William Martin

Bioinformatik – Eine Schlüsseltechnologie

In der deutschen Hochschullandschaft wie in der Wirtschaft hat in den letzten Jahren das Stichwort *Bioinformatik* für reichlich Pressewirbel gesorgt. Die Meldungen waren häufig mit eindrucksvollen Prognosen oder stolzen Investitionssummen verbunden, woraus man entnehmen konnte, die Bioinformatik müsse für Forschung, Lehre und Wirtschaft wichtig sein. Beispiele hierfür sind u. a. die Ausschreibung der Deutschen Forschungsgemeinschaft 1999 in Höhe von 25 Millionen € für die Einrichtung von Diplom- bzw. Bachelor-/Masterstudiengängen der Bioinformatik an deutschen Hochschulen, die Investition des deutschen Chemie-Riesen Bayer im Jahre 2000 in Höhe von 25 Millionen € in das Heidelberger Bioinformatik-Unternehmen Lion Bioscience AG, oder die im November 2001 verkündete strategische Zusammenarbeit zwischen Lion Bioscience und dem amerikanischen Computer-Giganten IBM, um „die Forschung und Entwicklungsprozesse der Pharmaindustrie grundlegend zu beschleunigen“[1]. Auch an der Heinrich-Heine-Universität Düsseldorf hat die Bioinformatik in jüngster Zeit als wichtiges, zwischen Biologie und Informatik angesiedeltes interdisziplinäres Fach an Bedeutung gewonnen. Seit dem Sommersemester 2000 können Studierende im Diplomstudiengang Biologie die Bioinformatik als Schwerpunkt wählen. Mit der Gründung der Wissenschaftlichen Einrichtung Informatik an der Math.-Nat. Fakultät der Heinrich-Heine-Universität entstehen weitere, an der späteren informatischen Berufspraxis orientierte Studiengänge, zu denen ebenfalls die Bioinformatik zählt. Aber dennoch blieb es vielen fachfremden sowie nicht wenigen fachnahen Beobachtern verborgen, was *Bioinformatik* überhaupt ist und – vor allem – wozu sie wichtig oder nützlich sein soll.

In wenigen Worten zusammengefasst kann man sagen, dass die Bioinformatik der wahre Schlüssel zum Erfolg der heutigen Genomsequenzierungs-Technologie ist, wobei die Genomsequenzierung auf überraschenden Wegen ein unerlässlicher Bestandteil der modernen Suche nach neuen Wirkstoffen – d. h. nach neuen patentierbaren Produkten – in den lebenswissenschaftlichen Anwendungsbereichen der chemischen Industrie geworden ist.

Damit wäre eigentlich schon alles gesagt. Aber wer nicht über konkrete, an Beispielen erläuterbare Vorstellungen dessen verfügt, was sich hinter den Begriffen eines *Genoms*, einer *Genomsequenzierung*, eines *Wirkstoffs* oder einer Produktentwicklung der chemischen Industrie verbirgt, weiß immer noch nicht *ganz genau*, worin die Bedeutung der Bioinformatik liegt. In diesem Aufsatz möchte ich daher in sehr allgemeinverständlicher Form das Fachgebiet der Bioinformatik, seine Entstehung im Gefüge der historisch gewachsenen Forschung, seine Einbindung in die moderne Technologie-Landschaft und seine wichtigsten wirtschaftlichen Anwendungbereiche skizzieren. Das Wesen und der Nutzen der Bioinformatik lassen sich am einfachsten anhand ihrer Bedeutung für die chemisch-phar-

[1] Siehe hierzu Presseberichte unter http://www.ibm.com/news/us/2001/11/28.html oder http://www.lionbioscience.com/press/.

mazeutische Industrie erklären – die überraschende Geschichte einer unvorhersehbaren Entwicklung und Ausgangspunkt einer kurzen Erzählung.

Die Suche nach Wirkung und Wirkstoffen

Sucht man nach einer einfachen Definition, ist die Bioinformatik nichts anderes als der fachspezifische Einsatz von Computertechnologie in den Lebenswissenschaften. Diese Definition trifft jedoch nicht exakt zu, weil Computer seit über 30 Jahren in den Biowissenschaften inklusive der Medizin eingesetzt werden. Dahingegen existiert der Begriff der Bioinformatik seit weniger als zehn Jahren. Man kann mit Recht behaupten, dass der Aufstieg und die Bedeutung des neuen, interdisiplinären Fachgebiets der Bioinformatik sehr eng mit dem rasanten Aufstieg der Genomforschung in den Jahren seit 1995 zusammenhängen. Um zu verstehen, warum die Bioinformatik wichtig ist, muss man also zunächst verstehen, was die Genomforschung ist und warum sie wichtig ist. Die meisten Leser wird es überraschen, dass aus volkswirtschaftlicher Sicht die Genomforschung in erster Linie für die chemisch-pharmazeutische Industrie wichtig ist, und zwar als Vehikel für die Entdeckung und Entwicklung neuer chemisch-pharmazeutischer Produkte. Warum das so ist, soll hier sehr verkürzt erzählt werden. Aber um zu verstehen, welche Vorteile die heutige Genomforschung für die Entwicklung neuer chemisch-pharmazeutischer Produkte bringt, muss man noch einen weiteren Schritt zurückgehen und im Groben verstehen, wie die chemische Industrie neue Produkte in der Zeit vor der Genomforschung bis zur Marktreife entwickelt hat.

Chemisch-pharmazeutische Konzerne bezahlen das Gehalt ihrer Mitarbeiter durch den Verkauf chemisch-pharmazeutischer Produkte. Im Bereich der Lebenswissenschaften sind das u. a. Präparate für das Gesundheitswesen, Produkte für die Tiermedizin und Pflanzenschutzmittel für den Agrarmarkt, um drei wichtige Beispiele zu nennen. Aber wie kommen Konzerne wie Bayer, Schering oder BASF zu neuen, patentierbaren Produkten, mit denen sie Geld verdienen können? In der Zeit vor der Genomforschung war die Vorgehensweise wenig elegant, aber sehr effektiv. Im Wesentlichen lief die Suche nach neuen Produkten so ab, wie im Folgenden skizziert:

Die Chemiker des Unternehmens synthetisierten neue chemische Verbindungen. Diese Verbindungen waren z. T. Variationen von bereits bekannten Stoffen, z. T. aber auch neue Stoffklassen samt Derivate. Neue Stoffklassen sind immer interessanter, weil sie sehr viel besser patentierbar sind. In einigen Fällen orientierte sich die Synthese an den Vorgaben bereits aus der Natur bekannter chemischer Wirkstoffe. Diese waren z. B. synthetische Varianten von krankheitshemmenden Antibiotika, welche ursprünglich als natürlich vorkommende Schutzsubstanzen in Mikroorganismen entdeckt worden waren, oder auch synthetische Varianten von entzündungshemmenden Sterioden, die natürlich vom menschlichen Körper produziert werden. Bei diesen Substanzen kannte man die Strukturen und in einigen Fällen die präzise Wirkungsweise der natürlichen Wirkstoffe. Daher hatte man bereits gewisse Erwartungen bezüglich der biologischen Wirkung solcher Verbindungen und folglich, wie man ihre Wirkung testen könnte (z. B. Wachstumshemmung von Krankheitserregern durch die Zugabe der neuen Substanz).

Aber bei den grundsätzlich neuen Klassen von chemischen Verbindungen, deren Synthese nicht von Kenntnissen der natürlichen Chemie gesteuert war, gestaltete sich die Suche nach einer möglichen Wirkung ganz und gar nicht so einfach. Jede neue Verbindung

hätte prinzipiell ein neues Produkt sein können, aber was für ein Produkt? Da sich die neusynthetisierte Struktur nicht nach einem natürlichen Vorbild richtete, war die mögliche Wirkung des neuen Stoffes sehr schwer bis überhaupt nicht vorhersagbar. Hatte man einen sehr wirksamen neuen Entzündungshemmer synthetisiert, ein gutes Mittel gegen Haarschuppen, ein Mittel zur Bekämpfung von Pilzkrankheiten im Weinanbau oder ein wirksames Mittel zur Eindämmung des Wachstums von Unkrautarten im Zuckerrübenfeld?

Weil die Wirkung neuer Substanzen und Substanzklassen kaum vorhersagbar ist, war (und ist) man in der chemischen Industrie auf gute Testsysteme angewiesen, mit deren Hilfe man die Wirkung neuer Substanzen identifizieren und nachweisen kann. Aus dem gleichen Grunde der Unvorhersagbarkeit der Wirkung mussten die neuen Substanzen aus dem Labor des Chemikers routinemäßig in nahezu allen Produktbereichen des Unternehmens zunächst auf eine Wirkung getestet werden. Diese Art der Suche nach Wirkung ist extrem aufwändig.

Wie funktionierte so ein Testsystem? Am einfachsten zu erklären ist dies am Beispiel der Unkrautmittel im Pflanzenschutz, deren weltweiter Umsatz mit rund 30 Milliarden € jährlich interessant ist. Die gesuchte Wirkung eines Unkrautmittels ist die Hemmung des pflanzlichen Wachstums. Chemiekonzerne mit Produkten im Pflanzenschutz unterhielten zu diesem Zweck hektarweise Gewächshausflächen, wo unterschiedlichste Nutzpflanzen und Unkrautarten in großen Mengen zur Verfügung standen. Die Substanzen aus dem chemischen Labor wurden in kleinen Probeflaschen angeliefert und es wurden ein paar Tropfen von jeder neuen Substanz auf die verschiedenen Pflanzen aufgetragen – Tausende von Pflanzen, Tausende von Substanzen und Millionen von Tropfen, alle von Hand aufgetragen. Konnte nach einigen Tagen eine Wachstumshemmung oder ein Absterben der Blattfläche beobachtet werden, so hatte der neue Stoff zunächst eine *Wirkung*, man war aber noch weit von einem martkfähigen Präparat mit einem patentierten *Wirkstoff* entfernt.

Solche Substanzen, die im Testsystem eine Wirkung zeigten, wurden in langwierigen und aufwändigen Arbeiten sehr viel näher untersucht. Es wurden strenge Kriterien herangezogen, wie Unschädlichkeit für Mensch und Umwelt, Abbaubarkeit durch die Pflanze, Spezifität ihrer Wirkung (z. B. gegen breitblättrige bzw. schmalblättrige Pflanzen), geringe Dosierung, praktikable Handhabung für den Landwirt, die Wirkung von chemischen Derivaten der ursprünglich identifizierten Substanz, den Mechanismus der Wirkung usw. Mit sehr viel komplizierteren Testsystemen, die stets über indirekte Verfahren auf Wirkung haben schließen lassen, wurden auch die neuen Wirkstoffe im medizinischen und in veterinärmedizinischen Bereich gesucht. Häufig wurde die eigentliche Wirkung einer neuen Substanz erst als Nebenwirkung entdeckt, so z. B. beim allseits bekannten Viagra. Dieses wurde bis in die letzten Testphasen hinein als Blutdruckmittel untersucht, die Testpatienten berichteten allerdings über unerwartete und für den Hersteller überaus kapitalisierbare Nebenwirkungen.

Die Suche wird zäh, neue Strategien werden notwendig

Es versteht sich von selbst, dass die Wahrscheinlichkeit, ein marktfähiges Pflanzenschutzmittel (oder einen anderen Wirkstoff) per Zufall zu synthetisieren, ziemlich gering ist. Aber mit einem derartigen Suchverfahren nach der Wirkung von Substanzen war es im-

merhin noch 1960 möglich, aus etwa 5.000 neu synthetisierten Substanzen einen neuen Wirkstoff zu finden, der letztendlich als patentiertes, gewinnbringendes (arbeitsplatzschaffendes) Pflanzenschutzprodukt zur Anwendung auf dem Markt zugelassen werden konnte. Seit 1960 hat sich jedoch viel geändert. Die Anforderungen an die Umweltverträglichkeit sind dramatisch gestiegen, und viele Substanzklassen wurden bereits erforscht – ein Industrie-Chemiker kann im Laufe seiner Karriere rund 10.000 neue Substanzen synthetisieren. Unzählige gute Substanzklassen wurden bereits patentiert, wobei die meisten dieser Patente inzwischen abgelaufen sind. Dies hat zur Folge, dass die betroffenen Substanzen von Billigherstellern, die aus Kostengründen keine nennenswerten Forschungs- und Entwicklungsaktivitäten unterhalten, zu sehr geringen Preisen auf den Markt gebracht werden können. Das hat geringfügige Preisvorteile für den Verbraucher, aber exorbitante Umsatznachteile für diejenigen Konzerne der chemischen Industrie zur Folge, welche die aufwändige, aber wichtige Forschung nach neuen Wirkstoffen gegen die diversen Leiden der Menschheit vorantreiben.

Konnte 1960 aus 5.000 neuen Verbindungen ein marktreifes Produkt im Beispiel-Bereich des Pflanzenschutzes entstehen, so waren es um 1985 schon 50.000; und heute müssen mehr als 100.000 neue Substanzen synthetisiert und untersucht werden, bis eine einzige als Produkt den Handel erreicht. Das hat zur Folge, dass die Forschungs- und Entwicklungskosten für jedes neue marktreife Produkt dieser Bereiche einige 100.000.000 € betragen – vor allem deshalb, weil die Kosten für diejenigen Substanzen, die den Markt nicht erreichen, auch gedeckt werden müssen. Das hat vor einigen Jahren die intensive Suche nach effizienteren Ansätzen bei der Identifizierung neuer Wirkstoffe ausgelöst. Das Ergebnis jener Suche ist heute bereits flächendeckend in der Praxis der chemischen Industrie realisiert – sie wird als molekulare Wirkstoff-Forschung bezeichnet. Wie wir sehen werden, hat die heutige molekulare Wirkstoff-Forschung sehr viel mit der Genomforschung zu tun, und die Bioinformatik spielt für ihre Realisierung in zweierlei Hinsicht die Schlüsselrolle.

Biologische Wirkung: das Schlüssel-Schloss-Prinzip

Die allermeisten im chemischen Labor synthetisierten Verbindungen haben überhaupt gar keine biologische Wirkung, weder positiv noch negativ. Die wenigen Substanzen, die eine biologische Wirkung ausüben, tun dies fast ausnahmslos nach dem gleichen Prinzip: Sie interagieren spezifisch mit den wichtigsten Funktionsträgern der Zelle – den Eiweißmolekülen, auch Proteine genannt. Die Zelle ist der Grundbaustein allen Lebens, von den Einzellern, wie den Bakterien, bis hin zum menschlichen Körper, der aus mehr als 10^{13} (10.000 Milliarden) einzelnen Zellen besteht. Ob menschliche Herzmuskelzelle oder Bakterienzelle, die Hauptarbeit der Zelle wird von den Proteinen geleistet. Die einfachsten Bakterienzellen enthalten etwa 1.000 bis 3.000 unterschiedliche Proteine, kompliziertere Bakterien enthalten etwa 5.000 unterschiedliche Proteine. Eine typische Zelle im menschlichen Körper oder in einer Pflanze ist sehr viel größer als eine Bakterienzelle und sehr viel komplizierter aufgebaut, sie enthält um die 10.000 bis 20.000 Proteine.

Um den Wirkmechanismus von chemischen Wirkstoffen zu verstehen, muss man das Wirkungsprinzip von Proteinen in der Zelle verstehen. Um die Sache sehr stark (aber nicht völlig unzutreffend) zu vereinfachen, kann die Zelle mit einer mikroskopisch kleinen, vollautomatisierten Autofabrik verglichen werden, die aus Tausenden von noch kleineren

Einzelmaschinen (den verschiedenen Proteinen) besteht. Diese haben klar definierte und hochspezialisierte Aufgaben (Funktionen) für die Zelle. Das reibungslose Funktionieren jedes einzelnen Proteins ist für das Gelingen des Produktionsablaufs erforderlich. Bleibt an unserem imaginären Auto-Fließband die Funktion auch nur einer Maschine (eines Proteins) aus, z. B. diejenige, die die Reifen auf die Räder montiert, dann kommt bald die gesamte Produktion ins Stocken, weil das Fließband (der zelluläre Stoffwechsel) nicht korrekt funktioniert – die Autos rollen nicht vom Band, sondern stauen sich karambolageartig an der defekten Stelle, und bald kommt die Produktion zum Stillstand. In unserer Analogie würde ein stillstehendes Fließband den Zelltod bedeuten.

Im Idealfall greifen Wirkstoffe ins zelluläre Geschehen dadurch ein, dass sie mit einem einzigen Protein der Zelle nach dem Schlüssel-Schloss-Prinzip spezifisch interagieren, und mit sonst keinem anderen Protein. Hier ein kleines Beispiel anhand des natürlichen Wirkstoffes Penicillin: Typische Bakterien brauchen etwa 100 Proteine, um ihre Zellwand zu synthetisieren. Können sie kein Zellwandmaterial synthetisieren, so können sich ihre Zellen nicht teilen und ihr Wachstum wird unterbunden. Penicillin unterbindet das Wachstum der Bakterien dadurch, dass es hochspezifisch mit einem Protein interagiert, das die Synthese eines einzigen – aber essentiellen – Bestandteils der Zellwand katalysiert.

Die mechanistische Basis dieser hochspezifischen Interaktion besteht darin, dass das Penicillin fast die identische molekulare Struktur hat wie der bakterielle Zellwandbestandteil, den das betroffene Protein als Substrat nach dem Schlüssel-Schloss-Prinzip natürlich zu verarbeiten hat. Das Penicillin wird aber nicht von dem Protein zur Zellwand verarbeitet, sondern es bleibt an dem Protein fest haften, und zwar am so genannten aktiven Zentrum des Enzyms.[2] In unserem imaginären Beispiel der Zelle als Autofabrik ist es so, als hätte Penicillin die gleiche Form wie ein normaler Reifen, jedoch mit fest haftendem Klebstoff beschichtet, so dass die Greifstelle für Reifen irreversibel besetzt und somit die Maschine für die Montage weiterer Reifen außer Gefecht gesetzt wird.

Im Wesentlichen funktioniert jeder Wirkstoff nach dem gleichen Prinzip, wie am obigen Beispiel des Penicillins geschildert: Der Wirkstoff hat eine molekulare Struktur, die der Struktur einer in der Zelle natürlich vorkommenden Substanz stark ähnelt. Aufgrund jener Strukturähnlichkeit bindet der Wirkstoff an dasjenige Enzym der Zelle, das die natürliche Substanz im Stoffwechsel umsetzt. Je fester die Bindung des Wirkstoffs zum Zielenzym (engl. *target*) und je spezifischer diese Wechselwirkung gegenüber anderen Enzymen ausfällt, desto besser ist in der Regel die Wirkung. Die chemisch-strukturelle Grundlage der Interaktion zwischen einem Wirkstoff und dessen Zielenzym ist zugleich der Mechanismus der biologischen Wirkung.

In der traditionellen Wirkstoff-Forschung war man durchaus auch am Mechanismus der Wirkung interessiert, aber der Mechanismus stand eher am Ende der Wirkstoff-Forschung: Man fing mit der Suche nach einer Wirkung an und tastete sich langsam vorwärts, bis man zum Schluss wusste, wo in der Zelle (an welchem Enzym) der Wirkstoff greift. In der molekularen Wirkstoff-Forschung steht hingegen die spezifische Interaktion zwischen Wirkstoff und Enzym am Anfang der Forschung und Entwicklung, nicht am Ende.

---

[2] Proteine, die ein Substrat katalytisch umsetzen, werden als Enzyme bezeichnet. Der Ort am Enzym, wo die Katalyse erfolgt, wird als aktives Zentrum bezeichnet.

Molekulare Wirkstoff-Forschung: Die Suche auf den Kopf gestellt

Die Erforschung der Interaktionen zwischen chemischen Verbindungen und Proteinen ist ein Kerngebiet der Biochemie, das im Laufe der vergangenen Jahrzehnte ein massives Wissen über den Stoffwechsel einzelner Organismen, aber auch über den Aufbau des Stoffwechsels als Ganzes geliefert hat. Wenn heute die Chemiker einen neuen Wirkstoff gegen einen bestimmten Krankheitserreger finden wollen, können sie aufgrund dieses *A-priori*-Wissens von vornherein in etwa sagen, wo im Stoffwechsel (d. h. an welchen Enzymen) sie eingreifen müssten, um das Wachstum des Krankheitserregers zu blockieren. Darüber hinaus wissen sie, wo es bereits bekannte und patentierte Wirkmechanismen gibt, und wo (d. h. bei welchen Enzymen) sie Hoffnung auf neue Mechanismen, Wirkstoff-Klassen und Patente schöpfen können.

Ende des vergangenen Jahrhunderts kamen zur gleichen Zeit viele chemische Unternehmen auf die gleiche Idee. Wäre man imstande, ein bestimmtes Zielenzym eines Krankheitserregers (oder eines Unkrauts) von den übrigen Enzymen der Zelle abzutrennen, und hätte man für das Enzym ein Testverfahren, mit dem man den Stoffumsatz des Enzyms messen könnte (ein Assay), so wäre es doch möglich, die Wirkung von neuen chemischen Verbindungen direkt am gewünschten Wirkort im Reagenzglas in Form einer Veränderung der Enzymaktivität zu messen. Und wäre man imstande, das Zielenzym auch noch in großer Menge zu gewinnen, so könnte man viele Substanzen jeweils einzeln im Reagenzglas testen. Wenn alles gut ginge, könnte man sogar einige 100.000 Substanzen jeweils einzeln im Reagenzglas testen . . .

Aber bevor sich so etwas realisieren lässt, gibt es noch konkrete technische Probleme zu lösen. Bei den meisten Krankheitserregern ist es z. B. nicht möglich, Enzyme in so großer Menge zu gewinnen, dass man auch nur zehn Substanzen testen könnte. Dies liegt daran, dass viele Krankheitserreger entweder gar nicht außerhalb des menschlichen Körpers wachsen, oder aber so langsam, dass man kein ausreichendes Ausgangsmaterial für die Enzymreinigung zur Verfügung hat, oder beides. Wie kann man Enzyme aus einem Organismus gewinnen, den man so gut wie nicht kultivieren kann? Eine Lösung zu diesem Problem bietet die Molekularbiologie und die Biotechnologie: Die Klonierung von Genen und die gentechnische Produktion der dazugehörigen Proteine, was im Folgenden kurz erläutert wird.

Ein Gen ist ein klar definierter Abschnitt der Erbinformation. Es enthält die vollständige Anleitung (Information), die für die Biosynthese eines Proteins erforderlich ist – pro Protein ein Gen. In unserer Analogie der Zelle als Autofabrik mit den Enzymen als Maschinen wären die Gene die Konstruktionspläne für die einzelnen Maschinen. Diese wichtigen Konstruktionspläne liegen nicht ungeordnet in der Zelle herum, sondern gut verwaltet in der Informationszentrale der Zelle, den Chromosomen (mehr dazu später).

Will man heute besonders große Mengen eines bestimmten Enzyms aus einem Organismus erhalten, so geht man nur noch selten den Weg der Enzymisolierung aus dem Organismus selbst, besonders dann nicht, wenn dieser schlecht wächst. Stattdessen nimmt man eine Abkürzung über dessen Gene, die man im Notfall aus einer einzigen Zelle isolieren kann. Man holt sich aus dem schlecht handhabbaren Organismus das Gen für das gewünschte Protein, überträgt das Gen auf einen anderen Organismus, der schnell wächst, und programmiert den neuen Genträger so um, dass er unter geeigneten Kulturbedingungen in wenigen Stunden das gewünschte Protein eimerweise produziert. Noch vor wenigen

Jahren war diese Standardtechnik in Deutschland umstritten, wie der Rechtsstreit der 80er Jahre um die Zulassung für die Produktionsanlage von gentechnisch hergestelltem Humaninsulin der Hoechst AG in Frankfurt belegt. (Heute sind gentechnische Anlagen weit weniger umstritten.)

Für eine solche gentechnische Enzymproduktion nimmt man sehr gerne das Haustier der Genetiker, das Bakterium *Escherichia coli*. *E. coli* wächst sehr gut im Labor, seine Zellen teilen sich unter optimalen Bedingungen etwa alle 20 Minuten. Eine einzige Zelle von *E. coli* wiegt nur ungefähr zwei Milliardstel Gramm (2×10^{-9} Gramm), aber durch die günstige Vermehrungsrate kann man unter Zufuhr ausreichender Nährlösung aus einer einzigen Zelle über Nacht schon viele Kilogramm *E. coli* Zellmasse gewinnen. Meistens reichen jedoch für die Enzymproduktion einige Gramm Zellmasse, wie man sie aus wenigen Litern Nährlosung binnen zwölf Stunden gewinnen kann. Interessanterweise wird das Wachstum von *E. coli* im Labor nur durch die zugeführte Menge an Nährlösung begrenzt. Hätte eine *E. coli*-Kultur unbegrenzt verfügbare Nährlösung, so würde sie aus einer einzigen Zelle binnen 48 Stunden zu einer Zellmasse heranwachsen, die mehr wiegen würde als unsere Erde. Diese Menge an Nährlösung erfordert jedoch so viel Kohlenstoff, wie es in unserem Sonnensystem gibt, und so besteht keine Gefahr, dass wir von *E. coli* überrannt werden.

Für die molekulare Wirkstoff-Forschung braucht man auch gentechnisch ausgebildete Biologen, die das Gen für das ausgesuchte Zielenzym aus dem gewählten Krankheitserreger oder Schädling isolieren. Sie übertragen das Gen auf *E. coli* (oder einen anderen geeigneten Enzym-Produzenten, wie z. B. die Bäckerhefe) und, sofern diese Biologen eine Ausbildung in der Biotechnologie haben, können so an einem Tag sehr viel mehr von dem Enzym aus einer Fermenter-Kultur von *E. coli* gewinnen als die Chemiker für deren Testsysteme benötigen. Das bedeutet, dass ausgebildete Biologen mit Hilfe eines geeigneten Messverfahrens neue Substanzen auf biologische Wirkung schneller testen können als die Chemiker neue Substanzen herstellen können. Streng genommen stimmt das nicht ganz, weil ein Biologe im Labor maximal etwa 1.000 Enzymmessungen (Substanztests) an einem Arbeitstag mit den eigenen zwei Händen pipettieren kann. Aber Roboter können schneller pipettieren, und große Roboter können noch schneller pipettieren . . .

In der chemischen Praxis testen heute PKW- bis LKW-große Pipettier-Roboter mehr als eine Million chemischer Substanzen anhand eines gentechnisch hergestellten Enzyms binnen weniger Tage. Ein solcher Test von vielen Hunderttausenden neuer Substanzen an einem gentechnisch hergestellten Enzym im robotisierten Messverfahren ist der wesentliche Arbeitsschritt der molekularen Wirkstoff-Forschung.

Diejenigen Substanzen, die eine Wirkung (d. h. eine Veränderung der Enzymaktivität) in einem solchen Test zeigen, sind vielversprechende Kandidaten für neue Wirkstoffe gegen das ausgewählte Enzym, und somit gegen den Krankheitserreger, z. B., aus dem das Enzym (bzw. dessen Gen) ursprünglich stammt. Diese Substanzen sind prinzipiell für die weitere Forschung und Entwicklung sehr interessant. Wie in der klassischen Wirkstoff-Forschung müssen die so identifizierten Substanzen auf Unweltverträglichkeit, Spezifität, Abbauverhalten, und, und, und untersucht werden, ehe sie den Weg bis zur Marktreife bestreiten können. Aber mit Hilfe der molekularen Wirkstoff-Forschung haben sich zwei Dinge gegenüber der klassische Wirkstoffsuche grundlegend geändert:

Erstens wissen die Chemiker mehr oder weniger gleich beim ersten Test (d. h. binnen weniger Tage statt nach einigen Jahren), wo im Stoffwechsel (an welchen Enzymen) ihre neuen Substanzen wirken. Darüber hinaus können sich die Chemiker die Enzyme aussuchen, gegen die sie neue Wirkstoffe spezifisch entwickeln möchten. Dieser Faktor verkürzt die Wege bis zur Marktreife eines neuen Produkts um weitere Monate bis Jahre. Aber um diese kürzeren Entwicklungszeiten zu realisieren, brauchen die Chemiker gentechnisch und biotechnisch ausgebildete Biologen, die die Gene klonieren, die Enzyme produzieren, und die Testsysteme bis zur Robotik-Reife etablieren.

Genau das ist die zweite grundlegende Änderung. Die Biologen mit ihren einst umständlichen, langsamen und primitiven biologischen Testsystemen für das Auffinden neuer Wirkstoffe (man denke an das einzelne Auftröpfeln von Millionen neuer Verbindungen auf die Blätter von so vielen verschiedenen Pflanzen im Gewächshaus und das Warten auf eine sichtbare Wirkung!) sind nicht mehr der limitierende Faktor für das Auffinden neuer Wirkstoffe. Auf einmal können die Biologen mit Hilfe der Gentechnik, der Biotechnologie und der Robotik sehr viel schneller neue Substanzen auf ihre Wirkung prüfen als die Chemiker neue Substanzen in ihren eigenen Syntheselaboratorien herstellen können. Die Biologen müssen auf einmal ungeduldig auf den langsamen Zustrom neuer Substanzklassen aus der chemischen Synthese warten. Im Gegenzug haben die Chemiker mittlerweile auch ihre eigenen Fortschritte in der beschleunigten, automatisierten Synthese neuer Verbindungen erzielt. Hier ist das Stichwort *kombinatorische Chemie* zu erwähnen, deren Erklärung jedoch den Rahmen dieses Beitrags deutlich sprengen würde.

Fakt bleibt, dass aufgrund der molekularen Wirkstoff-Forschung, die heute weltweit in allen chemischen Großkonzernen praktiziert wird, die Chemie in der Suche nach neuen Wirkstoffen auf eine High-Tech-Biologie angewiesen ist. Das hat es in den 100 Jahren der Großchemie vorher nie gegeben.

Wirkstoff-Forschung, Genomforschung und Bioinformatik – die Verbindung

Jetzt mag sich der Leser fragen, wo denn bei all dieser Technologie um die Entwicklung neuer biologisch wirksamer chemischer Produkte der Beitrag der Bioinformatik geblieben ist. Sie kommt jetzt auf den vorangegangenen Umwegen geradewegs ins Zentrum des Geschehens, und zwar bedingt durch die Genomforschung.

Die Erbinformation einer Zelle (die Gesamtheit der Bauanleitungen zur Synthese aller Proteine) ist in ihren Chromosomen gespeichert. Die Chromosomen bestehen aus der allseits bekannten Desoxyribonukleinsäure, kurz DNA. Seit den 50er Jahren weiß man, dass die DNA aus einer Doppelhelix von komplementären Strängen aufgebaut ist, wobei die zwei Stränge der Doppelhelix jeweils aus einem immer gleich aufgebauten Phosphorsäure-Zucker Rückgrat bestehen, an dem die vier Basen der DNA – Guanin, Adenin, Cytidin und Thymin – gekoppelt sind. Die Basen werden mit den Buchstaben G, A, C und T abgekürzt, wobei G auf dem einen Strang immer mit C auf dem anderen Strang paart, und A immer mit T paart. Seit den 60er Jahren weiß man, dass in der *Abfolge* der Basen der DNA die genetische Information enthalten ist. Die genetische Information ist somit vom Prinzip her sehr ähnlich aufgebaut wie die Information, die moderne digitale Medien in nahezu

endlosen Abfolgen aus 0 und 1 speichern. Der wichtigste Unterschied: digitale Medien sind breit und flach, DNA ist lang und dünn (mehr dazu später).

Erst in den späten 70er Jahren haben Biochemiker Techniken entwickelt, die es erlauben, die Abfolge der Basen – die *Sequenz* – einer DNA-Doppelhelix zu lesen. Seitdem sind Biologen fleißig dabei, die Abfolge der Basen von diversen klonierten Genen zu bestimmen, d. h. Gene zu sequenzieren. Diese Sequenzinformation ist unbedingt erforderlich, will man ein kloniertes Gen auf *E. coli* zwecks Produktion des im Gen kodierten Proteins übertragen (z. B. im Sinne der Wirkstoff-Forschung). Die DNA-Sequenzierungs-Technologie wurde seit ihrer Entdeckung stets verbessert. Konnte 1980 ein Biologe im Labor maximal etwa 500 Basen an neuer Sequenzinformation pro Woche erstellen, so konnte er 1990 schon etwa 1.000 Basen pro Tag – die Länge eines durchschnittlichen Gens – sequenzieren.

Während 1990 die Sequenzierung eines Gens nur einen Tag dauerte, so konnte die *Ab-initio*-Isolierung und Klonierung eines gesuchten Gens – eine Voraussetzung für deren Sequenzierung – mehrere Wochen, Monate oder gar Jahre dauern. Leider ist das auch bis heute noch der Fall. Es dauert um viele Größenordnungen länger, unter den 1.000 bis 30.000 Genen eines Organismus ein gesuchtes Gen gezielt zu isolieren, als es dauert, das gefundene Gen zu sequenzieren. Auf die Wirkstoff-Forschung bezogen bedeutet das, dass je nach Enzym der zeitlich limitierende Faktor bei der Entwicklung eines robotisierten Testsystems in der Genklonierung liegen kann.

Die eigentlichen Schwierigkeiten, die bei der Isolierung und Klonierung eines gesuchten Gens entstehen, sind viel zu langwierig und uninteressant, als dass sie hier behandelt werden könnten. Man kann aber sagen, dass die Schwierigkeiten im Wesentlichen darin liegen, dass sich die verschiedenen Gene (DNA-Abschnitte) in den Chromosomen eines Organismus nur in ihren Sequenzen unterscheiden, nicht aber in ihren chemischen oder physikalischen Eigenschaften. Man kann alle erdenklichen physikalischen Parameter an einem DNA-Abschnitt messen, aber es nutzt letztendlich nichts: Einzig und allein die Sequenz der DNA gibt direkt darüber Auskunft, welches Gen sie enthält und somit, welches Protein sie kodiert. Hier hilft eine gute Analogie zum Verständnis: Die Brockhaus Enzyklopädie hat etwa 16.000 Seiten, wie manche Organismen etwa 16.000 Gene haben. Die Seiten des Brockhaus werden aus dem Einband herausgetrennt und in zufälliger Reihenfolge sowie Orientierung wieder zusammengestellt, genau wie die zufällige Anordnung der Gene in einem Chromosom. Nun besteht die Aufgabe darin, diejenige Seite im Brockhaus gezielt zu identifizieren, auf der Karl der Große (hier das gesuchte Gen) behandelt wird, *ohne die Seiten lesen zu dürfen.*

Diese Aufgabe hört sich unlösbar an, ist sie aber nicht. In der molekularbiologischen Alltagspraxis macht man das so, dass man den Brockhaus kloniert; d. h., man überträgt die herausgetrennten Seiten auf *E. coli* so, dass jede von 16.000 verschiedenen *E. coli* Zellen genau eine Seite enthält. Dabei benutzt man einen absolut nichts wissenden, aber durchaus wissbegierigen sowie literaten *E. coli*-Stamm. Nachdem die Zellen ein paar Stunden Zeit gehabt haben, jeweils ihre Seite gründlich zu lesen, fragt man alle 16.000 emphatisch auf einmal: „*Wer hat die Sachsen romanisiert und wie war seine Einstellung zur Bildung?*“ Der korrekt antwortenden Zelle wird die Seite entnommen, und da hat man das Gesuchte. (Die zweite Hälfte der Frage ist wichtig, sonst antwortet auch die Zelle mit der Seite über die Sachsen.)

Im Jahre 1995 zeichnete sich eine sehr überraschende Lösung ab, was die aufwändige und zeitraubende Suche nach dem einzelnen Gen betrifft. Schon Ende der 80er Jahre hatten viele darüber nachgedacht, dass es möglich sein sollte, nicht nur einzelne Gene, sondern aus Tausenden von Genen bestehende ganze Genome zu sequenzieren. Diverse Forschergruppen schlossen sich zu Konsortien zusammen und schmiedeten Pläne, wie man eine vollständige Genomsequenzierung (heute sagt man ein Genomprojekt) organisieren und durchführen solle. Es gab heftige Strategie-Debatten und herbe Auseinandersetzungen darüber, welches Genom zuerst sequenziert werden sollte, aber alle waren sich über eines einig: Man müsse als erstes einen kompletten Satz von überlappenden Abschnitten des Genoms herstellen und diese in der absolut korrekten Anordnung vorliegen haben (man sagt, ein Genom *kartieren*), bevor man mit der Sequenzierung auch nur einer einzigen Base anfängt. Erst danach könne man sequenzieren, was aus technischen Gründen immer in Portionen zu etwa 800 aufeinander folgenden Basen erfolgt. Anschließend setze man die Teilsequenzen gemäß der Kartierung Stück für Stück der Reihenfolge nach zur Genomsequenz zusammen. In unserem Beispiel des Brockhaus hieße die Kartierung, erst einmal alle 16.000 losen und unnummerierten Blätter in der richtigen Orientierung und Reihenfolge hinzulegen, ehe man damit anfängt, den Text zu lesen. (So etwas ist mühsame Arbeit, aber es geht.) Und so fingen alle Konsortien an, ihre auserwählten Genome zu kartieren.

Ein Forscher war jedoch entschieden dagegen, die Kartierung als unabdingbare Voraussetzung der Genomsequenzierung zu akzeptieren – der ungeduldige und querdenkende Amerikaner Craig Venter. Sein Argument war es, dass die vorherige Kartierung nicht nur kostspielig und zeitraubend sei, sondern auch noch völlig überflüssig. Seine eigene Alternativstrategie wurde als absurd, unseriös, zu teuer und schlicht nicht machbar verteufelt. Dabei war die Grundidee seiner Strategie doch ganz einfach: Man kümmere sich gar nicht um die Anordnung der kleinen Abschnitte, die man sequenzieren will, sondern man sequenziert blindlings zufällig gewählte Abschnitte aus dem Genom und zwar so lange, bis man eine Menge an DNA sequenziert hat, die sechs bis zehnmal größer ist als das eigentliche Genom (sechs bis zehnmal mehr deshalb, weil alle zufällig sequenzierten Bereiche mit anderen überlappen müssen). Danach – und das war die entscheidende Idee – nehme man einen sehr großen Rechner und ein paar clevere Informatiker, die alle sequenzierten Kleinabschnitte lückenlos mittels der Überlappungen zur kompletten Genomsequenz zusammensetzen, woraus sich die Kartierung der Abschnitte von allein ergäbe. Das Prinzip dieser Strategie kann man anhand der Zusammensetzung eines in zufällig gewählten Auszügen gelesenen Textes verdeutlichen. Die folgenden Textstichproben setze man zum korrekten Text zusammen (Anfang und Ende des Textes überlappen, wie in einem zirkulären Bakterienchromosom).

meinen aufstand unter der führung des widukind die
slawen gesandtes fränkisches heer von den sach
lag der sachsen siebenhundert vierundsiebzig gegen fritz
on den sachsen am süntel vernichtet worden war und kar
tische kraftanstrengung des frankenreiches unter ka
ig ein gegen die slawen gesandtes fränkisches heer von
lar schritt er zur unterwerfung des volkes die in zwei feld
minsäule nach einem gegenschlag der sachsen sieben
d karl sächsische geiseln bei verden hatte hinricht
ches heer von den sachsen am süntel vernichtet worde
reiches unter karl siebenhundert zweiundsiebzig drang er i
den hatte hinrichten lassen kam es zu einem allgemeinen auf
dsiebzig drang er in sachsen ein eroberte die eresburg und
ierundsiebzig gegen fritzlar schritt er zur unter
er der führung des widukind die kriege mit den sachsen ware
n als aber siebenhundert zweiundachtzig ein gegen die sla
achsen ein eroberte die eresburg und zerstörte die irmi
eicht schien als aber siebenhundert zweiundachtzig ein ge

en die größte militärisch-politische kraftanstrengung des frank
ege mit den sachsen waren die größte militärisch-politi
assen kam es zu einem allgemeinen aufstand unter der fü
den war und karl sächsische geiseln bei verden hatte hin
fung des volkes die in zwei feldzügen erreicht schien als ab

Das obige Rätsel ist einfach zu lösen. Aber bei einer bakteriellen Genomsequenz hat man es mit einigen Millionen Buchstaben (Basen) und Zehntausenden von 800 Zeichen langen Textauszügen der vier Buchstaben GACT zu tun. Venters Gegner konstatierten, dass niemand die hierfür erforderlichen Rechneraufgaben bewältigen könne.

Venter bekam letztlich eine Finanzierung seines Vorhabens, und als die Konsortien noch mitten in ihren mühsamen Genomkartierungs-Vorarbeiten steckten, veröffentlichte Venters Team, damals am amerikanischen TIGR (The Institute for Genome Research), 1995 die erste Genomsequenz eines Bakteriums, das des humanpathogenen *Haemophilus influenzae*, mit rund 2.000.000 Basen und etwa 2000 Genen[3]. Damit war klar, dass Venters Strategie der Sequenzierung von zufällig ausgewählten DNA-Abschnitten (engl. *shotgun strategy*) mit der anschließenden Anordnung und Zusammensetzung am Rechner (Sequenz-Assemblierung) der schnellere und letztendlich auch kostengünstigere Weg war. In rascher Folge kamen weitere Genomsequenzen, die ersten alle von TIGR, immer schneller, immer größer, gekrönt von der Veröffentlichung der Sequenz der menschlichen Chromosomen in 2001[4]. Mit 2,9 Milliarden Basen ist das Humangenom mehr als tausendmal größer als das von *Haemophilus*, aber die Fertigstellung hat weniger als zehnmal so lange gedauert. Absurderweise wurde das Humangenom gleich zweimal sequenziert, einmal in öffentlich geförderter Arbeit und einmal von privater Hand finanziert durch die amerikanische Firma Celera, deren Chef Craig Venter heißt. Der entscheidende Schritt in 1994-95, die aufwändige Genomkartierung mit einer *A-posteriori*-Assemblierung der Genomsequenz durch biologisch versierte Informatiker (oder informatisch versierte Biologen) an Hochleistungsrechnern vorzunehmen, war nicht unbedingt die Geburtsstunde des Fachgebiets der *Bioinformatik*, weil Biologen schon seit Jahrzehnten am Rechner arbeiten. Aber es war wohl die Geburtsstunde des *Begriffs* der Bioinformatik, denn man erkannte auf einmal, dass mit bioinformatischer Technologie viele Monate oder Jahre der Laborarbeit eingespart werden können.

Mittlerweile wurden über 80 verschiedene Genome sequenziert,[5] über Bakterien, Pilze, Pflanzen und Tiere bis hin zum Menschen, und viele weitere werden folgen. Aber in den Datenbanken und in den Schubladen der chemischen Industrie liegen viele weitere unveröffentlichte Genomsequenzen. Niemand weiß genau, welche und wie viele es sind; dies sind gut gehütete Betriebsgeheimnisse.

Aber wieso sequenzieren Chemiekonzerne Genome?

Weil es mittlerweile sehr viel schneller geht und nicht viel teuerer ist, ein ganzes Genom mit 3.000 Genen zu sequenzieren, als ein einziges spezifisches Gen gezielt zu klonieren und zu sequenzieren. Und wenn man die Sequenz des ganzen Genoms kennt, kann man binnen Wochen je nach Strategie und Bedarf beliebig viele Zielproteine für Testverfahren in der molekularen Wirkstoff-Forschung heranziehen. Dies geht um Größenordnun-

---

[3] Fleischmann *et al.* (1995).

[4] International Human Genome Sequencing Consortium (2001); Venter *et al.* (2001).

[5] Siehe hierzu u. a. `http://www.tigr.org/tdb/` (TIGR) oder `http://www.jgi.doe.gov/JGI_microbial/html/` (DOE Joint Genome Insitute).

gen schneller als die gezielte Klonierung und Sequenzierung einzelner Gene. Mehr noch: Substanzen, die bei Enzym X keine Hemmwirkung zeigen, können durchaus bei Enzym Y hemmen; man kann die gleichen neuen Substanzen bei beliebig vielen verschiedenen Enzymen testen.

Diese Art der genomforschungsbasierten, robotikgestützten Suche nach biologischer Wirkung bei neuen Substanzen hat neben der Bezeichnung molekulare Wirkstoff-Forschung einen weiteren Namen: *high throughput screening*, abgekürzt HTS.

Fazit: Die Genomsequenzierung ist ein integraler Bestandteil der molekularen Wirkstoff-Forschung in den lebenswissenschaftlichen Produktbereichen der chemischen Industrie – dabei ist die Bioinformatik im wahrsten Sinne des Wortes der Schlüssel zum Erfolg der heutigen Genomsequenzierung.

Was macht eigentlich ein Bioinformatiker?

Die Genomsequenzierung ist binnen weniger Jahre eine kleine, sehr schlagkräftige Industrie geworden. Die Hochleistungs-Sequenzierzentren wie TIGR (Maryland, USA), das Kaszua Research Institute in Japan, das Sanger Centre in Großbritannien oder das US Department of Energy Joint Genome Institute in Kalifornien können sehr effizient sequenzieren. Konnte 1980 ein Biologe im Labor etwa 500 Basen an neuer Sequenzinformation pro Woche erstellen, so können die heutigen Genomzentren jeweils etwa 500 Basen pro Sekunde erzeugen; 24-7, wie die Amerikaner sagen (24 Stunden am Tag, sieben Tage pro Woche). Das entspricht etwa einer 600.000fachen Steigerung der Effizienz. Für die Sequenzierung des *Haemophilus*-Genoms bräuchte man heute nicht Jahre wie damals, sondern einen halben Tag (bei korrekter Finanzierung, versteht sich).

Dazu benötigt man gut ausgebildete Bioinformatiker, die alle Gene in der Sequenz identifizieren, den Namen des kodierten Enzyms in die Genom-Datei hineinschreiben und weitere wichtige Informationen aus der Genomsequenz extrahieren; dann entsteht aus einer endlosen Abfolge von Basensymbolen wie ACAGGACCCTTGTGTTGGACAGACAGCTA... eine echte, für den Biologen und Chemiker interpretierbare und nützliche *Information*.

Um die Arbeit der Bioinformatiker in einem Genomprojekt zu veranschaulichen, nehmen wir das obige Textbeispiel noch einmal. Die Rohsequenzen kommen etwa so aus den Sequenzierautomaten:

```
01 igeingegendieslawengesandtesfränkischesheervon
02 larschritterzurunterwerfungdesvolkesdieinzweifeld
03 minsäulenacheinemgegenschlagdersachsensieben
04 egemitdensachsenwarendiegrößtemilitärisch-politi
05 minsäulenacheinemgegenschlagdersachsensieben
06 reichesunterkarlsiebenhundertzweiun
07 dkarlsächsischegeiselnbeiverdenhattehinricht
08 chesheervondensachsenamsüntelvernichtetworde
09 endiegrößtemilitärisch-politischekraftanstrengungdesfrank
10 reichesunterkarlsiebenhundertzweiundsiebzigdrangeri
11 tischekraftanstrengungdesfrankenreichesunterka
12 denhattehinrichtenlassenkameszueinemallgemeinenauf
13 widukinddiekriegemitdensachsenware
14 igeingegendieslawengesandtesfränkischesheervon
```

```
15 larschritterzurunterwerfungdesvolkesdieinzweifeld
16 dsiebzigdrangerinsachseneinerobertedieeresburgund
```

Die überlappenden Sequenzen werden identifiziert, . . .

```
01 igeingegendieslawengesandtesfränkischesheervon
02 larschritterzurunterwerfungdesvolkesdieinzweifeld
03 minsäulenacheinemgegenschlagdersachsensieben
04 egemitdensachsenwarendiegrößtemilitärisch-politi
05 minsäulenacheinemgegenschlagdersachsensieben
06 reichesunterkarlsiebenhundertzweiun
07 dkarlsächsischegeiselnbeiverdenhattehinricht
08 chesheervondensachsenamsüntelvernichtetworde
09 endiegrößtemilitärisch-politischekraftanstrengungdesfrank
10 reichesunterkarlsiebenhundertzweiundsiebzigdrangeri
11 tischekraftanstrengungdesfrankenreichesunterka
12 denhattehinrichtenlassenkameszueinemallgemeinenauf
13 widukinddiekriegemitdensachsenware
14 igeingegendieslawengesandtesfränkischesheervon
15 larschritterzurunterwerfungdesvolkesdieinzweifeld
16 dsiebzigdrangerinsachseneinerobertedieeresburgund
```

. . . miteinander eingerastert[6] . . .

```
widukinddiekriegemitdensachsenware
             egemitdensachsenwarendiegrößtemilitärisch-politi
                          endiegrößtemilitärisch-politischekraftanstrengungdesfrank
                                                tischekraftanstrengungdesfrankenreichesunterka
                                                                                reichesunterkarlsiebenhundertzwe
```

. . . und anhand der Überlappungen zu einer einzigen Sequenz zusammengefügt. . . .

```
widukinddiekriegemitdensachsenwarendiegrößtemilitärisch-politischekraftanstrengungdesfrankenreichesunterkarlsiebenhundert
```

Dann werden die einzelnen Gene mit ihrem Anfang und Ende identifiziert, damit man die in der Sequenz kodierten Proteine erkennen kann . . .

```
   ... Widukind. Die Kriege mit den Sachsen waren die größte militärisch-politische Kraftanstrengung des Frankenreiches un-

ter Karl. Siebenhundert ...
```

. . . und schließlich annotiert, d. h. in einzelne Dateien bzw. Datenbanken geschrieben, die wesentliche Informationen zum Gen und zum kodierten Protein enthalten, damit Chemiker und Biologen mit der Information effizient arbeiten können:

```
*
Organismus: Brockhaus Wiesbaden, Stamm 1970
Position im Genom: Band 9, Seite 762, linke Spalte, 3. Abs., 1. Satz
Kenngrößen; 14 Wörter, 114 Zeichen, davon 100 Standardbuchstaben, 13 Leerzeichen, 1 Sonderzeichen.
Funktionsklasse: Politische Geschichte, Mitteleuropa, 6. - 10. Jahrhundert
Verknüpfunfgen: Frankreich; Sachsen; Widukind; Irminsäule; dt. Kaiser; Pippin; Ludwig der Fromme; Karolinger; Einhard
Synonyme Bezeichnungen: Karl I; Karl der Große; Carolus Magnus; Carlo Magno; Charlemange
Ähnlichkeit: eng verwandte Sätze in Meiers Konversionslexikon und Encyclopaedia Britanica vorhanden
Hinweis: Karl der Große wurde nicht in Aachen zum König gekrönt, sondern in der französischen Kleinstadt Noyon
Funktion: Herstellung kausalen Zusammenhangs zwischen der Ostpolitik Karl I. und Ressorcenverbrauch im Frankenreich.
..
—- Die Kriege mit den Sachsen waren die größte militärisch-politische Kraftanstrengung des Frankenreiches unter Karl. —-

//
```

[6] Statt miteinander „eingerastert“ sagt man (leider) auch miteinander „alignt“, ausgesprochen „alleint“, aus dem Englischen *alignment*=Einrasterung.

Selbstverständlich sieht eine wirkliche Gen-Annotierung etwas anders aus, so z. B.:

```
LOCUS AF216300 852 bp DNA BCT 16-FEB-2000
DEFINITION Escherichia coli 4-diphosphocytidyl-2C-methyl-D-erythritol kinase
           (ispE) gene, complete cds.
ACCESSION  AF216300
VERSION    AF216300.1 GI:6851368
KEYWORDS   isoprenoid non-mevalonate pathway; terpenoids
SOURCE     Escherichia coli; Bacteria; Proteobacteria; gamma subdivision; Enterobacteriaceae;
REFERENCE  1 (bases 1 to 852)
  AUTHORS  Luettgen,H., Rohdich,F., Herz,S., Wungsintaweekul,J., Hecht,S.,
           Schuhr,C.A., Fellermeier,M., Sagner,S., Zenk,M.H., Bacher,A. and
           Eisenreich,W.
  TITLE    Biosynthesis of terpenoids: YchB protein of escherichia coli
           phosphorylates the 2-hydroxy group of
           4-diphosphocytidyl-2C-methyl-D-erythritol
  JOURNAL  Proc. Natl. Acad. Sci. U.S.A. 97 (3), 1062-1067 (2000)
  MEDLINE  20122571
FEATURES             Location/Qualifiers
     source          1..852
                     /organism="Escherichia coli"
                     /db_xref="taxon:562"
     gene            1..852
                     /gene="ispE"
     CDS             1..852
                     /gene="ispE"
                     /note="enzyme involved in isoprenoid non-mevalonate
                     pathway"
                     /codon_start=1
                     /transl_table=11
                     /product="4-diphosphocytidyl-2C-methyl-D-erythritol
                     kinase"
                     /protein_id="AAF29530.1"
                     /db_xref="GI:6851369"
                     /translation="MRTQWPSPAKLNLFLYITGQRADGYHTLQTLFQFLDYGDTISIE
                     LRDDGDIRLLTPVEGVEHEDNLIVRAARLLMKTAADSGRLPTGSGANISIDKRLPMGG
                     GLGGGSSNAATVLVALNHLWQCGLSMDELAEMGLTLGADVPVFVRGHAAFAEGVGEIL
                     TPVDPPEKWYLVAHPGVSIPTPVIFKDPELPRNTPKRSIETLLKCEFSNDCEVIARKR
                     FREVDAVLSWLLEYAPSRLTGTGACVFAEFDTESEARQVLEQAPEWLNGFVAKGANLS
                     PLHRAML"
BASE COUNT     189 a 202 c 253 g 208 t
ORIGIN
       1 atgcggacac agtggccctc tccggcaaaa cttaatctgt ttttatacat taccggtcag
      61 cgtgcggatg gttaccacac gctgcaaacg ctgtttcagt ttcttgatta cggcgacacc
     121 atcagcattg agcttcgtga cgatggggat attcgtctgt taacgcccgt tgaaggcgtg
     181 gaacatgaag ataacctgat cgttcgcgca gcgcgattgt tgatgaaaac tgcggcagac
     241 agcgggcgtc ttccgacggg aagcggtgcg aatatcagca ttgacaagcg tttgccgatg
     301 ggcggcggtc tcggcggtgg ttcatccaat gccgcgacgg tcctggtggc attaaatcat
     361 ctctggcaat gcgggctaag catggatgag ctggcggaaa tggggctgac gctgggcgca
     421 gatgttcctg tctttgttcg ggggcatgcc gcgtttgccg aaggcgttgg tgaaatacta
     481 acgccggtgg atccgccaga gaagtggtat ctggtggcgc accctggtgt aagtattccg
     541 actccggtga tttttaaaga tcctgaactc ccgcgcaata cgccaaaaag gtcaatagaa
     601 acgttgctaa aatgtgaatt cagcaatgat tgcgaggtta tcgcaagaaa acgttttcgc
     661 gaggttgatg cggtgctttc ctggctgtta gaatacgccc cgtcgcgcct gactgggaca
     721 ggggcctgtg tctttgctga atttgataca gagtctgaag cccgccaggt gctagagcaa
     781 gccccggaat ggctcaatgg ctttgtggcg aaaggcgcta atctttcccc attgcacaga
     841 gccatgcttt aa
//
```

Die obige Annotierung wird dem Geisteswissenschaftler oder Physiker nicht sehr viel sagen, aber sie enthält wichtige und sehr interessante Informationen, die dem Chemiker und Biologen sofort sagen, was für eine chemische Reaktion das kodierte Protein in der Zelle katalysiert und was für eine biologische Funktion es ausübt. Der Chemiker weiß sogar gleich, was für neue Verbindungen er in letzter Zeit synthetisiert hat, die dieses Enzym vielleicht hemmen könnten und somit als neue Wirkstoffe in Frage kämen. Dahingegen weiß der Biologe gleich, in welchen Labor-Stämmen er die besten Aussichten auf eine gute Enzymproduktion hätte, wollte er viel von dem im Gen kodierten Enzym für ein Testsystem produzieren.

Einer der häufigsten Arbeitsschritte in der Alltagstätigkeit des Bioinformatikers ist der Vergleich einer neu ermittelten Sequenz mit den in den Gendatenbanken gespeicherten, bereits bekannten Sequenzen. Es kann z. B. sein, dass Gene mit ähnlicher oder gleicher Funktion aus anderen Organismen schon bekannt sind. Ist dies der Fall, dann sollte es eine Ähnlichkeit zwischen der neuen Sequenz und der Sequenz eines in der Datenbank existierenden, funktionsgleichen Gens geben. Der Sequenzvergleich mit der Datenbank ist ziemlich die allererste Handlung, die der Mensch im Allgemeinen beim Anblick einer

neuen Sequenz (fast reflexartig) unternehmen möchte. Die Gegenstände, die wir im Alltag wahrnehmen, werden anhand des Vergleichs mit bereits bekannten Gegenständen in die Kategorien des Gedächtnisses eingeteilt (sieht aus wie ein Buch, ist also wahrscheinlich ein Buch; hört sich an wie ein Hund, ist also wahrscheinlich ein Hund, etc.). Ganz analog ist die erste Ebene der Wahrnehmung im Umgang mit Sequenzen der Sequenzvergleich (ist in der Basenabfolge ähnlich zu einer Alkohol-Dehydrogenase, ist also wahrscheinlich eine Alkohol-Dehydrogenase).

In den öffentlich zugänglichen Gendatenbanken sind über zehn Millionen einzelner Gensequenzen mit vielen Milliarden von Basen in ihrer jeweiligen Basenabfolge gespeichert. Hat man im Labor eine neue Sequenz, z. B.

...TGCCCCTCAACGGCAAAACTTAACCTCTTTGTATAACTTACCAATCAG...

ermittelt, möchte man in der Regel als Erstes wissen, welches Protein die neue Sequenz kodiert. Dazu wird die Sequenz mit allen Sequenzen in der Datenbank verglichen, was an einem Großrechner ein paar Sekunden dauert. Als Ergebnis des Vergleichs sähe man u. a. eine Einrasterung (*Alignment*) der Suchsequenz (*Query*) mit einem gefundenen Treffer (*Subject*) aus der Datenbank, z. B.

```
Subject=AF216300
Query   ...TGCCCCTCTACGGCAAAACTTAACCTCTTTGTATAACTTACCAATCAG...
           || |||||| |||||||||||||| || ||| |||| ||||||| ||||
Subject ...TGGCCCTCTCCGGCAAAACTTAATCTGTTTTTATACCTTACCAGTCAG...
```

wobei die in beiden Sequenzen identischen Basen mit "|" gekennzeichnet sind. In diesem Beispiel ist die Sequenz bisher unbekannter Funktion im abgebildeten Bereich zu 85 Prozent identisch mit den Basen 13 bis 60 der Sequenz vom Lokus AF216300, deren Datenbankeintrag soeben abgebildet war. Erstreckt sich eine ähnlich hohe Identität über die gesamte Sequenz, so würde man mit großer Sicherheit sagen können, bei der neuen Sequenz handele es sich z. B. um ein Gen für eine 4-Diphosphocytidyl-2C-methyl-D-Erythritol Kinase. Und schon wüsste man das, was man stets als Erstes wissen will, nämlich, was für eine biologische bzw. chemische Funktion sich hinter der neuen Sequenz verbirgt.

Es kann aber auch sein, dass die neue Sequenz gar kein Protein kodiert. In diesem Falle würde man unter Umständen gar keinen Treffer beim Vergleich mit der Datenbank erhalten. Jeder Organismus hält nämlich einen gewissen Abstand zwischen den Genen in den Chromosomen. Diese Abstandhalter bestehen – wie die Gene – aus DNA, aber deren Sequenz kodiert kein Protein. Bei den Bakterien ist der Abstand zwischen den Genen sehr klein. Ist ein typisches Gen 1.000 Basen lang, so ist der Abstand zwischen den Genen im Bakteriengenom etwa 100 Basen; man spricht von nicht-kodierenden Sequenzen oder nicht-kodierender DNA. Bei den höher entwickelten Organismen wie dem Menschen kann jedoch der Abstand zwischen den Genen sehr groß ausfallen. Für jedes 1.000 Basen lange Gen beim Menschen gibt es ungefähr 100.000 Basen an nicht-kodierender DNA. Nur jede hundertste Sequenz aus dem Genom des Menschen hat etwas mit einem Gen zu tun, 99 Prozent des Humangenoms enthält gar keine richtige Information, es ist lediglich Füllmaterial!

Wozu die ganze nicht-kodierende DNA? Das ist eine gute Frage. Seit den 70er Jahren, also lange vor der Sequenzierung des Humangenoms, war die nicht-kodierende DNA beim Menschen und bei anderen Organismen bekannt. Man rätselte kräftig über ihre Bedeutung.

Ursprünglich war man der Meinung, das alles habe mit der Regulation der Gene zu tun. Gene werden in ihrer Ausprägung reguliert. Nicht jedes Gen soll in jeder Zelle angeschaltet und aktiv sein. Beispielsweise bestehen Haare aus einem Protein, Keratin genannt, das von einem Gen kodiert wird. Haare sollen auf dem Kopf wachsen; in den Zellen der Kopfhaut ist daher das Keratin-Gen angeschaltet. Haare sollen aber nicht auf den Zähnen wachsen; in den Zellen der Zähne ist das Keratin-Gen ausgeschaltet. Die Schaltstellen der Genregulation liegen vorwiegend in der nicht-kodierenden DNA. Sie sind aber insgesamt aus der Sicht ihrer spezifischen Sequenzlänge kürzer als die Gene selbst. Es gab daher auch andere Thesen, z. B., dass die nicht-kodierende DNA irgendwie vor Mutationen schützen könnte, als eine Art Puffersubstanz oder Schutzschicht für die echten Gene. Mittlerweile weiß man, dass die nicht-kodierende DNA nur deshalb da ist, weil sich manche DNA-Sequenzen im Genom von Natur aus vermehren können – man spricht von beweglicher oder *transponierbarer* (versetzbarer) DNA.

Transponierbare DNA kann man sehr gut mit Computerviren vergleichen. Computerviren vermehren sich auf der Festplatte dadurch, dass sie sich in alle beschreibbaren Bereiche der Festplatte hineinschreiben; manchmal beschädigen sie dabei wichtige gespeicherte Information. Sie können auch lange Zeit auf der Festplatte inaktiv bleiben, ehe sie ihre Vermehrungstätigkeit entfalten. Transponierbare DNA verhält sich sehr, sehr ähnlich. Gibt es eine Kopie eines *Transposons* im Genom, kann es sehr bald 10, 100 oder 1.000.000 davon geben. Sie haben von Natur aus die Fähigkeit, sich im Genom auszubreiten. Das menschliche Genom besteht beispielsweise zu fast elf Prozent aus einem einzigen solchen Transposon, dem so genannten ALU-Element, einer etwa 290 Basen langen Sequenz, die an 1.090.000 verschiedenen Stellen im Humangenom vorkommt. Weitere 17 Prozent des Humangenoms bestehen aus einem anderen Transposon, dem so genannten LINE1-Element, das im Durchschnitt rund 900 Basen lang ist und sich an 516.000 verschiedenen Stellen im Humangenom ausgebreitet hat. Solche Transposons sind von Natur aus weder gut noch schlecht. Sie sind ein natürlicher, beweglicher und dynamischer Bestandteil unseres Genoms. Allerdings weiß man, dass manche genetisch bedingten Krankheiten des Menschen darauf beruhen, dass sich im Laufe der Evolution ein ALU-Element oder LINE1-Element in ein wichtiges Gen hineinversetzt hat, dessen Funktion durch das Transposon empfindlich gestört wurde. Insgesamt besteht das Humangenom zu etwa 45 Prozent aus 18 verschiedenen Transposon-Familien, die jeweils an mehr als 2.000 verschiedenen Stellen im Genom vorkommen.

Die transponierbare DNA konfrontiert den Bioinformatiker bei der Assemblierung einer Genomsequenz im *Shotgun*-Verfahren mit ganz erheblichen, scheinbar unüberwindlichen Problemen. In einem vorangegangenen Abschnitt hatten wir gesehen, dass die Assemblierung der zufällig ermittelten Teilsequenzen zu einer zusammengesetzten Genomsequenz mittels der Einrasterung von Überlappungen geschieht. Allein die Sequenz eines ALU-Elements überlappt jedoch mit 1.090.000 verschiedenen Chromosomenabschnitten. Wie schafft man es, die richtigen Sequenzabschnitte zusammenzustellen? Als Craig Venter angekündigt hatte, das Humangenom im *Shotgun*-Verfahren sequenzieren zu wollen, glaubte fast niemand, dass sein Celera-Team mit diesem Problem fertig werden würde. Aber gerade Probleme dieser Dimension sind für Mathematiker und Informatiker interessant, und so hatte der amerikanische Bioinformatiker Eugene Myers bald eine elegante programmiertechnische Lösung gefunden. Um die Bonität dieser Lösung unter Beweis zu stellen, ent-

schied sich das Celera-Team, eine Art Vorlaufprojekt für die Humangenomsequenzierung einzuschieben, nämlich die Sequenzierung des 180.000.000 Basen umfassenden Genoms der Fruchtfliege *Drosophila melanogaster*,[7] das wie das Humangenom zu über 50 Prozent aus Transposonen mit extrem hoher Kopienzahl besteht. Die reine Sequenzierarbeit, d. h. die Erstellung der Rohsequenzdaten, dauerte gerade sechs Monate. Die automatische Assemblierung an Celeras Computer – einem der größten der Welt – dauerte gerade elf Tage, was allerdings sehr viel Rechenarbeit an einer solchen Mammut-Maschine bedeutet. Dass das nach elf Tagen gelieferte Ergebnis der Assemblierung richtig war, konnte man daran erkennen, dass die Genanordnung in den computerassemblierten Chromosomen exakt mit der Genanordnung übereinstimmte, die *Drosophila*-Genetiker in sieben Jahrzehnten klassischer Kreuzungsgenetik mühsamst ermittelt hatten. Heute leitet Eugene Myers die Bioinformatik bei Celera.

Der Trick und die erfinderische Höhe von Myers Lösung zum rechnerischen Umgang mit den lästigen Transposonen ist nicht ganz einfach zu erklären.[8] In grober Annäherung funktioniert es so, dass man zuerst nur solche Sequenzen assembliert, die nur einmal im Genom vorkommen, um ein fest verankertes Gerüst zu erhalten, und an dieses Gerüst werden die lästigen sich wiederholenden Sequenzen nach und nach angefügt, bis alles aufgeht. Im Detail ist das Verfahren jedoch erheblich komplizierter und fängt schon damit an, dass Celeras Biologen die zu sequenzierenden DNA-Abschnitte im Labor so klonieren mussten, dass sie nachher den Anforderungen seines Computerprogramms genügten – ein eindrucksvoller Beleg für die prioritäre Bedeutung der Bioinformatik in der Genomforschung.

Wie klein oder wie groß ist ein Genom?

Ein DNA-Molekül ist wie ein sehr langer, sehr dünner Faden aufgebaut. Wie lang? Wie dünn? Eine *E. coli* Zelle ist ungefähr zwei Mikrometer lang, dafür ist ihr einziges, ringförmiges Chromosom mit anderthalb Millimetern rund tausendmal länger als die Zelle. Eindrucksvoller wird die Betrachtung der menschlichen DNA: Eine einzige menschliche Zelle besitzt 46 lineare Chromosomen, die im ausgestreckten, aneinander gereihten Zustand eine Länge von ca. 3,4 Meter pro Zelle ergeben. Weil der menschliche Körper rund 10^{13} Zellen besitzt, enthält jeder Körper eine Gesamtlänge von DNA-Fäden in der Größenordnung von 34 Milliarden Kilometern. Wie viel sind 34 Milliarden Kilometer? Zum Vergleich ist die Entfernung zwischen der Sonne und der Erde bloß 0,15 Milliarden Kilometer, zwischen der Sonne und dem äußeren Planeten Pluto auch nur sechs Milliarden Kilometer.[9]

---

[7] Adams *et al.* (2000).

[8] Herr Myers kann es selber nicht mit ganz einfachen Worten in wenigen Minuten erklären, wie er mir letztens auf einer kleinen Tagung im Schwarzwald eingestand.

[9] Das Entscheidende an diesem Rechenbeispiel ist, dass Atome und Moleküle sehr klein sind und dass es sehr viele von ihnen gibt. Nimmt man z. B. die Eisenatome in einem Kubikmillimeter Eisen, also ein Stück so groß wie dieses „o“, und reiht man sie perlenschnurartig aneinander, so reicht die Perlenschnur 27 Mal von der Erde bis zum Mond und zurück, oder 55 Mal die einfache Strecke. Atome sind sehr klein, DNA-Fäden sind sehr lang und sehr dünn.

Dabei nimmt das *Volumen* der DNA-Fäden weniger als ein Prozent des Gesamtvolumens in unserem Körper ein. Man kann sich also die Gesamtmenge der DNA, die ein menschlicher Körper enthält, wie ein Wollknäuel in der Größe einer kleinen Melone vorstellen, dessen Fadenlänge allerdings *einhundertmal von der Erde bis zur Sonne und zurück* reicht.

Aber die Genomforscher haben nicht die Sequenz der DNA aller Zellen im menschlichen Körper ermittelt, sondern nur die repräsentative Sequenz einer *einzigen* Zelle, weil (fast) alle Zellen im Körper identische Chromosomen und identische DNA-Sequenzen enthalten. Es ist leicht gesagt, dass die Sequenz des menschlichen Genoms aus drei Milliarden Basen besteht. Aber wie viel ist das genau? Die nachfolgende DNA-Sequenz hat 6.000 Basen, ist sehr klein gedruckt und nimmt daher ca. eine halbe Seite ein:

TCCCCCCCCCCTTCTCTCTGAGAGGGGCTCCTCTCAAAACACACCGCGGTGTGTTGTGTGTTTCGCGCTCTGGACAGATGCAGATAGCGCTCGAGATCGTAGAGACAGATCGCGCTCGAG
ACAGATCGCGCTCGGACAGCTCGCAGATCGCTCGCTGACAGATAGCGCTCGCCACAGATAGAGACAGCTCGCGCTCGCCCAGATCAGATCCACGATAGCAGAGATCCAGTGTGACAGAGA
TTCACAGAGATAGGCGCTGAGAGTCGCGTAAGGCTGAGATCGGCTAGAGATCGGCGCTGAGTCTTACCTGTGTGCCCAAGATAGGGTTTCTCTCCTTCCCCCCCCCCTTCTCTCTGAGAG
GGGCTCCTCTCAAAACACACCGCGGTGTGTTGTGTGTTTCGCGCTCTGGACAGATGCAGATAGCGCTCGAGATCGTAGAGACAGATCGCGCTCGAGACAGATCGCGCTCGGACAGCGAGT
CGCTAGAGCTGTAGTACGATATAAGCGCTTATAGCGCGCTATACGCGCATATATAGCGCGCGCGCCCTCTCTCGGGAGAAATAATATAAAGGCTTCCTGGTGTCGCGATATCGCTATATA
GCGCTAGACGCTATATAGCGCTATACTATAGCGCATAAGACGCTATATATATAGAAAGGCGCGCCGCGCGCGGGTGTGTCCCCACAGAGATATCGTAGTCGGATAGGCGCTGAGAGTCGC
GTAAGGCTGAGATCGGCTAGAGATCGGCGCTGAGTCTTACCTGTGTGCCCAAGATAGGGTTTCTCTCCTTCCCCCCCCCCTTCTCTCTGAGAGGGGCTCCTCTCAAAACACACCGCGGTG
TGTTGTGTGTTTCGCGCTCTGGACAGATGCAGNTAGCGCTCGAGATCGTAGAGACAGATCGCGCTCGAGACAGATCGCGCTCGGACAGCTCGCAGATCGCTCGCTGACAGATAGCGCTCG
CCACAGATAGAGACAGCTCGCGCTCGCCCAGATCAGATCCACGATAGCAGAGATCCAGTGTGACAGAGATTCACAGATAGAGGCTCTCGCTGCAGATGACACAGATGCACAGATAGACAC
AGATTGGTCTCGCTCTGCGTCGCCACACCTCGCCCACAGATAGCTCGCTGACAGATGCGCTCGCACAGATAGGCTCGCACAGATGCGCTGACACAGATAGACACAGATGCGCTCGCGCTC
GCGCACAGATATAGACACAGATATAGGCGCTCTGGACAGATGCAGATAGCGCTCGAGATCGTAGAGACAGATCGCGCTCGAGACAGATCGCGCTCGGACAGCTCGCAGATCGCTCGCTGA
CAGATAGCGCTCGCCACAGATAGAGACAGCTCGCGCTCGCCCAGATCAGATCCACGATAGCAGAGATCCAGTGTGACAGAGATTCACAGAGATAGGCGCTGAGAGTCGCGTAAGGCTGAG
ATCGGCTAGAGATCGGCGCTGAGTCTTACCTGTGTGCCCAAGATAGGGTTTCTCTCCTTCCCCCCCCCCTTCTCTCTGAGAGGGGCTCCTCTCAAAACACACCGCGGTGTGTTGTGTGTT
TCGCGCTCTGGACAGATGCAGATAGCGCTCGAGATCGTAGAGACAGATCGCGCTCGAGACAGATCGCGCTCGGACAGCGAGTCGCTAGAGCTGTAGTACGATATAAGCGCTTATAGCGCG
CTATACGCGCATATATAGCGCGCGCGCCCTCTCTCGGGAGAAATAATATAAAGGCTTCCTGGTGTCGCGATATCGCTATATAGCGCTAGACGCTATATAGCGCTATACTATAGCGCATAA
GACGCTATATATATAGAAAGGCGCGCCGCGCGCGGGTGTGTCCCCACAGAGATATCGTAGTCGGATAGGCGCTGAGAGTCGCGTAAGGCTGAGATCGGCTAGAGATCGGCGCTGAGTCTT
ACCTGTGTGCCCAAGATAGGGTTTCTCTCCTTCCCCCCCCCCTTCTCTCTGAGAAAGGCTTCCTGGTGTCGCGATATCGCTATATAGCGCTAGACGCTATATAGCGCTATACTATAGCGC
ATAAGACGCTATATATATAGAAAGGCGCGCCGCGCGCGGGTGTGTCCCCACAGAGATATCGTAGTCGGATAGGCGCTGAGAGTCGGCATAGAGTCGCTAGAGCTGTAGTACGATATAAGC
GCTTATAGCGCGCTATACGCGCATATATAGCGCGCGCGCCCTCTCTCGGGAGAAATAATATAAAGGCTTCCTGGTGTCGCGATATCGCTATATAGCGCTAGACGCTATATAGCGCTATAC
TATAGCGCATAAGACGCTATATATATAGAAAGGCGCGCCGCGCGCGGGTGTGTCCCCACAGAGATATCGTAGTCGGATAGGCGCTGAGAGTCGCGTAAGGCTGAGATCGGCTAGAGATCG
GCGCTGAGTCTTACCTGTGTGCCCAAGATAGGGTTTCTCTCCTTCCCCCCCCCCTTCTCTCTGAGAGGGGCTCCTCTCAAAACANACCGCGGTGTGTTGTGTGTTTCGCGCTCTGGACAG
ATGCAGATAGCGCTCGAGATCGTAGAGACAGATCGCGCTCGAGACAGATCGCGCTCGGACAGCTCGCAGATCGCTCGCTGACAGATAGCGCTCGCCACAGATAGAGACAGCTCGCGCTCG
CCCAGATCAGATCCACGATAGCAGAGATCCAGTGTGACAGAGATTCACAGATAGAGGCTCTCGCTGCAGATGACACAGATGCACAGATAGACACAGATTGGTCTCGCTCTGCGTCGCCAC
ACCTCGCCCACAGATAGCTCGCTGACAGATGCGCTCGCACAGATAGGCTCGCACAGATGCGCTGACACAGATAGACACAGATGCGCTCGCGCTCGCGCACAGATATAGACACAGATATAG
ACACAGATAGACACAGAGATATTCTCGCTCGCTATAGGCGCTGAGAGTCGCGTAAGGCTGAGATCGGCTAGAGATCGGCGCTGAGTCTTACCTGTGTGCCCAAGATAGGGTTTCTCTCCT
TCCCCCCCCCCTTCTCTCTGAGAGGGGCTCCTCTCAAAACACACCGCGGTGTGTTGTGTGTTTCGCGCTCTGGACAGATGCAGATAGCGCTCGAGATCGTAGAGACAGATCGCGCTCGAG
ACAGATCGCGCTCGGACAGCTCGCAGATCGCTCGCTGACAGATAGCGCTCGCCACAGATAGAGACAGCTCGCGCTCGCCCAGATCAGATCCACGATAGCAGAGATCCAGTGTGACAGAGA
TTCACAGAGATAGGCGCTGAGAGTCGCGTAAGGCTGAGATCGGCTAGAGATCGGCGCTGAGTCTTACCTGTGTGCCCAAGATAGGGTTTCTCTCCTTCCCCCCCCCCTTCTCTCTGAGAG
GGGCTCCTCTCAAAACACACCGCGGTGTGTTGTGTGTTTCGCGCTCTGGACAGATGCAGATAGCGCTCGAGATCGTAGAGACAGATCGCGCTCGAGACAGATCGCGCTCGGACAGCGAGT
CGCTAGAGCTGTAGTACGATATAAGCGCTTATAGCGCGCTATACGCGCATATATAGCGCGCGCGCCCTCTCTCGGGAGAAATAATATAAAGGCTTCCTGGTGTCGCGATATCGCTATATA
GCGCTAGACGCTATATAGCGCTATACTATAGCGCATAAGACGCTATATATATAGAAAGGCGCGCCGCGCGCGGGTGTGTCCCCACAGAGATATCGTAGTCGGATAGGCGCTGAGAGTCGC
GTAAGGCTGAGATCGGCTAGAGATCGGCGCTGAGTCTTACCTGTGTGCCCAAGATAGGGTTTCTCTCCTTCCCCCCCCCCTTCTCTCTGAGAGGGGCTCCTCTCAAAACACACCGCGGTG
TGTTGTGTGTTTCGCGCTCTGGACAGATGCAGCTAGCGCTCGAGATCGTAGAGACAGATCGCGCTCGAGACAGATCGCGCTCGGACAGCTCGCAGATCGCTCGCTGACAGATAGCGCTCG
CCACAGATAGAGACAGCTCGCGCTCGCCCAGATCAGATCCACGATAGCAGAGATCCAGTGTGACAGAGATTCACAGATAGAGGCTCTCGCTGCAGATGACACAGATGCACAGATAGACAC
AGATTGGTCTCGCTCTGCGTCGCCACACCTCGCCCACAGATAGCTCGCTGACAGATGCGCTCGCACAGATAGGCTCGCACAGATGCGCTGACACAGATAGACACAGATGCGCTCGCGCTC
GCGCACAGATATAGACACAGATATAGGCGCTCTGGACAGATGCAGATAGCGCTCGAGATCGTAGAGACAGATCGCGCTCGAGACAGATCGCGCTCGGACAGCTCGCAGATCGCTCGCTGA
CAGATAGCGCTCGCCACAGATAGAGACAGCTCGCGCTCGCCCAGATCAGATCCACGATAGCAGAGATCCAGTGTGACAGAGATTCACAGAGATAGGCGCTGAGAGTCGCGTAAGGCTGAG
ATCGGCTAGAGATCGGCGCTGAGTCTTACCTGTGTGCCCAAGATAGGGTTTCTCTCCTTCCCCCCCCCCTTCTCTCTGAGAGGGGCTCCTCTCAAAACACACCGCGGTGTGTTGTGTGTT
TCGCGCTCTGGACAGATGCAGATAGCGCTCGAGATCGTAGAGACAGATCGCGCTCGAGACAGATCGCGCTCGGACAGCGAGTCGCTAGAGCTGTAGTACGATATAAGCGCTTATAGCGCG
CTATACGCGCATATATAGCGCGCGCGCCCTCTCTCGGGAGAAATAATATAAAGGCTTCCTGGTGTCGCGATATCGCTATATAGCGCTAGACGCTATATAGCGCTATACTATAGCGCATAA
GACGCTATATATATAGAAAGGCGCGCCGCGCGCGGGTGTGTCCCCACAGAGATATCGTAGTCGGATAGGCGCTGAGAGTCGCGTAAGGCTGAGATCGGCTAGAGATCGGCGCTGAGTCTT
ACCTGTGTGCCCAAGATAGGGTTTCTCTCCTTCCCCCCCCCCTTCTCTCTGAGAAAGGCTTCCTGGTGTCGCGATATCGCTATATAGCGCTAGACGCTATATAGCGCTATACTATAGCGC
ATAAGACGCTATATATATAGAAAGGCGCGCCGCGCGCGGGTGTGTCCCCACAGAGATATCGTAGTCGGATAGGCGCTGAGAGTCGGCATAGAGTCGCTAGAGCTGTAGTACGATATAAGC
GCTTATAGCGCGCTATACGCGCATATATAGCGCGCGCGCCCTCTCTCGGGAGAAATAATATAAAGGCTTCCTGGTGTCGCGATATCGCTATATAGCGCTAGACGCTATATAGCGCTATAC
TATAGCGCATAAGACGCTATATATATAGAAAGGCGCGCCGCGCGCGGGTGTGTCCCCACAGAGATATCGTAGTCGGATAGGCGCTGAGAGTCGCGTAAGGCTGAGATCGGCTAGAGATCG
GCGCTGAGTCTTACCTGTGTGCCCAAGATAGGGTTTCTCTCCTTCCCCCCCCCCTTCTCTCTGAGAGGGGCTCCTCTCAAAACACACCGCGGTGTGTTGTGTGTTTCGCGCTCTGGACAG
ATGCAGATAGCGCTCGAGATCGTAGAGACAGATCGCGCTCGAGACAGATCGCGCTCGGACAGCTCGCAGATCGCTCGCTGACAGATAGCGCTCGCCACAGATAGAGACAGCTCGCGCTCG
CCCAGATCAGATCCACGATAGCAGAGATCCAGTGTGACAGAGATTCACAGATAGAGGCTCTCGCTGCAGATGACACAGATGCACAGATAGACACAGATTGGTCTCGCTCTGCGTCGCCAC
ACCTCGCCCACAGATAGCTCGCTGACAGATGCGCTCGCACAGATAGGCTCGCACAGATGCGCTGACACAGATAGACACAGATGCGCTCGCGCTCGCGCACAGATATAGACACAGATATAG
ACACAGATAGACACAGAGATATTCTCGCTCGCTATAGGCGCTGAGAGTCGCGTAAGGCTGAGATCGGCTAGAGATCGGCGCTGAGTCTTACCTGTGTGCCCAAGATAGGGTTTCTCTCCT

Entsprechend der Realität enthält die obige Sequenz auch noch zwei „N“s, d. h. Positionen, an denen unsicher ist, ob dort ein G, ein A, ein C oder ein T hingehört, wie es etwa für jede dreitausendste Base in der Humangenomsequenz der Fall ist. Die richtige Identität dieser überaus lästigen Ns zu ermitteln erfordert leider fast so viel Aufwand wie die Erstellung der Genomsequenz selbst. Ferner besteht die obige Sequenz aus zwei identischen, aber ineinander gesetzten Kopien einer einzigen, 3.000 Basen langen Sequenz, wie es häufig bei Transposons vorkommt. Die 3.000er Einheiten haben auch noch jeweils ihre eigene innere Struktur von kürzeren Wiederholungen identischer Sequenzabschnitte, die ein geschultes Auge leicht erkennt.

Druckt man nun die Sequenz des menschlichen Genoms im obigen Textformat aus, so braucht man 125.000 vollständig und beidseitig bedruckte DIN-A4-Seiten. Aber im Gegensatz zum 16.000seitigen Brockhaus enthielte der 125.000seitige Text keine für den Menschen *unmittelbar* verständliche oder nützliche Information. Die erforderliche Übersetzungsarbeit haben Bioinformatiker geleistet.

Bioinformatiker haben aus unzähligen kleinen, zufällig sequenzierten Abschnitten zu je 800 Basen (ca. drei der o. a. Zeilen) die 125.000 Seiten des menschlichen Genoms ohne externe Orientierungshilfen korrekt zusammengesetzt – eine sehr beachtliche Leistung. Danach haben Bioinformatiker die in der Genomsequenz verschlüsselte Information so extrahiert und aufbereitet, dass jeder Forscher diese Information sofort versteht. Die öffentlich finanzierte Genomsequenz ist kostenlos in den Datenbaken verfügbar. Die von Celera kommerziell finanzierte Genomsequenz enthält weniger Fehler, ist vollständiger, wahrscheinlich besser annotiert und wird sehr teuer lizenziert. Die Formulierung „wahrscheinlich besser annotiert“ hat einen Grund: Anhand der Genomsequenz weiß man immer noch nicht, ob der Mensch 30.000, 40.000 oder noch mehr Gene besitzt. Es ist alles andere als trivial, Gene aus einer Genomsequenz korrekt zu identifizieren. Trotz der Veröffentlichung des menschlichen Genoms steht bei den Bioinformatikern noch sehr viel Arbeit mit der Analyse dieser Sequenzdaten an.

Industrielle Bioinformatik: Genomforschung, aber auch Strukturbiologie

Es gibt zwei große Anwendungsgebiete der Bioinformatik in der Industrie. Das eine Gebiet haben wir bereits gesehen, die Genomforschung. Das andere kommerziell wichtige biologische Anwendungsgebiet der Informatik in der Industrie ist auch in der chemisch-pharmazeutischen Industrie zu finden, auch im Bereich der Produktentwicklung, aber der Fokus liegt nicht auf der Genomforschung. Es ist die molekulare Strukturforschung, auch Strukturbiologie genannt. Hier soll dieses sehr wichtige Gebiet in aller Kürze vorgestellt werden.

Die Geschichte der Strukturforschung ist fast 100 Jahre alt. 1914 erhielt der deutsche Physiker Max von Laue den Nobelpreis für seine Entdeckung der Beugung von Röntgenstrahlen durch Kristalle.[10] Das war im Übrigen gerade 13 Jahre, nachdem der deutsche Physiker Wilhelm Conrad Röntgen den Nobelpreis für die Entdeckung der nach ihm benannten Strahlen erhalten hatte. Kristalle beugen die Röntgenstrahlen so, dass man aus dem Beugungsmuster mit recht einfachen Methoden auf die Anordnung der einzelnen Atome in dem bestrahlten Kristall schließen kann. Aus dem im Röntgenfilm festgehaltenen Beugungsmuster eines Kristalls eines chemisch einfach aufgebauten Minerals, wie z. B. dem Kochsalz, kann ein Studierender binnen weniger Stunden die präzise Lage und Anordnung der einzelnen Natrium- und Chloratome im Kristall mit Bleistift, Papier, Geometriedreieck und Rechenschieber bestimmen (vorausgesetzt, er oder sie weiß, wie das geht). Er oder sie kann dann aus Styroporkugeln o. ä. ein Modell des Kristalls bauen, das dem Kristallologen erlaubt, die atomare Organisation der Materie naturgetreu milliardenfach vergrößert in der greifbaren Dimension des makroskopisch Sichtbaren zu betrachten.

---

[10] Siehe http://www.nobel.se/physics/laureates/1914/.

Die Voraussetzungen für diesen Prozess der *Röntgenstrukturanalyse* sind ein Röntgenbeugungsapparat (den kann man kaufen) und Kristalle der Substanz, die man untersuchen möchte. Die Erzeugung von Kristallen erfordert wiederum, dass die zu kristallisierende Substanz in besonders reiner Form und in ausreichender Menge (z. B. ein paar Milligramm) vorliegt.

Proteine (Enzyme) kann man reinigen und kristallisieren. Deshalb kann man schon seit einigen Jahrzehnten auch die *atomare Struktur von Proteinen* mittels Röntgenstrukturanalyse sehr präzise ermitteln. Aber anders als bei Kochsalzkristallen, deren zwei verschiedene Atome besonders einfach angeordnet vorliegen, sind die Atome in einem Protein sehr viel komplexer angeordnet, und es gibt zudem sehr viele Atome pro Protein. Ein durchschnittliches Protein, das von einem 1.000 Basen langen Gen kodiert wird, enthält um die 15.000 Atome: in erster Linie Kohlenstoff, Sauerstoff, Stickstoff, viel Wasserstoff und etwas Schwefel, manchmal auch Metalle wie Eisen, Magnesium, Nickel oder Calcium, in seltenen Fällen auch exotische Elemente wie Wolfram, Vanadium, Kobalt oder Selen. Will man von dem komplexen Röntgenbeugungsmuster eines Proteinkristalls auf die feinaufgelöste, atomare Struktur des kompliziert gefalteten Proteinmoleküls schließen, ist es außerordentlich hilfreich, anstelle des Rechenschiebers einen Hochleistungscomputer zu besitzen und eine Person in der Nähe zu haben, die mit dem Computer zwecks Strukturaufklärung kompetent umgehen kann. Obwohl dieses Anwendungsgebiet der Informatik in der chemisch-biologischen Strukturforschung um Jahrzehnte älter ist als die Genomforschung, wird es heute zum allgemeinen Fachgebiet der Bioinformatik gezählt. Die Universität Düsseldorf kann im Übrigen durch die gemeinsamen Berufungen mit dem Forschungszentrum Jülich eine recht erfolgreiche Strukturbiologie vorweisen.

Die aufgeklärte Struktur eines Enzyms kann ein extrem wertvolles Hilfsinstrument für die Entwicklung neuer chemisch-pharmazeutischer Produkte sein. Der Grund dafür ist schnell erzählt. In den vorangegangenen Abschnitten zur molekularen Wirkstoff-Forschung haben wir gesehen, dass die mechanistische Grundlage der biologischen Wirkung einer neuen Substanz in der molekular-atomaren Interaktion zwischen der chemischen Substanz und ihrem Zielprotein nach dem Schlüssel-Schloss-Prinzip liegt. Dabei ist das aktive Zentrum des Proteins das Schloss, der Wirkstoff der Schlüssel. Kann man die atomare Struktur eines Zielproteins aus einem Röntgenbeugungsmuster am Computer aufklären, so gewinnt man einen direkten Einblick in die innere und äußere Gestalt des Schlosses. Mit dieser Information hat der Chemiker prinzipiell die Möglichkeit, *ab initio* eine neue Substanz zu synthetisieren, die wie ein maßgeschneiderter Schlüssel in das Schloss passt.

So etwas gelingt jedoch eher selten, weil Proteine keine starren, sondern dynamische, biegsame, geschmeidige Moleküle sind. Die eigentliche Gestalt (Konformation) des aktiven Zentrums (des Schlosses) am Protein wird in der Regel erst dann erreicht, wenn das natürliche Substrat oder ein Strukturanalogon dessen, z. B. ein Wirkstoff, gebunden ist (d. h., wenn der Schlüssel steckt). Man spricht dann von einer induzierten Konformationsänderung. Daher möchte man für die Produktentwicklung idealerweise das Zielprotein mit dem korrekt gebundenen Substrat oder mit einem korrekt gebundenen Wirkstoff zusammen kristallisieren, die so genannte Kokristallisation erreichen.

Gelingt dies, dann ist der synthetische Chemiker sehr glücklich, weil er ganz genau am Bildschirm oder am Modell sehen kann, welche reaktiven Atome oder chemische Bindungen in der Nähe des gebundenen Wirkstoffes angeordnet sind. Deren Eigenschaften

kann er wiederum ausnutzen, um mittels wissensgesteuerter, gezielter Modifikation des Wirkstoffes seine Bindung zum Protein maßgeschneidert zu optimieren. In dieser Phase der Produktentwicklung spricht man vom *molecular modelling* oder vom *drug design*. Hierin liegen ganz hervorragende Möglichkeiten, neue, biologisch wirksame Substanzen schnell und gezielt in der Produktentwicklung zur Marktreife zu bringen.

Zusammenfassend kann man festhalten, dass die zwei wichtigsten kommerziell-industriellen Anwendungsgebiete der Bioinformatik in der Genomforschung und in der Strukturbiologie liegen. Beide Anwendungsgebiete tragen entscheidend dazu bei, dass die Forschungs- und Entwicklungszeiten neuer biologisch wirksamer Produkte aus der chemischen Industrie verkürzt werden.

Ein neues, interdisziplinäres Fach mit hervorragenden Berufsaussichten

In der Zeit vor der Genomforschung gab es (auch) einige wenige Bioinformatiker. Die meisten davon waren Biologen, die durch Zweitstudium, Selbststudium oder *trial and error* den Umgang mit dem Computer, die Grundzüge einer Programmiersprache und ein paar einfache Anwendungen in der Verarbeitung von Sequenzdaten gelernt hatten. Informatiker, die ein Interesse für die Biologie entwickelten, waren sehr selten zu finden, weil sich ausgebildete Informatiker ihre Arbeitsstelle aussuchen konnten (und noch können). Warum sollte man mit (damals wertlosen) DNA-Sequenzendaten arbeiten, wenn man auch mit den Daten von Bankkonten, Versicherungssummen oder Fahrzeugproduktionsabläufen arbeiten kann? Die Folge dieser Entwicklung war, dass es mit dem Anbruch der Genomforschung vor sechs Jahren so gut wie keine ausgebildeten Spezialisten gab, um die anfallenden Berge an Rohdaten aus der Genomforschung zu verarbeiten, zu archivieren und in eine für die Biologie greifbare Information umzuwandeln. Aus diesem Grunde konnten sowohl a) Biologen mit fundierten Computerkenntnissen als auch b) Informatiker mit fundierten Kenntnissen der Molekularbiologie die positive Spannung zwischen Angebot und Nachfrage bei ihren Vorstellungsgesprächen erfahren. Auch heute ist es so, dass Hochschulabsolventen der oben genannten Berufsqualifikation mit einem Anfangsgehalt auf dem freien Arbeitsmarkt rechnen können, das 30 bis 50 Prozent über dem liegt, was Biologen ohne diese Zusatzqualifikation erwarten dürfen. Wie lange dieses Niveau anhalten wird, kann man nicht voraussagen, aber zurzeit sind die Aussichten der Bioinformatiker auf dem Arbeitsmarkt wirklich rosig.

Ist die Nachfrage zu groß, erhöht sich auf natürlichem Wege das Angebot. In den USA, in Japan und in Europa kann man jetzt vielerorts *bioinformatics*, *computational biology*, *biocomputing*, *bioinformation technology* o. ä. studieren. Weil sich kaum eine Universität einen eigenen Fachbereich für das sehr spezielle und noch nicht bewährte Fach *Bioinformatik* erlaubt, sind diese Studiengänge fast immer zwischen einer eigenständigen Informatik und einer eigenständigen Biologe angesiedelt. Und das ist auch richtig so. Was der Arbeitsmarkt auf absehbare Zeit verlangt, sind a) Biologen mit fundierten Computerkenntnissen oder b) Informatiker mit fundierten Kenntnissen der Molekularbiologie. Daher sind die diversen Studienangebote entweder in der Informatik fest angesiedelt, mit intensiver Zusatzausbildung durch die Biologie, oder umgekehrt. Beides führt zum gewünschten Ziel. In Deutschland konnte man Bioinformatik vor fünf Jahren gar

nicht studieren. Mittlerweile haben 40 verschiedene deutsche Hochschulen formalisierte Studiengänge, Hauptfächer, Nebenfächer, Spezialvertiefungen usw. als attraktives und breit gefächertes Ausbildungsangebot erarbeitet[11].

In Düsseldorf haben wir uns beim Aufbau der Bioinformatik an keinem existierenden Studiengang oder Modell orientiert. Das geht eigentlich auch gar nicht, weil jede Universität ihre infrastrukturellen Spezifika hat. Zudem geben die Anforderungen der bioinformatischen Berufspraxis und Grundlagenforschung einen so deutlichen Rahmen vor, dass binnen weniger Stunden ein halbes Dutzend einsatzfreudiger Informatiker, Mathematiker und Biologen ein sinnvolles Studienangebot zu Papier und auf den Weg bringen konnte. Das Düsseldorfer Lehrangebot in der Bioinformatik ist zurzeit ein sehr herausforderndes, aber breit von der Studierendenschaft angenommenes Nebenfach im Diplomstudiengang (künftig BSc/MSc) *Biologie*. Im Wesentlichen müssen die Düsseldorfer Biologen mit Nebenfach *Bioinformatik* erst zusätzliche Kenntnisse in der Mathematik als Voraussetzung erwerben, dann die gleichen herausfordernden Informatikveranstaltungen absolvieren wie die Mathematiker oder Physiker mit Nebenfach *Informatik* und dabei eine Programmiersprache lernen. Diese Lehrveranstaltungen werden zurzeit von der Mathematik und künftig von der neugegründeten WE Informatik an der Heinrich-Heine-Universität angeboten.

Die Berufsqualifizierung in der Informatik, die die Biologen dadurch erfahren, beinhaltet keinerlei Spezialisierung. Vielmehr erweben sie solide, fundierte Kenntnisse der Nutzung moderner Rechenleistung, die sich genauso gut in der Finanzverwaltung oder Autoindustrie einsetzen lassen wie in der Biologie. Was hierzulande fehlt, sind Hochschulabsolventinnen und -absolventen mit fundierten Kenntnissen der ganz normalen Informatik, was u. a. die neuerliche Green-Card-Debatte auf Ebene der Bundespolitik belegt.

Aus der Biologie heraus müssen dann Düsseldorfer Studierende der Bioinformatik in weiteren Veranstaltungen den konkreten, praxisorientierten Umgang mit biologischen Daten am Computer lernen; zum einen, was die Genomanalyse betrifft und zum anderen, was die Strukturforschung betrifft. Diese insgesamt hart erlernten Kenntnisse tragen die Biologen dann in ihr Hauptfach hinein, z. B. *Genetik*, *Molekularbiologie*, *Mikrobiologie* oder *Ökologie*, wobei die Computerkenntnisse unabhängig von ihrer Anwendung im Hauptfach eine signifikant erweiterte Berufsqualifizierung der Absolventinnen und Absolventen bilden. Künftig wird sich das Lehrangebot in der Düsseldorfer Bioinformatik – und damit die Möglichkeit einer Spezialisierung auf dem Fachgebiet – durch die Etablierung der WE Informatik und durch zwei gemeinsame Berufungen mit dem Forschungszentrum Jülich auf dem jungen, aber sehr stark wachsenden Spezialgebiet der Bioinformatik *sensu strictu* (je ein Lehrstuhl in der Strukturbiologie und in der Genomforschung) deutlich erweitern.

Düsseldorfs Profil in der Bioinformatik kann sich sehen lassen. Mit der gerade erfolgreich abgeschlossenen Berufung von Prof. Dr. Arndt von Haeseler vom Max-Planck-Institut für Evolutionäre Anthropologie in Leipzig und mit der Einrichtung zweier Juniorprofessuren für Bioinformatik sieht die Düsseldorfer Bioinformatik-Forschung guten Zeiten entgegen. Allerdings möchte ich hier noch etwas zum Stichwort *Profil*, wie es hochschulpolitisch verwendet wird, sagen. Es wird aus der Politik heraus nicht selten an uns als Hochschullehrer die Erwartung herangetragen, deutsche Universitäten sollen sich in ihrem Forschungsprofil mit Harvard, MIT, Stanford oder ähnlichen Eliteuniversitäten mes-

[11] Siehe hierzu http://www.bioinformatik.de/ ⇒ Research and Education ⇒ Universities in Germany.

sen lassen. Kein Problem. Verlangen wir von jedem Studierenden rund 100.000 € für das Recht zu studieren, so kämen an der Heinrich-Heine-Universität allein für die Biologie jedes Jahr 30 Millionen € für die Profilbildung zusammen. Aber Gott sei Dank will es die deutsche Gesetzgebung, dass auch jene Studierende eine universitäre Ausbildung genießen dürfen, deren Eltern solche Summen nicht aufbringen können. Daher würde ich mich sehr freuen, wenn wir mit Colorado State, Penn State, der University of Alabama, Georgia Tech oder Texas A&M verglichen würden. Das sind solide staatliche Universitäten, die geringe oder keine Studiengebühren verlangen.

Zum Thema Ausbildung möchte ich noch eines anmerken. Rund 40 deutsche Hochschulen können jetzt Ausbildungsprogramme in der Bioinformatik vorweisen, fünf davon haben vom BMBF für die Einrichtung von Studiengängen Summen in Millionenhöhe erhalten (Düsseldorf gehört nicht dazu). Man fragt sich nur, welche *Lehrbücher* sie in ihrer Lehre benutzen? Deutschsprachige Lehrbücher der Bioinformatik gibt es nämlich gar nicht bzw. gab es nicht, bis zum vergangenen Sommer. *Bioinformatik – ein Leitfaden für Naturwissenschaftler* ist im Juli im Birkhäuser Verlag erschienen. Es kostet 22 € und bietet eine sehr konkrete, praxisorientierte Einführung in die Verarbeitung von Gensequenzdaten am Computer. Verfasserin des Buches ist Andrea Hansen, eine überaus fähige und stark motivierte Doktorandin, die gerade meiner Abteilung durch ein hervorragendes Angebot eines süddeutschen Bioinformatik-Unternehmens entrissen wurde. Frau Hansen hat unser Düsseldorfer Praktikumsskript ergänzt, professionell formatiert und nun erfolgreich vermarktet. Profil kann man sowohl in der Forschung als auch in der Lehre schaffen.

Grundlagenforschung und Evolution

Würde man die Bioinformatik für die Wirkstoff-Forschung als „sehr wichtig“ bezeichnen, so müsste das Adjektiv für ihre Beziehung zur Grundlagenforschung „unabdingbar“ lauten. Der fachspezifische Einsatz von Computertechnologie in den Biowissenschaften ist bei genauer Betrachtung eigentlich kein Fachgebiet, sondern eher eine Basistechnologie der Biowissenschaften, wie das Telefonieren im Bereich der Kommunikation. Zurzeit erleben wir einen wahren Boom in der Bioinformatik. Die Goldgräberstimmung wird irgendwann verfliegen, Angebot und Nachfrage werden sich auf ein ausgewogenes Niveau einpendeln. Es fragt sich nur, wann? Ein Blick auf die Entwicklung in der Informatik gibt Anlass zur Annahme, dass sich die guten Berufsperspektiven für Bioinformatiker sehr lange halten werden. Informatik kann man schon seit Jahrzehnten vielerorts studieren, aber trotzdem kann der Bedarf an ausgebildeten Informatikern auf dem deutschen Arbeitsmarkt nicht gedeckt werden. Deshalb werden vermutlich viele Studierende, die sich heute für Bioinformatik entscheiden, in ihrem späteren Berufsalltag sehr wenig mit Biologie zu tun haben. Wer sich breit und umsichtig qualifiziert, hat nachher die besseren Chancen. Wenn sich die Nachfrage an Spezialisten im engeren Gebiet der Bioinformatik einpendelt, wird dies aus einem einfachen Grund auf hohem Niveau geschehen. Eine grundsätzliche Abkehr von der Bioinformatik zurück zu einer biowissenschaftlichen Forschungslandschaft, die weder genetische Sequenzinformation verarbeitet noch Proteinstrukturen aufklärt, ist für die kommenden 30 Jahre völlig undenkbar.

In der bioinformatischen Grundlagenforschung selbst gibt es nahezu unerschöpfliches Forschungspotential. Das fängt mit ganz einfachen praktischen Dingen wie der Entwicklung kleiner Programme für den Forschungsalltag an, insbesondere verbesserte Verfah-

ren zur effizienteren, empfindlicheren Datenbanksuche. Das bisher am häufigsten benutzte Verfahren, BLAST[12], hat unübersehbare Nachteile, aber es gibt bis heute keine bessere Alternative. Die Originalveröffentlichung von BLAST gehört im Übrigen zu den meistzitierten wissenschaftlichen Arbeiten der letzten 20 Jahre überhaupt. Schaut man von der Gegenwart in die Zukunft, ist es auch auffällig, dass man heute die Verwandtschaft von Proteinen fast ausschließlich mit dem Maß der Sequenzähnlichkeit ihrer Gene definiert. Dabei weiß man schon lange, dass viele Proteine, deren Gene gar keine Sequenzähnlichkeit aufweisen, trotzdem miteinander verwandt sind, weil deren dreidimensionale Formen aus der Sicht der Röntgenstrukturanalyse nahezu identisch sind. Solche Proteine sind ohne Zweifel stammesgeschichtlich auf einen gemeinsamen Vorfahren zurückzuführen, wobei die natürliche Selektion lediglich ihre Strukturen konserviert hat, nicht die dahinter liegende Sequenz. Daher kann man sich eine Zeit in nicht allzu ferner Zukunft vorstellen, in der Datenbanksuchen nicht auf der Ebene der Sequenzähnlichkeit, sondern auf der Ebene der räumlichen Struktur durchgeführt werden. Das würde unsere Sicht der Gene und Proteine noch einmal gründlich schärfen. Auch die gezielte Nutzung der kürzlich aus der Genomforschung hervorgegangenen *DNA-Chip Technologie*, die grundsätzlich neue Türen in der klinischen Diagnose und in der Grundlagenforschung eröffnet, ist eine der verheißungsvollsten künftigen Anwendungsgebiete der Bioinformatik. Was auch immer die künftige biologische Forschung aufdeckt, der Computer wird dabei sein, völlig unabhängig davon, ob Bioinformatik als einschlägiger Begriff eines Tages verblasst oder nicht.

Ich möchte ausdrücklich betonen, dass die Bioinformatik weitaus mehr als ein wichtiges Werkzeug für das Sammeln neuer *Informationen* ist, wie hier am Beispiel der Wirkstoff-Forschung erläutert. Sie hat direkt zu einer Vielzahl neuer *Erkenntnisse* geführt. Durch die Erkenntnis werden nicht neue Informationen erzeugt, sondern vorhandene Informationen neu verbunden. Bei einer großen Erkenntnis werden sehr viele Informationen sehr plötzlich und sehr grundlegend neu verbunden, was stets für viel Aufregung, aber auch für viel Widerstand sorgt.

Ganz ohne Frage ist die größte Erkenntnis, welche die Biowissenschaften je hervorgebracht haben, die DARWINsche Evolutionstheorie. Sie ist bis heute das Fundament und das gemeinsame Dach der mannigfaltigen Spezialdisziplinen der Biologie. Es ist auch daher weder umstritten noch überraschend, dass die *meisten* neuen Erkenntnisse – nicht notwendigerweise die *wichtigsten* –, welche die Bioinformatik hervorgebracht hat, auf dem Gebiet der Evolutionsforschung liegen. Warum das so ist, soll zum Schluss kurz erläutert werden.

Gene sind nicht unveränderlich, sie befinden sich seit dem Beginn des Lebens im ständigen Wandel durch den so unberechenbaren wie unaufhaltsamen Prozess der Mutation. Die kleinste Einheit der Mutation ist der Austausch einer Base in der DNA gegen eine andere. Mutation erfordert Zeit. Je mehr Zeit vergangen ist, seitdem sich zwei Arten während der Evolution getrennt haben, umso mehr Mutationen sammeln sich jeweils in ihren Chromosomen an. Vergleicht man zum Beispiel die DNA-Sequenzen von verschiedenen Menschen, so sind sie im Durchschnitt zu 99,9 Prozent identisch; d. h., sie weisen 0,1 Prozent Unterschiede auf. Vergleicht man Mensch und Schimpanse, zeigen die Gensequenzen ca. ein Prozent Unterschiede, im Vergleich Mensch *vs.* Orang-Utan schon drei Prozent

[12] Altschul *et al.* (1997).

Unterschiede usw. durchs Tierreich. Beim Vergleich der Gene von so unterschiedlichen Organismengruppen wie Tieren, Pflanzen und Pilzen finden sich nur noch sporadisch in den Chromosomen erkennbare Sequenzähnlichkeiten, weil der Mutationsprozess während der ca. eineinhalb bis zwei Milliarden Jahre, die diese Gruppen trennen, die meisten Gene bis zur Unkenntlichkeit verändert hat. Aber auch über solche Zeiträume bleibt das gleiche Grundprinzip gewahrt: Je enger verwandt zwei Organismen miteinander sind, umso ähnlicher sind ihre Gensequenzen. Aber viel wichtiger ist der Umkehrschluss: Je ähnlicher die Gensequenzen sind, umso enger verwandt sind die Organismen. Daher ist es möglich, mit Hilfe eines leistungsfähigen Computers aus Gensequenzen Stammbäume zu konstruieren. Und so haben sich in den letzten Jahren die Evolutionsbiologen in großen Schritten auf die Bioinformatik zubewegt, die Technologie genutzt und sie teilweise sogar sehr stark verbessert. Das wichtigste Ergebnis der Auseinandersetzung der Evolutionsbiologen mit der Bioinformatik ist, dass wir Erkenntnisse über alle Phasen der Naturgeschichte des Lebens gewonnen haben: zum Beispiel über die Stammesgeschichte der Mikroorganismen, deren fossile Überlieferung kaum zu deuten ist, über den Übergang zum Landleben vor 450 Millionen Jahren, über die Besiedlung von abgelegenen Inseln, über die Ausbreitung von Arten im Zuge der Kontinentalverschiebungen oder im Gefolge der Eiszeiten, über die Auswanderungen der Urmenschen aus Afrika vor 100.000 Jahren, sogar über die Evolution von Düsseldorfs berühmtestem Nachbarn – dem Neandertaler –, dessen Knochen noch DNA enthalten, deren Sequenz seine Stellung im Familienstammbaum der Menschen geklärt hat.

Für die Genomforschung, für die Strukturforschung, für die molekularbiologische Grundlagenforschung und für die Evolutionsforschung ist die Bioinformatik eine unerlässliche Schlüsseltechnologie geworden. Das trifft auch für die Erforschung der globalen Biodiversität zu, eine der nächsten großen Herausforderungen am naturwissenschaftlichen Horizont.

Nachdem die Menschen in den 70er Jahren gelernt haben, Gene zu sequenzieren, begann Bioinformatik als spielerisches Hobby einiger neugieriger Biologen, die mehr über die in den Gensequenzen enthaltene Information wissen wollten. Als Ergebnis ihrer Neugierde gaben sie eine Technologie an die Fachwelt zurück, die die Sequenzierung ganzer Genome ermöglicht. Die chemisch-pharmazeutische Industrie erkannte sehr schnell, dass man mit Hilfe der Bioinformatik durch Genomforschung und Strukturbiologie schneller und effizienter als bisher neue Wirkstoffe finden und somit neue Produkte auf den Markt bringen kann. Daher braucht dieser Industriezweig auf längere Sicht mehr fachkompetente Bioinformatiker als die Universitäten zur Zeit hervorbringen können. Auf diese rasche und unvorhersehbare Entwicklung haben die Universitäten reagiert, auch die Universität Düsseldorf. In ihrer noch jungen Geschichte hat die Bioinformatik eine stürmische Entwicklung durchlaufen – wir dürfen gespannt sein, wie die Entwicklung weitergeht.

Bibliographie

ADAMS, M. D. *et al.* „The genome sequence of *Drosophila melanogaster*“, *Science* 287 (2000), 2185-2195.

ALTSCHUL, S. F., T.L. MADDEN, A. A. SCHAFFER, J. H. ZHANG, Z. ZHANG, W. MILLER und D. J. LIPMAN. „Gapped BLAST and PSI-BLAST: a new generation of protein database search programs“, *Nucleic Acids Research* 25 (1997), 3389-3402.

FLEISCHMANN, R. D. *et al.* „Whole-genome random sequencing and assembly of *Haemophilus influenzae*“ *Science* 269 (1995), 496-512.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. „Initial sequencing and analysis of the human genome“, *Nature* 409 (2001), 860-921.

VENTER, J. C. *et al.* „The sequence of the human genome“, *Science* 291 (2001), 1304-1351.