Entschlüsselung von Proteinfunktionen mit Hilfe des Computers: Erkennung und Interpretation entfernter Sequenzähnlichkeiten

Peer Bork

EMBL, 6900 Heidelberg
und
Max-Delbrück-Centrum für Molekulare Medizin, 1115 Berlin-Buch

Zusammenfassung

Anhand verschiedener Beispiele wird versucht, die Möglichkeiten der Sequenzanalyse bei der Erklärung von molekularer Proteinfunktion aufzuzeigen. Den Hauptanteil machen dabei die Homologiesuchen aus, die auf heuristischen Methoden basieren. Schon heute sind sie unverzichtbarer Bestandteil in allen an Genomprojekten beteiligten Labors. Doch eine sensitive Auswertung der Sequenzdaten erfordert eine Kombination vieler zusätzlicher Methoden, wie Aminosäurekompositions- Stammbaum-, Muster- und Strukturanalysen. Trotz erstaunlich guter Ergebnisse bei Testbeispielen ist einerseits eine Automatisierung bei komplexen Aufgaben, wie der Analyse eines ganzen Chromosomes, andererseits eine Erhöhung der Sensitivität bei Detailproblemen wie Bindungsstellenvorhersage nötig.

1. Einleitung

In den letzten Jahrzehnten hat sich der Zugang zu einem Datenmassiv eröffnet das entscheidend zumVerständnis molekularbiologischer Prozesse beitragen könnte - das in Textform vorliegende genetische Material. Große Hoffnungen verknüpfen sich mit der Entschlüsselung genetischer Information, die z. B. Aufschluß über Erbkrankheiten ermöglicht. Es wurden deshalb vor einigen Jahren Genomsequenzierungsprojekte für einige Organismen initiiert (s. z.B. Tab.1, die als Modelle für weitere Vorhaben dienen sollen. Durch diesen Übergang zur "Massenproduktion" beträgt der Anteil der innerhalb von Genomprojekten publizierten Sequenzdaten schon jetzt ca. 10%.

Tab1. Zusammenstellung und Stand einiger Genomprojekte. Weitere Projekte für Spezies wie Maus, Kresse oder Mycoplasma wurden bereits initiert.

	Anzahl sequenzierter Gene	Zu anderen Genen verwandt	Gesamt-anzahl	Voraussichtl. Komplettierungs-datum
Genomprojekte				
C.elegans			*≈15000*	*2000*
Chromosom III (Teil)	32	14(44%)		
Hefe			*≈7000*	*2002*
Chromosom III	176	67(38%)	*176*	*1992*
Chromosom IX (Teil)	46	15(33%)		
Bibliotheken expressionierter Gene				
Mensch			*≈50000*	*2010*
Gehirn	≈1400	406(30%)		
Caenorhabditis elegans			*≈15000*	*2000*
St.Louis-Cambridge	1517	512(34%)		
NIH	585	210(36%)		
E.coli	≈2000	≈800(40%)	*≈4000*	*1996*

Eine Datenflut ist absehbar, doch daß diese schneller als erwartet auf uns zukommen kann, verdeutlichen die kürzlich veröffentlichten Genkarten zweier menschlicher Chromosomen (Y und 21; [1,2]). Die überlappenden DNA Stücken wurden mit Hilfe sogenannter künstlicher Hefechromosomen (YAC: yeast artificial chromosome) konstruiert. Somit können schon jetzt die direkten Sequenzierungsarbeiten beginnen, die nach der Erstellung solcher genetischer Genkarten den zweiten, entscheidenden Schritt in einem Genomprojekt darstellen. Man rechnet nach diesem unerwartet schnellen Fortschreiten der Arbeiten nunmehr mit einer vollständigen Genkarte des Menschen in spätestens 5 Jahren, womit sich der in Tabelle 1 angegebene Zeitpunkt der Vollendung noch erheblich nach vorne verschieben dürfte. Mit diesem Tempo der Datenproduktion können sowohl die biochemische Charakterisierung als auch die 3D-Strukturaufklärung von Proteinen trotz immer besser werdender Methodik nicht mehr mithalten, was zu immer mehr Rohdaten führt, über die immer weniger bekannt ist.

Hier ist klar die Struktur- und Funktionsanalyse von Sequenzdaten gefordert. Es gilt also möglichst viel, der in diesen Daten enthaltenen Information zu entschlüsseln (d.h. z.B. Aufzudeckung extrem entfernter Verwandtschaften (Homologien) oder auch Vorhersage der 3D-Struktur), um die

molekulare Funktion der entsprechenden Proteine verstehen zu lernen und Aussagen auf anderen Ebenen (wie z.B über genetisch-evolutionäre Mechanismen) machen zu können. Dabei tritt natürlich die Frage auf, ob die derzeitigen Methoden mit der zu erwartenden Datenflut zurechtkommen (siehe Abschnitt 2). Berücksichtigt werden müssen aber auch bei dem Sequenzvergleich viele Verkomplizierungen, wie zum Beispiel der modulare Aufbau vieler Proteinen, der zu nicht eindeutigen Funktionszuordnungen führt (siehe Abschnitt 3). Wie durch Sequenzanalyse gezeigt werden konnte, kann es sogar zum horizontalen Austausch von Proteinen oder Proteinteilabschnitten kommen, d.h. Organismen scheinen in der Lage zu sein fremdes genetisches Material in die eigene Vererbungsmaschinerie zu implementieren (siehe Abschnitt 4). Basierend auf den derzeitigen Erfahrungen in der Sequenzanalyse werden einige Erfordernisse in der Methodenentwicklung aufgezeigt (siehe Abschnitt 5), um die Funktionszuordnung auf der Basis von Sequenzvergeichen zu effektivieren.

2. Sequenzanalyse des kompletten Hefe Chromosoms III

Um den derzeitigen Entwicklungsstand der Sequenzanalyse und die Möglichkeiten der Struktur- und Funktionszuordnung einschätzen zu können entwickelten wir ein Netzwerk aus Computermethoden, daß bestehend aus Standardprogrammen und Eigenentwicklungen, von uns am ersten vollständig aufgeklärten eukaryotischen Chromosomen (Hefe Chromosom III [3]) getestetet und optimiert wurde [4]. Die Auswertung dieser Fallstudie ergab (Abb.l), daß für mehr als 40% der wahrscheinlich 176 Proteine dieses Chromosoms eine definierte Funktion vorhergesagt werden kann, für weitere 20% sind zumindestens Einschränkungen des möglichen Funktionsspektrums erfolgt (z.B. Transmembranprotein, Lokalisation im Nucleus, ER-Durchquerungssignal etc.). Interessanterweise konnten für fast 15% aller ORF's (open reading frames; offene Leserahmen) dieses Hefechromosoms Ähnlichkeiten zu Proteinen mit bekannter Raumstruktur festgestellt werden (Abb.l), was mit impliziten 3D-Strukturvorhersagen einhergeht. Das Wissen über eine Protein-3D-Struktur ermöglicht wiederum sensitivere Homologiesuchtechniken, die einen Informationstransfer (Struktur und Funktion!) auf extrem entfernte Verwandte ermöglicht [5].

Dieser erstaunlich hohe Prozentsatz an klassifizierbaren Primärstrukturen resultierte nicht zuletzt aus der sorgfältigen Analyse sehr entfernter Ähnlichkeiten die von Standardhomologiesuchprogrammen übersehen werden, deren Signifikanz aber mit verschiedenen Methoden nachgewiesen

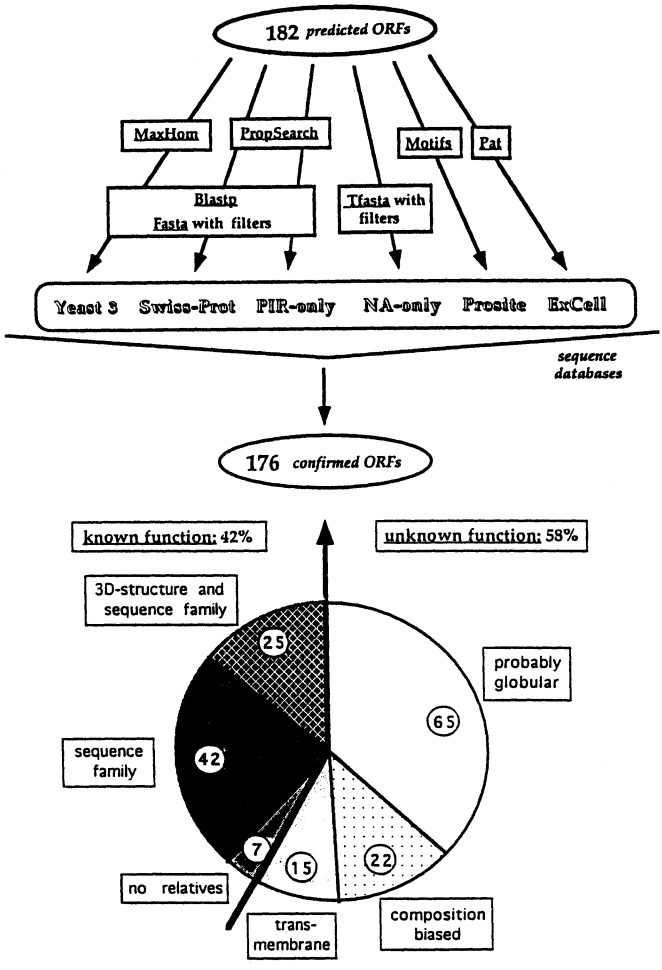

Abb.1 a) Methoden- und Datenbankeinsatz zur Sequenzanalyse des Hefechromosoms III. Derzeitigen Standarddatenbanksuchprogrammen wie Blastp [9] und Fasta [7] wurden zusätzlich nach verschiedenen Kriterien gefiltert [16]. Des weiteren wurden Profil- und Mustersuchen eingesetzt, wenn eine Erstzuordnung erfolgen konnte. Verschiedene Datenbanken wurden benutzt. Eine eigens erstellte Datenbank von Nukleinsäuresequenzen, die noch nicht in Proteindatenbanken übersetzt wurden, half zum Beispiel bei der Erkennung von 6 offenen Leserahmen (ORFs), die eindutig regulatorische DNA-Elemente darstellen und nicht codiert werden. PROSITE und EXCELL sind Musterdatenbanken, in denen markante (Signatur-) Regionen aus bekannten Protein (Domänen) -familien gespeichert sind [15, 12]. **b)** Anteil an Funktions- und Strukturzuordnung der wahrscheinlichen Proteine des Hefechromosoms III. Für mehr als 50% dieser Proteine können bislang kaum Vorhersagen getroffen werden [4]. Obwohl z.B. die Identifizierung von Transmembranregionen das Funktionsspektrum erheblich einengen, bleibt die eigentliche Aufgabe (Transporter, Rezeptor, Adhäsionsmolekül?) immer noch unerkannt. Das Verständnis der molekularen Funktionen erfordert das Wissen der 3D-Struktur, die in nur 15% aller Fälle und auch nur indirekt angenommen werden kann.

werden kann [6].Das folgende Beispiel (Abb.2) zeigt ein sogenanntes "multiples alignment" eines der unbekannten ORFs aus dem Hefechromosom III mit verschiedenen Methyltransferasen.

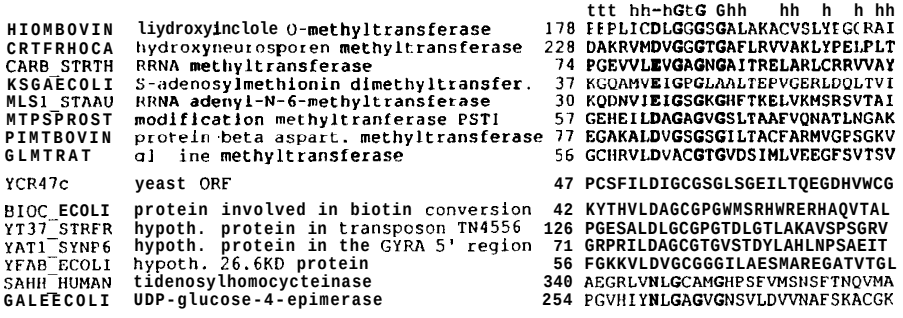

Abb. 2 Übereinanderlagung konservierter Bereiche in Methyltransferasen (oben) mit dem zu studierenden Hefeprotein (Mitte). Durch die Charakterisierung dieser konservierten Region lassen sich auch für weitere Proteine aus der Datenbank mit einem ähnlichen Muster (unten) Funktionsvorhersagen treffen. "Ähnlichkeit beruht hier weniger auf den in Buchstabencode dargestellten Aminosäuren als auf sich dahinter verbergenen sterischen und physikochemischen Eigenschaften, die in verschiedenen "Buchstaben" versteckt sein können.

Die Ähnlichkeit bezieht sich nur auf eine beschränkte Region und auch dort sind nur wenige Reste komplett in allen diesen Proteinen erhalten. Wir haben das Ergebnis einer Datenbanksuche mit einem Standardprogramm (FASTA [7]) nach verschiedenen Parametern gefiltert und Teilsegmente der Suchsequenz (des ORFs), die immer wieder eine lokale Ähnlichkeit zu anderen Proteinen aufwiesen extrahiert. Im paarweisen Vergleich würde eine solche schwache, lokale Ähnlichkeit keinem Signifikanztest standhalten, doch man kann gezielt positionsabhängig Eigenschaften mit Mustererkennungsprogrammen beschreiben und z.B. von dem in Abb.2 dargestellten Alignment ein Profil erzeugen und dieses zur erneuten Datenbanksuche verwenden. In Falle einer eindeutigen Diskriminierung zwischen den "Lernsequenzen" des Alignments und einiger neuer Kandidaten einerseits und dem "Hintergrundrauschen" nicht verwandter Proteinsequenzen andererseits, können diese Kandidaten den Lernsatz iterativ verbessern. Bei Konvergenz ergibt sich ein spezifisches Muster (s. Abb.2) das in einer abgegrenzten Sequenzfamilie funktionelle und/oder strukturelle Bedeutung hat

[8]. In diesem Fall ist bekannt, daß diese Sequenzregion in die Übertragung von Methylgruppen involviert ist. Solche und auch andere Beispiele zeigen, daß trotz zunehmender Automatisierung in der Homologiesuche menschliches Wissen eingebracht werden muß um Grenzfälle (Sequenzähnlichkeiten unterhalb bestimmter Signifikanzabschätzungen richtig zu deuten. Dies führt zu der Frage der Geschwindigkeit solcher Analysen angesichts großer Datenmengen (Tab.1). Im Falle des Hefechromosomes III fielen 'nur' 182 offene Leserahmen an, für deren Analyse wir immerhin 14 Tage benötigten [6]. Auch wenn in Zukunft also das menschliche Expertenwissen das Nadelöhr sein mag - die geschwindigkeitslimitierenden Schritte im Analyseprozess sind zur Zeit immer noch die Datenbanksuchen (Tab.2)

Tab.2 Homologiesuche von 182 Proteinsequenzen gegen verschiedene Datenbanken mit derzeitigen Standardmethoden.

Programm	time	Computer*	Datenbankgröße
BlastP [9]	3h	Silicon Graphics 4D/480	35000 Sequenzen
Fasta [7]	90h	VAX 6040	35000 Sequenzen
Fasta [7]	15h	Alliant FX 2800	35000 Sequenzen
TFasta [7]	23d	Silicon Graphics 4D/480	300000 Sequenzen
Extrapolation für die Auswertung eines kleinen menschlichen Chromosomes (ca. 5000 Sequenzen) in vielleicht schon 3 Jahren			
TFasta [7]	5000d	Silicon Graphics 4D/480	3000000 Sequenzen
Blaze*	160h	Maspar MP1	3000000 Sequenzen

bezogen auf 1 CPU, mit Ausnahme der Maspar MP14K-Prozessoren

* In der Entwicklung befindliches kommerzielles Produkt, das seine Geschwindigkeit durch Parallelisierung erhält.

Natürlich können verschiedene Aufgaben parallel abgearbeitet werden. Trotzdem gibt es für viele Problemstellungen innerhalb der Ähnlichkeitssuchen noch keine Lösungen. Es zeichnet sich ab, daß bestimmte funktionelle Merkmale anders als durch positionabhängige Textanalyse prognostiziert werden müssen Beispiele sind Funktionen, die auf einer Häufung von bestimmten Aminosäuren basieren, nicht aber auf positionsabhängige Wechselwirkungen beruhen. Solche 'ungewöhnlichen' Aminosäurezusammensetzungen stellen erhebliche Probleme bei der Signifikanzabschätzung gefundener Ähnlichkeiten dar (für einen Überblick derzeit angewandter mathematisch-statistischer Modelle siehe Referenz [10]). Ein weiteres Erschwernis bei der Homologiesuche ist zum Beispiel auch die durch genetische Mechanismen bedingte

Durcheinandermischung ganzer Genabschnitte ("exon shuffling"), die dann als funktionell und strukturell unabhängige Bausteine (Module) in unterschiedlichsten Proteinen zum Einsatz kommen [ll]. Das Ergebnis solcher Prozesse sind modulartig aufgebaute Proteine, die nur partielle Ähnlichkeit zu anderen Moleküle aufweisen (für einen Überblick bisher bekannter Module siehe Referenz [12])

3. Entschlüsselung der modularen Architektur "moderner "Proteine

Als "moderne" Proteine werden hier solche bezeichnet, die nur in höherentwickelten (mehrzelligen) Organismen vorkommen, und die dementsprechend nur in bestimmten Prozessen wie Differenzierung oder Zell-Zell-Wechselwirkungen eine Rolle spielen (Abb.2). Die Bausteine solcher modularen Proteine lassen sich mit derzeitigen Standardverfahren nur schlecht nachweisen, da es sich immer nur um Teilabschnitte handelt, die zu dem auch noch sehr in der Sequenz variieren.

Eine effiziente Methode zur Beschreibung von Struktur- und Funktionsparametern, die auch in sehr entfernt verwandten Proteinen Gültigkeit besitzen, wurde schon erwähnt: Sequenzkonsensusmuster. Module können derzeit oft nur durch sehr flexible Konsensusmethoden beschrieben werden. Wir haben ein solches sensitives Verfahren entwickelt [8] und bereits an mehr als 100 Modulen getestet. Alle diese Domänen (Module) werden über ihre spezifischen Konsensusmuster in einer Datenbank erfaßt und stellen einen neuen Zugang zur Homologiesuche von Domänen dar: Neu sequenzierte Proteine werden mit der Musterdatenbank verglichen und entsprechende Module werden sofort ermittelt. Der Anteil dieser Module am Proteinbestand ist nicht unbeträchtlich allein das Modul, welches zuerst in Immunoglobulinen gefunden wurde schätzt man heute als Bestandteil von mindestens 5% aller bekannter Proteine. Weitere weitverbreitete Module wie EGF ("epidermal growth factor"; siehe Fig.3) oder auch eine Domäne die zuerst in dem Matrixprotein Fibronektin identifiziert wurde, kommen in ca. 2-3% aller Primärstrukturen vor [13]. Trotz dieser hohen Prozentzahlen wurden solche Module weder in Pflanzen, noch in Hefe gefunden, wohl aber in einigen Bakterien, die ja offensichtlich in der Evolution viel weiter von den Tieren entfernt sind als Pflanzen oder Hefe. Auch hier kann die Erkennung, aber auch die Interpretation von Sequenzähnlichkeiten Aufschluß über mögliche Gründe geben.

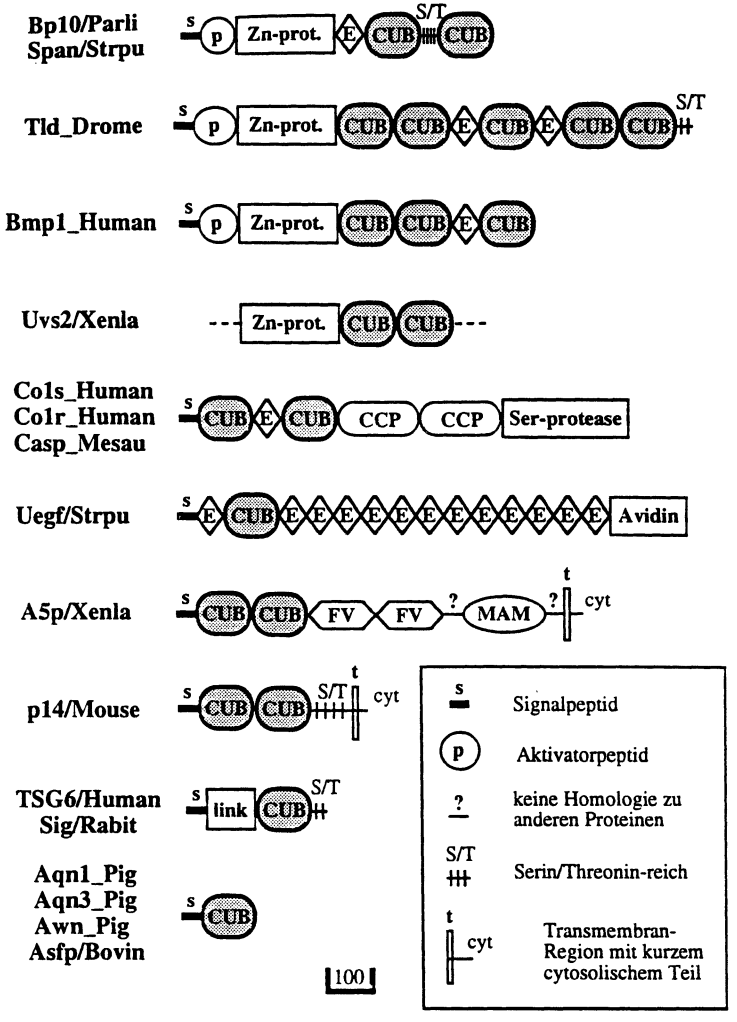

Abb.3 Modularer Aufbau einiger Proteine, deren einzige Gemeinsamkeit oftmals nur der CUB Baustein ist (dunkel), der auch mehrmals in einem Protein vorkommen kann (interne Duplikation von genetischem Material). Die paarweisen Ähnlichkeiten innerhalb verschiedener CUB Module sind oftmals unterhalb jeglicher Signifikanzabschätzung. Alle Mitglieder dieser Familie konnten dennoch mittels Mustersuchen eindeutig identifiziert werden. Mit diesen Homologien im Hintergrund konnten nun durch Analogieschlüsse auch Funktionsvorhersagen gemacht werden. Da für die meisten dieser Proteine bereits eine Rolle in Entwicklungsprozessen (Organogenese, Embryogenese) experimentell nachgewiesen wurde, liegen analoge Aufgaben für die restlichen Mitglieder dieser CUB-Familie nahe. Auf molekularer Ebene scheint der CUB-Baustein eine gezielte Carbohydratbindung innerhalb einer Signalweiterleitungskette zu realisieren.

4. Warum Bakterien Proteindomänen stehlen

Die Aufdeckung von "evolutionären Unregelmäßigkeiten" setzt neben der Erkennung von entfernten Verwandschaften eine Clusteranalyse voraus, mit der dann Dendrogramme (oder evolutionären Stammbäumen) aus einem multiplen Alignment ähnlicher Sequenzen heraus berechnet werden können. Eine sorgfältige Phylogenie-Analyse kann z.B. horizontalen Genaustausch aufdecken, d.h. den Erwerb fremden genetischen Materials (z.B. durch Plasmide oder Viren). Dies soll hier am Beispiel des Fibronektin Typ III Modules (eines Bausteines von ca. 90 Aminosäuren Länge, ähnlich den Immunoglobulindomänen) erläutert werden. Unter den über 300 Bausteinen dieses Types, die mit unserer Mustererkennungsmethode in Sequenzdatenbanken identifiziert wurden befanden sich auch 13 dieser Module in 7 verschiedenen bakteriellen Enzymen (Abb.4).

Aus dem multiplen Alignment aller Fibronektin Typ III Domänen wurden Dendrogramme konstruiert, die eindeutig die Abstammung der bakteriellen Domänen voneinander verdeutlichen (Abb.5). Aus zwei Phänomenen kann man nun den Erwerb dieser bakteriellen "Urdomäne" von einem eukaryotischen Genom ableiten. **1.** Alle bakteriellen Module sind viel ähnlicher zu bestimmten eukaryotischen Sequenzen als diese untereinander. 2. Das Vorkommen der prokaryotischen Module entspricht nicht der bakteriellen Phylogenie: Taxonomisch entfernte grampositive und gramnegative Bakterien besitzen sehr ähnliche Fibronektin Typ III Module, aber diese untereinander sehr unähnliche. Der horizontale Austausch von genetischem Material innerhalb von Bakterien durch Plasmide ist bereits bekannt (Alle diese hier beschriebenen Bakterien coexistieren in oberen Erdbodenschichten!). Warum sollten Bakterien Proteinabschnitte von höheren Eukaryoten (Tieren) übernehmen? Die biologische Zusammenhänge bieten eine Erklärung an: Alle diese Enzyme spalten Carbohydrate, die als Energiequelle dienen. Fibronektin Typ III Module sind verschiedentlich als Carbohydratbindungsdomäne beschrieben. Besonders gut ist die Heparinbindungsstelle (einem Carbohydrat) im Fibronektin selber charakterisiert, an der die Module beteiligt sind, die im Dendrogramm den bakteriellen am ähnlichsten sind (Abb.5). Diese sind offenbar zur Affinitätssteigerung gegenüber den bakteriellen Substraten (Carbohydraten) in die Enzyme eingebaut worden [13]. Der Mechanismus des nachgewiesenen horizontalen Gentransfers (Eukaryot-Prokaryot) bleibt allerding nach wie vor im Unklaren.

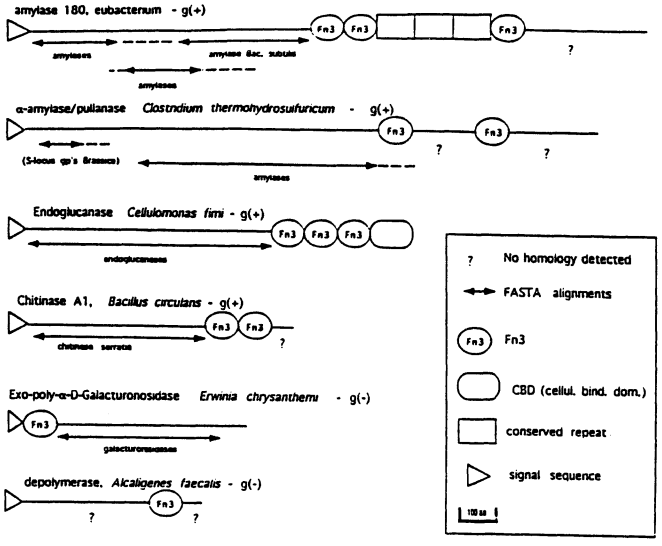

Abb.4 Verteilung von Fibronektin Typ III Domänen (FnIII) in verschiedenen prokaryotischen Enzymen. Alle diese Domänen sind um die eigenlichen Enzyme gruppiert, wahrscheinlich um die Affinität zu den jeweiligen Substraten (Carbohydrate) zu erhöhen.

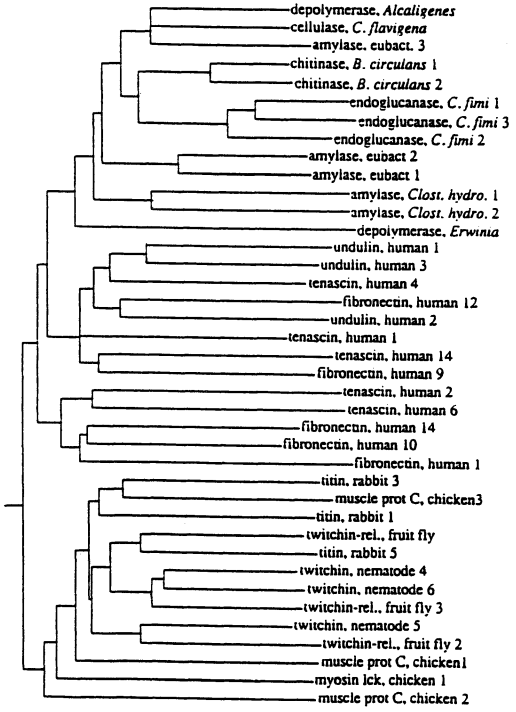

Abb.5 Dendrogramm eines Bausteins (Fibronektin Typ III Modul) aus ausgewählten Spezies einschließlich 7 verschiedener Carbohydrat-spaltender bakterieller Enzyme. Dieses Modul konnte bisher in über 300 Kopien in den verschiedensten tierischen Proteinen identifiziert werden [13]. Das Dendrogramm wurde mit dem Programm PAPA [17] erstellt

5. Offene Fragen und weitere Forschung

Die in den vorangegangenen Abschnitten behandelten Beispiele zeigen einerseits Möglichkeiten andererseits aber auch noch Schwächen derzeitiger Methoden auf. Bei den Homologiesuchverfahren handelt es sich durchweg um heuristische Methoden, z.B. basieren sämtliche Signifikanzabschätzungen zur Untermauerung gefundener Homologien letztendlich auf vorhandenen Daten. Das Ähnlichkeitsmaß zweier Sequenzen - die Anzahl identischer Aminosäuren ist unterhalb von 25-30% verrauscht. Viele Ähnlichkeitsmatrizen von Aminosäuren wurden entwickelt, doch ist es bislang nicht möglich beim paarweisen Vergleich unterhalb von 25% Sequenzidentität klare Signifikanzaussagen zu treffen. Wie soll man dann die Signifikanz von multiplen Alignments (z.B. Abb.2) abschätzen Schwache Ähnlichkeiten gewinnen an Stärke wenn bestimmte "Muster" in mehreren Sequenzen enthalten sind (weil sie z.B. für eine Funktion benötigt werden, Abb.2., siehe für einen Überblick charakterisierter Sequenzmuster die Musterdatenbank PROSITE [14]). Ein altes aber immer noch aktuelles Problem beim Sequenzvergleich - die Behandlung von Insertionen oder Deletionen - sei hier nur am Rande erwähnt Schwierigkeiten bereiten derzeit auch noch Sequenzabschnitte, die in einem Protein dupliziert worden sind. Derzeitige Datenbanksuchalgorithmen konzentrieren sich nur auf das beste Alignment. Suboptimale Alignments werden unterdrückt und die sich wiederholenden Abschnitte übersehen. Ein weites Methodenspektrum eröffnet sich aus der Verbindung von Primär und Tertiärstruktur. Man nimmt an, daß ca. 50% aller 3D-Faltungsmotive schon aufgeklärt worden sind [15] - schon bald wird man zu vielen Sequenzfamilien mindestens einen "Strukturprototypen" haben. Die Strukturinformation kann dann auf verschiedene Wege wieder zur sensitiven Sequenzsuche verwandt werden (siehe z.B. [5]). Die Analyse von Dendrogrammen oder Stammbäumen (siehe Abb.5) kann auch zur Funktionsbestimmung in Teilfamilien herangezogen werden, indem man diejenigen Positionen bestimmt, die für die Anordnung im Stammbaum verantwortlich sind. Da dies Positionen sind, die innerhalb von Teilfamilien konserviert, gegenüber anderen Teilfamilien aber variieren, stellen sie potentielle Kandidaten für eine spezifische Funktion dar. Kennt man nun noch die 3D-Struktur eines Mitgliedes der Sequenzfamilie, können solche Positionen in ihrer räumlichen Anordnung dargestellt und, auf mögliche Bindungstellen hinweisende 3D-Cluster charakterisiert werden. Bei der Analyse des Hefechromosoms III zeigte sich eine ungenügende Vernetzung vorhandener Methoden. Da an den Schnittstellen verschiedener Programme immer noch manuelle Eingriffe nötig sind (um z.B. Signifikanzfragen mit Hilfe funktioneller Zusatzinformation zu klären) müssen möglichst viele

Informationsquellen schnell bereit stehen. Eine benutzerfreundliche Integration verschiedenster Programme und Informationssysteme ist hier nötig. Da die experimentelle Charakterisierung von Proteinen hinter der stark zunehmenden Sequenzierung zurückbleibt, häufen sich schon jetzt die Fälle in denen Sequenzfamilien, basierend auf Homologien, zusammengefaßt werden können, die genaue Funktion aber nicht bekannt ist. Für diese, in Zukunft sehr oft auftretenden Fälle ist eine kombinierte Funktionsanalyse denkbar: Ein Expertensystem, daß die Kombination von Computeranalyse und standartisierten experimentellen Tests beinhaltet.

Danksagung

Mein Dank gilt den Kollegen aus der Protein Design Gruppe am EMBL und aus dem Forschungsschwerpunkt Genetische Information am Max-Delbrück-Centrum für die Unterstützung bei den angeführten Beispielen. Chris Sander und Reinhard Schneider möchte ich für Diskussionen zu diesem Beitrag danken.

Literatur

1. Foote, S., Vollrath, D., Hilton, **A.** und Page, D. **(1992)** *Science* 258, 60-66.
2. Chumakov et al. **(1992)** *Nature* 359, 380-387.
3. Oliver, S.G. et al. **(1992)** *Nature* 357, 38-46.
4. Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R. und Sonnhammer, E. **(1992a)** *Nature*, 358, 287.
5. Bork, **P.,** Sander, C., Valencia, **A. (1992c)** *Proc.Nafl.Acad.Sci.USA* 89**,** 7290-7294.
6. Bork, **P.,** Ouzounis, C., Sander, C., Scharf, M., Schneider, R. und Sonnhammer, E. **(1992b)** *Prot.Sci.,* im Druck.
7. Pearson W.R. und Lipman, D. J. **(1988)** *Proc.Natl.Acad.Sci.USA* 85, 3338-3342.
8. Rohde, K. und Bork, P. **(1993)** *Cornp.Appl.Biosci.,* eingereicht.
9. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. und Lipman, D.J. **(1990)** *J.Mol.Biol.* 215, 403-410.
10. Karlin, S. und Brendel, V. **(1992)** *Science* 257, 39-49.
11. Bork, P. **(1992)** *Curr.Opin.Struct.Bio1.* 2,413-421.
12. Bork, P. **(1991)** *FEBS Lett.* 286,47-54.
13. Bork, P., Doolittle, R.F. **(1992)** *Proc.Natl.Acad.Sci.USA* 89, 8990-8994.
14. Bairoch, **A. (1992)** *Nucl.Ac.Res.* 11, 2013-2018.
15. Blundell und Doolittle, R.F. **(1992)** *Curr.Opin.Struct.Bio1,* 2,
16. Sander, C. und Schneider! R. **(1991)** *Proteins* 9,56-68
17. Doolittle, R.F. und Feng, D.F. **(1990)** Methods in Enzym. 183, 659-669