WILEY-VCH

Reinhard Rauhut

Bioinformatik

Sequenz-Struktur-Funktion

Reinhard Rauhut

Bioinformatik

Bioinformatik

Sequenz – Struktur – Funktion

von
Reinhard Rauhut

WILEY-VCH

Weinheim – New York – Chichester – Brisbane – Singapore – Toronto

Autor:

Priv.-Doz. Dr. Reinhard Rauhut
Max-Planck-Institut für Biophysikalische Chemie – Abt. Zelluläre Biochemie
Am Faßberg 11
D-37077 Göttingen

Der Autor ist Privatdozent des Fachbereiches Biologie der Justus-Liebig-Universität Gießen, Institut für Biochemie

e-mail: rrauhut@gwdg.de

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Ein Titeldatensatz für diese Publikation ist bei Der Deutschen Bibliothek erhältlich

Printed in the Federal Republic of Germany.

Gedruckt auf säurefreiem Papier.

Satz Hagedorn Kommunikation, Viernheim
Druck Druckhaus Darmstadt GmbH, Darmstadt
Bindung J. Schäffer, Grünstadt

ISBN 3-527-30355-3

Vorwort

Betrachtet man das Bild der modernen Biologie, wie es sich in diesen Tagen in den Medien präsentiert, so fragt sich der Beobachter bisweilen, ob denn die Zukunft der Biologie eher an der Börse oder aber im Labor liege. Der Wirtschaftsteil berichtet ebenso oft über Biologisches wie der Wissenschaftsteil. Dieses plötzliche wirtschaftliche Interesse an den Biowissenschaften ist zu einem Gutteil auch dem jungen Wissenschaftsgebiet der Bioinformatik zu verdanken. Ich sage verdanken, da sich Wissenschaft neben einem Erkenntniszuwachs, einer Umsetzung von intellektuell aufregenden Ideen in neue innovative Produkte des biomedizinischen Sektors, neuer diagnostischer Ansätze und Produktionsverfahren nicht schämen muß. Biologie als Wachstumsbranche und Hoffnungsträger, – es bleibt zu hoffen, daß einer ganzen Generation hervorragend ausgebildeter Biochemie- und Biologiestudenten in Kürze einmal ein freundlicherer Arbeitsmarkt beschieden sei, als dies bisher der Fall war. Die Frage, was von all den Börsengängen bleiben wird, ist noch nicht zu beantworten, die Bioinformatik wird jedoch mit Sicherheit die biologischen Wissenschaften nachhaltig revolutionieren. Es ist dabei ganz und gar kein Zufall, daß die Geburt der Bioinformatik mit der Entwicklung des Internets in den 90er Jahren einherging und durch öffentliche Datenbanken sowie die Benutzung internet-basierter Software gekennzeichnet ist. Der experimentell arbeitende Biologie muß in den Zeiten des Internets auf eine ganz neue Weise lernen zu „wissen, wo es steht". Wo kann ich Informationen und Hilfsmittel zu meinem konkreten Laborproblem im Internet finden, wie kann ich das Maximum an Informationen erhalten, die zu meinem Protein, zu meiner Sequenz in Beziehung stehen, wie erkenne ich den maximalen Informationsgehalt meiner eigenen Daten?

Das Buch ist aus einer einsemestrigen Vorlesung Bioinformatik für Biologen und Biochemiker entstanden. Den Teilnehmern sollten im Rahmen dieser Veranstaltung die Möglichkeiten und Quellen der heutigen Bioinformatik vorgestellt werden, so daß sie für die eigene Arbeit im Labor, für die eigenen Experimente, die richtigen Entscheidungen treffen können. Zudem sollte dem Hörer klar werden, in welcher Richtung sich die modernen biologischen

Wissenschaften ändern werden, eine für den Studenten nicht unwichtige Fragestellung, geht es doch auch um sein zukünftiges Arbeitsgebiet.

Das Buch soll sich also vornehmlich an den experimentell tätigen Biochemiker und Biologen wenden, dessen Ausbildung künftig Bioinformatikwissen enthalten muß. Die Mathematik, die hinter bestimmten Bioinformatikprogrammen steht, wird hier nur ansatzweise verfolgt. Bisher ist nur eine sehr begrenzte Anzahl von Bioinformatik-Monographien erschienen, von denen die meisten für den Studenten und auch für den experimentell tätigen Wissenschaftler wenig hilfreich sind, da sie sich zumeist allzusehr mit dem mathematischen Innenleben von Bioinformatikanwendungen beschäftigen, also mehr auf der Entwickler- als auf der Anwenderseite beheimatet sind. Vorbild bei der Planung des Buches war eigentlich nur das 1998 von Baxevanis und Ouellette herausgegebene Buch *Bioinformatics* (Wiley, New York), das Anfang 2001 in der zweiten Auflage erschienen ist. Dem Informatiker, der neue Datenbankstrukturen entwickelt, Algorithmen entwirft oder Software schreibt, kann im vorliegenden Werk aber sicherlich eine Menge der Biologie vermittelt werden, die hinter den Daten steht.

Der Text verzichtet auf eine allzu bemühte Verdeutschung von Bioinformatik-Begriffen, da dies der Wiederauffindbarkeit in realen Websites eher abträglich ist. Ich habe versucht, die Linkinformationen auf dem neuesten Stand zu halten. Der Benutzer wird merken, daß gerade die besten Websites einem ständigen raschen Wandel unterliegen. Perfekte Lehrbücher entstehen nicht in der ersten Auflage, sie wachsen vielmehr durch das Feedback der Leser. Verlag und Autor erhoffen sich für zukünftige Auflagen reichlich Kommentare und Anregungen zu möglichen Verbesserungen, Fehlern, Unklarheiten, oder Aspekten, die keine Berücksichtigung gefunden haben.

Für zahlreiche Anregungen zum Thema Bioinformatik möchte ich Dr. Gerd Helftenbein, Heidelberg und Dr. Markus Sauerborn, Berlin danken, sowie in Gießen dem Kollegen Prof. Dr. Alfred Pingoud. Dem Verlag Wiley-VCH und seinem Projektverantwortlichen Dr. Hans-Joachim Kraus sei gedankt, daß dieses Buchprojekt so zügig auf den Weg gebracht und mit Elan durchgeführt werden konnte.

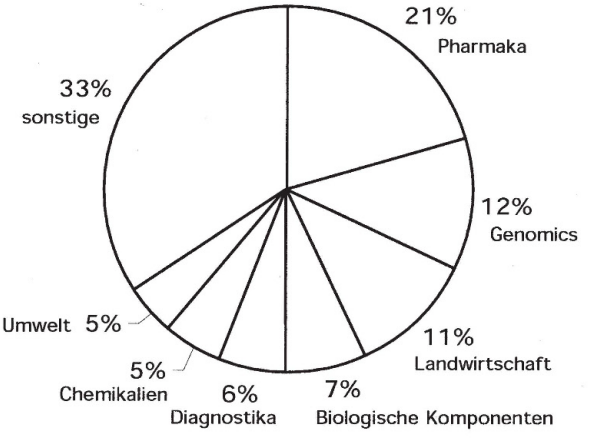
Göttingen – Gießen, August 2001 Reinhard Rauhut

Inhaltsverzeichnis

Einleitung

Bioinformatik – Biologische Wissenschaften im 21. Jahrhundert

Man hat, wer sich erinnert, als experimentell tätiger Biochemiker und Biologe eigentlich erst zu Beginn der 90er Jahre vermehrt die Erfahrung gemacht, daß das rasche Wachstum der Datenbankeinträge tatsächlich einen Einfluß auf den Laboralltag haben kann. War die Situation bis zu diesem Zeitpunkt eher so, daß man zunächst experimentell arbeitete, um eine biologische Funktion z. B. durch Proteinaufreinigung und -charakterisierung sowie Klonierung des zugehörigen Gens zu beschreiben und man dann an den Computer ging, um die Resultate mit anderen Ergebnissen zu vergleichen, so ist es heute, nach mehr als zehn Jahren raschen Wachstums der Datenmengen, nach dem Erscheinen von *Proteomics*, *Genomics* und *high-throughput-research*, oft so, daß man zuerst am Computer arbeitet und dann eine *in silico* geborene Idee experimentell verfolgt und bestätigt. Man muß aber zunächst akzeptieren, daß die Entdeckung und Definition lohnender targets für experimentelle Ansätze in der explodierenden Datenmenge nur durch automatisierte, sensitive Verfahren des Erkennens von zusammengehörenden Einzelfakten, von Sequenz- und Regulationsmustern möglich ist. Dies ist ein fundamentaler Beitrag der Bioinformatik. Bei allen Teildisziplinen des biomedizinischen Sektors und vielen Anwendern chemischer Produkte ist ein reges Interesse an der Bioinformatik vorhanden (Abb. E.1). Bioinformatik und der Computer werden aber das Experiment auch in Zukunft nicht ersetzen, ganz im Gegenteil. Genomprojekte, die enorm fortgeschrittenen Techniken der Strukturaufklärung biologischer Makromoleküle, die Erstellung komplexer Datensets mit Chiptechnologien führen seit den 90er Jahren zu einer immer rasanteren Zunahme des biologischen Wissens. Allein die bloße Menge existierender Daten machte spezielle Methoden zu ihrer Erschließung nötig. Entdeckungen sind heute möglich, indem man die bereits existierende Datenmenge genau analysiert. Bioinformatik schafft die Ordnungskriterien, die zur Bewältigung der Datenmenge notwendig sind. Und wir werden sehen, daß sich die Vielfalt der



E.1 Eine Zusammenstellung der technologischen Sektoren, die gegenwärtig Bioinformatik Ressourcen benutzen. (nach Saviotti et al., *Nature Biotech* 2000, 18: 1247-1249)

beobachteten Lebensformen und Biomakromoleküle auf ein relativ begrenztes Set evolutionären „Spielmaterials" zurückführen läßt.

Dies sind die rein quantitativen Zwänge für das Entstehen einer spezialisierten Form von Biologie (bzw. Informatik) wie sie die Bioinformatik darstellt. Wir haben es aber nicht mit einem bloßen quantitativen Phänomen zu tun. Der vergleichende Blick auf ganze Genome, Proteome und Transkriptome erlaubt es seit wenigen Jahren, experimentelle Ansätze zu verfolgen, die so zuvor überhaupt nicht denkbar waren. Hier ist offensichtlich eine neue Qualität der biologischen Forschung möglich geworden, die sowohl Fragestellungen der evolutionsorientierten Forschung, der Evolution von Proteinstrukturen und des Sequenz-Struktur-Funktions Zusammenhanges, als auch Fragen der komplexen Regulation großer Genverbände oder sogar ganzer Genome einschließt.

Jede Hypothese, die unter Zuhilfenahme des Bioinformatik-Instrumentariums formuliert wird, bedarf des nachfolgenden experimentellen Beweises. Ich werde versuchen klarzumachen, wie sehr die Bioinformatik hilft, neue Experimente gezielter und aussagekräftiger zu gestalten, oft sogar erst den ersten Hinweis darauf gibt, welche Experimente überhaupt möglich und angebracht sind.

Nur sechs Jahre nach der Veröffentlichung des ersten komplett sequenzierten mikrobiellen Genoms (Abb. 1.9 und 1.11) leben wir bereits in dem, was man gemeinhin die „post-genomische" Phase nennt, ein Begriff, unter dem die neuen Techniken zusammengefaßt werden, die unter Verwendung von Genomdaten den Zusammenhang von Sequenz, Struktur und Funktion im Regelwerk einer Zelle untersuchen. Gerade die Proteinforschung erlebt durch die post-Genom-Phase eine wahre Renaissance.

Die Geschwindigkeit bei der Erarbeitung neuer Erkenntnisse wird enorm zunehmen. So werden medizinisch-pharmazeutisch orientierte Laboratorien bei der molekularen Beschreibung von Krankheitsbildern, bei der Identifizierung neuer therapeutischer Targets und der Targetvalidierung sehr viel schneller arbeiten können. Es ist daher nicht verwunderlich, daß es gerade die Ergebnisse des *high-throughput-research* (HTR) sind, die einer Industrialisierung geradezu bedürfen. Nur so kann das in den Datenmengen enthaltene Potential ausgeschöpft werden und zur Entwicklung von HTR-gestützten Assays führen. Neue molekulare Ätiologien bisher diffuser Krankheitsbilder machen Hoffnung, daß auch in solchen Fällen neue diagnostische Marker und therapeutische Targetklassen definiert werden können und der biomedizinischen Forschung neue Erfolge in der Bekämpfung von Krankheiten, die sich bisher einer Therapie widersetzten, beschieden sind.

Ist Bioinformatik nun eine spezialisierte Form von Biologie oder von Informatik? Die Rolle des experimentell tätigen oder Experimente planenden Biologen wird zumeist die eines Benutzers von Bioinformatik-Hilfsmitteln sein. Bioinformatik ist für die Weiterentwicklung der biologischen Wissenschaften so wichtig, daß sie in ihren Grundzügen Teil einer jeden Ausbildung zum Biologen oder Biochemiker werden muß. Es soll daher hier der Stoff behandelt werden, der jedem Studenten der Biowissenschaften und jedem aktiven Biowissenschaftler geläufig sein sollte. Im Mittelpunkt soll also der Anwender stehen. Es wird natürlich, wie in jeder arbeitsteiligen Struktur, auch in der Bioinformatik zur Ausbildung eines Spezialistentums kommen. Die gegenwärtigen Gründungsinitiativen für Studiengänge der Bioinformatik belegen dies. Die Anwender-spezifische Entwicklung von Software erfordert einen anderen, mehr Informatik-orientierten Studiengang, dessen Absolventen sicherlich in einschlägigen Start-Up Firmen gesucht sind. Die Realität der Bioinformatik ist derart, daß die Programmentwicklung und Ausformulierung international gültiger Datenformate für den akademischen Bereich in den Händen spezialisierter, zumeist Datenbank-assoziierter Forschungsgruppen liegt (z. B. NIH, EMBL, Swiss Institute for Bioinformatics).

In der Zukunft wird es sicherlich verstärkt einen Markt für spezialisierte kommerzielle biologische Software geben, wie z. B. integrierte Formen des *data-mining* mit benutzerfreundlichen Programm-Suiten und Software für die Analyse laborintern erstellter Expressionsdaten. Im Rahmen dieses Buches werde ich kommerzielle Software allerdings nur kurz berühren, das Schwergewicht liegt vielmehr in der Verwendung frei zugänglicher internetbasierter Software. Das Datensuchen und -analysieren wird zunehmend so komplex, daß es gerade für Großfirmen notwendig sein wird, damit eine spezielle Abteilung und entsprechende Fachkräfte zu beschäftigen, während kleinere Betriebe vielleicht die externe Bearbeitung durch spezielle Service-Provider vorziehen werden. Vielleicht kann dieses Buch auch dem einen oder anderen Börsenanalysten ein Hilfsmittel sein, wenn er über den nächsten Startup zu entscheiden hat.

Es wird für eine künftige Ausbildung von „hauptamtlichen" Bioinformatikern wichtig sein, eine gesunde Kombination von biologischem und mathematischem Wissen zu vermitteln. Da Bioinformatik aber die tägliche Arbeit eines jeden Biowissenschaftlers betrifft, sollte jeder mit den grundlegenden Ansätzen selbst vertraut sein, sollte die wichtigsten Hilfmittel, die ihm das Internet kostenlos zur Verfügung stellt, selbst nutzen und die Limitationen gängiger *tools* abschätzen können. Man sollte sich bei Fragen, die zum Tagesgeschäft gehören, nicht unnötig in die Abhängigkeit von Spezialisten begeben, denen man sich huldvoll nähern muß, damit sie einmal einen Blick auf das Problem werfen, ein Phänomen, das man im Zusammenhang mit Computern sicherlich in vielen Labors kennt. Die Bedeutung der Bioinformatik liegt nicht in ihrer Rolle für nur einige wenige Spezialisten, sie liegt vielmehr darin, daß sich in absehbarer Zeit das Instrumentarium und die Forschungsplanung eines jeden Naturwissenschaftlers in einer biologischen Disziplin ändern wird und daß ein jeder sich um diese neuen Entwicklungen wird kümmern müssen, allein schon im Interesse einer gesicherten Forschungsfinanzierung.

Die unterschiedlichen Bioinformatik-Bedürfnisse lassen sich an zwei Äußerungen verdeutlichen, wie sie in *Nature* (15 Feb 2001) aus Anlaß der Veröffentlichung des menschlichen Genoms gemacht wurden. Ein so bedeutender Biologe wie Leroy Hood fordert, daß man Bioinformatik auf das engste mit der Ausführung von Experimenten verknüpfen muß, daß ein Biologe Kenntnisse der Bioinformatik besitzen muß, da er nur so in der Lage ist, im Labor „hypothesis driven research" zu betreiben. Ein Vertreter eines führenden Software-Anbieters für Bioinformatik äußert sich dagegen dahingehend, daß der ideale Firmenmitarbeiter ein Programmierer mit biologischer Nachschulung ist.

Es gibt für die Bioinformatik noch keinen festen Kanon von Lehrinhalten. Gedruckte Informationen sind sehr weit verstreut und bei schlechter Bibliotheksversorgung kaum zugänglich. Ich werde daher sehr oft Originalveröffentlichungen heranziehen, um eine bestimmte Problematik zu verdeutlichen. Dies gilt z. B. für solche Techniken wie datenbankgestützte Sequenzierung und DNA-Chip Technologie, die hier als Teile der Bioinformatik aufgefaßt und präsentiert werden. Wir werden wichtige Websites besuchen, es soll aber darauf verzichten werden, dort, wo ausführliche Online-Manuals zugänglich sind, diese noch einmal in ganzer Breite zu wiederholen. Das Buch soll nicht nur Anleitung sein, wie ich Bioinformatik-Ressourcen erschließe, es soll auch die durch die Bioinformatik bereits gewonnenen neuen Einsichten in das Werden und Funktionieren von Organismen vorstellen. Das Konzept verfolgt also keinen engen Bioinformatik-Begriff, sondern will auch die dazugehörige neue Biologie ansatzweise vorstellen.

Es wird im Rahmen dieses Buches nicht möglich sein, auch nur annähernd alle Webressourcen vorzustellen, die der Kategorie Bioinformatik zuzurechnen sind, da es für nahezu jede Ausrichtung der Biologie, Molekularbiologie und Biochemie, für jede Molekülklasse eine spezielle Datenbank gibt. Einen sehr

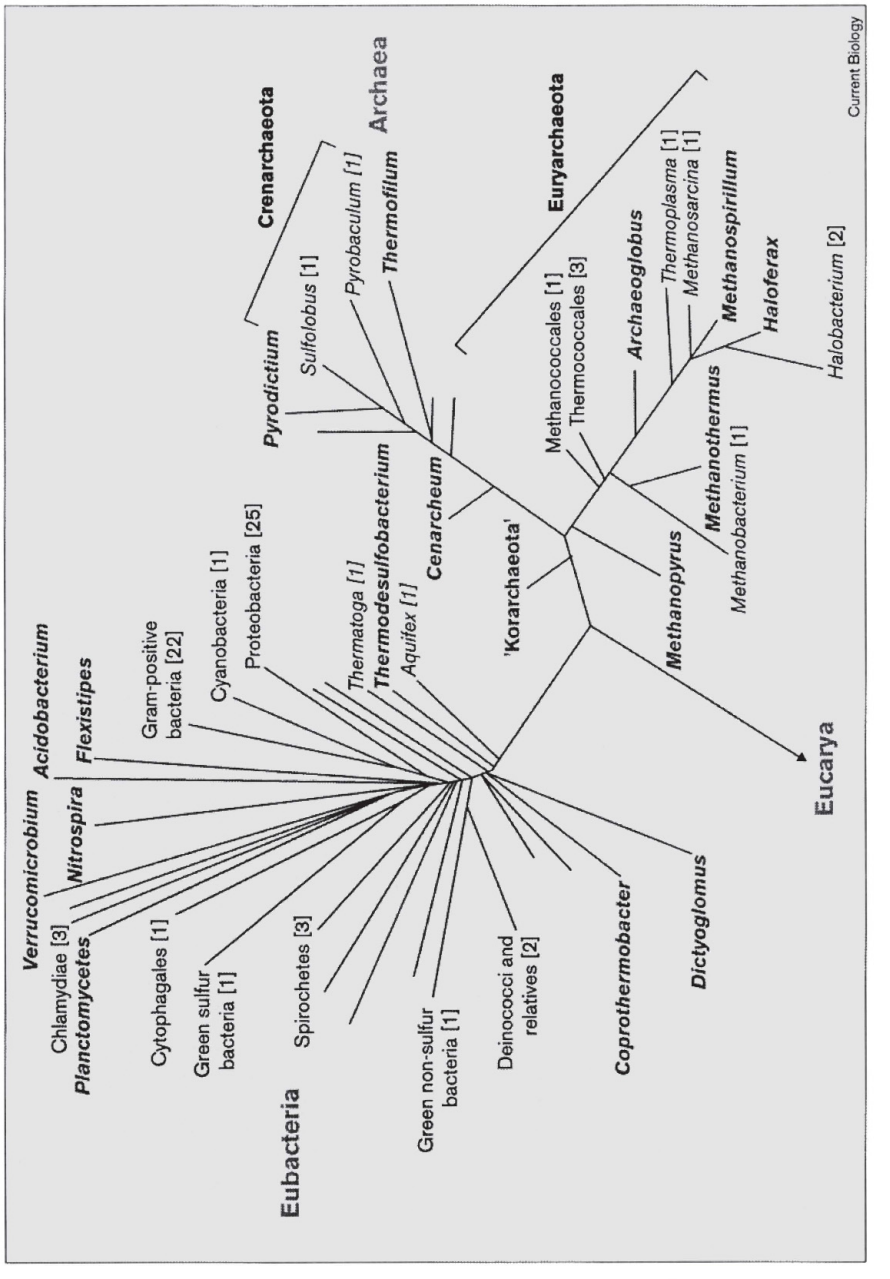
guten Überblick über alle Datenbanken gibt die jährliche Datenbank-Sondernummer von *Nucleic Acids Research*. Die Ausgabe vom Januar 2001 enthält vollständige Beschreibungen für 95 Datenbanken. Außerdem ist eine online frei zugängliche Kompilation von Baxevanis enthalten [http://nar.oupjournals.org], die insgesamt 281 Datenbanken in einer Liste aktiver Links vereinigt.

Wir wollen lernen, welche Erkenntnisse man aus der gewaltig zunehmenden, aber zunächst gestaltlosen Masse an Primärdaten (Sequenzen) gewinnen kann, wenn man die entsprechenden Methoden kennt. Bioinformatik ersetzt nicht Experimente, sondern hilft beim Design intelligenter Experimente. Wir müssen also wissen, wo man Daten findet, was man überhaupt finden kann, wir müssen die Prinzipien verstehen, die z. B. hinter einem Alignmentprogramm, einem Homologiesuchprogramm stehen. Ein Verständnis dessen, was im Hintergrund abläuft, wenn man ein solches Programm anwendet, ist natürlich wünschenswert, nur so kann man auch die Limitierungen abschätzen. Eine vollständige Durchdringung des zugrunde liegenden mathematischen Konzepts von Sequenzalignments ist nicht intendiert, da man hier sehr schnell in den Bereich einer hochspezialisierten Wahrscheinlichkeitsmathematik, von Stochastik, formaler Logik und quasimathematischer Linguistik gerät, der stets weit jenseits des Horizontes eines normalen anwendenden Naturwissenschaftlers liegen wird.

Das Interesse, das Wechselspiel von Funktion und Struktur eines biologischen Makromoleküls zu verstehen, kennzeichnet die moderne Biochemie und Molekularbiologie. In einem eher klassischen Ansatz wird man dazu versuchen, eine funktionelle Mutante zu charakterisieren, das Gen zu identifizieren, oder ein Protein zunächst unter Verwendung eines spezifischen Assay aufzureinigen, biochemisch zu charakterisieren, eine partielle Aminosäuresequenz zu erstellen und nach Überexpression des zugehörigen Gens eine Strukturanalyse z. B. durch Kristallisation durchzuführen. Alle diese experimentellen Techniken wird man auch in Zukunft anwenden, aber man wird im Vorfeld weitaus mehr Zeit darauf verwenden, das wirklich lohnende Target für diese Arbeiten auszuwählen. Und man wird in der Bewertung der Resultate sehr viel Zeit aufwenden, diese mit anderen Sequenzen zu vergleichen. Über Struktur und Funktion hinaus ist es gerade die *Regulation* auch komplexer Molekülverbände und Reaktionsfolgen, die mit den neuen Techniken der *functional genomics* und der Bioinformatik analysiert werden können. Diesen Techniken ist ein Kapitel mit exemplarischen Beispielen gewidmet.

Datenbanken für Primärsequenzen und die Suche in diesen werden uns daher zunächst beschäftigen. Insbesondere werden wir uns dem Problem widmen müssen, zwei oder mehrere Sequenzen, die eventuell eine evolutionäre Beziehung zueinander haben, miteinander zu vergleichen (Problematik paarweiser oder multipler *Sequenzalignments*).

Die evolutionsorientierte biologische Forschung ist seit etwa 1980 durch die Verwendung von 16 und 23 S rRNA Sequenzen und die Propagierung



E.2 Der auf rRNA Sequenzen basierende *universal phylogenetic tree* in seiner Form ohne *root* (s. auch Abb. 6.6). Einführung und Durchsetzung dieses Konzeptes in den frühen 80er Jahren waren vornehmlich das Verdienst von Carl Woese. Die Zahlen geben die Anzahl der abgeschlossenen bzw. in Arbeit befindlichen Genome wieder (Stand 1998; für einen Überblick des jeweils neuesten Standes siehe die Websites von TIGR und des NCBI). (aus Woese, *Curr Biol* 1998, 8: R781-783; Abdruck mit Genehmigung von Elsevier Science)

des Archaea-Konzeptes durch Woese auf eine solide Basis gestellt worden (Abb. E.2). Mit der steigenden Anzahl von Gesamtgenomen ist jetzt die Möglichkeit gegeben, die hier gewonnenen Schlüsse auf genomischer Ebene zu überprüfen und neue verbesserte Konzepte zum Evolutionsverlauf zu entwickeln. Einige wichtige Konzepte bei der Darstellung evolutionärer Beziehungen werden im Kapitel Evolution vorgestellt.

Die ständig zunehmende Menge an Proteindaten (Primärsequenzen und 3D Strukturen) erlaubt neue Erkenntnisse bei der Klassifizierung von Proteinen, ihrer Zusammenfassung zu Familien und Superfamilien. Da solche Klassifizierungen genomweit durchgeführt werden können, wird dabei ein großer Teil des erlaubten Protein 3D-Raums einbezogen. Protein-Evolution kann daher heute viel globaler analysiert werden, als das auf der Basis einzelner Proteinfamilien jemals möglich war. Struktur- und Motivdatenbanken für Proteine wird daher ein eigener Abschnitt gewidmet sein. Die Beziehung zwischen Struktur, Sequenz und Funktion wird dabei in einem veränderten Licht erscheinen. Vielfalt wird hier durch die Verwendung eines relativ beschränkten Sets von Bausteinen erreicht, ein weiteres Beispiel für die Allgegenwart des kombinatorischen Prinzips der zu selbstreplizierenden Systemen organisierten Materie.

Empfohlene Literatur

- Mount: *Bioinformatics – Sequence and Genome Analysis* (Cold Spring Harbor Press, New York, 2001). Gerade bei Abschluß der Arbeiten zum vorliegenden Band erschienen, bietet dieses vorzügliche Buch einen umfassenden und anwenderorientierten Überblick aller Aspekte der Bioinformatik.
- Baxevanis, Ouellette, eds.: *Bioinformatics* (Wiley, New York, 2001, 2. Auflage). Dieser Band ist für den normalen Bioinformatik-Nutzer einer der nützlichsten auf dem Markt.
- Gibas, Jambeck: *Developing Bioinformatics Computer Skills* (O'Reilly, Sebastopol, CA; 2001). Dieser gerade erschienene Band geht für den Nicht-Informatiker auf sehr ansprechende, verständliche Weise auf Unix- und Scripterfordernisse der Bioinformatik ein.
- Eine weitaus stärker theoretisch-mathematische Ausrichtung haben Setubal/Meidanis: *Introduction to Computational Molecular Biology* (PWS Publ., Boston, 1997) und Durbin, Eddy, Krogh, Mitchison: *Biological Sequence Analysis* (Cambridge University Press, 1998)
- Methods in Enzymology, Vol 266: *Computer Methods for Macromolecular Sequence Analysis* (R. F. Doolittle, ed., Academic Press, 1996); Methods in Enzymology, Vol 183: *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences* (R. F. Doolittle, ed., Academic Press, 1990). Wenn auch etwas in die Jahre gekommen, bieten beide Titel noch viel Wissenswertes.

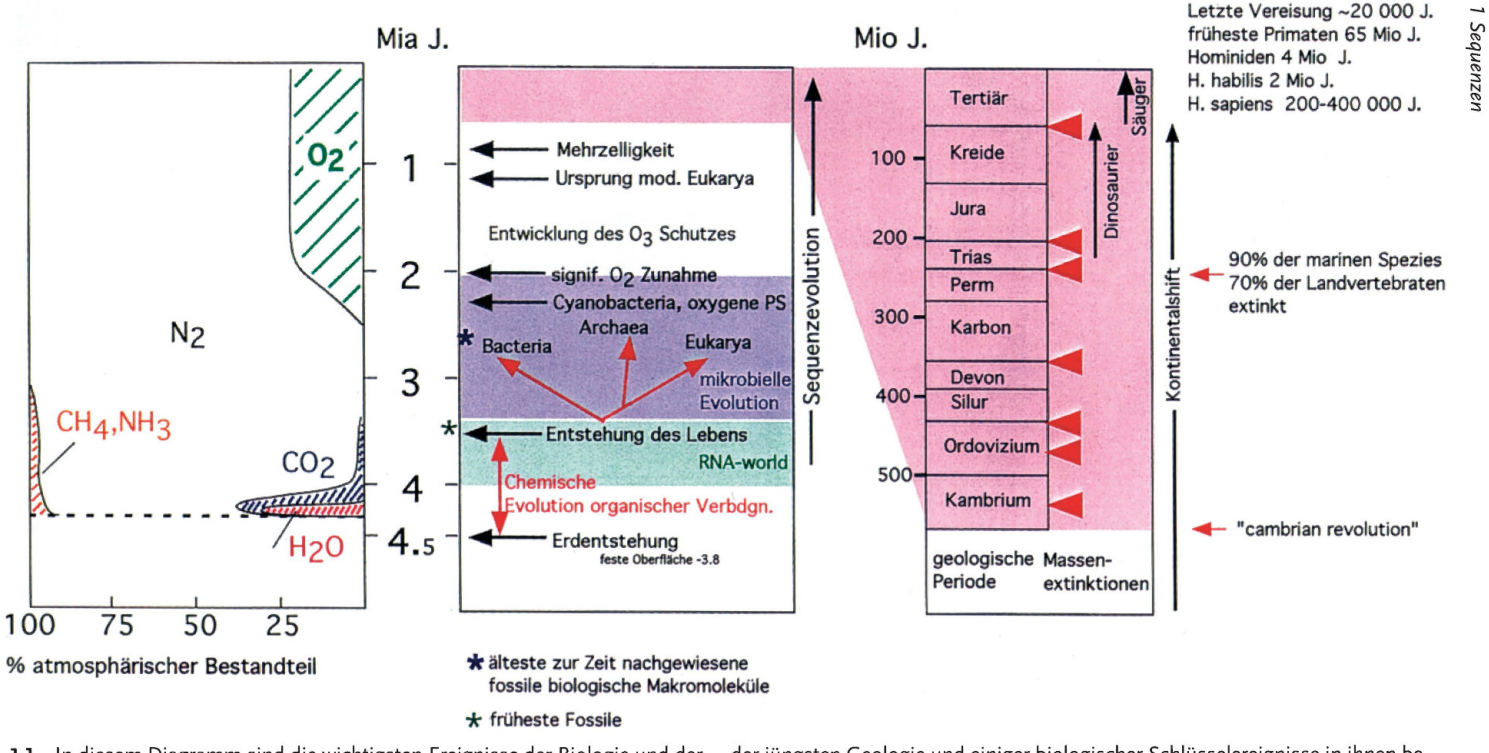
- Als vorzüglichen Überblick über das weite Feld von allgemeinen und speziellen Datenbanken und Bioinformatik-Anwendungen, sei auf das jährliche Januar *Nucleic Acids Research* Sonderheft hingewiesen. Hier lassen sich neben Kurzbeschreibung einer Datenbank oder Website auch die aktuellen Web-Adressen entnehmen.
- Einen ansprechenden Kurzüberblick über die Bioinformatik gibt das *TIBS* Supplement 1998: Trends Guide to Bioinformatics.
- Saenger: *Principles of Nucleic Acid Structure* (Springer, New York – Berlin, 1983). Immer noch der führende Titel auf diesem Gebiet.
- Branden, Tooze: *Introduction to Protein Structure* (Garland Publ., New York, 1998, 2. Auflage). Der führende Titel zum Verständnis von Proteinstrukturen.

1
Sequenzen

1.1
Der Evolutionsverlauf des Planeten Erde, die molekulare Evolution biologischer Systeme und die Suche nach Ähnlichkeiten

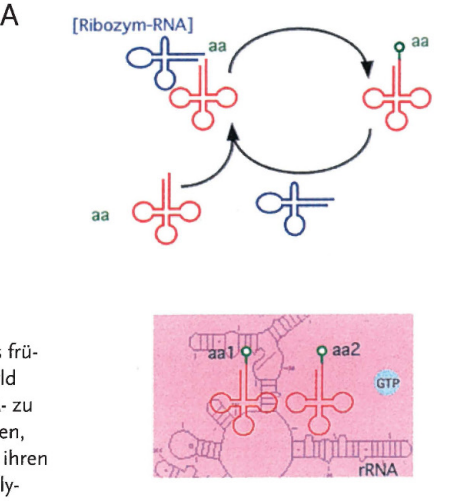
Die komparative Analyse ist in der Biologie ein seit langem eingesetztes Mittel, Entdeckungen zu machen. Wurden anfangs Morphologien ganzer Organismen verglichen, vergleichen wir heute Sequenzen. Das Ergebniss einer Suche nach Ähnlichkeiten zwischen zwei oder mehreren Sequenzen, nach Homologien, wird gewöhnlich in Form eines „sequence alignment" dargestellt. Dabei wird eine distinkte Beziehung zwischen den Positionen zweier oder mehrerer Nukleinsäure- bzw. Proteinsequenzpositionen hergestellt, die untereinander im Alignment stehen (siehe z. B. Abb. 1.59). Die auf diese Weise erkennbar gemachten Ähnlichkeiten bzw. Abweichungen lassen dann Schlüsse auf strukturelle, funktionelle und evolutionäre Beziehungen zu. Ein Alignment hat also das Ziel, erkennbar zu machen, ob zwei Sequenzen hinreichend ähnlich sind (Ähnlichkeit, *similarity*, ist eine quantifizierbare Größe, z. B. ausgedrückt als % Identität zweier Sequenzen), so daß man das Vorliegen einer Homologie annehmen kann. (*homology* ist also der Schluß, der aus dem Vergleich der beiden Sequenzen gezogen wird.) Zwei Gene sind entweder homolog, oder sie sind es nicht. Korrekt gesprochen, gibt es Grade von Ähnlichkeit (similarity) aber nicht von Homologie (homology). Hinter „Alignments" steht also der Gedanke, daß evolutionär verwandte Proteine Sequenzähnlichkeit zeigen. Inwieweit dies dann auch für Struktur und Funktion gilt, wird im folgenden zu diskutieren sein.

Zunächst müssen wir uns ansehen, wie der Evolutionsverlauf auf dem Planeten Erde aussah und in welchen Zeitdimensionen Sequenzen evolvierten (Abb. 1.1). Bemerkenswert ist, daß distinkte Organismenformen sich bereits zu einem Zeitpunkt von −3,5 Milliarden Jahren nachweisen lassen, also zu einem Zeitpunkt, der tief in die Geschichte des jungen Planeten zurückreicht und weit vor den klassischen geologischen Epochen liegt (siehe Webversion [http://www.sciencemag.org] von A. H. Knoll, A new molecular window



1.1 In diesem Diagramm sind die wichtigsten Ereignisse der Biologie und der Chemie belebter Materie in den 4,5 Milliarden Jahren der Existenz des Planeten Erde zusammengefaßt. Das linke Diagramm beschreibt die Entwicklung der Atmosphäre, das mittlere die Evolution der chemischen und biologischen Vorgänge, die zu den heute beobachteten Organismen führten, während rechts die Phasen der jüngsten Geologie und einiger biologischer Schlüsselereignisse in ihnen beschrieben sind. Signifikant ist das extrem frühe Erscheinen komplexer biologischer Systeme in der Erdgeschichte, entsprechend lang ist die Vorgeschichte biologischer Makromoleküle.

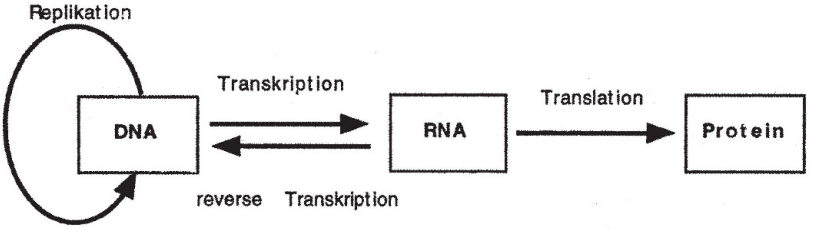
on early life. *Science* 1999, 285: 1025–1026). Vorläufer in Form von mehr oder weniger effizienten selbstreplizierenden Molekülsystemen müssen daher bereits viel früher vorhanden gewesen sein. Selbstreplizierende Molekülsysteme sind vielleicht bereits 300.000 Jahre nach Ausbildung einer festen Planetenoberfläche entstanden. Diese frühen Formen von Leben bestanden, wie eine attraktive Theorie annimmt, aus reinen RNA Systemen, in denen RNA sowohl Informationsmolekül als auch katalytisch kompetentes Molekül war (Abb 1.2). Walter Gilbert prägte 1986 hierfür den Begriff der *RNA World*. Es war stets eine wichtige Annahme bei der Modellbildung einer frühen RNA-Evolution, daß Translation ein RNA-katalysierter Prozeß ist. Gerade diese Annahme wurde durch Nissen et al. im Jahre 2000 belegt (*Science*, 289: 920–930). Diesen Autoren gelang es, den lange vermuteten Ribozym-Charakter des Ribosoms nachzuweisen. Vielleicht waren Protoribosomen in einer frühen Evolutionsphase reine RNA-Körper. Wiederum ein Hinweis auf die inhärente Eigenschaft von Materie, sich als selbstreplizierendes Informations/Katalyse-System zu organisieren.



1.2 Darstellung eines denkbaren Ablaufs früher chemischer Evolution in der RNA World Phase. Auch der Übergang von einer RNA- zu einer RNA-Protein Welt läßt sich so erklären, wenn man die ribosomale Translation bzw ihren evolutionären Vorläufer als eine RNA katalysierte Reaktion begreift. A) Zunächst unterliegt eine selbstreplizierende Ribozym-RNA (blau) einer darwinistischen Evolution, in deren Verlauf sie die Fähigkeit entwickelt, eine Aminosäure kovalent an eine Art tRNA-Vorläufer (rot) zu koppeln. Dabei entsteht ein Aminoacyl-Ribozym. B) Die Kopplung zweier oder mehrerer Aminosäuren führt dann zu Peptid-RNA-Komplexen und Proteinen. Die Transpeptidierung wie wir sie auch in „modernen" Ribosomen beobachten, erfordert außer GTP keine zusätzliche Energie, da die Aminoacylester bereits energiereich sind.

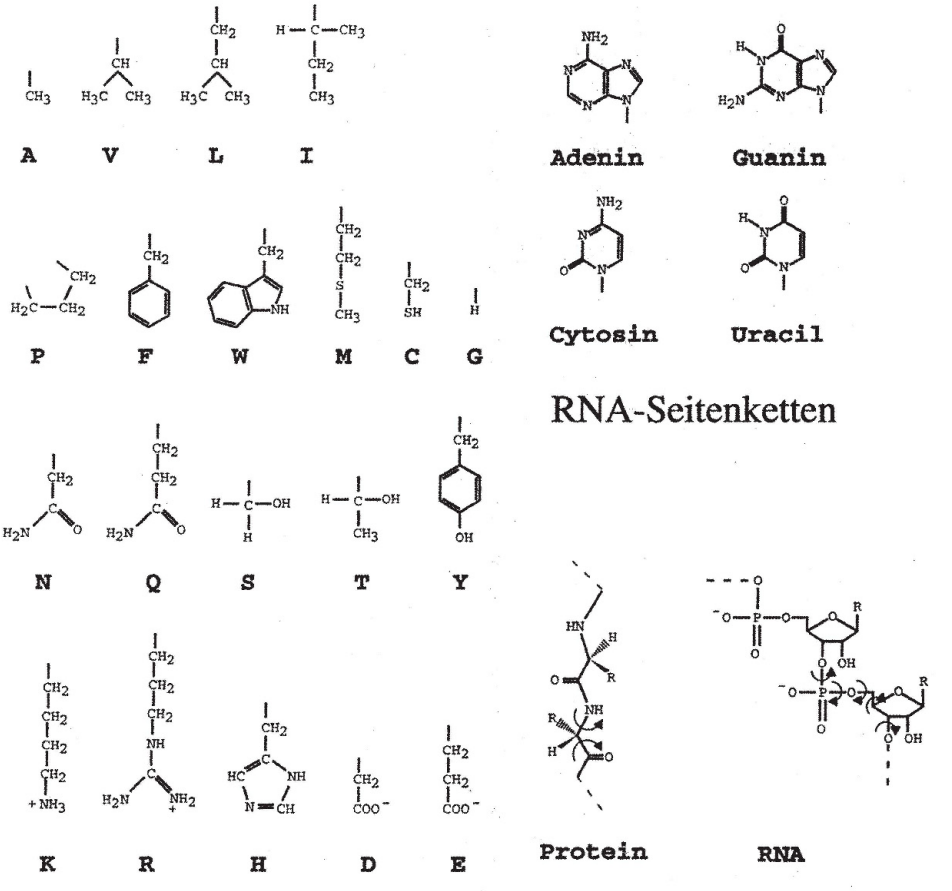
folgender Verlauf ist denkbar:

1. **Selbstreplizierende Ribozym-RNA, die einer darwinistischen Evolution unterliegt.**
2. **Aminoacyl-Ribozym**
3. **Peptid-RNA-Komplex**
4. **Protein**



Informationsfluß in selbstreplizierenden Systemen

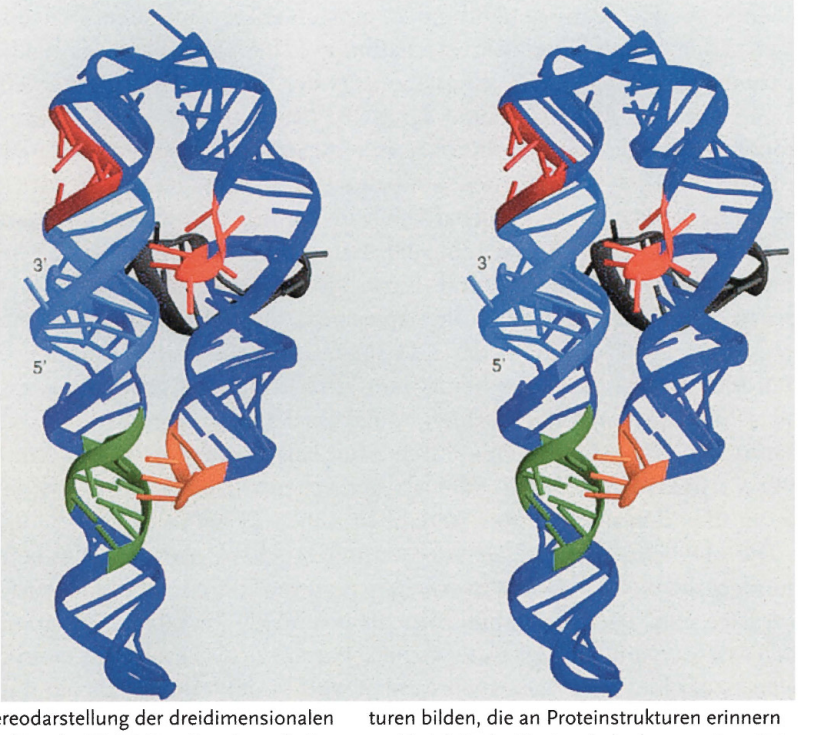
1.3 Moderne biologische Systeme benutzen DNA als Informationsspeicher und Proteine für die katalytischen Aufgaben.



1.4 Proteine besitzen eine weit höhere Seitenkettenvielfalt als RNA. Ungeachtet dessen und trotz hoher Ladungsdichte und Beweglichkeit des Backbone (siehe Pfeile), kann aber auch RNA mit der Unterstützung von Metallionen komplexe Strukturen und katalytische Zentren formen.

Moderne biologische Systeme benutzen im biologischen Informationsfluß fast immer den Informationsspeicher DNA (Abb. 1.3) und haben die meisten Struktur- und Katalysefunktionen der Stoffklasse der Proteine anvertraut. Die Vorschrift, nach der die Information des DNA-Informationsmoleküls in Proteine umgesetzt wird, ist der genetische Code, die vermittelnden Moleküle sind messenger und transfer RNA. DNA und RNA sind Biopolymere, die ein Kodierungsalphabet von vier Buchstaben besitzen. Beide sind auf Grund ihres Aufbaus aus Nukleotidbausteinen 5′→3′ gerichtete Moleküle. Proteine reichen in ihrer Evolutionsgeschichte also weit in den Raum jenseits der 3 Milliarden Grenze zurück. Der molekulare Evolutionsverlauf ist in seinen Details in verschiedenen evolutionären Phasen stets unterschiedlich, da eine darwinistische Evolution von Molekülpopulationen stets von der Fehlerrate des evolvierenden Systems abhängt. Siehe hierzu auch die Gedanken in Kapitel 6.

Proteine bestehen aus den 20 proteinogenen Aminosäuren (Abb 1.4). Wie Nukleinsäuren (5′→3′), so sind auch Proteine gerichtete Biopolymere (N Terminus → C Terminus). Auf Grund der zahlreichen verschiedenen Seitenketten



1.5 Stereodarstellung der dreidimensionalen RNA-Struktur der P4-P6 Domäne des selbstspleißenden Gruppe I Introns aus *Tetrahymena*. Trotz eines limitierten Sets an Bausteinen kann RNA mit Hilfe von Metallionen kompakte Strukturen bilden, die an Proteinstrukturen erinnern und katalytische Zentren beherbergen. (aus Cate et al., *Science* 1996, 273: 1678–1685; Abdruck mit Genehmigung der American Association for the Advancement of Science)

können Aminosäuren eine große Vielfalt von Strukturen bilden. RNA muß insbesondere wegen der fehlenden hydrophoben Seitenketten andere Lösungen finden, um hydrophobe Taschen zu bilden. Gerade die Struktur der P4/P6 Domäne der group I-selfsplicing RNA hat auf eindrucksvolle Weise gezeigt, welche reichen Strukturmöglichkeiten auch RNA zur Verfügung stehen, um eine dichte Raumpackung zu erreichen und Wasser aus einem Faltungs-"innenraum" zu verdrängen, auch wenn dazu nur 4 verschiedene Nukleotide und Metallionen zur Verfügung stehen (Abb. 1.5). Dieser Umstand verleiht RNA die Fähigkeit, katalytische Zentren auszubilden.

1.2 Sequenzdatenbanken

Zunächst ein Blick auf die historische Entwicklung, die uns zu den heutigen Primärsequenz- und Strukturdatenbanken führte. Genomische Sequenzdaten werden heute in großen Mengen durch die zahlreichen laufenden Genomprojekte erstellt. Die meisten dieser Projekte arbeiten an prokaryontischen Organismen (vielfach Pathogenen) und an einigen eukaryontischen Modellorganismen. Einen guten Überblick verschaffen z. B. die Homepage des Institute for Genomic Research, TIGR, eines Pioneers der Sequenzierung kompletter Genome, oder das ENTREZ Portal des NCBI (Abb. 1.6 und 1.7). Genomprojekte konzentrieren sich auf evolutionär interessante Organismen, auf molekularbiologische Modellorganismen, auf pathogene Mikroorganismen, auf Organismen mit beträchtlicher wirtschaftlicher Bedeutung und natürlich das menschliche Genom (Abb. 1.8 und 1.9). Abb. 1.10 läßt erkennen, wie sehr Wachstum des biologischen Wissens und Entwicklung der Computertechnik einhergehen. Dazu kam natürlich die explosionsartige Entwicklung des Internets im Verlauf der 90er Jahre. Abb. 1.11 zeigt, daß die ersten 40 Jahre der modernen Molekularbiologie von heroischen Einzelresultaten geprägt waren, die in oft jahrelanger Arbeit einzelnen Molekülen abgerungen wurden. Ab 1990 dann die Datenexplosion, die durch neue Labortechnologien möglich wurde. Der Charakter des Jahres 1990 als Schwellenjahr wird besonders deutlich, wenn man den quantitativen Verlauf in Abb. 1.12 verfolgt.

Die Möglichkeit, ganze Genome zu untersuchen (*genomics*), hat bereits begonnen, die biologischen Wissenschaften zu revolutionieren. Man wird z. B. in der Lage sein, ganze Proteinfamilien als potentielle Therapietargets in pathogenen Mikroorganismen zu untersuchen. Der spezielle Bereich der *functional genomics*, der hier auch vorgestellt werden wird, bedeutet mehr als ein nur quantitativer Fortschritt in unseren Verständnismöglichkeiten von biologischen Systemen. Die Möglichkeit, den Einfluß eines bestimmten Makromoleküls auf die gesamte Expressionssituation und alle Regelkreise in einem Organismus zu untersuchen, war so vor dem Ereignis ganzer Genome nie gegeben.

TIGR Databases

What's New

About TIGR

TIGR Faculty

TIGR Gene Indices

Conferences, Education & Training

TIGR Software

Career Opportunities

Related Links

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

TIGR Databases

The TIGR Databases are a collection of curated databases containing DNA and protein sequence, gene expression, cellular role, protein family, and taxonomic data for microbes, plants and humans. Anonymous FTP access to sequence data is also provided. Please read the disclaimer regarding use of data. The TIGR clone distribution policy is available for viewing.

Comprehensive Microbial Resource (CMR) Please forward any questions/comments/broken links to cmr@tigr.org.

The TIGR Microbial Database provides links to world-wide genome sequencing **projects completed** and **projects underway**, including the completed TIGR genomes:

Archaeoglobus fulgidus	*Methanococcus jannaschii*
Borrelia burgdorferi	*Mycobacterium tuberculosis*
Chlamydia pneumoniae	*Mycoplasma genitalium*
Chlamydia trachomatis	*Neisseria meningitidis*
Deinococcus radiodurans	*Thermotoga maritima*
Haemophilus influenzae	*Treponema pallidum*
Helicobacter pylori	*Vibrio cholerae*

New! Run a BLAST search on our unfinished genomes, or subscribe to get the unfinished genomic data in flatfile format.

The TIGR *Arabidopsis thaliana* Database provides access to genomic sequence data and annotation generated at TIGR and assemblies of *Arabidopsis* ESTs from world-wide sequencing projects.

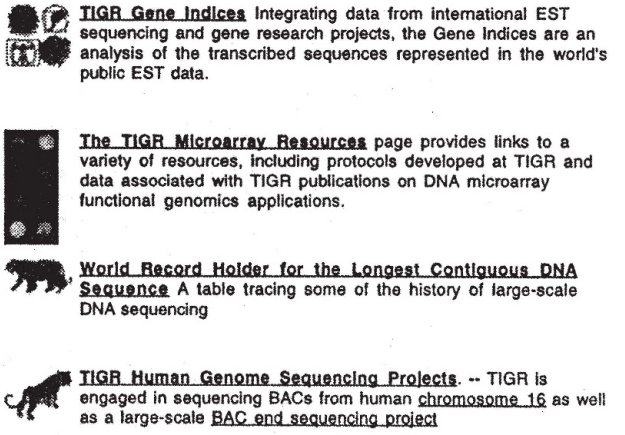
The TIGR Rice Database provides links to the USDA/NSF/DOE-funded rice genome project at TIGR and includes sequence data, annotation, and links to the *Oryza sativa* Gene Index.

Potato Functional Genomics Project provides links to the NSF-funded potato genome project at TIGR and includes sequence data, annotation, and links to the *Solanum tuberosum* Gene Index.

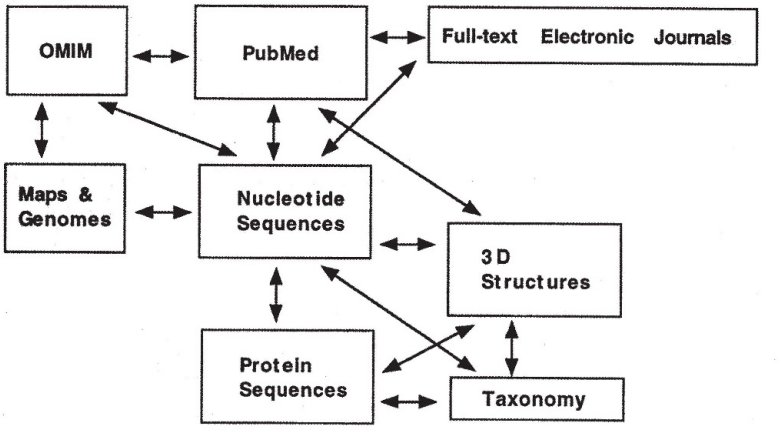
The TIGR Parasites Database provides links to TIGR sequencing projects completed and underway as well as links to related world-wide sequencing efforts: *Trypanosoma brucei, Trypanosoma cruzi, Plasmodium falciparum, Plasmodium yoelii,* and *Entamoeba histolytica*

TIGRFAMs are protein families based on Hidden Markov Models or HMMs.

TIGR Viral Genome Sequencing Project In collaboration with the Max Planck Institute for Biochemistry, TIGR has sequenced the 40 Kb genome of the Sulfolobus islandicus filamentous virus.

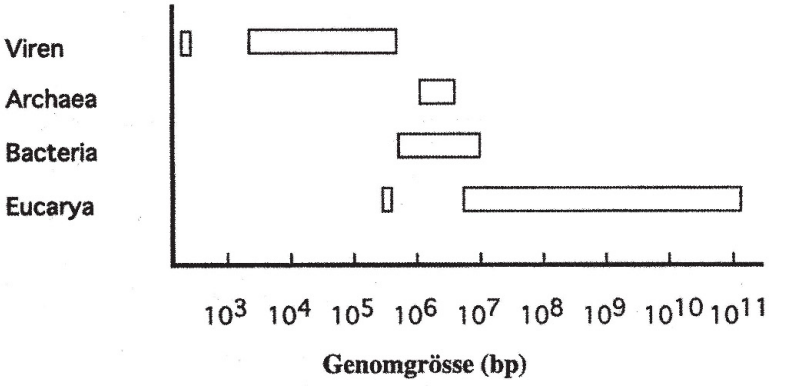


1.6 TIGR Database Homepage (The Institute for Genomic Research; [http://www.tigr.org/tdb/]). Dieses Institut veröffentlichte 1995 das erste komplette Genom eines Mikroorganismus, *M. jannaschii*.



1.7 Das integrierte ENTREZ Search and Retrieval System. Entrez bietet den Einstiegspunkt zu allen Informationsbereichen des NCBI. Die Pfeile beschreiben die Vernetzung dieser Bereiche, die durch aktive Links in den individuellen Einzeleinträgen hergestellt wird.
Die Nukleotidsequenzen entsprechen dem Kernbereich *GenBank*. Proteinsequenzen stammen aus anderen Datenbanken, die Proteinsequenzen enthalten, wie PIR, PRF und PDB, oder sie sind von DNA-Sequenzen in GenBank oder RefSeq abgeleitet. 3D Structures bietet online-Ressourcen zum Verständnis biologischer Strukturen. Die enthaltenen MMDB Strukturfiles stammen aus der PDB Datenbank. PubMed ermöglicht die Suche nach biomedizinischer Literatur (seit1962) und stellt gegebenenfalls die Verbindung zum Volltext einer gefundenen Referenz her. Diese technische Möglichkeit wird bei weiterer Entwicklung des *electronic publishing* an Bedeutung gewinnen. Taxonomy bietet spezielle Informationen und online-Ressourcen zu wichtigen Modellorganismen. Maps & Genomes präsentiert mehr als 600 virale Genome, die komplettierten bakteriellen, archaealen und eukaryontischen Genome, sowie einzelne Chromosomen und Organellengenome. OMIM (für Online Mendelian Inheritance of Man) ist ein Katalog menschlicher Gene und ihrer möglichen Defekte.

A



B

H. sapiens	2.91x10^9 bp	~30 000 Gene	2001***
Drosophila	120x10^6 bp*	14 200 Gene	2000
C. elegans	~10^8 bp	18 400 Gene	1998
S. cerevisiae	13x10^6 bp	~ 6 000 Gene	1997
E. coli	4.6x10^6 bp	4 405 Gene	1997
Arabidopsis**	125x10^6 bp	25 500 Gene	2000

1.8 A) Verteilung der Genomgrößen in Vertretern der drei evolutionären Primärreiche. War für lange Zeit das einfache cirkuläre Chromosom das Standardmodell für bakterielle genomische Organisation, wurde im Verlauf der Untersuchungen klar, daß hier eine beträchtliche inter- und intra-Spezies Variabilität existiert. Die komplette genomische Ausstattung ist in ein oder mehreren linearen oder cirkulären Chromosomen, in freien oder integrierten Plasmiden und Prophagen, Pathogenitätsinseln und anderen kleinen beweglichen genetischen Elementen untergebracht.
B) Bereits publizierte Genome wichtiger Modellorganismen. Man beachte besonders die disproportionale Beziehung zwischen Genomgröße und Genzahl.

* nur Euchromatin.
** Da das *Arabidopsis* Genom Bereiche extensiver Verdopplungen zeigt, liegt die Zahl individueller Gene bei <15.000.
*** Zwei draft-Sequenzen des menschlichen Genoms (also vorläufige Entwürfe, die jeweils mehr als 90 % fertig erstellte Sequenz beinhalten) wurden im Februar 2001 von der private Forschungsgruppe um Craig Venter (Celera) und dem öffentlich finanzierten International Human Genome Sequencing Consortium veröffentlicht. Celera benutzte whole genome shotgun Sequenzierung, während die internationale Gruppe BAC basiertes, hierarchisches shotgun-Sequenzieren benutzte. Zwischen 26.000 und 38.000 Gene sind vorhergesagt. (J. C. Venter et al., *Science* 2001, 291: 1304–1351; E. S. Lander et al., *Nature* 2001, 409: 860–921)