erfolge bioinformatischer Natur sind. Der wesentliche Schritt war, neben einem Fortschritt bei sogenannten Sequenzierungsrobotern, die für das Projekt einen finanziell überschaubaren Rahmen gewährleisteten, die Entwicklung eines Assemblieralgorithmus. Dieser „Assemblierer" muss die mit dem für die Sequenzierung benutzten Shotgun-Verfahren erzeugten kleinen Teilstücke der genomischen Sequenzen trotz hoher Fehlerraten in der richtigen Reihenfolge zur Gesamtsequenz zusammensetzen.

Anders als oft in den Medien dargestellt ist mit der Erstsequenzierung des menschlichen Genoms noch lange nicht das Ende der biotechnologischen Entwicklung erreicht. Die Sequenz alleine bedeutet noch keinen entscheidenden Erkenntnisgewinn. Vielmehr ist das Finden von Genen, die Bestimmung der Struktur und Funktion von Proteinen, das Erkennen und Verstehen von metabolischen Netzwerken nötig, also die systematische Aufklärung molekularer Mechanismen komplexer Lebensvorgänge.

Die Bioinformatik, als ein neues, eigenständiges, interdisziplinäres Arbeitsfeld, in dem Methoden der Mathematik, Informatik und Statistik benutzt werden, wird dabei auch weiterhin ein zentraler Baustein sein.

Am ZAIK wird seit 1996 an Fragestellungen der Bioinformatik geforscht. Inzwischen bestehen Kooperationen mit dem Los Alamos National Laboratory, den Instituten für Biochemie (AG Prof. Schomburg) und Genetik (AG Prof. Tautz) an der Universtität zu Köln, dem Max-Planck-Institut für Züchtungsforschung, sowie dem Bioinformatik-Startup Science Factory, ebenfalls in Köln.

In Lehre und Ausbildung wird durch Seminare und Vorlesungen zur Bioinformatik, durch enge, regelmäßige Kontakte auf Mitarbeiterebene und die von Mitarbeitern der Arbeitsgruppe ins Leben gerufene Bioinformatik-Interessen-Gruppe, einem Zusammenschluss von Studenten, Firmenangestellten und Wissenschaftlern, ein multidisziplinäres Umfeld für weitere Erfolge in der Bioinformatik aktiv gestaltet.

3.1 Ein neues Verfahren zum Lernen von Hidden-Markov-Modellen

Hidden-Markov-Modelle (HMM) sind eine Klasse von stochastischen Modellen, die schon für eine Vielzahl von Anwendungsgebieten, von der Spracherkennung bis zur Simulation biochemischer Prozesse in Zellmembranen, erfolgreich eingesetzt wurden. Am ZAIK werden z.B. HMM auch für die Analyse und Simulation ökonomischer Zeitreihen, genauer von Bausparverträgen, benutzt.

HMM bestehen aus einer Markov-Kette, also einer „gedächtnislosen" Folge von Zufallsvariablen, den sogenannten Zuständen, die aber nicht direkt zu beobachten sind. Weiterhin gibt es für jeden dieser Zustände eine diskrete oder kontinuierliche Verteilung, anhand derer dann die beobachtbaren Ausgaben erzeugt werden. Ein HMM impliziert eine Wahrscheinlichkeitsverteilung auf der Menge aller möglichen Sequenzen von Ausgaben.

Für DNA-Sequenzen, also Abfolgen der Buchstaben A, C, G und T, kann in einem einfachen Beispiel ein Zustand des HMM jeweils einer funktional ausgezeichneten Region des DNA-Moleküls entsprechen (coding vs. non-coding). Diese unsichtbaren Zustände zeichnen sich

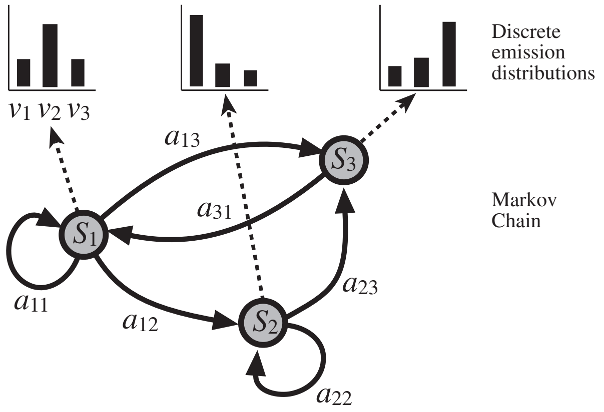

Abbildung 3.1: *Ein HMM mit drei Zuständen als gerichteter Graph mit diskreten Ausgabeverteilungen. Die Kantengewichte sind als Übergangswahrscheinlichkeiten zwischen Zuständen zu interpretieren.*

z.B. durch unterschiedliche Sequenzkompositionen, also Unterschiede in der beobachteten relativen Häufigkeit der einzelnen Buchstaben A, C, G und T zwischen den verschiedenen Regionen aus. Ein mit entsprechendem Datenmaterial trainiertes HMM, kann dann z.B. für die Erkennung von kodierenden Regionen benutzt werden.

Weitere Anwendungsfelder von HMM in der Bioinformatik sind das Finden von Genen, Strukturvorhersagen, Verfahren zur Motivsuche, Finden entfernt homologer Proteinsequenzen, Multiple Alignments, Klassifizierung von DNA/Proteinsequenzen (Transkriptionseinheiten) und Modellierung von Proteinfamilien.

In der Literatur zu HMM, insbesonders in dem bekannten Tutorial-Artikel von Rabiner, werden drei Probleme in Bezug auf HMM erwähnt:

1. Wie kann man die Wahrscheinlichkeit ausrechnen, dass ein bestimmtes Modell eine gegebene Ausgabensequenz ausgerechnet hat?

2. Gegeben ein Modell und eine Ausgabensequenz, wie kann man die „dazugehörende“ Sequenz an „versteckten“ Zuständen rekonstruieren?

3. Wie kann man die Parameter eines HMM so trainieren, dass die Wahrscheinlichkeit, eine bestimmte Ausgabensequenz zu erzeugen, maximiert wird?

Da die von Rabiner betrachteten Anwendungen in der Spracherkennung auf natürliche Art und Weise ein geeignetes HMM nahelegten, ist es verständlich, dass eine viel grundlegendere Frage nicht gestellt wurde. Wie wählt man die Struktur oder Topologie eines HMM, also die Anzahl der Zustände und der zwischen ihnen erlaubten bzw. verbotenen Übergänge?

Die Frage ist aus zweierlei Gründen von Bedeutung. Zum einem ist die Anzahl der zu trainierenden Parameter, vereinfacht gesprochen, quadratisch in der Anzahl der Zustände. Bei einer fest vorgegebenen, beschränkten Datenmenge, wie es in der Praxis fast immer der Fall ist, ergeben sich damit Trainingsfehler für die Parameter, die mit der Modellgröße zunehmen. Des weiteren läuft man mit einer zu großen Anzahl an Parametern Gefahr, das Modell „überzutrainieren“, also Generalisierungsfähigkeit zu verlieren. Zum anderen geht die Anzahl der Zustände auch quadratisch in die Laufzeit der Algorithmen ein, die z.B. für die Datenbanksuche nach verwandten Proteinen benutzt werden.

Sofern das Anwendungsproblem nicht klare Vorgaben lieferte, wurden bisher weitgehend adhoc Verfahren, z.B. die sogenannte Model-Surgery, zur Auswahl einer geeigneten Topologie verwendet. Im Rahmen einer Promotion wurde am ZAIK ein Bayes'scher Ansatz entwickelt, der nach Vorgabe einer Prior-Distribution, die als erwartete Fehlerverteilung der Daten zu interpretieren ist, auf der Basis eines Clusterverfahrens Topologie und Modellparameter lernt.

Mit diesem Verfahren können, ohne Bias für bestehende Problemstellungen, Modelle mit einer geringeren Anzahl an Zuständen und einer größeren Generalisierungsfähigkeit gelernt werden, die darüberhinaus für die typischen Bioinformatik-Anwendungen einen klaren Laufzeitvorteil bieten.

3.2 Analyse von Proteinsequenzen für die Strukturvorhersage von Proteinen

Proteine stehen im Mittelpunkt jedes biologischen Prozesses. Sie katalysieren als Enzyme einen komplexen Ablauf biochemischer Reaktionen, die in ihrer Gesamtheit „das Leben“ ausmachen. Um die molekularen Mechanismen der enzymkatalysierten Reaktionen zu verstehen und beispielsweise in der Medikamentenentwicklung (drug design) in sie eingreifen zu können, ist es notwendig, die 3D-Struktur der beteiligten Proteine zu kennen.

Die experimentelle Bestimmung dieser Struktur ist jedoch wesentlich zeitaufwendiger und teurer als die Bestimmung

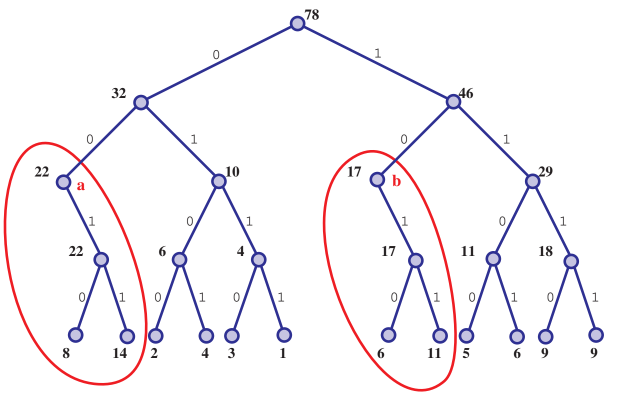

Abbildung 3.2: *Sequenzen werden in einen sog. Prefix-Baum eingefügt. Kanten sind mit Ausgabesymbolen gekennzeichnet. Die Knotengewichte entsprechen den Häufigkeiten von Prefixen. Z.B. bedeutet das Knotengewicht 22 des Knotens* a*, das der Prefix* `00` *22-mal in den Eingabesequenzen vorkam. In diesem Baum werden die Knoten geclustert, wobei die Cluster dann Zuständen des HMM entsprechen. Die für das Clusterverfahren benutzten Abstände ergeben sich durch Betrachtung der implizit durch die Knotengewichte gegebenen relativen Häufigkeiten (vgl. Wurzel und Blätter der beiden ausgezeichneten Teilbäume).*

von Proteinsequenzen: Verdeutlicht wird dies 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 Protein-Sequenz, eines der zentralen Probleme der Bioinformatik.

Die Strukturvorhersage funktioniert derzeit allenfalls 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 Kooperation der Arbeitsgruppe Schomburg, Institut für Biochemie, und der Arbeitsgruppe Faigle/Schrader, ZAIK, wird daher versucht, das Auffinden entfernt homologer Proteine zu verbessern.

Das Prinzip des Ansatzes beruht auf der allgemein akzeptierten Annahme, dass zwei Sequenzen mit ausreichend hoher Sequenzähnlichkeit auch eine ähnliche Struktur besitzen. Für die Qualität der Sequenz–Struktur–Zuordnung ist die Sequenzidentität ein entscheidendes Maß. Aber selbst bei geringer Sequenzähnlichkeit können zwei Proteine homolog sein. Eine Eigenschaft der Homologie ist ihre Transitivität: 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.

Da Heuristiken zum Sequenzvergleich, wie z.B. Blast und FASTA, bei geringer Sequenzähnlichkeit große Fehler produzieren, haben wir uns entschieden, den rechenaufwendigen Algorithmus nach Smith-Waterman zu verwenden. Für die 86.654 Proteinsequenzen in SwissProt, Release 39, ergibt sich damit ein Gesamtrechenaufwand von über 1.000 CPU-Tagen (UltraSparc CPU). Da das Problem jedoch perfekt zu verteilen ist, konnten wir sämtliche Workstations (ca. 25 CPUs) der Arbeitsgruppe als ein Rechencluster benutzen, und so über einen Zeitrahmen von 18 Wochen die Berechnung durchführen.

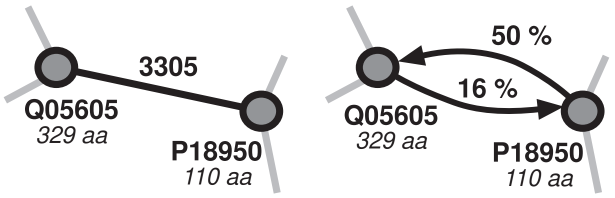

Abbildung 3.3: *Vor dem Clustern wird eine ungerichtete Kante, die mit dem Smith-Watermann Alignment-Score gewichtet ist (links) ersetzt durch zwei gerichtete Kanten (rechts). Durch eine Skalierung mit der Sequenzlänge werden unterschiedliche Prozent-Ähnlichkeitswerte auf diesen beiden Kanten erreicht.*

Die Ergebnisse der Sequenzvergleiche werden als Eingabe für ein graphenbasiertes Cluster–Verfahren benutzt. Ziel ist dabei herauszufinden, ob sich mittels der resultierenden Cluster tatsächlich mehr entfernt homologe Sequenzen auffinden lassen als bisher. Dies wird durch den Vergleich mit bestehenden Strukturdatensätzen untersucht. Dabei stellen sog. Multidomänenproteine ein Problem dar, dem wir durch Übergang zu einem gerichteten Graphen begegnen konnten.

Die erste Phase dieses Projekts wurde im Rahmen einer gemeinsam von der Arbeitsgruppe Schomburg und Arbeitsgruppe Faigle/Schrader betreuten Diplomarbeit (Eva Bolten, „Eine graphenbasierte Clustermethode zur Detektion entfernt homologer Proteinsequenzen“) durchgeführt. Der Ansatz wird durch die Verwendung der gefundenen Cluster als Trainingsmenge für Profile-basierte Verfahren zum Finden entfernt homologer Sequenzen im Rahmen

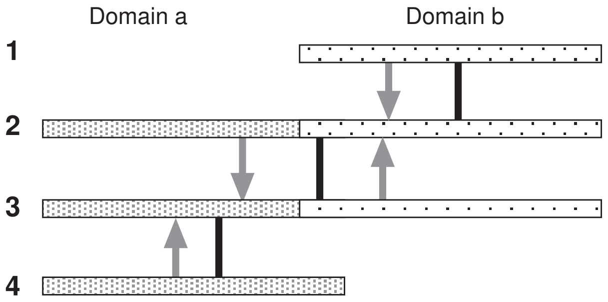

Abbildung 3.4: *Das durch Multidomänenproteine entstehende Problem: Auf einem ungerichteten Graphen (schwarze Kanten) sind Proteine #1 und #4 inkorrekterweise durch einen Pfad verbunden. Durch den Übergang zu einem gerichteten Graphen (graue Kanten) und einer von der Länge der Proteinsequenzen abhängige Skalierung der Smith-Waterman-Alignmentscores werden solche Kanten vermieden.*

einer zweiten gemeinsam betreuten Diplomarbeit ausgebaut. Weiterhin soll der Zugriff auf die Cluster über eine entsprechende Weboberfläche ermöglicht werden.

3.3 Auswahl spezifischer Proben für DNA Arrays

Sowohl Medizin als auch Biologie benötigen effiziente diagnostische Verfahren, um das genetische Erbgut zu analysieren. Die Verfügbarkeit kompletter genomischer Sequenzen ermöglicht es bereits heute, interessante Fragen auf chromosomaler Ebene zu stellen und zu beantworten. Traditionelle Verfahren der Genanalyse wie etwa die Polymerase-Kettenreaktion (PCR), die verschiedenen Blotting-Techniken sowie automatisierte Sequenzierungsverfahren sind jedoch eher sequentieller Natur und für Versuche auf chromosomaler Ebene – bei ca. 3 Milliarden Basenpaare im menschlichen Genom – oder zur schnellen Verarbeitung vieler Proben zu ineffizient und teuer.

Mittels sogenannter DNA-Arrays bzw. DNA-Chips ist es mittlerweile möglich, durch Parallelisierung der Experimente erhebliche Kosten- und Geschwindigkeitsvorteile zu erzielen. Dazu werden einige tausende kurze DNA-Stücke auf vorbestimmten Positionen, sogenannten Spots, eines Glasträgers immobilisert. Die zu analysierenden (wesentlich längeren) DNA- oder RNA-Sequenzen werden fluoreszent markiert und mit dem Chip zur Reaktion gebracht. Komplementäre Stränge in der zu analysierenden Menge lagern sich dabei auf dem Chip an. Nachdem überschüssige Polymere durch Waschen des Chips entfernt wurden, kann anhand der Floureszenz der einzelnen Spots auf dem Chip bestimmt werden, welche Reaktionen stattgefunden haben. Damit ist es möglich, Informationen über die zu diagnostizierenden Zielsequenzen zu gewinnen.

Das Probe Design Problem

Ein großes Problem beim Design von DNA-Chips ist die Auswahl geeigneter Teilstücke für den Chip. Durch die parallelen Reaktionen auf dem Chip ist es erforderlich, die auf dem Glasträger zu immobilisierenden Stücke so auszuwählen, dass anschließend auch tatsächlich Rückschlüsse auf die Zielsequenzen möglich sind. Dazu sind im wesentlichen drei Bedingungen zu erfüllen:

1. *Spezifität:* Die Proben auf dem Chip sind so zu wählen, dass sie nur mit dem gewünschten komplementären Strang in der Menge der Zielsequenzen binden. Dazu ist für jede Zielsequenz ein kurzes Teilstück auszuwählen, das diese eindeutig charakterisiert, d.h. das in keiner anderen Sequenz vorkommt. Das Problem wird weiter dadurch verkompliziert, dass auch nicht perfekt komplementäre Stränge stabile Bindungen eingehen können. Entsprechende Teilstücke sind keine guten Kandidaten für Proben.

2. *Sensitivität:* Oftmals soll nun mit einem Spot auf dem Chip nicht nur eine einzelne Sequenz erkannt werden, sondern eine ganze Familie verwandter Sequenzen. In diesem Fall muss die Chip-Probe mit jeder Sequenz dieser Familie binden (*muss sensitiv sein*), darf aber mit keiner anderen Sequenz reagieren (*soll spezifisch sein*).

3. *Gleiche Hybridisierungsbedingungen:* Auf der Chipoberfläche müssen einige hundert bis tausend solcher Reaktionen gleichzeitig ablaufen, wobei alle diese Reaktionen unter gleichen Bedingungen, d.h. unter gleichem Salzgehalt in der Reaktionslösung, bei gleichem pH-Wert und vor allem bei gleicher Temperatur stattfinden müssen.

Alle diese Probleme hängen voneinander ab, wodurch ein kompliziertes kombinatorisches Problem entsteht. Im Rahmen einer Diplomarbeit (Lars Kaderali, „Selecting Target Specific Probes for DNA-Arrays“) wurde am ZAIK ein Algorithmus entwickelt, der es dem biologischen Anwender erlaubt, geeignete Kandidaten für das Chipdesign auszuwählen.

Thermodynamisches Modell

Die Stabilität einer Bindung zwischen zwei DNA– bzw. RNA–Stücken kann im Hinblick auf das Chip-Experiment am besten durch die Temperatur ausgedrückt werden, bei der sich die beiden Teilstücke trennen. Sie kann (mittels eines Nearest–Neighbor–Modells) approximativ berechnet werden, unter der Annahme dass

$$T_M = \frac{\Delta H}{\Delta S + R \log C_T}$$

wobei T_M die gesuchte Temperatur, ΔH die Enthalpie-Änderung, ΔS die Entropie-Änderung bei Bindungsreaktion und C_T die Konzentration der DNA bzw. RNA bezeichnet. Damit lässt sich das Probenauswahl-Problem (zunächst für einzelne Sequenzen, unter Vernachlässigung der Familien-Sensitivitäts-Problematik) wie folgt formalisieren: Gegeben n Sequenzen $s_1, s_2, \ldots, s_n$, bestimme eine Temperatur T und n Proben $p_1, p_2, \ldots, p_n$ so dass

$$T_M(s_i, p_i) > T > T_M(s_i, p_k) \quad \text{für alle } k \neq i$$

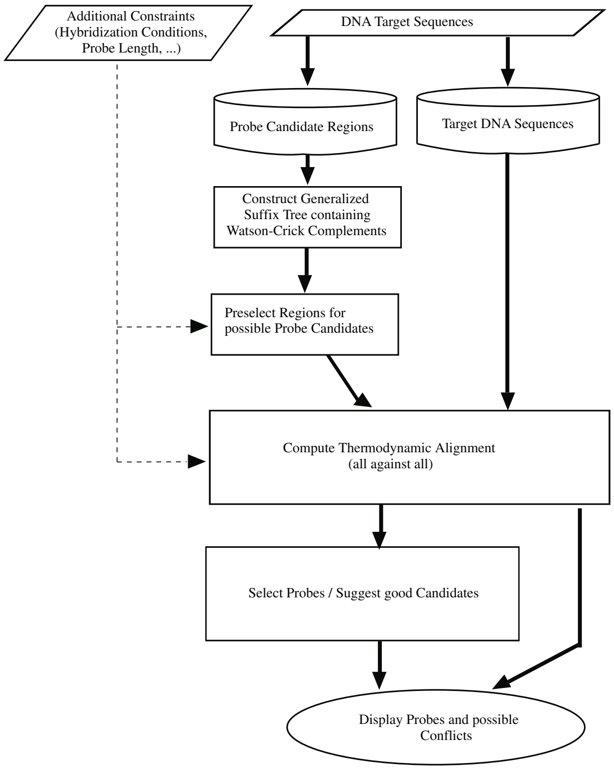

Abbildung 3.5: *Schematischer Ablauf des Verfahrens zur Probenauswahl.*

Thermodynamisches Alignment

Das Problem erfordert die Bestimmung der Schmelztemperatur jeder Zielsequenz mit jedem invers-komplementären Teilstück (als Probe). Dabei ist denkbar, dass die beiden nicht linear hybridisieren, sondern einzelne Basen ausgelassen werden und Schleifen oder andere Sekundärstrukturformen bilden. Die Berechnung kann mittels dynamischer Programmierung erfolgen, ähnlich wie in dem bekannten Algorithmus von Needleman-Wunsch bzw. Smith-Waterman, wobei jedoch anstatt der dort üblicherweise verwandten Kostenfunktion hier thermodynamische Parameter verwandt werden. Der Rechenaufwand ist allerdings enorm.

Durch eine Filterung ungeeigneter Teilsequenzen vor Beginn der eigentlichen Rechnungen kann bereits im Vorfeld eine dramatische Reduktion des Rechenaufwandes erreicht werden. Filterkriterien sind z.B. die Eindeutigkeit des gewählten Teilstückes, aber auch die Schmelztemperatur des zugehörigen Duplexes sowie die Länge der Probe. Erst im Anschluss daran beginnt die Berechnung aller möglichen Reaktionen. Abbildung **??** zeigt schematisch den Ablauf dieses Prozesses.

Genomische Sequenzen enthalten viele repetitive Motive. Dadurch kommt es relativ häufig vor, dass Proben gemeinsame Präfixe haben. Da alle Proben gegen die gleichen Zielsequenzen aligniert werden müssen, kann viel Zeit gespart werden, wenn dann solche Zwischenergebnisse gespeichert werden und bei ihrem erneuten Auftreten nicht neu berechnet werden müssen.

Eine geeignete Datenstruktur zum Erkennen solch gemeinsamer Präfixe sind Suffixbäume. Ein Suffixbaum repräsentiert sehr kompakt alle Teilsequenzen eines Strings. Dabei sind die Kanten des Baumes mit Teilsequenzen beschriftet, und der Weg von der Wurzel zum i-ten Blatt des Baumes liefert genau den i-ten Suffix der ursprünglichen Sequenz. Gemeinsame Präfixe von Suffixen zeichnen sich durch einen gemeinsamen Pfad von der Wurzel aus. Suffixbäume können kanonisch so generalisiert werden, dass sie gleichzeitig die Suffixe mehrerer Sequenzen enthalten.

Die Idee des am ZAIK entwickelten Verfahrens besteht nun darin, einen generalisierten Suffixbaum aus den Zielsequenzen aufzubauen. Dieser wird zunächst anhand oben genannter Kriterien gefiltert, um ungeeignete Proben auszuschliessen. Anschliessend wird der Baum in Tiefensuche durchlaufen, und für alle Knoten wird ein Alignment mit den Zielsequenzen berechnet. Dabei müssen Alignments für gemeinsame Präfixe nur einmal berechnet werden.

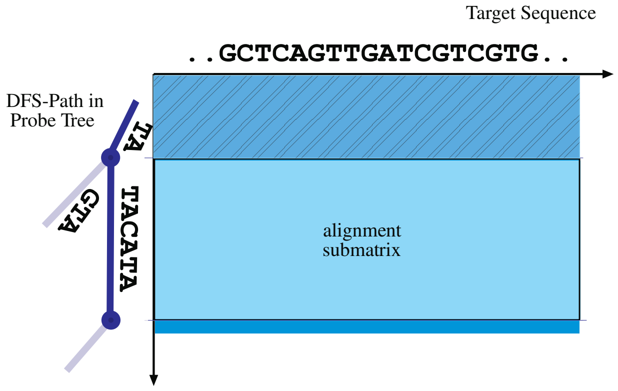

Abbildung 3.6: *Thermodynamisches Alignment mit Hilfe von Bäumen.*

Aus den so berechneten Schmelztemperaturen kann anschließend mit Hilfe von im wesentlichen einfachen Sortierverfahren eine Probenmenge für den Chip ausgewählt werden. Dabei ist es möglich, sowohl im Pre-Processing, d.h. bei der Auswahl von Kandidaten, als auch bei der abschliessenden Auswahl Proben zu bestimmen, die sensitiv für eine ganze Familie von DNA-Sequenzen sind.

Weiterführende Arbeiten

- Die Qualität der durch das Programm bestimmten Proben für tatsächliche Chip-Experimente werden in Zusammenarbeit mit Professor Tautz (Institut für Genetik, Universität zu Köln) bestimmt.
- Sofern keine eindeutigen Proben gefunden werden können, ist es ein einfaches Optimierungsproblem, minimal-konfliktäre Proben zu bestimmen.
- Gegeben eine minimal konfliktäre Probenmenge, muss bei der Chipdaten-Auswertung eine Rückrechnung auf die Daten der einzelnen Proben erfolgen. Dies kann mittels eines statistischen Auswertungsverfahrens, das schon erfolgreich bei statistischen Gruppentests eingesetzt wurde, unter Anwendung von Markov-Chain-Monte-Carlo-Methoden erfolgen.

3.4 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.

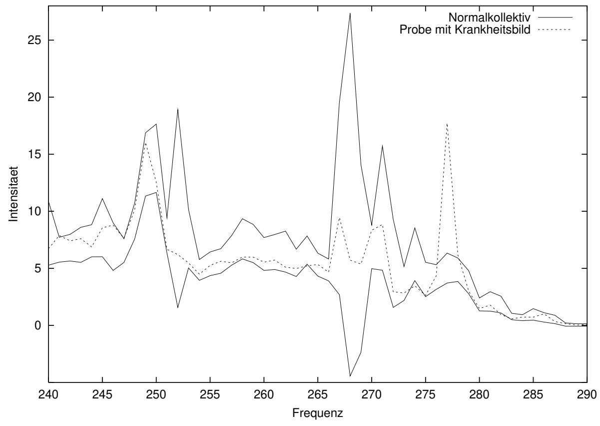

Abbildung 3.7: *Das Normalkollektiv ist durch den Bereich zwischen den durchgezogenen Linien gekennzeichnet. Das Spektrum der Patientenprobe weicht von diesem signifikant ab.*

3.5 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 Forschungsaufenthalte eines ZAIK-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

Gruppentests lassen sich 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 den Umgang mit Fehlern und sind damit beim Einsatz in der experimentellen Praxis meist unumgänglich. Gegeben sind hierbei Wahrscheinlichkeiten für das Auftreten falsch negativer Pooltests, also einem negativen Testergebnis trotz positiver Objekte im Pool, und ebenso Wahrscheinlichkeiten für falsch positive Testergebnisse. Häufig hängen diese Wahrscheinlichkeiten von der Anzahl der positiven (negativen) Objekte ab, bei denen man ein falsch negatives (positives) Testergebnis erwarten kann. Je mehr positive Objekte existieren desto unwahrscheinlicher ist ein falsch negativer Pooltest.

Weiterhin unterscheidet man zwischen adaptivem und nicht-adaptivem Gruppentesten: Während bei 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.

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 und 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.

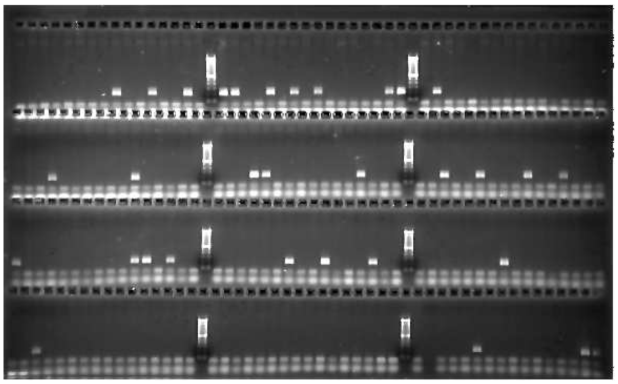

Abbildung 3.8: *Das Ergebnis einer PCR-Reaktion wird durch Gel-Elektrophorese sichtbar gemacht. Die hellen Stellen sind genomische DNA.*

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. Darüberhinaus ist es 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. Diese 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 Klone rekonstruiert werden. Hierzu werden sogenannte Sequence-Tagged-Sites (STS) Marker benutzt. Alle Klone, die denselben 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 zu begrenzen.

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 Klone 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.

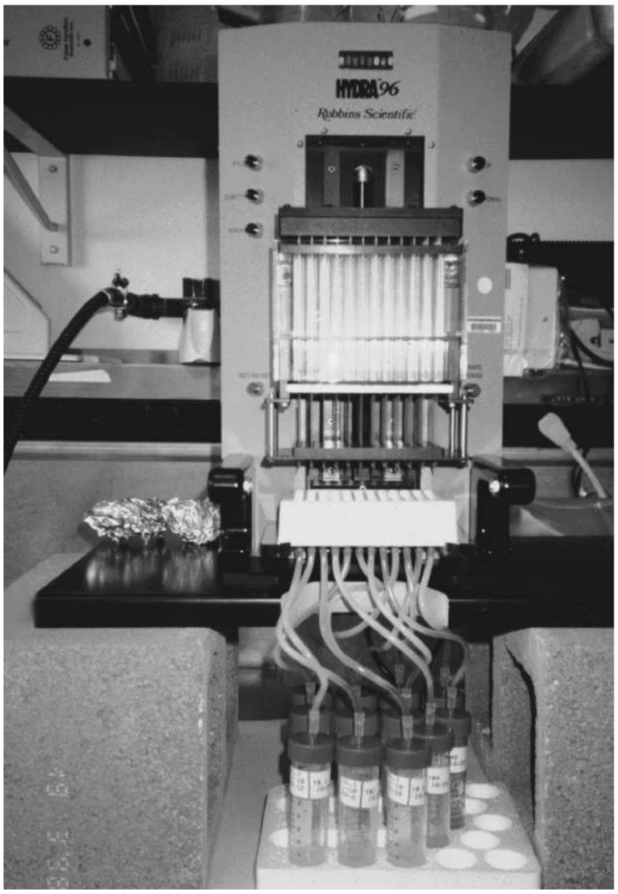

Abbildung 3.9: Mit der Pipettiermaschine *HYDRA* wurden große Teile der Pools erstellt.

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 Klone 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 Klonen 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.

Die Auswertungstrategie lässt sich auch auf Problemstellungen bei DNA-Chips anwenden, und zwar wenn keine eindeutigen sequenzspezifischen Proben gefunden werden können. In dieser Hinsicht laufen Untersuchungen in Zusammenhang mit dem Projekt unserer Arbeitsgruppe zur Selektion von Proben für DNA-Chips.

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).

3.6 Modellierung metabolischer Netzwerke

Aufgrund der stetig anwachsenden Menge an genetischer Information durch die Entschlüsselung kompletter

Genome entsteht in vielen Bereichen der Biologie eine Vielfalt neuer Fragestellungen. Hierbei betreffen viele offene Probleme den Metabolismus einer Zelle. Unter dem Metabolismus versteht man die Gesamtheit der chemischen Reaktionen, die für die Bereitstellung der freien Energie verantwortlich sind, damit für Erhaltung, Teilung etc. sorgen und somit wiederum die Ausübung ihrer Funktionen sichern.

Nahezu alle metabolischen Reaktionen sind enzymatische Reaktionen und werden so durch zelleigene Proteine gesteuert, womit der Zusammenhang Metabolismus-Genom hergestellt ist. Wichtige Subsysteme des Metabolismus sind sogenannte metabolische Pfade. Ein metabolischer Pfad ist eine Abfolge von Reaktionen, so daß jeweils das Endprodukt der vorhergehenden Reaktion das Ausgangsprodukt der nächsten ist (inklusive Verzweigungen oder auch Kreisen). Auf diesen Wegen werden zellwichtige Zwischen- und Endprodukte synthetisiert.

Im Rahmen der Promotion eines ZAIK-Mitarbeiters werden neue Verfahren zum Schaffen von “Substraträumen” für enzymatische Reaktionen entwickelt. Diese Substrat-Räume enthalten eine möglichst umfassende Auswahl von Substanzen, die von einem Enzym prozessiert werden bzw. das Enzym blockieren und somit die enzymatische Reaktion inhibieren. Wir versehen unsere Substraträume mit einem geeigneten Abstandsmaß und bekommen so mit Hilfe bereits gut bekannter Substrate ein Maß dafür zurück, wie wahrscheinlich noch unerforschte Stoffe von einem Enzym prozessiert werden bzw. das Enzym inhibieren. Auf ähnlichem Wege schaffen wir “Produkt-Räume”, die eine möglichst umfassende Auswahl von Substanzen enthalten, die als Produkt einer enzymatischen Reaktion in Frage kommen. Durch Iterieren, d.h. durch Bilden von Durchschnitten von Substrat- und Produkträumen lassen sich ganze Kaskaden enzymatischer Reaktionen durchspielen. Damit lassen sich die Mechanismen noch unerforschter Substanzen wie z.B. neuer Medikamente simulieren bzw. deren Risiken abschätzen.

3.7 Bioinformatik Interessengruppe BIG

Zur Vernetzung und Förderung des wissenschaftlichen Austausches in der Bioinformatik wurde im vergangenen Jahr die Bioinformatik-Interessen-Gruppe (“BIG”) gegründet. BIG steht allen aus der Region Köln-Bonn offen, die ein Interesse an Methoden, Verfahren, Problemen und Aufgabenstellungen der Bioinformatik haben, die sich mit Gleichgesinnten austauschen möchten, oder einfach nur einen ersten Einblick in das Gebiet bekommen möchten. Dabei richtet sich BIG sowohl an Studenten und Mitarbeiter universitärer Institute als auch an Angestellte kommerzieller Unternehmen, um einen möglichst praxisnahen und engen Austausch zwischen Bildung, Forschung und Industrie zu fördern.

In regelmäßigem Abstand veranstaltet BIG dazu Treffen, die jeweils einem aktuellen Thema gewidmet sind. So wurden bei vergangenen Treffen unter anderem DNA-Chips, Homologien bei Proteinsequenzen oder Data Mining Probleme vorgestellt und diskutiert.

Um den Vernetzungsgedanken weiter zu unterstützen, wurden ausserdem zwei öffentliche Mailinglisten eingerichtet. Über *big-announce@zpr.uni-koeln.de* können Termine zu Veranstaltungen zum Thema Bioinformatik bekanntgegeben werden. Über diese Mailingliste werden auch die regelmässigen BIG-Treffen angekündigt. Über die Liste *big-discussion@zpr.uni-koeln.de* können inhaltliche Diskussionen geführt und Fragen erörtert werden, die sich auf Verfahren, Probleme, Tools und Aufgabenstellungen der Bioinformatik beziehen oder sonst von allgemeinem Interesse sind.

Weitere Informationen über BIG und über die Mailinglisten (und wie man sich für diese anmeldet) können im Internet unter **http://www.zaik.uni-koeln.de/~big** abgerufen werden, oder Kontaktaufnahme per E-Mail an **big@zpr.uni-koeln.de**.

Kontakt: `bioinformatik@zpr.uni-koeln.de`