Sequenzvergleiche in der Genomanalyse

Peer Bork[1,2], Christos Ouzounis[1], Georg Casari[1], Chris Sander[1], Pat M. Gillevet[3]

[1]EMBL
Meyerhofstraße 1, 69117 Heidelberg

[2]Max-Delbrück-Centrum für Molekulare Medizin
Robert-Rössle-Straße 10, 13125 Berlin-Buch

[3]George Mason University
Fairfax, Va., USA

⇒ Karemann Tierecki

Ein Hauptziel der verschiedenen Genomsequenzierungsprojekte ist die vollständige Beschreibung aller Proteinfunktionen der jeweiligen Organismen. Für die meisten Proteine, die im Rahmen von Genomprojekten sequenziert werden, ist allerdings eine experimentell ermittelte Funktion noch nicht verfügbar. Es können jedoch evolutionäre Verwandschaften ausgewertet und die Funktion von vielen Proteinen aus der Aminosäuresequenz vorhergesagt werden. Computergestützte Sequenzanalyse umfaßt eine Vielzahl von Techniken, in denen Sequenzvergleiche und Homologiesuchen im Mittelpunkt stehen. Diese Techniken werden ständig verbessert und automatisiert und sind derzeit essentieller Bestandteil der Genomprojekte.

Ausgehend von einem Überblick der derzeit laufenden internationalen Genomsequenzierungsprojekte [1] sollen hier einerseits die Nutzung von Sequenzvergleichen bei der Datenkorrektur (Fehlererkennung) aufgezeigt und andererseits die Möglichkeiten computergestützter Funktionsvorhersagen untersucht werden.

Trotz immer besser werdender Sequenzierungsmethoden befinden sich in den produzierten Rohdaten oftmals noch Fehler, d. h. Insertionen/Deletionen oder falsche Basenpaare in der DNA-Sequenz. Durch Homologiesuchen können solche Fehler zumindest in kodierenden Regionen in vielen Fällen ermittelt und berichtigt werden. Anhand von Rohdaten aus dem Mycoplasma-Genomprojekt (P. Gillevet, NIH; unveröffentlicht) wurde eine implementierte Fehlererkennungsmethode getestet und führte in fast allen Fällen zur Datenkorrektur. Die Analyse von 500kb (Kilobasen) aus *E. coli* deckte ebenfalls das Vorhandensein verschiedenster Fehler auf.

Der Schwerpunkt bei der Nutzung von Sequenzvergleichen liegt allerdings nach wie vor bei der Funktionsvorhersage aus der Aminosäuresequenz [2], die die experimentelle Funktionsermittlung komplementiert. Die Möglichkeiten der Funktionsvorhersage wurden an mehreren Datensätzen aus Genomprojekten getestet. Der Anteil der Sequenzen mit funktionell charakterisierten Homologen in Sequenzdatenbanken beträgt zwischen 33% (*C. elegans*, Chromosom III) und 75% (*Mycoplasma capricolum*, ca. 1/3 des Genomes). In den untersuchten Sequenzen können ca. 70–80% der Homologien mit automatisierten Standardverfahren ermittelt werden – allerdings werden oftmals „falsch positive" Treffer registriert. Somit sind immer noch nichtautomatisierte Kontrolluntersuchungen notwendig.

Um den Einfluß von Datenbankwachstum und Methodenentwicklung auf die computergestützte Funktionsvorhersage abschätzen zu können, verfolgen wir kontinuierlich den Anteil von funktionell charakterisierten oder vorhergesagten Proteinen in Hefechromosom III [3,4], dem ersten vollständig sequenzierten eukaryotischen Chromosom (komplettiert Anfang 1992). Der Anteil an funktionell charakterisierten Proteinen liegt gegenwärtig nocht unter 60%; Homologien wurden aber für fast 65% der 171 wahrscheinlichen Genprodukte von Hefechromosom III ermittelt.

Die Funktionsvorhersage durch Sequenzvergleich umfaßt verschiedene Präzisionsgrade, die unbedingt abgeschätzt werden müssen. Die Vorhersagen können sehr genau sein (z. B. Identifizierung des äquivalenten Genes in einem anderen Organismus) oder aber sehr vage (z. B. „ATP-bindendes Protein"). Oft beschränken sich Vorhersagen nur auf einen Teil des Proteins (Domäne), der eine bestimmte Subfunktion ausübt. Solche mobilen Domänen, die in verschiedene Proteine „kopiert" worden sind, können sogar in phylogenetisch alten Enzymen auftreten [5,6]. Ein Beispiel dazu wird vorgestellt.

Trotz genannter Limitierungen bei der Funktionsvorhersage [1] können (mit Verfügbarkeit des Genomes) Metabolismus und Proteinevolution ganzer Organismen entschlüsselt werden. Um entsprechende Methodiken zu entwickeln, sind Fallstudien unumgänglich. Deshalb haben wir mit einem der kleinsten lebenden Organismen begonnen: dem bakteriellen Parasit *Mycoplasma capricolum* mit einer Genomgröße von wahrscheinlich weniger als 1000kb. Die Computeranalyse von 214kb des Genomes (P. Gillevet et al., unveröffentlicht) ergab Ähnlichkeiten zu 220 verschiedenen Proteinen, mehr als einem Drittel des geschätzten Gesamtproteingehalts. Verwandschaftsbeziehungen zu *E. coli*-Proteinen (105 der 220 Treffer) wurden quantifiziert, um die Präzision von Vorhersagen bei entfernteren Sequenzähnlichkeiten besser abschätzen zu können. Der Mittelwert zu erwartender Sequenzidentität zwischen beiden Species beträgt 40% auf Proteinebene und stellt eine Art Eichkurve für Mycoplasma dar. Durch derartige Quantifizierungen konnten Hinweise auf horizontalen Gentransfer festgestellt werden. Die Identifizierung spezieller Enzyme durch Sequenzähnlichkeit gibt wertvolle Hinweise auf das Vorhandensein bestimmter Stoffwechselwege und ermöglicht die Rekonstruktion des Metabolismus. Es wurden z. B. verschiedene DNA-Reparatursysteme gefunden, die biochemischen und molekularbiologischen Untersuchungen trotz jahrelanger Suche bisher entgangen waren.

Um den Anforderungen vergleichender Genomuntersuchungen gerecht zu werden, versuchen wir, ein Analysepaket aufzubauen, das möglichst viele Schritte automatisiert und die Daten für die Beantwortung biologischer Fragestellungen aufbereitet. Eine vollständige Automatisierung erscheint auf Grund der Komplexität biologischer Sachverhalte derzeit noch nicht möglich.

Literatur

[1] Bork, P., Ouzounis, C. & Sander, C. (1994) *Curr. Opin. Struct. Biol.* 4, in press
From genome sequences to protein function

[2] Bork, P. (1993) Funktionsvorhersage mit Hilfe des Computers. In: Informatik aktuell – Informatik in den Biowissenschaften, (Hrsg. R. Hofestädt, F. Krückeberg, T. Lengauer) Springer Verlag, 67–78

[3] Bork, P. Ouzounis, C., Sander, C., Scharf, M., Schneider, R., Sonnhammer, E. (1992) *Nature* 358, 287
What's in a genome?

[4] Koonin, E. V., Bork, P. & Sander, C. (1994) *EMBO J.*, 13, 493–503
Yeast chromosome III: new gene functions

[5] Doolittle, R. F. & Bork, P. (1993) Spektrum der Wissenschaften, Dez., 40–46
Mobile Protein-Module: evolutionär alt oder jung?

[6] Bork, P. & Doolittle, R. F. (1994) *J. Mol. Biol.* 236, 1277–1282
Drosophila kelch motif is derived from a common enzyme fold