Bioinformatik des Genoms von A. thaliana

H.W. MEWES, FORSCHUNGSZENTRUM FÜR UMWELT UND GESUNDHEIT, MAX-PLANCK-INSTITUT FÜR BIOCHEMIE, MARTINSRIED

Arabidopsis thaliana hat sich in den letzen Jahren als der pflanzliche Modellorganismus schlechthin etabliert. Neben den für die praktische Handhabung wichtigen Eigenschaften kurzer Generationszeiten, Robustheit und kleinem Wuchs zeichnet sich Arabidopsis auch durch ein sehr kompaktes Genom aus. Fünf Chromosomen enthalten eine Sequenz von etwa 120 Millionen Basenpaaren, die für etwa 25.000 Gene kodieren. Das Heterochromatin weist eine Gendichte von 4.6 kBasen/Gen auf und ist damit äußerst kompakt im Vergleich zu anderen pflanzlichen Genomen (z.B. Reis mit ca. 500 Mbasen, entsprechend etwa 15-20 kB/Gen).

Im Rahmen einer internationalen Kooperation, an der Arbeitsgruppen aus den USA, Japan und Europa teilnehmen, wurde das Genom von Arabidopsis vollständig sequenziert und liefert damit die Grundlage für die Aufklärung der individuellen Funktionen aller Gene. Mit der Sequenz der Chromosomen II (White, O. et al., Nature, 1999) und IV (Mayer, K. et al., Nature, 1999) wurden Ende 1999 bereits mehr als 7000 Gene des Arabidopsisgenoms veröffentlicht. Die Sequenzierung wird im Spätsommer 2000 abgeschlossen sein, die annotierte Sequenz des gesamten Genoms wird zum Ende des Jahres erwartet.

Die Sequenz als Reihenfolge der Nukleinsäurereste im Genom selbst enthält nur unzureichend interpretierbare Information (Vorhersage der Sekundärstrukturen, Signalpeptide, Membransegmente). Erst ihre detaillierte Analyse mit den Methoden der Bioinformatik erlaubt es, Relationen zu verwandten und bereits in ihrer Funktion charakterisierten Genen oder regulatorischen Elementen zu identifizieren. Jede Sequenz wird dabei in den Kontext des biologischen Wissens homologer Sequenzen gestellt, die Interpretation genomischer Information baut auf den bereits an ähnlichen Sequenzen gefundenen Eigenschaften auf (z.B. Proteinfamilien, Domänen, Sequenzmotive).

Die meisten vollständig sequenzierten Organismen sind Prokaryonten (derzeit ca. 25 Genome, siehe http://pedant.mips.biochem.mpg.de). Hinzu kommen die eukaryontischen Genome von S. cerevisiae (6000 Gene, 1996), C. elegans (ca. 20.000 Gene, 1998), und D. melanogaster (ca. 12.000 Gene, 2000). Das Genom von Reis wurde ebenso wie das Humangenom bereits zu über 90% sequenziert, die Daten sind jedoch nicht uneingeschränkt öffentlich zugänglich. Im Gegensatz zu Drosophila, Reis und dem Humangenom wurde A. thaliana aus einer das Genom vollständig abdeckenden BAC-Bibliothek sequenziert, die im Gegensatz zur shot-gun Strategie nicht nur die eindeutige Zuweisung von Repeats (z.B. Transposons) zu den jeweiligen Loci, sondern auch die Konstruktion vollständiger Contigs der 10 chromosomalen Arme erlaubte.

Die systematische Genomanalyse umfasst zunächst die Identifikation der genetischen Elemente wie der codierenden Regionen (CDS), der regulatorischen Elemente, der SnRNAs (small nucleolar RNAs), der tRNAs oder der repetetiven Elemente wie Transposon oder LTR Sequenzen. Während die Genvorhersage in Prokaryonten oder niederen Eukaryonten, in denen nur ein geringer Anteil der Gene durch Intronsequenzen unterbrochen ist, mit hoher Zuverlässigkeit gelingt, stehen bisher keine Algorithmen zur Verfügung, die auch nur befriedigend übereinstimmende Genmodelle generieren können. Daher ist vor der detaillierten Charakterisierung der Gene die individuelle, manuelle Bearbeitung der Genmodelle erforderlich, die zwar die Verhersagequalität verbessern kann, aber wegen der individuellen Interpretation

der Daten keine statistisch zuverlässigen Aussagen über die Wahrscheinlichkeit liefern, mit der ein bestimmtes Genmodell korrekt ist.

Nur 30-40% aller Gene lassen sich unmittelbar mit den Methoden der vergleichenden Sequenzdatenanalyse charakterisieren, etwa 60% zeigen signifikante Ähnlichkeiten zu Genen anderer Organismen, die aber keine ausreichende funktionelle Zuweisung erlauben. Wie bereits der grobe Vergleich eukaryontischer Genome zeigt (Cherwitz et al., Science, 1999), variiert der Konservierungsgrad der einzelnen Funktionsklassen stark. So sind die Gene der Proteinbiosynthese zwischen Hefe und C. elegans weitgehend ortholog, d.h. die gesamte Gruppe lässt sich paarweise darstellen, zu jedem Gen in Hefe existiert ein hoch konserviertes orthologes Gen in C. elegans, während die strukturellen Gene des multicellulären Wurms sich stark vom unizellulären Pilz unterscheiden.

Für die funktionelle Charakterisierung codierender Regionen eukaryontischer Genome müssen im wesentlichen 3 Schritte durchgeführt werden: (1) semimanuelle Genvorhersage, (2) automatische Annotation aller identifizierter Gene durch Kombination einer ganzen Reihe von Algorithmen zur Sequenzhomologie, Funktions- und Strukturvorhersage, (3) manuelle Annotation und Interpretation aller Gensequenzen. Für das Arabidopsis Genom haben für die Bioinformatik das TIGR (The Institute for Genome Research, Bethesda USA) die Bearbeitung der Chromosomen I und II übernommen, während die Arbeitsgruppe MIPS die Chromosomen III, IV und V annotiert.

Die automatische Annotation der individuellen Gene erfolgt durch das gemeinsam mit der Firma Biomax entwickelte PEDANT-System. PEDANT leistet nicht nur die systematische Analyse jeder einzelnen Sequenz durch Suche nach Sequenzhomologen (BLAST), der Klassifikation in Proteinfamilien und Superfamilien (Protfam), die Identifizierung der Sequenzmotive und Domänen (Prosite und Interpro), sondern auch die Zuordnung der funktionellen und strukturellen Kategorien. Die Ergebnisse der PEDANT Analyse werden in einer relationalen Datenbank zugänglich gemacht (siehe auch Frishman et al., Bioinformatics, 2000).

Die intergenomische Analyse des detailliert annotierten Genoms von *A. thaliana* mit *C. elegans* und *D. melanogaster* erlaubt es zum ersten Mal systematisch 3 multizelluläre Eukaryonten zu vergleichen. Aus diesem Vergleich können weitreichende Schlüsse über die den Eukaryonten gemeinsamen metabolischen und regulatorischen Netzwerke einerseits und die im Laufe der Evolution differenziell ausgeprägten Genfamilien gezogen werden.

Mit der Verfügbarkeit des vollständigen Genoms beginnt eine neue Phase der Genomforschung in Pflanzen. Die experimentell gewonnenen Daten müssen in Relation zur Sequenz dargestellt werden, d.h. Informationen über die Funktion und Interaktion der Gene aus Arabidopsis in der Datenbank des Arabidopsisgenoms ständig aktualisiert werden. Im Rahmen des vom BMBF geförderten Genomanalyseprojekts in Pflanzen (GABI) wird MIPS ein Ressourcenzentrum etablieren, das die genomische Information um die Ergebnisse der funktionellen Analyse ergänzt.