



Bioinformatik

Jäger, Sammler und Forscher
im Daten-Dschungel der Molekularbiologie

Dirk Evers
Robert Giegerich

Technische Fakultät
Arbeitsgruppe Praktische Informatik



Die Bier- oder Bäckerhefe, auch *Saccharomyces cerevisiae* genannt, ist kein unbescholtenes Naturprodukt. Die heutigen Hefestämme sind ein Ergebnis von einigen tausend Jahren der Züchtung durch den Menschen. Die Wahl der Hefe als Modellorganismus zur Genomanalyse ist daher nicht unumstritten.

Die genetische Information eines Lebewesens ist zwar aus nur vier Buchstaben aufgebaut, aber deren Sequenz kann ungeheuer lang sein. Sie zu entziffern – ein Genom zu entschlüsseln – ist schwierig und trotz ausgefeilter Techniken erst bei wenigen Lebewesen gelungen. Welche Bedeutung die Informationen in den einzelnen Abschnitten eines Genoms für den Stoffwechsel dieses Lebewesens besitzen, ist für die Zellen dieses Lebewesens ohne weiteres verständlich, den Wissenschaftlern aber noch keineswegs klar. Hier hilft die Informatik, die an zahlreichen Stellen mit der Molekularbiologie kooperieren kann und die durch die Probleme, die der Biologe formuliert, zur Weiterentwicklung ihrer Methoden beständig herausgefordert wird, auch und nicht zuletzt im Bereich der multimedialen Repräsentation genetischer Daten. Die Bioinformatik ist ein neues Studienfach geworden, das weltweit erst von einigen Universitäten angeboten wird und außerordentlich günstige Berufschancen eröffnet.

Der große Lauschangriff auf das Leben wurde seit Mai 1986 vorbereitet. Damals begann man in den USA, eine der „Grand Challenges“ ernsthaft ins Auge zu fassen: Die Entschlüsselung des menschlichen Genoms. Dies bedeutet zunächst nicht mehr, als die genaue Anordnung der immer gleichen Bausteine A, C, G und T (sie stehen für die Nukleotide Adenin-, Cytosin-, Guanin- und Thymin-Triphosphat) zu bestimmen. Und es bedeutet nicht weniger als 3,3 Milliarden solcher Bausteine, die da ermittelt werden müssen. Durch das vollständige Entziffern der Erbanlagen des Menschen und einer Reihe weiterer Modellorganismen sollten Biologie und Medizin, Biotechnologie und Pharmazie auf eine neue Grundlage gestellt werden. Welch' tragende Rolle der Informatik in dieser Unternehmung zuwachsen würde, war aber damals nur wenigen Beteiligten klar.

12 Jahre später: Der größte Teil des menschlichen Genoms ist noch nicht sequenziert. Die bisherige Zeit hat man mit vorbereitenden Arbeiten und der Verbesserung der Sequenzieretechniken verbracht. Viele kleinere Genome liegen dagegen bereits vor. Man schätzt, daß etwa zwanzig Genome von Mikroben komplett sequenziert sind – nicht alle davon sind öffentlich zugänglich.

Ein besonderer Meilenstein war die Entzifferung des Hefegenoms im März 1996. Damit war der erste höhere Organismus vollständig entschlüsselt – ein Genom von immerhin 14 Millionen Bausteinen. Schätzt man grob die Anzahl der Buchstaben ab, entspricht dies einer Neuauflage des Brockhaus mit 70 Bänden. Man kann daraus hochrechnen, daß das menschliche Genom eine ganze Regalwand füllen würde.

Dieser sehr beliebte Vergleich hinkt gewaltig, was sich am einfachsten an einem Vergleich zweier korrespondierender Textstellen nachvollziehen läßt. So findet man im Brockhaus, 19. Auflage, Band 1, Seite 375:

alkoholische Gärung, unter dem Einfluß von Hefezy-
men ablaufender Abbau von Glucose (u.a. Mono- und
Disacchariden) zu → Athanol (Gärungsäthanol) und
Kohlendioxid . . .

Saccharomyces cerevisiae, die Bier- oder auch
Bäckerhefe, schreibt zum gleichen Thema (Chromo-
som XV, Nukleotid-Position 159550 ff) etwas lako-
nisch:

```
ATGTCATCCAGAACTCAAAAAGGTGTTATCTTCTA
CGAATCCACGGTAAGTTGAATACAAAGATATTCCA
GTTCCAAAGCCAAAGGCCAACGAATTGTTGATCAACG
TTAAATACTCTGGTGTCTGCACACTGACTGCACGCT
TGCCACGGTGACTGGCCATTGCCAGTTAAGCTACCAT
TAGTCGGTGGTCACGAAGGTGCCGGTGTCTGTGTCG
GCATGGGTGAAAACGTTAAGGGCTGGA ...
```

Die Hefezellen kommen mit dieser Auskunft ohne weiteres klar. Aber wir? Um also unseren Vergleich geradezurücken, könnte man sagen: Die Enzyklopädie der Hefe liegt gedruckt vor – aber lesen können wir sie nicht. Nicht erst hier, aber hier ganz besonders, kommt die Informatik ins Spiel.

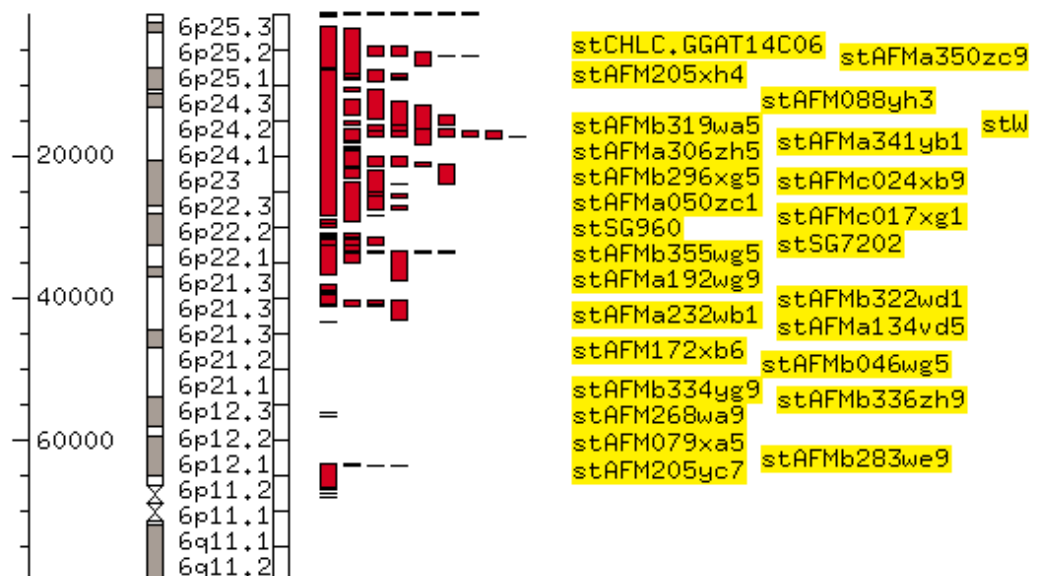
■ Datengewinnung in Sequenzierprojekten

Die große Crux der Genomsequenzierung liegt darin, daß die momentan verfügbaren Sequenzierautomaten im besten Fall Abschnitte von etwa 1000 Nukleotiden „lesen“ können. Das Genom muß also (z. B. enzymatisch oder mit Ultraschall) in kleine Abschnitte zerlegt werden, die dann einzeln analysiert werden. Bei der Hefe käme man so auf ca. 14 000 solcher Abschnitte – was aber bei weitem nicht ausreicht. Denn durch das Auftrennen geht die natürliche Anordnung der Abschnitte verloren. Um sie später rekonstruieren zu können, muß das Genom in mehrfacher Überdeckung durch überlappende Abschnitte sequenziert werden. Die Überlappungen dienen dann nicht nur zur Rekonstruktion der Gesamtsequenz, sondern auch zur Kontrolle von Lesefehlern.

Entscheidend für den Erfolg eines Sequenzierprojektes sind die Sequenzierstrategie und die Kontrolle des Datenflusses. Zunächst werden die Abschnitte experimentell markiert, um eine große „Karte“ (Map) ihrer Anordnung zu gewinnen. Die Mapping-Software wertet die experimentellen Daten aus und hilft beim Vorschlag einer Auswahl von Abschnitten, die eine einerseits gleichmäßige, andererseits minimale Überdeckung der Gesamtsequenz bilden. Vollständig automatisieren läßt sich diese Aufgabe heute noch nicht, so daß ein Teil dieser Auswahl „von Hand“, d.h. interaktiv mit dem Mapping-Programm bestimmt werden muß.

Moderne Sequenziergeräte, wie sie etwa in Bielefeld am Lehrstuhl für Genetik der Fakultät für Biologie der Universität Bielefeld zu finden sind, schaffen etwa 30 000 Nukleotide pro Tag. Bevor diese Daten-

Ein Schnappschuß aus der Sequenzierung des menschlichen Chromosoms 6. Gezeigt ist ein Ausschnitt des Chromosoms zwischen den Positionen 10 000 und 80 000. Durch die Marken (gelb) wurde bereits eine Anordnung der Abschnitte bestimmt. Die Farbe der Abschnitte (hier rot) gibt an, in welchem von 10 Verarbeitungsschritten sich diese Daten zur Zeit befinden (<http://www.sanger.ac.uk/HGP/Chr6/>).



schleudern angeworfen werden, muß der Prozeß der Weiterverarbeitung – Datensicherung, Fehlerabgleich, Sequenzrekonstruktion – wohlorganisiert und weitgehend automatisiert sein. Sonst hat das Projekt schnell einen wesentlichen Teil des Budgets in den Sand gesetzt.

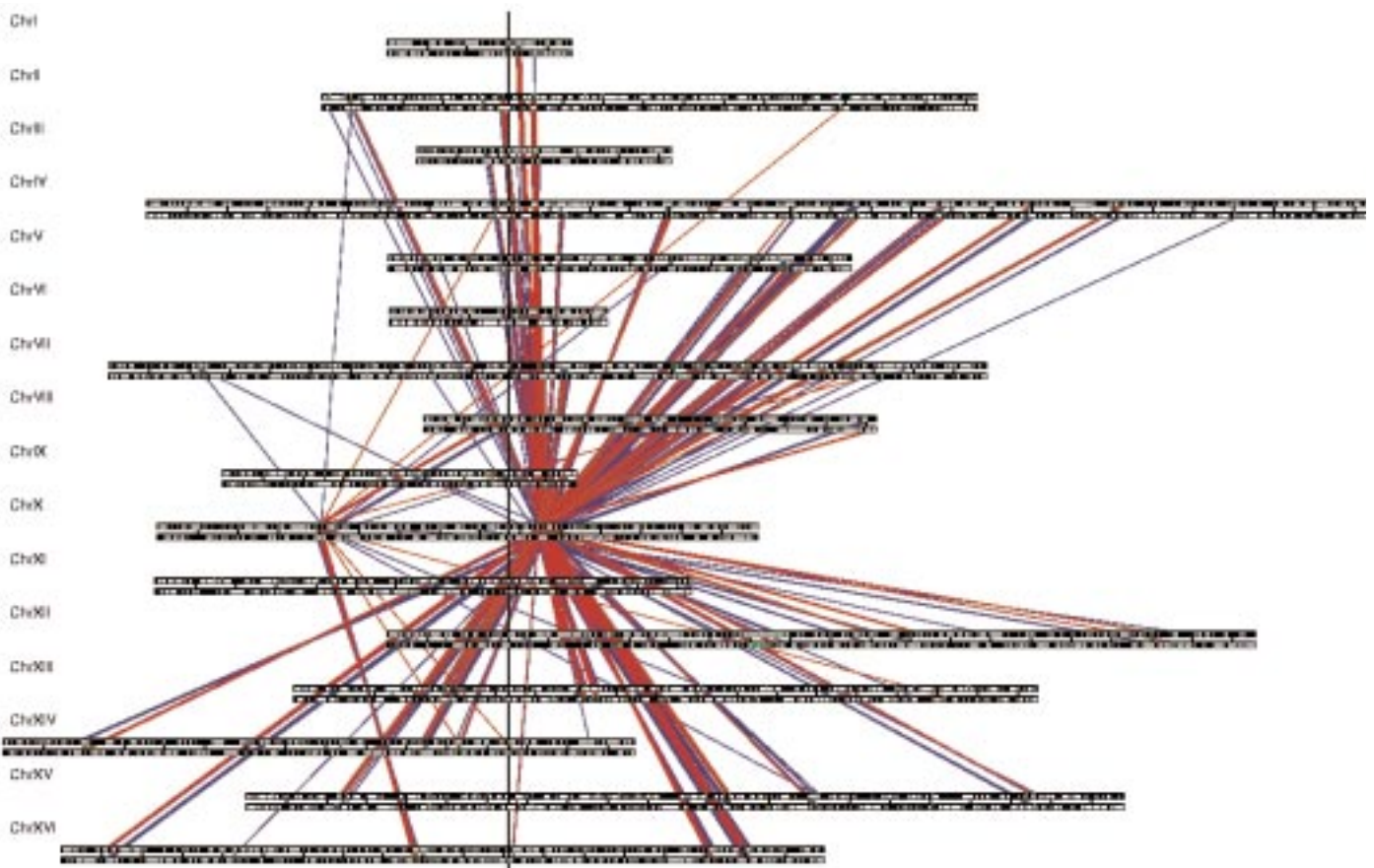
■ **Weltweite Verfügbarkeit dank Gendatenbanken**

Die gewonnenen und geprüften Informationen landen in Datenbanken verschiedenster Arten:

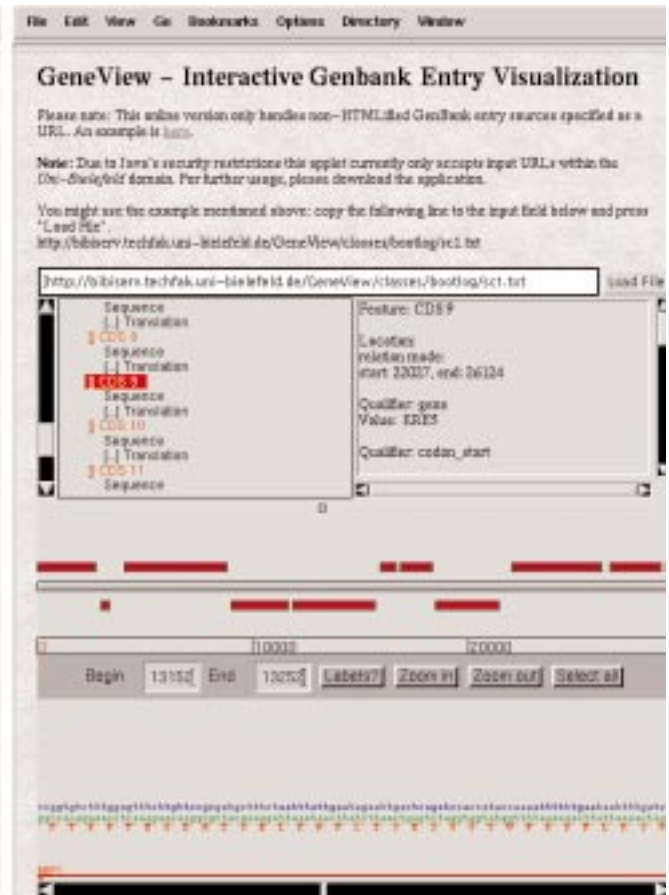
- Gendatenbanken beinhalten die oben beschriebenen Sequenzdaten – keinesfalls nur Gene, sondern auch alles andere, was ein Genom darüber hinaus enthält.
- Proteindatenbanken enthalten die Aminosäuresequenzen von Proteinen, die heute in der Regel nicht mehr experimentell bestimmt werden. Statt dessen wird das Gen, das ein Protein beschreibt, entsprechend den bekannten Regeln des genetischen Codes übersetzt.

- Strukturdatenbanken enthalten Strukturinformationen von Proteinen oder RNA-Molekülen, da deren Struktur der wesentliche Schlüssel zu ihrer Wirkungsweise in der Zelle ist.
- Weitere Datenbanken fassen Informationen über spezifische Phänomene oder Zelltypen zusammen, etwa über bekannte Mechanismen der Genregulation.

Das Betrachten dieser Informationen durch den Menschen ist natürlich wichtig, trotzdem aber die Ausnahme. Den systematischen Erkenntnisgewinn liefern hier Suchprogramme, die quer über alle beteiligten Organismen hinweg Ähnlichkeiten in Sequenz und Struktur ermitteln. So werden an einer Stelle experimentell aufgeklärte Wirkungszusammenhänge mit einer Vielzahl von ähnlichen Situationen in Verbindung gebracht. Eine sehr hohe Homologie auf Sequenz- oder Strukturebene gilt heute als fast sicherer Hinweis auf vergleichbare Funktion.



Duplikationen im Hefegenom. Das Chromosom 10 wurde in kurze Abschnitte von je 500 Nukleotiden zerlegt. Diese Abschnitte wurden mit allen Abschnitten auf den anderen Chromosomen verglichen. Die Linien zwischen den Chromosomen zeigen an, wo Abschnitte mit großer Ähnlichkeit (>50%) gefunden wurden. So suggestiv diese Graphik sein mag – man darf nicht den Schluß ziehen, die duplizierten Abschnitte wären sternförmig vom Chromosom 10 aus in das gesamte Genom migriert: Wählt man andere Chromosomen als Ausgangspunkt, entstehen ähnlich sternförmige Bilder (<http://speedy.mips.biochem.mpg.de/mips/yeast/index.htmlx>).



Ein Eintrag der Genbank, einmal mit dem nackten Auge, einmal durch die Brille eines Java-Applets betrachtet. Das Applet bietet eine Übersicht über den Eintrag auf verschiedenen Detaillierungsstufen und verbindet interessante Abschnitte per Mausklick mit zusätzlich verfügbarer Information. Das Applet wurde von Studierenden im Studiengang „Naturwissenschaftliche Informatik“ entwickelt (<http://bibiserv.techfak.uni-bielefeld.de/GeneView/>).

■ Automatische Annotation

Zwanzig oder mehr Analyseprogramme kommen routinemäßig zum Einsatz, um den frischgebackenen Sequenzdaten die ersten Informationen zu entlocken. Einerseits sollten die automatisch abgeleiteten Ergebnisse der Kontrolle eines Experten unterliegen – andererseits wäre ein menschlicher Datenverarbeiter ein Flaschenhals in der fließbandartigen Datenproduktion. Daher gibt es seit kurzem Programme zur automatischen Annotation. Sie benutzen diverse Analyseprogramme, bewerten die Signifikanz der Ergebnisse und fassen sie zusammen. Oberstes Gebot ist dabei, daß alle Beobachtungen, die getroffen, und die Schlüsse, die automatisch daraus gezogen werden, im Prinzip rekonstruiert werden können – auch wenn eine Überprüfung durch Experten erst sehr viel später geschieht. Die Automatisierung der Annotation löst gleichzeitig ein weiteres Problem: Die annotierten Informationen können veralten. Betrachten wir eine typische Aussage: „Abschnitt A, 900 Nukleotide lang, hat wesentliche Merkmale eines Gens (Start- und Stopcodon, typischer Codongebrauch usw.). Er codiert für ein hypothetisches Protein H. Proteine mit hoher Ähnlichkeit zu H sind jedoch nicht bekannt.“ Diese Information ist aus

diversen Datenbanken extrahiert, die ihrerseits ständig wachsen. Schon morgen kann ein weiteres hypothetisches Protein H' oder gar ein reales (in der Zelle nachgewiesenes) Protein P auftauchen. Natürlich will man den Eintrag von A mit H' bzw. P verknüpfen. Die automatische Annotation erleichtert die Aktualisierung von Querbezügen ganz erheblich.

■ Modellierung von Stoffwechselkreisläufen

Bisher haben wir einen datenorientierten Standpunkt eingenommen: Hier sind die genomischen Daten, was haben sie uns zu bieten? Dieser Standpunkt kann nur ein vorübergehendes Moment sein, sozusagen unser Jäger- und Sammler-Stadium im Informations-Dschungel der Molekularbiologie. Denn die Wissenschaft und die Anwendung (und im übertragenen Sinne auch die Natur) gehen ja von einem problemorientierten Standpunkt aus. Welche Enzyme und Signalmoleküle bewirken die embryonale Zelldifferenzierung? Und wie? Wenn wir schon wissen, daß sich eine unschädliche und eine tödliche Variante des CMV-Virus nur in zwei Nukleotiden unterscheiden, wie können wir diese Wirkung erklären? Wie lösen wir das Paradox der Entstehung des Lebens, daß DNA und Proteine einander wechselseitig voraussetzen?

Dazu ist es nötig, Wissen über metabolische Zyklen zu erarbeiten, den Zusammenhang von Sequenz, Struktur und (pathogener) Funktion von RNA-Viren zu klären, oder Modelle der präbiotischen molekularen Evolution zu entwickeln. Hier beginnt die eigentliche Arbeit, die die Wissenschaftler noch auf Jahrzehnte hinaus beschäftigen wird. Auch wenn die Phase der Datensammlung durch Genomprojekte im wesentlichen abgeschlossen sein wird, so werden diese kompletten Baupläne des Lebens die Grundlage aller weiterführenden Untersuchungen sein. Die Verbindung von Molekularbiologie und Informatik wird dadurch nicht einschlafen, sondern eher noch enger werden. Heute schon kann man beobachten, daß praktisch jede neue Fragestellung, die die geschaffene Datenbasis nutzt, ihre spezifischen Modelle, Algorithmen und Software-Werkzeuge braucht. Zwei Beispiele solcher Fragestellungen und Lösungswege werden in den nächsten Abschnitten vorgestellt.

■ RNA-Movies:
Darstellung dynamischer Vorgänge

Die Ribonukleinsäure (RNA) ist das Bindeglied zwischen dem Genom und den Proteinen, den Hauptfunktionsträgern eines lebenden Organismus. Die RNA ist in der Lage, durch komplexe Rückfaltungen Strukturen zu erzeugen, die Funktionen ermöglichen. Dabei gelten einfache Regeln: A paart mit U, G mit C und zusätzlich nochmals U mit G. Uracil nimmt in der RNA die Rolle des Thymins ein.

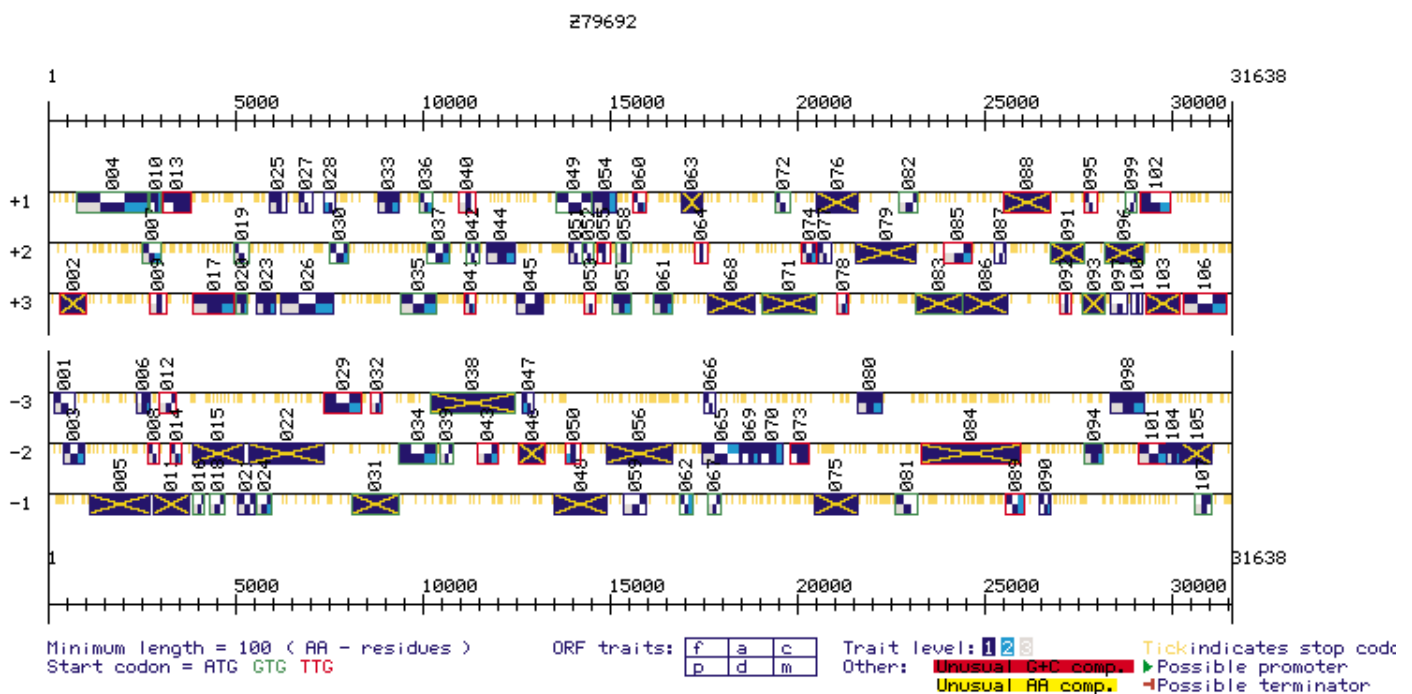
Diese einfachen drei Regeln führen dazu, daß ein Molekül der Länge x ungefähr 2^x Möglichkeiten hat, sich zu strukturieren. Wie findet man in einer sich mit jedem Nukleotid verdoppelnden Vielfalt von Strukturen die sprichwörtliche Nadel im Heuhaufen?

Die Lage ist zum Glück nicht so völlig aussichtslos, da dieser kombinatorische Strukturraum durch die biophysikalischen Eigenschaften der RNA eingeschränkt wird. Kettenmoleküle wie die RNA nehmen gemäß der Thermodynamik die Struktur mit dem niedrigsten Energieniveau ein.

Nun könnte man meinen, daß nur eine übrig bleibt, die man durch Nachbildung der Energieregeln bestimmen kann. Leider ist die Situation nicht so einfach, da die Umgebung eines Moleküls Einfluß auf seine Struktur hat. Typischerweise ist ein RNA-Molekül in einer Zelle von einer Vielzahl von Proteinen und Salzen in wäßriger Lösung umgeben.

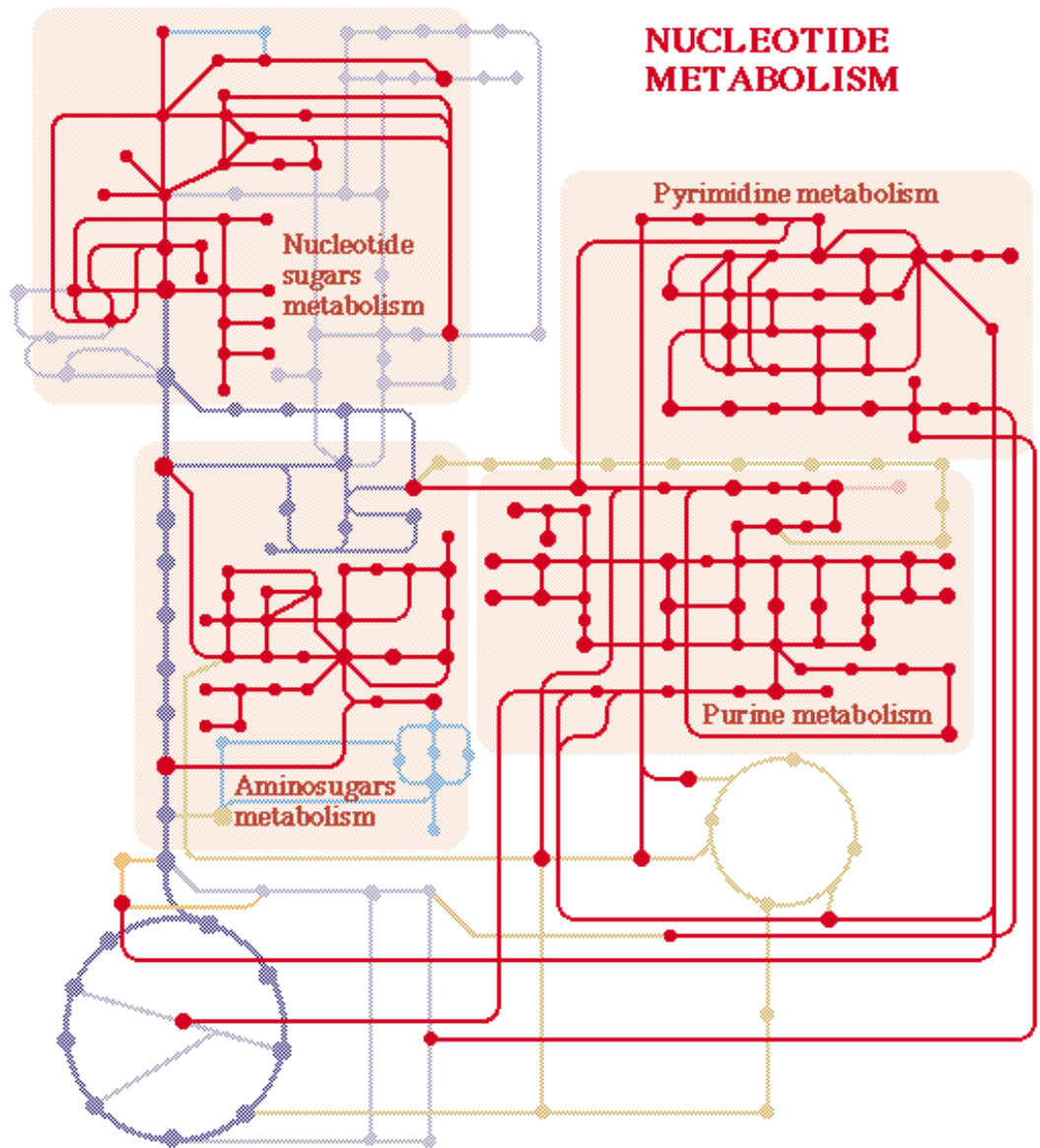
Die Vorgehensweise des Molekularbiologen bei der Strukturanalyse gleicht der eines Detektivs. Viele kleine Fakten und Hinweise führen zu einem Gesamtergebnis.

Eine Übersicht über eine Vielzahl von Analysen liefert das Annotierungsprogramm Magpie (<http://www-c.mcs.anl.gov/home/gaasterl/magpie.html>). Magpie wurde von Terry Gaasterland und Christoph Sensen entwickelt. Es werden mögliche Gene in den sechs verschiedenen Leserastern eines (doppelsträngigen) DNA-Abschnitts angezeigt. Wir sehen potentielle Gene für Proteine einer Mindestlänge von 100 Aminosäuren. Ein Farbcode zeigt den Grad der Zuverlässigkeit der abgeleiteten Information an. Durch Anklicken eines Kästchens erhält man die mit ihm verknüpften Beobachtungen. Durchgekennzeichnete Boxen sind von einem Experten manuell als korrekt annotierte Gene gekennzeichnet worden.



Die Darstellung metabolischer Zyklen in Form der Boehringer-schen Wandtafeln sind allen Biologen und Biochemikern vertraut.

Es liegt auf der Hand, daß dieses Wissen künftig elektronisch repräsentiert, dynamisiert und mit den existierenden Sequenz- und Strukturdatenbanken verknüpft werden muß. An diesem Thema arbeitet z.B. die KEGG-Gruppe am Institut für chemische Forschung an der Universität Kyoto. Klickt man einzelne Punkte oder Pfade in der Graphik an, erhält man eine Darstellung im nächsthöheren Detailgrad (<http://www.tokyo-center.genome.ad.jp/kegg/>).



01140 10/17/97

Die Königsdisziplin, die Röntgenstrukturanalyse, ist in der Lage, die Position einzelner Atome zu bestimmen – leider nur bei Molekülen einer begrenzten Größe. Mit anderen Verfahren ist es möglich zu bestimmen, ob ein Element der Nukleotidkette gepaart ist oder nicht. Zusätzliche Hinweise erhält man durch den Vergleich mit den analogen Molekülen verwandter Organismen. Mit diesen Informationen und den thermodynamischen Paarungsregeln ist es möglich, Hypothesen über die tatsächliche Struktur des RNA-Moleküls mit Hilfe eines Computerprogramms zu berechnen.

Trotz all dieser Zusatzfakten bleibt also eine Menge potentieller Kandidaten übrig, die der Experte sichten und für weitere Experimente auswählen muß. Häufig sieht man dabei Molekularbiologen einen Stapel Papier durchblättern, auf dem alle Kandidaten abgebildet sind. Eine scherzhafte Bemerkung über

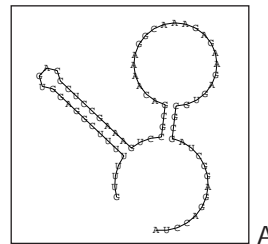
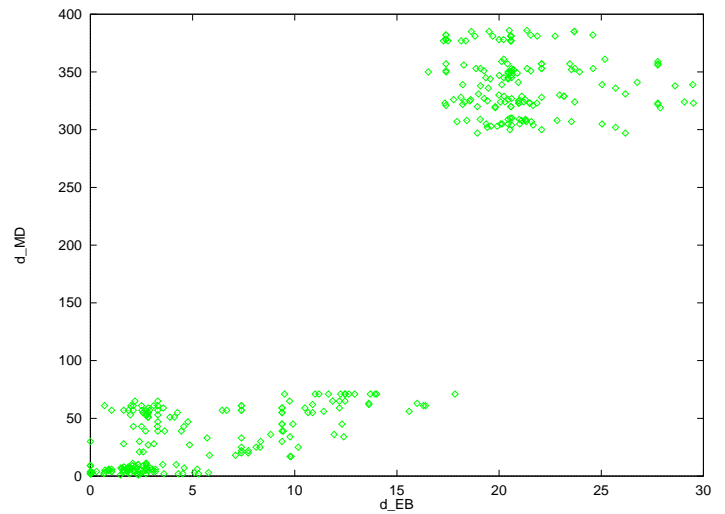
dieses „Daumenkino“ führte zum Projekt *RNA-Movies*. *RNA-Movies* ist ein Programm, daß sich genauso verhält wie ein gängiger Multimedia-Movie-player, wie er inzwischen auf jedem Rechner zur Verfügung steht.

Ein Drehbuch für *RNA-Movies* besteht aus einer Folge von Strukturbeschreibungen. Daraus wird eine animierte Graphik erzeugt – sozusagen ein Film –, in dem wir zusehen können, wie ein Molekül seine eigenen Faltungsmöglichkeiten erkundet. Die Animation lenkt unser Augenmerk automatisch auf die Unterschiede in einer Folge ähnlicher Strukturen, und man sieht auf einen Blick, wo sich etwas bewegt und wo Ruhe herrscht. Die *RNA-Movies* sind daher geeignet, alternative Strukturvorschläge oder auch Strukturübergänge zu untersuchen, sind aber selbst ein reines Werkzeug zur Visualisierung.

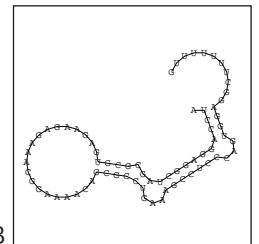
■ Die Vorhersage molekularer Schalter

Mag es schon aufwendig sein, die korrekte Struktur eines RNA-Moleküls zu bestimmen – eine Stufe schwieriger wird dies noch im Falle molekularer Schalter. Als solche bezeichnet man Moleküle, die zwei verschiedene Strukturen ausbilden, die jeweils eine andere Funktion haben. Sogenannte Attenuatoren z.B. sind RNA-Abschnitte vor einem Gen, die seine Expression an- oder abschalten können. Schaltende Strukturen sind schwer zu bestimmen, unter anderem deshalb, weil die Methoden zur Strukturklärung widersprüchliche Ergebnisse liefern können.

Das Werkzeug *paRNAss* versucht dem Experimentator Hinweise auf das mögliche Vorliegen eines strukturellen Schalters zu geben. Es beruht auf der Überlegung, daß der Strukturraum eines schaltenden Moleküls besondere Merkmale haben muß: Es gibt zwei (aber nicht mehr) wesentlich verschiedene Strukturen, die nahe am Energieminimum liegen, aber durch eine gewisse Energiebarriere getrennt sind. Um diese Eigenschaft zu prüfen, wird eine große Anzahl alternativer Strukturen berechnet und nach unterschiedlichen Distanzmaßen verglichen.

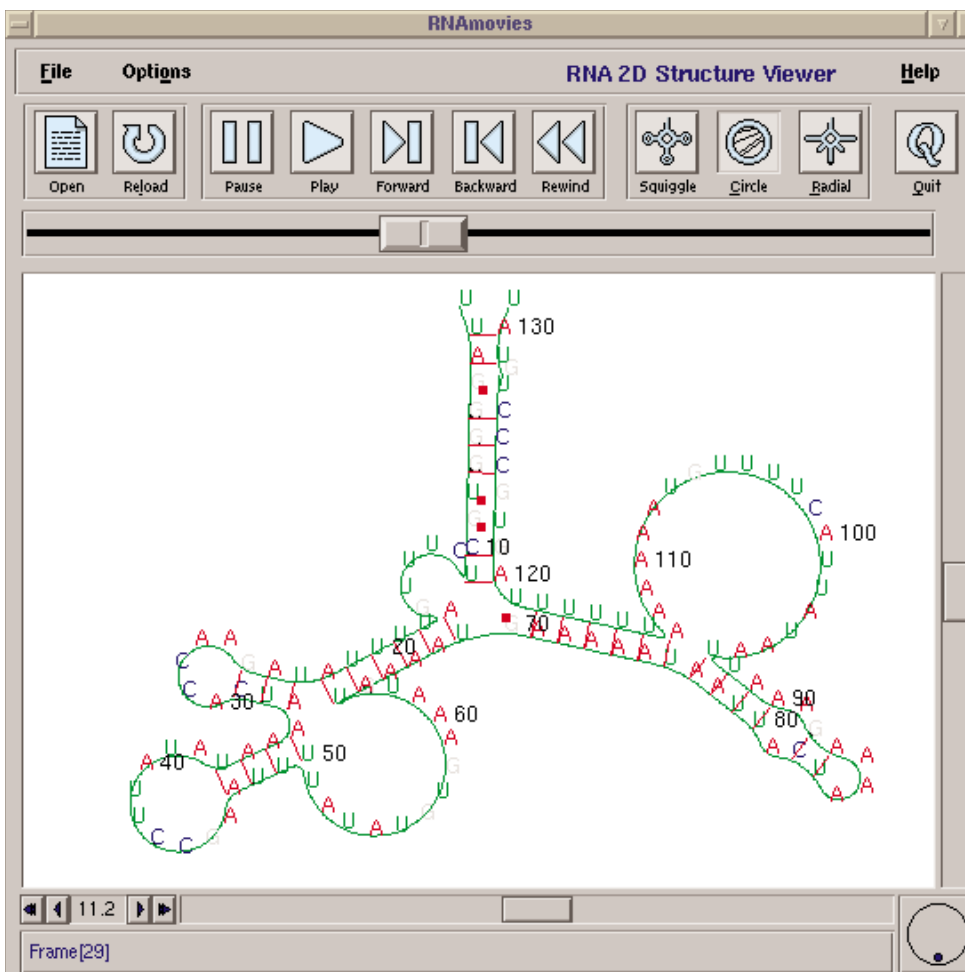


A



B

Ein Molekül – zwei Strukturen, zusammen mit der Information, die zu ihrer Auswahl führte. Für jedes Paar von möglichen Strukturen werden zwei verschiedene Eigenschaften (Distanzmaße) berechnet. Zufällig ausgewählte Strukturen werden paarweise verglichen. Ein Diagramm mit zwei deutlich ausgeprägten Punktwolken entsteht, wenn der Strukturraum gerade zwei Strukturfamilien A und B enthält: Sind beide Vergleichskandidaten aus der gleichen Familie (A:A) oder (B:B), ist ihre Distanz gering und der Meßwert fällt in die untere Wolke. Bei einem Vergleich (A:B) oder (B:A) wird in etwa der typische Abstand der beiden Familien gemessen (obere Wolke).



Der RNA-Movieplayer ermöglicht die Darstellung mehrerer RNA-Sekundärstrukturen in Folge. So lassen sich viele verschiedene Strukturen schnell sichten und vergleichen oder dynamische Vorgänge wie der Übergang von einer RNA-Schalterposition in die andere animieren.

Spricht das Ergebnis für einen möglichen Schalter, so werden zwei Strukturen (als wahrscheinliche Schalterstellungen) vorgeschlagen. Diese müssen dann experimentell überprüft werden, denn ob ein schaltfähiges Molekül tatsächlich in der Zelle eine Schalterfunktion wahrnimmt – diese Frage bleibt weiterhin der Biologie allein überlassen.

■ Eine Herausforderung auf breiter Front

Für die Informatik ist die Kooperation mit der Molekularbiologie eine Herausforderung von bemerkenswerter Bandbreite. Viele klassische Disziplinen der Informatik werden mit neuen Problemen konfrontiert, deren Lösung auf dem gegebenen Stand der Technik oft noch gar nicht möglich ist.

Effiziente Algorithmen werden benötigt zum Vergleich von Sequenzen und Strukturen sowie zur Strukturberechnung. Immer komplexere Muster werden eingesetzt, die Sequenz-, Struktur- und sonstige Merkmale verknüpfen. Musterbeschreibungssprachen für Suchwerkzeuge sind Beispiele *deklarativer Programmiersprachen*, während bei der Sequenzannotation die *logische Programmierung* effektiver eingesetzt werden kann. Sucht man nach Mustern, deren Beschreibung man (noch) nicht kennt, kommen *künstliche neuronale Netze* zum Einsatz. Datenbanken sind allgegenwärtig, wobei das Problem föderierter Datenbanken, die getrennt entwickelt werden, aber trotzdem kooperativ nutzbar sind, noch lange nicht gelöst sein wird. *Bildverarbeitung* wird heute schon umfassend eingesetzt, um experimentelle Daten zu erfassen. *Rechnernetze* sind der eigentliche Lebensraum der Bioinformatik, weil alle wichtigen Datenbanken und Werkzeuge auf Bioinformatik-Servern im Internet zur Verfügung stehen. (Den Bielefelder Bioinformatik-Server findet man unter <http://bibiserv.techfak.uni-bielefeld.de/>.)

Die *Technische Informatik* kommt ins Spiel, wenn Prozesse der Zellkultur und Fermentation zu überwachen und zu steuern sind. Last not least ist die *Visualisierung* eine wesentliche Teilaufgabe in allen Anwendungsgebieten. Dies zeigt schon ein Blick auf das Bildmaterial in diesem Aufsatz: Alle Bilder sind vom Computer erzeugte Graphiken – nur der gute alte Hefezopf sorgt noch selbst für sein Erscheinungsbild.



Prof. Dr. Robert Giegerich (links) studierte Informatik an der Technischen Universität München und an der Stanford University, USA. Er promovierte in München mit einer Arbeit zur Übersetzung von Programmiersprachen. Nach einer Zwischenetappe an der Universität Dortmund wurde er 1989 auf eine Professur für Praktische Informatik (Fachgebiet Programmiersprachen und Übersetzerbau) an der Technischen Fakultät der Universität Bielefeld berufen. Motiviert durch die Entwicklung des interdisziplinären Studiengangs „Naturwissenschaftliche Informatik“ begann er 1993 mit dem Aufbau der Bioinformatik als Forschungs- und Lehrgebiet.

Dipl.-Inform. Dirk Evers (rechts) studierte Naturwissenschaftliche Informatik an der Universität Bielefeld und am Trinity College in Dublin. Er ist Mitglied des Forschungsschwerpunkts Mathematisierung-Strukturbildungsprozesse der Universität Bielefeld. Seit 1997 promoviert er als Stipendiat des Graduiertenkollegs „Strukturbildungsprozesse“ der Deutschen Forschungsgemeinschaft zum Thema „Algorithmen zur Analyse und Visualisierung von RNA-Sekundärstrukturräumen“ an der Technischen Fakultät.

■ Bioinformatik als Studienfach

Die meisten der heutigen Bioinformatiker sind Naturwissenschaftler, die durch ihre Forschungsarbeit dazu gezwungen waren, sich mehr und mehr Informatikkenntnisse anzueignen. Allerdings hat solch autodidaktisch erworbenes Wissen angesichts der raschen Innovationszyklen in der Informatik eine relativ kurze Halbwertszeit. Der Bioinformatiker wie er heute gefordert wird, hat idealerweise eine solide Grundlagenausbildung in (Molekular-) Biologie und Informatik sowie spezielle Kenntnisse über die wichtigsten algorithmischen Methoden in der Molekularbiologie. Erfahrung in der interdisziplinären Zusammenarbeit ist ebenfalls ein entscheidendes Qualifikationsmerkmal.

Studiengänge, die Bioinformatiker heranbilden, kann man heute noch an einer Hand abzählen.

Die Universität Bielefeld kann hier eine gewisse Pionierrolle beanspruchen, da der bereits 1989 etablierte Studiengang „Naturwissenschaftliche Informatik“ seit 1995 einen kleinen, aber begehrten Strom von Absolventen [1] mit dem obigen Qualifikationsprofil hervorbringt. Die Universität von Pennsylvania hat 1994 ein gemeinsames *PhD*-Programm des Biology Departments und des Computer-Science Departments eingerichtet. Die Boston University bereitet ein Bioinformatik-Programm vor, das 1998 anlaufen soll. Universitäten in London, Nottingham und Manchester haben vor kurzem ein *Master-of-Science*-Programm in Bioinformatik etabliert. Einzelne Kurse zur interdisziplinären Spezialisierung von Biologen und Informatikern gibt es an zahlreichen US-amerikanischen Universitäten; in Europa an den Universitäten Stockholm und Bergen, in Utrecht, Dublin und München (Ludwigs-Maximilians-Universität) sowie am Pasteur-Institut in Paris und am Humangenom-Zentrum in Hinxton, UK. Mehr Informationen zu den hier genannten Kursen findet man unter <http://www.techfak.uni-bielefeld.de/bcd/ForAll/Econom/study.html>.

Wem kein lokales Ausbildungsangebot in der Bioinformatik zugänglich ist, der sollte sich im Internet umsehen. Die ohnehin netzbasiert arbeitende Bioinformatik spielt nämlich auch eine wichtige Rolle bei der Entwicklung interaktiver netzbasierter Lehr- und Lernformen. Das Birbeck College der Universität London bietet Kurse zur Proteinstrukturanalyse gegen Entgelt an. Die Biocomputing Division der Virtual School of Natural Sciences bietet seit 1995 Bioinformatik-Kurse im Internet an. Für deren Besuch wird zwar keine Gebühr erhoben, aber es wird erwartet, daß die Teilnehmer je nach spezifischer Qualifikation sich an der Weiterentwicklung des Lehrmaterials beteiligen. Die Vielfalt an Material, die hier in weltweiter Kooperation entstanden ist, findet man unter <http://www.techfak.uni-bielefeld.de/bcd/>.

■ Bioinformatiker – verzweifelt gesucht

Qualifizierte Bioinformatiker sind rar, während die Nachfrage wächst. Die Zeitschrift *Nature* verfolgt diese Entwicklung seit längerem [2] und hat in der Ausgabe vom 25.9.97 eine Artikelserie diesem Thema gewidmet [3], inklusive 13 Seiten mit Stellenanzeigen.

Die großen pharmazeutischen und gentechnischen Firmen bauen Bioinformatik-Abteilungen auf, während zugleich kleinere Firmen entstehen, die sich explizit als Bioinformatik-Dienstleister verstehen. So beschäftigt z.B. einer der Marktführer, Smith Kline Beecham, eine 50köpfige Bioinformatik-Abteilung in den USA und will eine weitere Gruppe in England aufbauen. Eine deutsche Neugründung, die LION Bioscience AG in Heidelberg, die sich auf die kommerzielle Gewinnung und Interpretation genomischer Daten spezialisiert, erwartet zehn Neueinstellungen im Bioinformatikbereich im laufenden Jahr. Wer die Stellenangebote auf dem Server des Europäischen Bioinformatik-Instituts in Hinxton studieren will (<http://www.ebi.ac.uk/>), sollte sich ruhig einen ganzen Nachmittag Zeit nehmen.

Auch ein Blick ins persönliche e-mail-Archiv der Autoren bestätigt dieses Bild. Dort finden sich aus den Monaten Mai bis Oktober 1997 insgesamt 16 Anfragen mit insgesamt 36 Stellenangeboten, nicht selten verbunden mit einer direkten Frage nach Absolventen des Bielefelder Studienganges.

Der Nachfrage-Überhang macht sich bei den Gehältern bemerkbar. In den USA kann ein fertiger Bioinformatiker (Master of Science) ein deutlich höheres Gehalt erwarten als ein promovierter Biologe. Von einer Firma wie Smith Kline Beecham erfährt man ohne Umschweife, daß ein promovierter Bioinformatiker mit dem Äquivalent eines vollen Professorengehaltes rechnen darf.

Deutsche Firmen geben sich da eher bedeckt. Bioinformatiker in industriellen Forschungsprojekten dürfen eine Bezahlung nach BAT IIa erwarten (ca. 65 000 DM pro Jahr). Am unteren Ende der Skala rangieren Universitäten und Großforschungseinrichtungen, die – etwa in Forschungsprojekten, die von der Deutschen Forschungsgemeinschaft gefördert werden – nur Promotionsstellen nach 1/2 BAT IIa oder noch darunter anbieten.

So dringend die Industrie gut ausgebildete Bioinformatiker braucht, so wenig nimmt sie auf die Belange der Ausbildung Rücksicht. Viele universitäre Gruppen klagen, daß Mitarbeiter sogar aus laufenden Projekten in die Industrie abwandern. Symptomatisch die Klage eines Kollegen an einer Medizinischen Hochschule in Texas: Dort wurde für den Sommer 1997 ein Bioinformatik-Kurs vorbereitet. Er mußte ausfallen, weil alle drei Kollegen, die für diesen Kurs als Dozenten in Frage kamen, innerhalb weniger Wochen von der Industrie abgeworben worden waren.

Literatur dazu:

- [1] M. Strobl: Germany on the trail of the Americans, *Nature* 389: 421, 1997.
- [2] D. Gershon: The Boom in Bioinformatics, *Nature* 375: 262, 1995.
- [3] D. Gershon, B.W.S. Sobral, B. Merlon, P. Wickware, H. Gavaghan, M. Strobl: Bioinformatics in a post-genomics age, *Nature* 389: 417-421, 1997.