Von Genomsequenzen zu Proteinfunktionen

Peer Bork, Europäisches Laboratorium für Molekularbiologie, Heidelberg

Dr.-habil. Peer Bork studierte von 1983 bis 1988 Biochemie und wurde 1990 über die Mustererkennung in Sequenzen promoviert. 1995 habilitierte er sich in theoretischer Biophysik. Seit 1990 ist er Wissenschaftler am MDC für Molekulare Medizin in Berlin und seit 1991 Gastwissenschaftler am Europäischen Laboratorium für Molekularbiologie (EMLB) in Heidelberg. Dr. Bork beschäftigt sich hauptsächlich mit der vergleichenden Sequenz- und Genomanalyse, molekularer Evolution sowie mit Funktions- und Stoffwechselwegvorhersagen.

Mit dem Vorliegen kompletter Genome in Textform bricht für die Molekularbiologie und Molekularmedizin ein neues Zeitalter an, da das Erbmaterial im Prinzip alle Informationen für Lebenszyklus und Vermehrung von Zellen und Organismen enthält. Auf dem Weg zur Entschlüsselung dieser Information müssen zuerst die Informationsträger, die Gene, erkannt und die entsprechenden Funktionsträger, die Genprodukte (Proteine) charakterisiert werden. In diesem Beitrag wird in vergleichende sequenzanalytische Methoden im Prozeß der Funktionsvorhersage von Proteinen eingeführt und ihre Stärken und Schwächen beleuchtet.

From Genom Sequences to Protein Functions

. The successful sequencing of complete genomes is the beginning of a new age in Molecular Biology and Molecular Medicine. Encoding the information of a complete genome should enable us to understand fundamental processes of life. A first step in the understanding of these molecular processes is the identification of the genes and their gene products, the proteins, as well as their functional characterization. This introduction into the prediction of protein function using comparative sequence analysis will also deal with the powers and pitfalls of the currently used methodology.

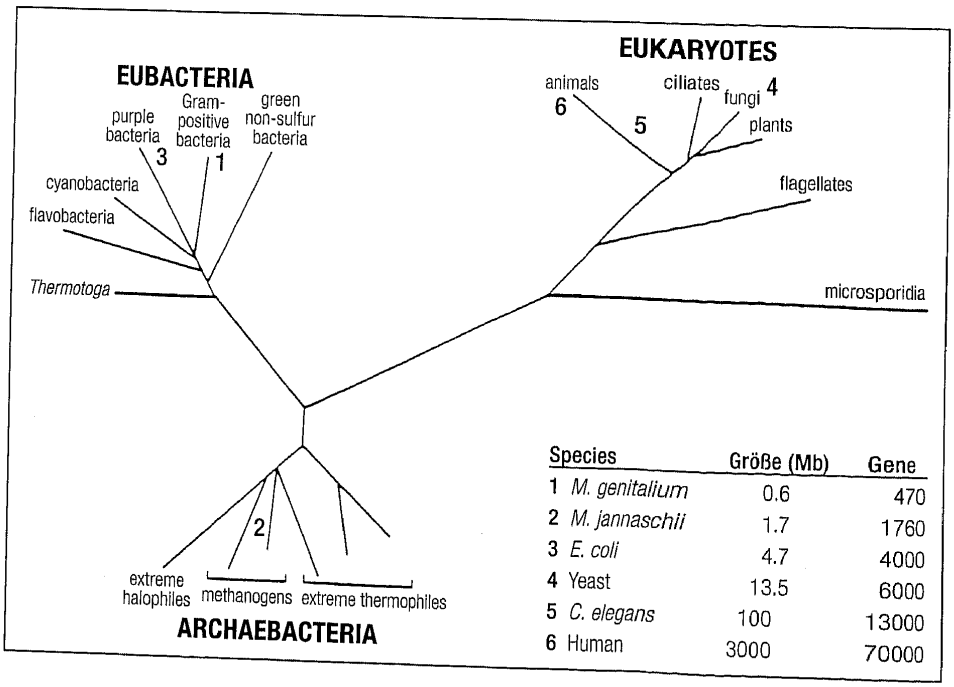
1 Einleitung

1.1 Die „Postgenom"-Ära in der Biologie

In der Molekularbiologie ist eine neue Ära angebrochen – die der komplett sequenzierten Genome zellulärer Organismen. Am 28. Juli 1995 wurde das Erbmaterial des parasitisch lebenden Bakteriums *Haemophilus influenza* veröffentlicht [1]. Neben den Genomen einiger Eubakterien liegt nun seit 1996 auch von Archaebakterien und mit der Bäckerhefe auch von Eukaryoten das komplette Erbmaterial in Form von 4-Buchstabentexten vor (für die 4 Basen **A**denine, **G**uanin, **C**ytidin und **T**hymidin). Schon bald werden komplexere Genome wie die des Nematoden *C.elegans* (im Jahre 1998) und auch des Menschen (derzeitige Schätzungen gehen von 2003 aus) folgen (Bild 1). Damit eröffnet sich ein völlig neuer Zugang für das Verständnis der Grundstrukturen des Lebens, den Zellen [2].

Das Erbgut (die komplette DNA eines Organismus) sollte alle Informationen enthalten, die eine Zelle entstehen, sich entwickeln, mit ihrer Umwelt interagieren, und sie als ein offenes Fließgleichgewicht bis zu ihrem vorprogrammierten Tod erhalten lassen. Die Entschlüsselung dieser Information ist eine gewaltige Aufgabe, bei der wir erst am Anfang stehen (Bild 2). Wie kann man aus der Erbin-

Bild 1: Speciesevolution und Genomdaten. Ein Stammbaum basierend auf RNA-Daten (aus [19]) ist mit Proteinanzahl und Genomgrößen von Modellorganismen verglichen, die bereits oder in naher Zukunft komplett vorliegen werden.



Species	Größe (Mb)	Gene
1 *M. genitalium*	0.6	470
2 *M. jannaschii*	1.7	1760
3 *E. coli*	4.7	4000
4 Yeast	13.5	6000
5 *C. elegans*	100	13000
6 Human	3000	70000

formation auf den Lebenszyklus einer Zelle oder gar eines multizellulären Organismus mit seinen speziellen Umweltbedingungen schließen? Wie erkennt man Dysfunktionen, d.h. krankhaft veränderte Gene? Wie kann man eine Zelle simulieren und bestimmte Eigenschaften vorhersagen?

Dies sind Fragen, die man in naher Zukunft sicher nicht so einfach beantworten kann – trotzdem ist es an der Zeit, sich solchen Aufgaben zu stellen. Der **Bioinformatik** wird hierbei eine Schlüsselrolle zugeordnet. Dies drückt sich nicht nur in sprunghaft gestiegenen Forschungsaktivitäten aus, sondern auch in neuen Konzepten und Investitionen in der pharmazeutischen Industrie und in einer neuen Generation von Biotechnologiefirmen.

1.2 Bioinformatik von Genomanalyse bis Gewebemodellierung

Das Erbgut besteht aus DNA (Desoxyribonucleinsäure), dem „Informationsträger" der Zelle. Zelluläre Prozesse werden aber hauptsächlich von Proteinen, den „Funktionsträgern" gesteuert. Diese Steuerung realisiert sich erst im gefalteten Zustand über die dreidimensionale Struktur und Dynamik von Proteinen und ihren Reaktionspartnern. Die Zelle besteht auch aus anderen Molekülen, z.B. Fetten, Zuckern, Wasser, Metallionen etc., die durch ihre Wechselwirkungen die Zelle am Leben erhalten. Neben den im Ergbut kodierten Informationen hängt der Zustand einer Zelle aber auch von äußeren Faktoren, wie dem Nahrungs- und Energieangebot ab, und ist je nach Zell-Zyklus-Stadium verschieden. In komplexeren Organismen sind Zellen spezialisiert und bilden gemeinsam mit ähnlich spezialisierten Zellen übergeordnete Struktur- und Funktionseinheiten, z.B. die Gewebe. Die **Bioinformatik** ist in allen diesen verschiedenen Ebenen (von der DNA bis zur Dynamik eines Organismus) von enormer Bedeutung.

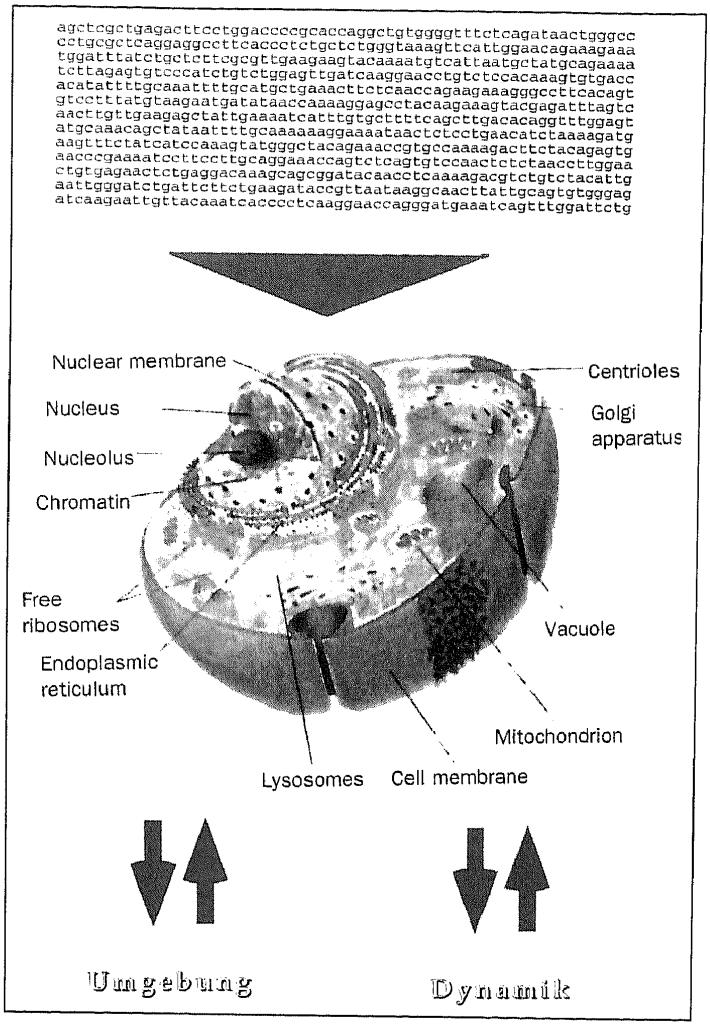


Bild 2: Ein Bruchteil der Nukleotidsequenz des Brustkrebsgenes BRCA1 ist dargestellt (einige Hundert Basen; das menschliche Genom enthält 3000 Megabasen dieser Buchstaben), um die Aufgabe zu verdeutlichen, Information aus der DNA zwecks Vorhersage zellulärer Funktionen zu extrahieren. Die Zellen (Abbildung einer Zelle mit Erlaubnis aus [20]) sind allerdings keine statischen Gebilde – sie wechselwirken mit der Umgebung (z.B. Nahrungsaufnahme) und durchlaufen verschiedene Entwicklungsstadien.

2 Proteinfunktion: Was wollen wir woraus vorhersagen?

2.1 Was ist Proteinfunktion?

Hier soll in den Teil der Bioinformatik eingeführt werden, der sich mit der Genomanalyse und Proteinfunktionsvorhersage beschäftigt, dem gegenwärtigen Hauptziel bei der Aufarbeitung der genomischen Daten [2]. Computergestützte Funktionsvorhersagen erfolgen gewöhnlich mit Hilfe von Proteinsequenzvergleichen. Proteinsequenzen, die durch den genetischen Code bestimmte lineare Anordnung von 20 verschiedenen Aminosäuren, sind durch ihr größeres Alphabet (20 Buchstaben) für die Erkennung entfernter Ähnlichkeiten besser geeignet als DNA (4-Buchstaben-Alphabet). Normalerweise wird eine Zielsequenz mit Tausenden anderen, in Datenbanken gespeicherten Sequenzen verglichen. Ein einfacher Fall ist eine hohe Ähnlichkeit mit einem funktionell beschriebenen Protein, dessen Funktion dann auf das Zielprotein übertragen werden kann.

Leider ist das Wort „Proteinfunktion" nicht sehr gut definiert, da die Funktion im Zusammenspiel mit anderen Molekülen ausgeübt wird, sich je nach Status der Zelle verändern kann und man funktionelle Details noch von keinem Protein genau kennt. Proteinfunktionsvorhersagen können sich aber nur auf Information stützen, die zusammen mit den Sequenzen in Datenbanken abgespeichert ist. Selbst für die am besten charakterisierten Proteine, die metabolischen Enzyme, deren Stoffwechselwege bekannt und von denen Raumstruktur und Substrat-Bindungsverhalten aufgeklärt sind, bleiben noch viele Fragen offen: Wie finden auf der molekularen Ebene Regulationen statt? Bilden die Enzyme eines Stoffwechselweges einen Komplex oder werden die Substrate in Lösung weitergereicht? Bei den mei-

sten Proteinen ist die sogenannte molekulare Funktion aber viel schlechter beschrieben, z.B. „Dieses Protein bindet ATP". Wenn man nun über Ähnlichkeitssuchen in Datenbanken versucht, eine molekulare Funktion vorherzusagen, kann man für das vorherzusagende Protein im besten Fall diese sehr vage Aussage übernehmen. Leider sind in den Sequenzdatenbanken aber molekulare Funktionen mit Funktionsaussagen auf anderen Ebenen vermischt, z.B. mit phänotypischen Aussagen wie „Dieses Protein spielt eine Rolle in der Entwicklung von Flügeln" oder zellulären Erkenntnissen wie z.B. „involviert in Proteinfaltung". Viele Aspekte der Funktion, wie z.B. die zeitliche gesteuerte Expression des entsprechenden Proteins, sind überhaupt noch nicht berücksichtigt. Hier muß also noch erhebliche Arbeit in die Annotation der Datenbanken investiert werden, um die durch komplizierte Methoden gewonnenen Ähnlichkeitsaussagen zur Funktionsvorhersage nutzen zu können.

2.2 Wie gut sind die Sequenzdaten und wie verläßlich die Datenbanken?

Ein weiteres Problem, das ständig berücksichtigt werden muß, ist die Qualität der Daten und ihrer Annotation [3]. Obwohl die 4-Buchstabentexte (DNA) oder das 20-Buchstabenalphabet (Proteine) eindeutig interpretierbar sind, ist mit einer hohen Fehlerrate zu rechnen. Bedingt durch den genetischen Code kann das Fehlen einer einzigen Base in der DNA zu Leserahmenverschiebungen bei der Übersetzung in Proteinsequenzen führen und die korrespondierende Proteinsequenz wird nach der fehlerhaften Stelle total verschieden von der wirklichen Sequenz sein. Bei Stichprobenuntersuchungen in dem Bakterium *E.coli* fanden wir z.B., daß fast 10% der Gene (für Proteine codierende DNA Bereiche) zu Leserahmenverschiebung führende Fehler enthielten. Bei höheren Organismen, in denen Gene durch nicht-codierende Einschübe (Introns) unterbrochen sind, ist die Übersetzung in Proteinsequenzen komplizierter. In derzeit fast 30% aller Fälle werden die codierenden Abschnitte des Genes (Exons) fehlerhaft zusammengefügt [4], d.h. trotz korrekter DNA-Sequenz ist die Proteinsequenz mit groben Fehlern behaftet. Zu den durch technische Schwierigkeiten bei der Sequenzierung bedingten Datenfehlern kommen noch Fälle, in denen durch experimentelle Probleme DNA von einem falschen Organismus sequenziert wird (ein bekanntes Beispiel ist eine kontaminierte menschliche DNA-Bibliothek der Firma Genethon, in der viele Hefesequenzen enthalten sind, die auch heute noch in Datenbanken als „vom Menschen stammend" geführt werden). Eine weitere häufig vorkommende fehlerhafte Annotation ist eine funktionelle Beschreibung, die für die gegebene Sequenz nicht zutrifft – Ursachen können im Experiment, aber auch in der halbautomatischen Annotation selber liegen. Durch die tägliche Datenflut ist manuelle Annotation nicht mehr in vollem Umfange möglich. Da Fehler in den Daten, der Interpretation der Daten und ihrer Annotation akkumulieren, ist bei Genomanalysen Vorsicht in der Interpretation der Resultate geboten, und Rückschlüsse auf einzelne Gene oder Proteine bedürfen deshalb derzeit noch der manuellen Kontrolle. Das Ausmaß zweifelhafter molekularbiologischer Daten wurde kürzlich anhand der weitverbreiteten Proteinstrukturdatenbank (PDB) ermittelt: über eine Million wahrscheinlicher Fehler unterschiedlichster Ursache wurden Anfang 1996 in den nur knapp 3500 verschiedenen Datenbankeinträgen ermittelt [5].

3 Computermethoden in der Funktionsvorhersage

3.1 Erstanalyse genomischer Sequenzen

Der erste Schritt bei der Funktionsvorhersage ist die Analyse genomischer Sequenzen, hauptsächlich zur Identifizierung der Gene. Da in der Praxis ein Genom in der Größenordnung von 3000 Megabasen (z.B. Mensch) nicht in einem Stück sequenziert wird, sondern über Jahre hinweg kleine Bruchstücke von 300–600 Basen mit (je nach Methodik) einer 3–10-fachen Redundanz (zur Reduzierung der Fehler) erzeugt und gesammelt werden, benötigt man Computermethoden schon für das Zusammensetzen („assembly") und Korrigieren der Sequenzen. Da im Erbgut oftmals identische oder fast identische Wiederholungen vorkommen, ist selbst diese mathematisch triviale Aufgabe nicht einfach zu bewältigen. Die Methoden sind recht komplex, da nicht nur durch Überlappungen Sequenzregionen aneinandergefügt werden, sondern auch externe Information herangezogen wird, z.B. Marker (kurze unterscheidbare Sequenzstücke, deren Reihenfolge bekannt ist, und mit denen längere Genomstücken vor der Sequenzierung kartiert werden). Oftmals werden bestimmte Regionen von verschiedenen Arbeitsgruppen unabhängig sequenziert und es gilt, diese Fremdinformation mitzubenutzen, um das Zusammensetzen zu beschleunigen.

Ein nächster Schritt ist die Identifizierung der Gene, die die Information für die Proteinsequenz enthalten. Dies ist speziell bei höheren Eukaryoten ein schwieriges Problem, da der Gehalt an Genen teilweise weniger als 3% des gesamten Erbgutes ausmachen kann und die Protein-kodierenden Regionen der Gene, die Exons, durch die nicht-kodierenden Introns unterbrochen sind. Die Zellen produzieren ihre Proteine, indem sie eine Kopie der gesamten Genregion erstellen und in mehreren späteren Schritten die Exons zusammenfügen („splicing"). Die Stellen, an denen das „splicing" erfolgt, enthalten schwache Signale; viele Computerprogramme versuchen, diese bei der Genvorhersage mitzubenutzen. Erschwerend kommt allerdings hinzu, daß viele Proteine in Varianten existieren, je nach Zellstatus also aus demselben Pool von Exons verschiedene mRNAs (die von Introns befreiten Kopien der Genabschnitte der DNA) geschaffen werden („alternative splicing"), die dann die eigentliche In-

formation zur Proteinproduktion enthalten. Das Hauptsignal in der Genidentifizierung ist aber der Gebrauch der Codons (DNA Triplett, das für eine bestimmte Aminosäure kodiert). Häufig benutzte Aminosäuren werden durch mehrere Codons kodiert, jedoch nutzen die Organismen nicht alle Varianten gleichmäßig und jeder Organismus hat ein bestimmtes Erwartungsmuster für die Benutzung dieser Codons. Dieses Signal ist aber nicht stark genug, was sich darin zeigt, daß derzeit neuronale Netze bessere Resultate liefern als regelbasierte Methoden (für einen Überblick gängiger Methoden und ihrer Stärken und Schwächen siehe [4]). Die Vorhersagegüte derzeitiger Methoden ist aber trotz besser klingender Angaben der jeweiligen Autoren immer noch im 70%-Bereich [4], was nicht ausreichend ist in der automatischen Analyse großer Genomdatenmengen.

In zunehmendem Maße werden schon in die Genvorhersage Sequenzdatenbanksuchen integriert: Ähnlichkeiten zu charakterisierten Proteinen sind ein starkes Indiz für das Vorhandensein von Exons. Leider bietet die Biologie wie immer viele Ausnahmen, die bedacht sein müssen, wie z.B. Pseudogene (duplizierte Gene, die nicht exprimiert werden und oftmals durch Leserahmenverschiebungen und eine hohe Mutationsrate charakterisiert sind). Neuere Überlegungen gehen davon aus, Genvorhersagen mit der Erkennung anderer Elemente in der DNA zu koppeln, z.B. mit RNA-Vorhersage und der Identifizierung von nichtkodierenden Wiederholungen (im menschlichen Erbmaterial befindet sich z.B. eine Vielzahl von sogenannten „Alu repeats"). Die Erkennung solcher Elemente schließt das Vorhandensein kodierender Bereiche aus. Gene sind umgeben von regulatorischen nichtkodierenden Elementen, z.B. Ribosomenbindungsstellen und Promotoren. An letztere binden regulatorische Proteine, um die Genregion zum Kopieren freizugeben. Auch diese Elemente tragen zur Signalverschärfung bei. Deren Erkennung erfordert aber andere Methoden – auch hier liegen neuronale Netze derzeit vorn. Die zur Zeit am häufigsten benutzten Genvorhersageprogramme sind GRAIL-Varianten [6] (ursprünglich ein reines neuronales Netz, später aber mit zahlreichen Expertenregeln ergänzt), GENEFINDER (P. Green, unveröffentlicht), das in verschiedenen Sequenzierungszentren eingesetzt wird, und GENEMARK [7], ein auf Markov-Modellen basierendes System, das sich bei Prokaryoten bewährt hat. Trotz des Nachholbedarfes in der Analyse nichtkodierender DNA sind wissensbasierte Systeme im Vormarsch, die Einzelmethoden für die Erkennung verschiedener DNA Elemente kombinieren.

3.2 Analyse intrinsischer Eigenschaften von Proteinen

Mit der Identifizierung eines Genes und der Vorhersage der Exons ist die Übersetzung der DNA in eine Proteinsequenz mittels des genetischen Codes möglich. Obwohl die erfolgversprechendste Methode zur Funktionsvorhersage eine Ähnlichkeitssuche in Datenbanken ist, sollten möglichst unabhängige Informationsquellen erschlossen werden, die auf verschiedenste funktionelle und strukturelle Eigenschaften des Zielproteins hinweisen können. Einige dieser Eigenschaften beeinflussen entscheidend die Parameter von Datenbankähnlichkeitssuchen und müssen deshalb zuerst ausgewertet werden [2; 8]. Ein wichtiger Test ist die Analyse der Aminosäurezusammensetzung von Proteinen. Nicht alle der 20 Aminosäuren kommen gleichhäufig vor – bestimmte Abweichungen von der Normalverteilung geben schon erste funktionelle Hinweise. Es gibt viele biologisch relevante Regionen, die mit einer Reduktion des 20-Aminosäuren-Alphabets auskommen [8; 9]. Ein typisches Beispiel sind membrandurchdringende Abschnitte, die wegen der Lipidumgebung meist hydrophob sein müssen und somit über mehr als 20 Positionen nur hydrophobe Aminosäuren enthalten – ein Charakteristikum, das bei normaler löslicher Umgebung unmöglich ist. Zur Ermittlung solcher Transmembranregionen gibt es mittlerweile Dutzende Programme, basierend auf neuronalen Netzen, sowie regelbasierten, statistischen und informationstheoretischen Ansätzen. Obwohl sich fast genausoviele Aminosäuren in Regionen mit einer anderen ungewöhnlichen Zusammensetzung befinden, den „Tripelhelices" („coiled coil"), dominiert hier derzeit nur ein Programm für die Erkennung solcher Regionen [10], die außer einer Reduzierung des Alphabetes zudem noch eine Positionsabhängigkeit bestimmter Aminosäuren aufweisen. Obwohl von vielen Extremen der Aminosäurekomposition die Funktion und Struktur noch weitgehend unbekannt ist, kann die einfache Ermittlung solcher Abweichungen zumindest bei Datenbanksuchen behilflich sein, da diese Regionen die Bewertungschemata von Ähnlichkeitssuchen unterlaufen und, wenn nicht vorher „maskiert", zu falschen Ergebnissen führen. Inzwischen gibt es mehrere Algorithmen, die versuchen, solche Regionen zu erkennen. Am häufigsten wird zur Zeit das Programm „SEG" eingesetzt, das verschiedene Komplexitätsmessungen beinhaltet [9].

Eine weitere Eigenschaft von vielen Proteinen sind interne Duplikationen, d.h. nicht-identische, oftmals nur sehr entfernt ähnliche, sich wiederholende Abschnitte. Obwohl hier Algorithmen von Laplacetransformationen bis zu Markovmodellen reichen, sind Programme zur Identifizierung solcher „repeats" nicht weitverbreitet, da ihre Identifizierungsrate noch relativ gering ist.

Eine weitere Kategorie von Signalen in Proteinsequenzen, die nicht durch Ähnlichkeitssuchen erfaßt werden können, sind posttranslationale Modifikationen, d.h. kurze Abschnitte, die auf eine Modifizierung einer Aminosäure hinweisen z.B. die kovalente Bindung von Zuckerresten. Diese Modifizierungen sind im 20-Buchstabenalphabet nicht direkt sichtbar, sind oftmals aber entscheidend für die Funktion des Proteins. Inzwischen gibt es Datenbanken solcher Modifikationen und Sammlungen kurzer Aminosäurekonsensusmuster, die auf bestimmte Modifikationen hin-

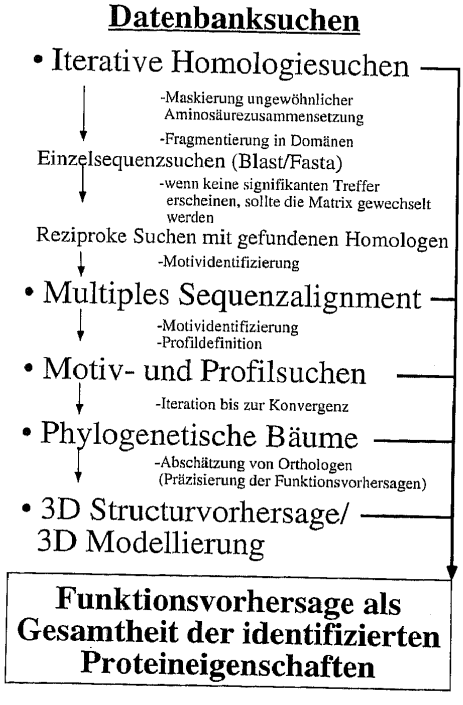
weisen. Leider sind diese in den meisten Fällen nur hinreichend, aber nicht notwendig – im Kontext anderer Eigenschaften eines Proteines verschärft sich aber das Signal. So kommen bestimmte Glykosylierungen nur in extrazellulären Proteinbereichen vor und spielen erst eine Rolle, wenn für das Zielprotein die zelluläre Lokalisierung bekannt ist. Für die Bestimmung der Lokalisierung eines Proteines (s. Bild 2) existieren verschiedene Verfahren. Allerdings ist bei allen Verfahren die Vorhersagegüte unter 80%, d.h. für das einzelne Protein können keine sicheren Angaben gemacht werden. Auch diese Eigenschaft muß in einem automatischen System zur Funktionvorhersage mit Wahrscheinlichkeiten behandelt werden. Der Fall vom Brustkrebsgen BRCA1 macht die Notwendigkeit guter Vorhersagen deutlich: Keine klaren Funktionsaussagen sind möglich; nicht einmal die Lokalisierung ist bekannt, um weitere Experimente planen zu können. Zwischen Ende 1995 und Mitte 1996 haben mehrere Gruppen in führenden Journalen Experimente publizert, die gegensätzliche Aussagen liefern und BRCA1 sowohl im Zellkern als auch im Cytoplasma oder im extrazellulären Raum ansiedeln. Die vergleichende Sequenzanalyse konnte auch hier einen entscheidenen Beitrag zur Klärung dieser Frage liefern und darüberhinaus funktionelle Eigenschaften von BRCA1 vorhersagen: Es ist mit hoher Wahrscheinlichkeit ein nukleares Protein, das einen Zellzyklus-Kontrollpunkt darstellt (s. [11] und Zitate in [11]).

3.3 Funktionsvorhersage durch vergleichende Sequenzanalysen

Wie kommt man vom Sequenzvergleich zu diesen funktionellen Aussagen? In den meisten Fällen wird eine Datenbanksuche durchgeführt (Bild 3) und die Funktion vom ähnlichsten Protein übernommen. Dies ist aber ein gefährliches Unterfangen, da es in einem Organismus viele Proteine gibt, die ähnlich zueinander sind, aber unterschiedliche Funktionen haben. Es ist noch nicht klar, welche Proteinfamilie im Menschen am größten ist, aber allein zwei aussichtsreiche Kandidatenfamilien (Proteinkinasen und an G-Proteine gekoppelte Rezeptoren) schätzt man auf ungefähr je 1000 Mitglieder. Obwohl diese bestimmte Teilfunktionen gemeinsam haben, hat jedes Mitglied einer Proteinfamilie eine spezielle Aufgabe. Somit ist die Vorhersage der Funktion durch Übernahme der Beschreibung eines ähnlichen Proteins, so wie es derzeit noch üblich ist, in den meisten Fällen nicht korrekt. Eine weitere Komplikation ist der modulare Aufbau vieler Proteine. Es kann eine kurze Region (Baustein, Domäne) einer Datenbanksequenz durchaus ähnlich zum Zielprotein sein, aber der überwiegende Teil der beiden Proteine ist völlig verschieden und dementsprechend auch ihre Funktion (obwohl die ähnliche Domäne eine gemeinsame Teilfunktion ausüben kann, aber nicht muß). Solche Überlegungen muß man sich bei der Analyse ganzer Chromosomen oder Genome immer vor Augen halten, genauso wie die Stärken und Schwächen der Datenbanksuchmethoden. Auch hier kann es leicht zu Überinterpretationen kommen, d.h. aus zufälligen Ähnlichkeiten werden evolutionäre Verwandtschaften, Homologien, gedeutet. Die Erkennung von Homologien zwischen zwei Sequenzen, d.h. der Annahme einer gemeinsamen Vorläufersequenz, bildet die Grundlage der Datenbanksuchen. Im Normalfall ist die gemeinsame Wurzel zweier Sequenzen viele Millionen Jahre alt und Mutationen haben die beiden zu untersuchenden Sequenzen stark verändert – ihre Buchstabenabfolge läßt also kaum noch Ähnlichkeiten erkennen. Trotzdem ist die Information über ähnliche Raumstruktur und Funktion in den Buchstabenketten verschlüsselt und kann über Mustererkennungsverfahren oftmals noch identifiziert werden. Eine Hilfe dabei sind Ähnlichkeitsmaße zwischen den einzelnen Buchstaben, den Aminosäuren, die teilweise ähnliche sterische und physikochemische Eigenschaften besitzen und somit nicht unabhängig voneinander sind. Allerdings ist es nach wie vor ein noch nicht vollständig geklärtes Problem, Sequenzen optimal gegeneinander auszurichten, so daß die Ähnlichkeit maximiert wird, da die Komplexität durch Einschübe (eine Sequenz ist länger als die andere) noch erhöht wird. Die Behandlung von Einschüben erfordert sogenannte dynamische Programmierung, kostet Rechenzeit und wird angesichts der riesigen Datenmengen derzeit in der Genomanalyse nur in Einzelfällen angewandt.

International werden schnelle Programme der Blast-Serie [8] am häufigsten benutzt, die im ersten Schritt nur Teilstringe nach Identitäten durchsuchen, danach mit Hilfe von Ähnlichkeitsmaßen und Statistiken die Länge der Treffer optimieren und sie nach bestimmten Kriterien sortieren. Gerade in großen Sequenzierungsprojekten wird Blast oft als einziges Datenbanksuchverfahren eingesetzt, teilweise noch durch den Vergleich mit in speziellen Datenbanken abgespeicherten Sequenzmustern für charakteristische funktionelle Regionen erweitert. Diese Muster sind meist konservierte Regionen (d.h. bestimmte Buchstabenfolgen oder unterliegende Aminosäureei-

Bild 3: Iterative Datenbanksuchen und daran ansetzende Methoden zur Funktionsvorhersage.



genschaften sind in sonst stark unterschiedlichen Sequenzen invariant), die durch Studium einzelner Proteinfamilien charakterisiert werden konnten.

Der Erwartungswert für einen Treffer bei einer Blast-Datenbanksuche mit einem signifikanten Schwellenwert hängt von der Species ab: bei Bakterien ist er im Bereich von 80%, bei Hefe um 70% und beim Menschen bei 50%. Hierbei muß aber berücksichtigt werden, daß viele der Datenbanksequenzen nicht oder nur ungenügend funktionell charakterisiert sind, d.h. der Prozentsatz für Funktionsvorhersagen ist gerade bei menschlichen Sequenzen noch ungenügend. Die Trefferausbeute kann aber durch Anwendung von komplexeren Suchverfahren und sensitiveren Methoden um 10-20% gesteigert werden [12, 13]. Weiterhin kann auch die Anzahl der funktionellen Vorhersagen prozentual gesteigert und präzisiert werden, wenn eine bessere Analyse der gefundenen Ähnlichkeiten erfolgt [13]. Die Ausbeuterate verbessert sich auch mit der Zeit, da der „Sequenzraum" immer dichter und die funktionelle Charakterisierung immer besser wird.

Ein typischer Ablauf einer iterativen Suche (Bild 3) ist der Start mit einer Standarddatenbanksuche (z.B. Blast), und bei fehlenden Treffern die Variation der unterliegenden Ähnlichkeitsmatrizen. Ein nächster Schritt ist die Identifizierung von Mustern im Grauzonenbereich der Standardsuchprogramme, d.h. der ausführlichen Analyse von Treffern, die nicht als signifikant eingestuft werden können, aber trotzdem Kandidaten für entfernte Homologien darstellen. Die erkannten Muster (gleiche Abfolgen von Buchstaben sind identisch in mehreren, nicht klar verwandten Proteinen) können dann für Motivsuchverfahren eingesetzt werden, in denen man sich nur auf diese lokalen Muster konzentriert [12]. Oftmals sind solche Muster der einzige sichtbare Hinweise auf eine funktionelle Region, die einer Proteinfamilie gemeinsam ist und auf die man aus der (derzeit noch manuellen) Analyse der biochemischen Charakterisierung einzelner Mitglieder schließen kann. Um die Proteinfamilie möglichst komplett zu erkennen, sind Iterationen nötig: neu gefundene Mitglieder müssen in die Berechnung von Matrizen/Mustern für eine erneute Datenbanksuche einfließen. Reziproke Suchen mit gefundenen Kandidaten sind essentiell. Oftmals müssen auch noch Profilsuchen mit globaleren Ausgangsregionen durchgeführt werden [12].

Das ganze Arsenal von Methoden, Parametern und die Reihenfolge ihrer Anwendung reicht für wochenlange Beschäftigung mit einer einzigen Sequenz aus, die sich allerdings in besonderen Fällen (z.B. Krankheitsgenen) auszahlen kann. Im Zeitalter der Genomsequenzierung sollten aber Tausende Sequenzen in wenigen Tagen analysiert werden. Hier müssen Kompromisse in der Senistivität gemacht werden und es Bedarf einer genauen Kenntnis der Prozedur, um trotzdem eine hohe und korrekte Funktionsvorhersagerate zu erreichen. Wir haben z.B. 1992 [14; 15] mit der Analyse des ersten komplett sequenzierten eukaryotischen Chromosomes (Hefechromosom III [16]) begonnen, ein automatisiertes Verfahren zur Funktionsvorhersage zu entwickeln. Trotz vieler Mannjahre Arbeit, die in ein erfolgversprechendes System, GENEQUIZ, mündeten [13], sind noch nicht alle der genannten Punkte integriert und besonders die letzten Schritte der präzise Vorhersage von Funktionen aus ermittelten homologen Proteinen in der Datenbank bereiten noch Schwierigkeiten – Sequenzdatenbanken sind noch zu inhomogen und funktionelle Informationen zu ungenau gespeichert.

3.4 Funktionsvorhersagen durch Analyse kompletter Genome

Mit dem Vorhandensein kompletter sequenzierter Genome eröffnen sich aber weitere indirekte Möglichkeiten zur Funktionsvorhersage: die Ausnutzung von Wissen über Stoffwechselwege. Wenn z.B. bestimmte metabolische Reaktionsketten hinreichend bekannt sind und durch Homologiesuchen einige Mitglieder der Kette zweifelsfrei identifiziert worden sind, müssen entweder die weiteren Mitglieder der Kette auch vorhanden sein, oder es können in limitiertem Umfange Abweichungen des Stoffwechselweges auftreten, die mit Zusatzinformation über benachbarte Wege vorhersagbar sind [17]. Auch hier gibt es natürlich verschiedene Schwierigkeiten zu überwinden. Durch die Analyse kompletter Genome konnte kürzlich gezeigt werden, daß in verschiedenen Organismen unterschiedliche, nicht-ähnliche Gene für die gleiche Funktion kodieren und daß dieses Phänomen relativ häufig vorzukommen scheint [18]. Das Nichterkennen eines Proteins einer Kette bedeutet somit noch lange nicht, daß die Kette nicht geschlossen und funktionstüchtig ist.

Weitere Aussagen können aus der Anordnung von Genen im Genom getroffen werden. In Bakterien sind z.B. häufig Stoffwechelwege in einer Einheit, dem Operon, kodiert, d.h. sie werden als ein gemeinsames Stück von der DNA abgelesen und dann erst später in die eigentlichen Gene zerlegt. In Eukaryoten sind solche Abhängigkeiten kaum noch zu finden, doch sind z.B. Untereinheiten eines Proteines (unabhängige Proteinketten, die aber nur im Komplex ihre Funktion ausüben können) oft noch in benachbarter Position im Genom.

3.5 Methodenentwicklung und Datenverwaltung im Internet

Es ist in diesem Artikel bewußt wenig auf spezielle Methoden eingegangen worden, da sich das gesamte Gebiet rasant verändert. Durch das Medium Internet werden fast täglich neue Programme für Teillösungen angeboten, die wieder ein paar der Probleme reduzieren. Neue, besser annotierte Spezialdatenbanken entstehen, die je nach Aufgabenstellung mitgenutzt werden können. Neben der Datenexplosion gibt es also auch eine Methodenexplosion, und durch die freie Verteilung im Internet ist es oftmals schwer, die Güte der Methoden und ihre Kompatibilität genau zu bestimmen. Da zur

Zeit gerade die ersten kompletten Genome fertig sequenziert werden, ist ein gewisser Zeitdruck für deren Auswertung gegeben. Oftmals leidet darunter die Qualität der Analysen und fehlerhaft interpretierte Information wird in die Datenbanken eingespeist.

Gegenwärtige Methoden richten sich auf die Verbindung zwischen Datenbanken, man kann sich manuell seine Informationen in einem Netz von mehreren hundert Spezialdatenbanken zusammensuchen. Trotz „links", d.h. einer Verbindung von Datenbanken, ist derzeit jedoch kaum eine gezielte Extraktion unterschiedlicher Daten möglich. Viele der sogenanntemn Datenbanken sind eigentlich nur Datensammlungen – selbst die häufig benutzten Sequenzdatenbanken stellen erst jetzt auf Datenbanksysteme um. Besonders schwierig erweist sich die erwähnte schlecht strukturierte funktionelle Information, „data mining"-Technologien werden aber hoffentlich schon bald auch diese Lücke schließen

Ein weiteres Problem sind die oftmals unzureichenden Netzwerkverbindungen, bei denen hinsichtlich detaillierter Analyse oft Kompromisse gemacht werden müssen. Da es sich aber in der Genom und Sequenzanalyse vorwiegend um lineare Information handelt, machen diese Daten nur einen Bruchteil des internationalen Datentranfers aus. Rechnerleistungen und Speicherkapazitäten sind aus ähnlichen Gründen ebenfalls keine limitierenden Faktoren.

4 Wo sind die Engpässe?

Viele der Algorithmen zur Lösung von Teilproblemen scheinen an ihre Grenzen gekommen zu sein. So sind existierende Ähnlichkeitssuchen schon sehr ergiebig und die vielen neuen Daten machen die Suchen auch einfacher. Wenige Lösungen sind für die Automatisierung verschiedener Schritte in der Funktionsvorhersage vorhanden. Die Datenerzeuger haben meist ihre eigene Software und entwickeln sie weiter, und das Problem ist zu komplex, als daß es von kleineren Arbeitsgruppen allein bewältigt werden kann. Ein Hauptproblem besteht in der Extraktion von Daten aus existierenden, nicht immer gut strukturierten Datensammlungen und Datenbanken. Hier sind eindeutig neue Technologien gefordert. Integrierte Systeme mit anspruchsvollen und dennoch nutzerfreundlichen grafischen Oberflächen, in denen nicht nur alle Daten verfügbar gemacht, sondern auch aufbereitet werden, sind für die nahe Zukunft zu erwarten. Hier macht besonders die Aufarbeitung experimenteller Arbeiten Schwierigkeiten, da fehlerhafte Information bisher kaum abgeschätzt werden kann und strukturierte Informationsaufarbeitung von funktionellen Daten bislang kaum möglich ist. Selbst bei Details wie Gen- oder Proteinnamen gibt es nur wenige Konventionen und keine Systematik.

Eine gute Funktionsvorhersage ist natürlich Voraussetzung für die Lösung von globaleren Aufgaben wie der Identifizierung von Stoffwechselwegen und der Modellierung von Regulationsmechanismen. Hier wird sich ein weiterer Schwerpunkt in der Bioinformatik herausbilden.

Literatur

[1] *Fleischmann et al.*: In: Science 269 (1995), S. 496–512.

[2] *Bork, P., Ouzounis, C. Sander, C.*: Curr. Opinion. Struct.Biol. 4. (1994) S. 393–403.

[3] *Bork, P., Bairoch, A.*: Trends Genet. (1996), Oktober-Ausgabe.

[4] Burseé, M., *Guigo, R.*: Genomics 34 (1996) S. 353–367.

[5] *Hooft, R. W. W., Vriend, G., Sander, C., Abola, E. E.*: Nature 381 (1996) S. 272.

[6] *Xu, Y., Einstein, J. R., Mural, R. J., Shah, M., Uberbacher E. C.*: ISMB 2 (1994), S. 376–384.

[7] *Borodovsky, M., McIninch, J. D., Koonin, E. V., Rudd, K. E., Medigue, C., Danchin, A.*: Nucl.Acid. Res. 23 (1995), S. 3554–3562.

[8] *Altschul, S. F., Boguski, M. S., Gish, W., Wootton, J. C.*: Nature Genet. 6 (1994), S. 119–129.

[9] *Wootton. J. C.*: Curr. Opin. Struct. Biol. 4 (1994), 413–421.

[10] *Lupas, A., Van Dyke, M., Stock, J.*: Science 252 (1991), S. 1162–1164.

[11] *Koonin, E. V., Altschul, S. F., Bork, P.*: Nature Genet. 13, (1996), S. 266–268.

[12] *Bork, P., Gibson, T.*: Meth. Enzymol. 266, (1996), S. 162–184.

[13] *Casari, G., Andrade, M., Bork, P., Boyle, J., Daruvar, A., Ouzounis, C., Schneider, R., Tamames, A., Valencia, A., Sander, C.*: Nature 376 (1995), S. 647–648.

[14] *Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., Sonnhammer, E.*: Protein Science 1 (1992), S. 1677–1690.

[15] *Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., Sonnhammer, E.*: Nature 358 (1992), S. 358.

[16] *Oliver, S. G. et al.*: Nature 357 (1992), S. 38–46.

[17] *Tatusov, R. L., Mushegian, A. R., Bork, P., Brown, N. P., Hayes, W. S., Borodovsky, M., Rudd, K. E., Koonin, E. V.*: Curr. Biol. 6 (1996), S. 279–291.

[18] *Koonin, E. V., Mushegian, A. R., Bork, P.*: Trends Genet. (1996), September-Ausgabe.

[19] *Woese, C. R.*: Microbiol. Rev. (1987), S 221–271.

[20] *Voet, D., Voet, J.*: Biochemie; Weinheim, New York, Basel, VCH 1992.

Dr. habil. Peer Bork
Max-Delbrück-Centrum für Molekulare Medizin, 13122 Berlin-Buch und EMBL, Meyerhofstr. 1, 69012 Heidelberg,
Email: bork@embl-heidelberg.de