IV. DNS-Computer

Alles Leben beruht auf der Informationsverarbeitung in den Zellen eines Organismus. Diese arbeitet mit komplexen chemischen Reaktionszykeln, die bei weitem noch nicht alle verstanden sind. Das wichtigste Speichermedium ist die Desoxyribonucleinsäure, kurz DNS, der Zelle.

Die Informationsdichte dort ist ungeheuer hoch: In trockenem Zustand benötigt ein Bit gerade einmal ein Volumen von $1\,\mathrm{nm}^3$, in der Zelle etwa $100\,\mathrm{nm}^3$. Für einen Kubikzentimeter trockener DNS kommt man somit auf eine Speicherkapazität von 10^{21} Bit, das ist eine achtel Billion Gigabyte; für DNS in der Zelle kommt man mit einem Kubikzentimeter immerhin noch auch 125 Milliarden Gigabyte. In einem Computer faßt ein Speicherriegel der Größe $15\,\mathrm{cm} \times 3\,\mathrm{cm} \times 1\,\mathrm{cm}$ gerade einmal ein halbes Gigabyte, und auch bei CDs braucht man größenordnungsmäßig fast eine Billion Stück, um die in einem Kubikzentimeter DNS enthaltene Information zu speichern.

Hinzu kommt, daß die Information in der DNS massiv parallel verarbeitet werden kann mit einem Energieaufwand pro Operation, der ungefähr zehn Größenordnungen unter dem heutiger Supercomputer liegt.

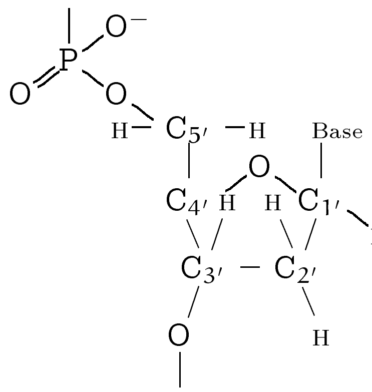
Wenn man dieses Potential für klassische Probleme der wissenschaftlichen oder gewerblichen Informationsverarbeitung nutzbar machen könnte, wäre dies eine Revolution, die auch die Kryptographie massiv verändern würde.

Seit 1994 gibt es einen ersten Hinweis darauf, daß dies vielleicht nicht völlig unrealistisch ist: Damals löste der Mathematiker und Informatiker LEONARD M. ADLEMAN (der hier bereits im Zusammenhang mit dem RSA-Verfahren erwähnt wurde) in einem molekularbiologischen Labor ein (ziemlich einfaches) graphentheoretisches Problem mit Hilfe von DNS.

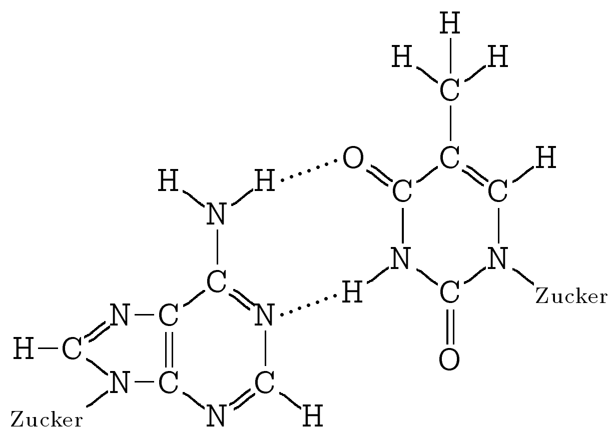
Um seine Vorgehensweise und das Potential der Methode zu verstehen, müssen wir zunächst kurz den Aufbau der DNS betrachten.

1. Die Desoxyribonucleinsäure

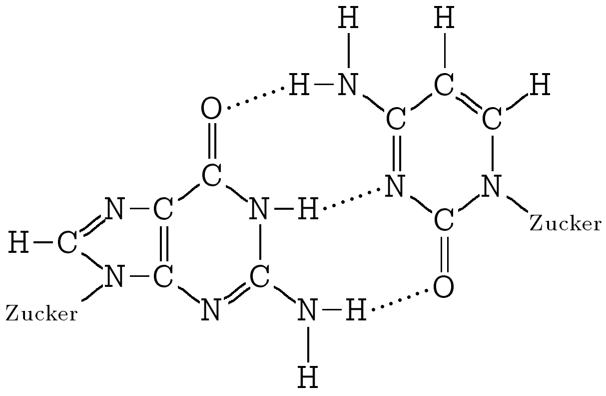
Die Desoxyribonucleinsäure hat ihren Namen vom hier abgebildeten Zuckermolekül Desoxyribose, aus dem ihr Gerüst aufgebaut ist.



Die fünf Kohlenstoffatome werden mit $1'$ bis $5'$ bezeichnet; in der Abbildung ist dies als Index eingetragen. An der Phosphatgruppe von $5'$ und der Hydroxygruppe von $3'$ fehlt jeweils ein Wasserstoffatom: Hier werden die Zucker zu einer Kette zusammengesetzt, wobei stets auf die $5'$-Phosphatgruppe eine $3'$-Hydroxygruppe folgt. Am Kohlenstoffatom $1'$ steht *Base;* hier wird eine jener vier Basen eingebaut, die die eigentlichen Informationsträger der DNS sind: Adenin, Guanin, Thymin und Cytosin.



Adenin Thymin



Guanin Cytosin

Wie man an den Abbildungen sieht, können sich zwischen Adenin und Thymin sowie zwischen Guanin und Cytosin die gestrichelt eingezeichneten Wasserstoffbrücken ausbilden; diese sorgen dafür, daß DNS nur selten in einzelnen Strängen vorkommt: Meist kombinieren sich zwei Stränge zu einem Doppelstrang, wobei jedem Adenin ein Thymin und jedem Guanin ein Cytosin gegenübersteht. Man bezeichnet diese Basen deshalb als zueinander WATSON-CRICK-komplementär.

Aus geometrischen Gründen hat ein DNS-Doppelstrang die berühmte Form der Doppelhelix, die wiederum selbst weitere geometrische Strukturen bildet. Für die Informationsverarbeitung in der Zelle sind alle diese geometrischen Strukturen sehr wichtig; die entsprechenden Mechanismen sind aber noch nicht so gut verstanden, daß man sie auch in Laborexperimenten ausnutzen könnte, so daß diese Strukturen für dieses Seminar irrelevant sind.

2. Die Polymerase-Kettenreaktion

Ein wesentlicher Bestandteil des molekularen Rechnens ist der Aufbau von DNS-Sequenzen sowie die Vervielfältigung von Sequenzen mit erwünschten Eigenschaften. Das wichtigste Hilfsmittel dafür ist die 1987 von KARY B. MULLIS entwickelte Polymerase-Kettenreaktion.

Polymerasen sind Enzyme, die in der Zelle die Duplizierung der Information aus DNS-Strängen steuern; sie sind je nach Lebewesen verschieden, und auch innerhalb derselben Zelle gibt es oft mehrere Polymerasen mit verschiedenen Aufgaben. Das besonders gut erforschte Bakterium *Escherichia coli* etwa enthält drei DNS-Polymerasen; die am stärksten vertretene DNS-Polymerase I ist eine Kette mit 928 Aminosäuren. Es handelt sich hier also um sehr komplexe Enzyme, die deshalb auch heute noch nicht synthetisiert werden, sondern auch für die Laborarbeit von Bakterien produziert werden.

Polymerasen werden zusammen mit zwei sogenannten *Primern* eingesetzt, d.h. mit zwei Folgen von DNS-Basen; diese sorgen dafür, daß selektiv ein DNS-Strang produziert wird, der WATSON-CRICK-komplementär ist zum Stück zwischen den beiden Primern auf einem vorhandenen DNS-Strang. Insbesondere muß DNS also einsträngig vorliegen, bevor eine Polymerase aktiv werden kann. In der lebenden Zelle sorgen Enzyme dafür, daß benötigte Teilstränge zur Verfügung stehen; im Labor geht man brutaler vor und teilt den ganzen Doppelstrang durch Erhitzen auf eine Temperatur von $90° - 95°$ C.

Da diese Temperatur praktisch alle Enzyme zerstört, arbeitet man heute nicht mehr mit der Polymerase von *Escherichia coli,* sondern meist mit der sogenannten Taq-Polymerase des Bakteriums *Thermus aquaticus,* das in heißem Wasser lebt und daher auch eine hitzebeständige Polymerase hat, die ihre optimale katalytische Wirkung bei etwa 75° C entfaltet.

Zur Vervielfältigung von DNS geht man folgendermaßen vor:

Man gibt die DNS zusammen mit den beiden Primern, der Taq-Polymerase und den vier Nucleotiden, aus denen die DNS aufgebaut ist, in ein Gefäß und denaturiert zunächst die DNS durch Erhitzen, d.h. man zerlegt sie in ihre Einzelstränge. Sodann kühlt man ab auf etwa 55° C; dies führt dazu, daß sich die Primer an die passenden Stellen der DNS-Stränge angliedern. Eine Erhöhung der Temperatur auf etwa 75° C aktiviert die Polymerase, die nun dafür sorgt, daß (ausgehend von den $3'$-Enden der Primer) WATSON-CRICK-komplementäre Nucleotide an die DNS-Stränge angelegt werden.

Damit hat man die gewünschten DNS-Sequenzen verdoppelt, was im allgemeinen nicht ausreicht; deshalb wird das Verfahren etwa 25–35 Mal wiederholt, was zu einer etwa millionenfachen Vermehrung führt. Das Verfahren kann in sogenannten Thermocyclern automatisch durchgeführt werden, da es nur auf die zyklische Temperatursteuerung ankommt.

3. Adlemans Experiment

Gegeben seien eine gewisse Anzahl von Städten sowie Straßen, die Städte gewissen anderen Städten verbinden. Man finde eine Straßenverbindung, die eine vorgegebene Stadt mit einem vorgegebenen Ziel so verbindet, daß jede Stadt genau einmal besucht wird.

Dies ist eine von vielen Varianten eines graphentheoretischen Problems, das für große Anzahlen von Städten und Verbindungsstraßen rechnerisch nur mit großem Aufwand gelöst werden kann. (Es ist NP-vollständig.) ADLEMAN beschränkte sich allerdings auf nur sieben Städte und fand heraus, daß der durchschnittliche menschliche Betrachter etwa 54 Sekunden braucht um sein spezielles Problem zu lösen. Interessanter ist aber die Art und Weise, wie er diese Lösung statt durch Hinschauen durch sieben Tage Arbeit im Labor produzierte.

Die Grundidee ist folgende: Jede Stadt wird durch eine Folge von zwanzig DNS-Basen kodiert; als Lösung soll jene Folge von 140 DNS-Basen produziert werden, die den (in ADLEMANs Beispiel einzigen) Lösungsweg angibt.

Die Herstellung kurzer DNS-Sequenzen mit vorgegebener Basenfolge gehört heutzutage zu den Standardtechniken der Molekularbiologie; große Labors haben Automaten, die solche Sequenzen (über die Polymerase-Kettenreaktion) erzeugen, kleinere kaufen sie von den großen: Für 25$ kann man innerhalb von wenigen Tagen ein Reagenzglas bekommen, das etwa 10^{18} DNS-Stränge mit einer vorgegebenen Folge von zwanzig Basen enthält.

Für jede der sieben Städte wurde also ein solches Reagenzglas benötigt, dazu aber auch noch für jede Straße zwischen zwei Städten. Natürlich müssen die Basenfolgen für Städte und für Straßen aufeinander abgestimmt sein: ADLEMAN faßte die Basenfolge für die i-te Stadt auf als ein Paar (x_i, y_i) aus zwei Basenfolgen der Länge zehn, und wenn es eine Straßenverbindung zwischen i-ter und j-ter Stadt gibt, kodierte er diese durch die Basenfolge $(\overline{y_i}, \overline{x_j})$, wobei die Überstreichung hier für das WATSON-CRICK-Komplement stehen soll.

Der erste Schritt der eigentlichen Berechnung bestand darin, alle diese Sequenzen (jeweils etwa 10^{14} Moleküle) in Wasser aufzulösen und Ligase dazuzugeben. (Ligasen sind Enzyme, die kurze DNS-Sequenzen zu langen zusammenfügen.) Dazu kommt noch etwas Puffer und ähnliches, und ungefähr eine Sekunde später sind im Reagenzglas viele doppelsträngige DNS-Sequenzen zu finden, die möglichen Wegen durch den Graph entsprechen.

Unter diesen Sequenzen sind mit praktisch 100%-iger Sicherheit auch solche, die der Lösung des Problems entsprechen; das Problem besteht genau wie bei den überlagerten Zuständen in Quantencomputern darin, diese herauszufiltern. Ein Vorteil gegenüber Quantencomputern ist allerdings, daß man beim molekularen Rechnen durch Messungen üblicherweise nichts zerstört.

Im zweiten Schritt ging es darum, Sequenzen zu finden, die zu Wegen mit dem Richtigen Anfangs- und Zielort gehören. Finden bedeutete in diesem Fall, daß diese Sequenzen stark vermehrt werden sollten, und das ist ein klarer Fall für die Polymerase-Kettenreaktion: Man nehme die WATSON-CRICK-Komplemente des Start- und des Zielorts als Primer und lasse die Reaktion laufen. Danach sind alle Sequenzen, bei denen Anfangs- und Zielort stimmen, etwa millionenfach verstärkt; die, bei denen nur einer der beiden Orte stimmt, etwa dreißigfach, und der Rest überhaupt nicht. Entnimmt man also eine kleine Probe zur Weiterverarbeitung, so enthält diese kaum noch Sequenzen, bei denen einer der beiden Orte falsch ist.

Im dritten Schritt ging es darum, alle Sequenzen auszusondern, die eine falsche Länge haben. Ein Weg, der jede Stadt genau einmal berührt, entspricht einer Folge von sieben Städten und damit einem DNS-Strang der Länge 140. Zur Isolation dieser Stränge führt eine Art chromatographisches Verfahren, die *Gel-Elektrophorese.*

Hierbei wird ausgenutzt, daß die betrachteten Moleküle negativ geladen sind (man achte auf das O^- in der Desoxyribose). Bringt man die Lösung also auf einen Gel auf, meist Agarose oder Polyacrylamid, und legt ein elektrisches Feld an, so hängt die Wanderungsgeschwindigkeit ab sowohl von der elektrischen Ladung als auch der Molekülgröse und (in geringerem Ausmaß) einigen anderen Gegebenheiten. Die Methode ist aber jedenfalls trennscharf genug, um Sequenzen, deren Länge sich um zwanzig Nucleotide unterscheidet, zuverlässig zu trennen; nachdem die Moleküle einige Zeit gewandert sind, entnimmt man die aus der 140er-Region und verwirft den Rest.

Unter den noch verbleibenden Sequenzen befinden sich immer noch zahlreiche Nichtlösungen, denn eine Folge von sieben Städten erhält man auch, wenn man einige Städte ausläßt und andere dafür mehrfach besucht. Es muß also entweder sichergestellt werden, daß keine Stadt mehrfach besucht wurde, was mit molekularbiologischen Methoden sehr schwer sein dürfte, oder aber, daß jede Stadt mindestens einmal besucht wurde.

Letzteres ist für den Anfangs- und den Zielort bereits bekannt; bleiben also noch fünf Städte. Für diese verwendete ADLEMAN eine Kombination aus molekularbiologischer und physikalischer Vorgehensweise: Um diejenigen Moleküle auszusondern, die die Sequenz für eine gegebene Stadt nicht enthalten, heftete er WATSON-CRICK-Komplemente dieser Stadt an kleine Eisenkügelchen mit einem Durchmesser von etwa 1 μm, gab dies zu der denaturierten (d.h. wieder durch Erhitzen in Einzelstränge zerlegten) noch verbliebenen Lösung und lies die Stränge sich wieder kombinieren. Dabei paarten sich die Sequenzen an Eisenkugeln mit Teilsequenzen jener

Moleküle, die einen Weg durch die betrachtete Stadt beschreiben. Die so entstandenen DNS-Sequenzen waren also mit Eisenkugeln verbunden und konnten so mit einem Magneten am Rand des Reagenzglases festgehalten werden, während der Rest des Glases ausgeschüttet wurde.

Danach wurde frisches Wasser zugegeben und das ganze wieder erhitzt zur Denaturierung; jetzt aber wurde der Magnet an das erhitzte Reagenzglas gehalten, so daß der die Sequenzen mit Eisenkügelchen festhielt. Der Rest wurde umgegossen in ein anderes Reagenzglas, wo dann das gleiche Spiel für die nächste Stadt wiederholt werden konnte, bis alle fünf Städte abgearbeitet waren.

Falls danach noch Moleküle übrig waren, konnte es sich nur um Lösungen handeln. Da es aber wahrscheinlich nur noch relativ wenige waren, vermehrte sie ADLEMAN zunächst durch eine Polymerase-Kettenreaktion und überzeugte sich durch Gel-Elektrophorese davon, daß seine Lösung wirklich Sequenzen der Länge 140 enthielt. Zur vollständigen Lösung des Problems mußten diese dann nur noch analysiert werden.

Auch hierzu dient wieder die Polymerase-Kettenreaktion in Verbindung mit Gel-Elektrophorese: Für jeden Zwischenort führte ADLEMAN eine Polymerase-Kettenreaktion durch, wobei er als Primer die WATSON-CRICK-Komplemente von Anfangsstadt und Zwischenort nahm. Dadurch wurde die Teilsequenz bis zu diesem Zwischenort stark vermehrt, und durch Gel-Elektrophorese läßt sich feststellen, wie lange sie ist. Diese Länge, geteilt durch zwanzig, ist die Position der Stadt im Lösungsweg.

4. Wie geht es weiter?

ADLEMAN hat mit seinem Experiment zwar kein neues Problem gelöst, aber er hat gezeigt, daß molekularbiologische Verfahren zumindest grundsätzlich zur Lösung rechnerischer Probleme benutzt werden können – bis zu ihrem effizienten Einsatz ist es sicherlich noch ein langer Weg.

Das durchgeführte Experiment überprüft mit brutaler Gewalt (fast) *alle* möglichen Wege durch den Graphen, was schon bei 200 Städten eine DNS-Menge verlangen würde, die mehr wiegt als die Erde. Mit klassischen Computern und den besten derzeit bekannten Algorithmen lassen sich dagegen auch Probleme mit einigen Tausend Städten behandeln.

Auch der letzte Schritt des Experiments funktionierte nur deshalb, weil das Problem eine eindeutige Lösung hatte. Dies ist allerdings kein großes Problem, denn alternative Methoden zur Bestimmung der Basenfolge in einem DNS-Strang gibt es natürlich: Schließlich wurde inzwischen sogar das gesamte menschliche Genom entschlüsselt.

Um ein Gefühl für die mögliche Relevanz des molekularen Rechnens in der Zukunft zu bekommen, insbesondere auch in Bezug auf die Kryptologie, ist es vielleicht ganz nützlich, zum Vergleich ein völlig anderes Thema zu betrachten, die bisherige Geschichte der Faktorisierung ganzer Zahlen.

Das einfachste Verfahren zur Faktorisierung einer ganzen Zahl N ist das systematische Durchprobieren aller Zahlen $d \leq \sqrt{N}$; dieser Algorithmus ist vergleichbar mit ADLEMANs oben beschriebener Methode.

Wie wir bei der Diskussion von SHORs Algorithmus gesehen haben, liefert der Ansatz von FERMAT ein alternatives Verfahren; berechnet man, wie es FERMAT tat, für $a = 1, 2, 3, \ldots$ systematisch die Zahlen $N + a^2$ in der Hoffnung, darunter ein Quadrat zu finden, so lassen sich zumindest Zahlen mit zwei nahe beieinanderliegenden Faktoren deutlich schneller zerlegen.

Das neunzehnte Jahrhundert war von der Mechanisierung geprägt; in der ersten Hälfte des zwanzigsten Jahrhunderts erreichte diese auch die Zahlentheorie, als D.H. LEHMER eine Siebversion der FERMAT-Methode mit Zahnrädern und Fahrradketten zur Faktorisierung verwendete.

Bei den elektrischen und elektronischen Rechner griff die neue Technik schneller auf die Zahlentheorie über: D.H. LEHMER hatte bereits auf den ersten Computern seiner Universität Hintergrundprogramme laufen, die sich mit der Faktorisierung von Zahlen beschäftigten, wenn es sonst nichts zu tun gab.

In den Siebzigerjahren, als Computer leistungsfähig genug waren, um in Zahlbereiche vorzudringen, die vorher jenseits praktisch durchführbarer Rechnungen lagen, war wieder die Mathematik gefragt, die in den letzten dreißig Jahren immer neue Faktorisierungsalgorithmen entwickelte. Diese mußten auf den jeweils aktuellen Computern implementiert werden, wobei in den Achtzigerjahren vor allem die CRAY-Computer eine sehr große Rolle spielten und aufgrund ihrer Pipeline-Architektur zur Entwicklung und Optimierung neuer Programmiertechniken zwangen.

In den Neunzigerjahren brachte das Internet die Möglichkeit zum verteilten Rechnen; 1994 wurde dadurch jene berühmte 129-stellige Zahl faktorisiert, die 1978 bei der Vorstellung des RSA-Systems als Beispiel diente und von der man damals überzeugt war, daß sie auch in hundert Jahren noch sicher sei. (Heute hält das Bundesamt für Sicherheit in der Informationstechnik Zahlen mit mindestens 1024 Bit für sicher; ab 2005 verlangt es die doppelte Länge.)

Für Fortschritte bei der Faktorisierung waren also jeweils drei Aspekte maßgeblich:

- Bessere mathematische Algorithmen
- Bessere Maschinen
- Besseres Verständnis des Umgangs mit diesen Maschinen.

Auch die weitere Entwicklung des molekularen Rechnens wird wohl von drei ähnlichen Aspekten abhängen.

Sechs Jahre nach ADLEMANs Experiment kann man versuchen, zumindest ein bißchen über jeden dieser drei Aspekte zu spekulieren.

Die weitaus größte Zahl von Publikationen im Umkreis der DNS-Computer beschäftigt sich mit theoretischen Algorithmen sowie mit abstrakten Maschinen und formalen Sprachen zur Modellierung des molekularen Rechnens. Viele dieser Arbeiten kommen von theoretischen Informatikern ohne Chemieausbildung und dürften wohl selbst innerhalb der Informatik schon in wenigen Jahren vergessen sein; bis zu einer Erprobung im Labor dürften es kurzfristig auch vom verbleibenden Rest kaum eines bringen. Langfristig allerdings könnten einige wenige dieser Ansätze durchaus zu brauchbaren Verfahren weiterentwickelt werden.

Von besseren *Maschinen* dürfte der Fortschritt des molekularen Rechnens angesichts der hochentwickelten verfügbaren chemischen Labortechnik in den nächsten Jahren wohl kaum abhängen; vielmehr geht es darum, zunächst zu erforschen, welche primitiven Grundoperationen wie durchgeführt werden können, um effizient molekular zu rechnen. Mit dieser Frage beschäftigen sich eine ganze Reihe molekularbiologischer Laboratorien, die beispielsweise mit an Oberflächen angehefteten Molekülen und ähnlichem arbeiten und teilweise auch schon erste Erfolge melden können: So gelang inzwischen die molekulare Addition zweier (kleiner) Binärzahlen. Im Augenblick ist noch nicht abzusehen, ob sich mehrere Grundparadigmen durchsetzen werden, oder ob die Entwicklung auf *den* DNS-Computer zusteuert.

Der dritte Aspekt, der bessere Umgang mit der Maschinerie, kann wohl erst dann zum Tragen kommen, wenn die Entwicklungen beim Algorithmenentwurf und bei den molekularbiologischen Ansätzen anfangen, gegeneinander zu konvergieren, wenn also die entsprechenden Experten anfangen, Notiz voneinander zu nehmen. Dies setzt einerseits voraus, daß die Labors stabile und zuverlässige Verfahren entwickeln. Vor allem aber wird ein wirklicher Fortschritt erst dann möglich sein, wenn es eine nennenswerte Anzahl von Wissenschaftlern gibt, die sowohl in der Molekularbiologie als auch in der Algorithmik zumindest eine gewisse Mindestqualifikation haben.

Der Zeitrahmen dafür hängt wesentlich davon ab, ob es gelingt, ausreichend Studenten für eine solche duale Qualifikation zu begeistern und sie ihnen auch zu ermöglichen, oder ob aber die Studenten angesichts des Booms sowohl in der Informationstechnik als auch in der Molekularbiologie vollauf mit der Qualifikation in einem dieser Gebiete zufrieden sind.

An diesem Punkt geraten die Spekulationen ins Gebiet der Politik, wo Spekulationen nur selten zu etwas vernünftigem führen; daher soll dieses Seminar besser hier enden.

4. Literaturhinweise

Die Originalarbeit von Adleman ist

Leonard M. Adleman: Molecular computation of solutions to combinatorial problems, *Science* **266** (1994), S. 1021–1024

Eine erste Diskussion, bestehend aus mehreren Leserbriefen und einer Antwort Adlemans ist in *Science* **268** (1995), S. 481–484 zu finden. Eine gut lesbare Darstellung der wesentlichen Punkte des Experiments ist

Leonard M. Adleman: Rechnen mit DNA, *Spektrum der Wissenschaft,* November 1998, S. 70–77

Einen Überblick über die weitere Entwicklung geben beispielsweise die Tagungsbände *DNA Based Computers* I-V; die ersten drei wurden von der American Mathematical Society als Bände 27 (1996), 44 und 48 (beide 1999) der *Series in Discrete Mathematics and Theoretical Computer Science* veröffentlicht, die beiden weiteren habe ich noch nicht im Druck gesehen. Ein ähnlicher Tagungsband ist

Gheorghe Răun [Hrsg:] *Computing with bio-molecules,* Springer, 1998

Eine Übersicht über die molekularbiologischen Methoden, die dem molekularen Rechnen zugrunde liegen, findet man beispielsweise in den Büchern

Bernard L. Glick, Jack J. Pasternak: *Molekulare Biotechnologie,* Spektrum Akademischer Verlag, 1995 *und*

Rolf Knippers: *Molekulare Genetik,* Thieme, ^{7}1997,

von denen das erste mit etwas anschaulicheren Darstellungen arbeitet und das zweite stärker auf den biochemischen Hintergrund eingeht. Detailierte Laboranweisungen zu den Grundtechniken findet man etwa in

Hans Günther Gassen, Gangolf Schrimpf [Hrsg.]: *Gentechnische Methoden,* Spektrum Akademischer Verlag, ^{2}1999