Bioinformatische Analyse und Annotation eines eukaryotischen Genoms am Beispiel der Pflanze *Arabidopsis thaliana*

CHRISTINE SCHÜLLER, BIOMAX INFORMATICS, MARTINSRIED
KLAUS MAYER, GSF FORSCHUNGSZENTRUM FÜR UMWELT UND GESUNDHEIT, MARTINSRIED
HANS-WERNER MEWES, GSF FORSCHUNGSZENTRUM FÜR UMWELT UND GESUNDHEIT, MARTINSRIED

Abstract
The small weed Arabidopsis thaliana is the first plant whose genome will be completely deciphered. Since 1996, an international consortium of scientists has been working on the sequencing and the subsequent bioinformatic analysis and annotation of the genomic data. Genome analysis includes the identification of all genes and other elements on the genomic DNA and the elucidation of possible biological roles of the encoded proteins. The development of appropriate tools for the analysis and construction of databases to store, handle and display the huge amount of information is a necessary prerequisite for such publicly funded genome projects as well as other large scale sequencing efforts in academics or the biotech industry.

1 Einführung

Arabidopsis thaliana, ein Unkraut aus der Familie der Kreuzblütler, hat eine lange Geschichte als bevorzugter Modellorganismus in der molekularen Pflanzenbiologie. Da *Arabidopsis* darüber hinaus ein für Pflanzen relativ kleines Genom von ca. 130 Mb besitzt, war es das ideale Objekt für die erste vollständige Entschlüsselung der Erbinformation einer Pflanze. Ziel des Projektes ist es, alle Gene von *Arabidopsis* zu verstehen. Dies ist die einzige Möglichkeit herauszufinden, was eine Pflanze zu einer Pflanze macht. Ein tieferes Verständnis der Biologie höherer Pflanzen ist dringend erforderlich, um den Herausforderungen, vor die welche Landwirtschaft in den kommenden Jahren gestellt ist, wie verbesserte Pflanzenproduktivität, Anpassungsfähigkeit und Verarbeitbarkeit, gewachsen zu sein. *Arabidopsis* eignet einerseits als Modell für das Verständnis der Biologie von Pflanzen im allgemeinen und als Referenzorganismus für die vielfältige Gruppe der dikotylen Blütenpflanzen im besonderen, insbesondere der verwandten Brassica-Arten wie z.B. Soja, die als landwirtschaftliche Nutzpflanzen große Bedeutung haben.

2 Das *Arabidopsis thaliana* Genomprojekt

Früher als in anderen wirtschaftlich wichtigen Pflanzen, wie z.B. Mais, Reis, Weizen oder Soja wurden von dem vergleichsweise kleinen *Arabidopsis* Genom genetische und physikalische Karten erstellt und das gesamte Genom repräsentierende Klon-Bibliotheken etabliert. Dies war eine Voraussetzung für den Beginn der systematischen Sequenzierung. 1994 startete ein europäisches Pilotprojekt und 1996 wurde die weltweite „Arabidopsis Genome Initiative" (AGI) gegründet, an der sich Institute in Japan, den USA und Europa beteiligten. Das europäische ESSA-Konsortium („European Scientists Sequencing Arabidopsis") umfasst mehr als 20 Labors in denen die Sequenzierung durchgeführt wurde, während die bioinformatische Sammlung und Auswertung der Daten zentral bei MIPS („Munich Information Center for Protein Sequences") am MPI für Biochemie in München erfolgte.

3 Spezifische Herausforderungen bei der Genomanalyse von höheren Eukaryonten

Die Genomanalyse umfasst die Identifizierung und Charakterisierung aller auf der genomischen DNA codierten genetischen Elemente, wie regulatorische Sequenzen, strukturelle Elemente und, als wichtigsten Bestandteil, alle Gene, die daraus resultierenden Proteine und deren potentielle Funktion. Prinzipiell gibt es dafür zwei Ansatzpunkte, einerseits die Identifizierung über Homologien zu bereits bekannten Genen und Proteinen (extrinsische Methode) und zum anderen die Anwendung von Algorithmen, die in der Lage sind, codierende Bereiche in der DNA zu erkennen (intrinsische Methode). Dabei ist zu beachten, dass anders als bei niederen Organismen (Bakterien, Viren), die Gene höherer Organismen häufig in Einzelteile aufgespalten sind, was die korrekte Vorhersage der Gasamtstruktur erheblich erschwert.

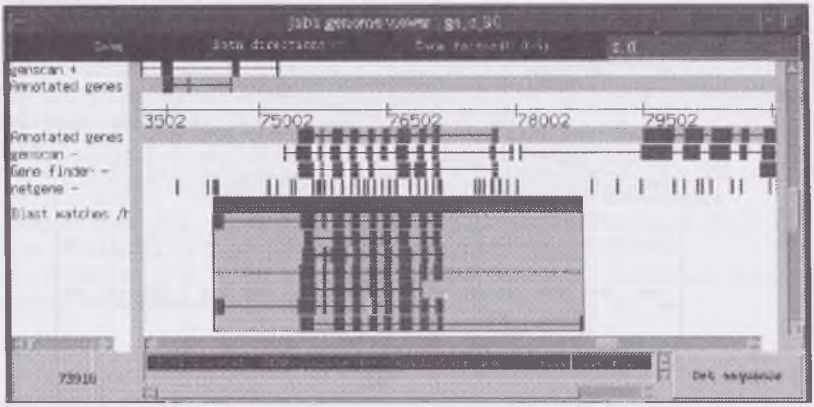

Abb. 1: Darstellung eines Bereichs des Arabidopsis Genoms in einem zur Annotation verwendeten, interaktiven „viewer". Gezeigt sind die Ergebnisse verschiedender Genvorhersageprogramme („genscan, Gene finder, netgene") und die Bereiche des Gens, die Homologien zu bekannten Proteinen aufweisen („Blast matches").

4 Verwaltung und Darstellung der Daten

Neben der strukturellen und funktionellen Analyse ist die Speicherung, Aufbereitung und Zugänglichkeit der gesammelten Information für die Nutzer von entscheidender Bedeutung. Geeignete Datenbanken müssen dafür entwickelt werden, die einerseits einen schnellen Zugriff auf die gespeicherten Sequenzdaten erlauben und die andererseits eine komfortable und benutzerfreundliche Darstellung der annotierten Information unterstützen. Die Vernetzung der Information via Hyperlinks zu den verschiedensten biologisch relevanten Datenbanken (z. Zt. mehr als 100 und einer exponentiellen Wachstumsrate für Sequenzdatenbanken) ist dabei selbstverständlich. Die graphische Darstellung komplexer Zusammenhänge erleichtert die Verständlichkeit erheblich (Abbildung 2).

5 Einige Ergebnisse von *Arabidopsis*

Bisher wurden zwei der fünf Chromosomen mit einer Länge von ca. 20 Mb (Chr.2) bzw. 17 Mb (Chr.4) entziffert. Dabei wurden ca. 8000 der geschätzten 22000 bis 25000 Gene von *Arabidopsis* identifiziert. Neben den ca. 10% bekannten Genen konnte für ca. 50% der Gene auf Grund von Homologievergleichen die vermutliche Struktur und Funktion ermittelt werden. Weitere 40% sind neue Gene, deren wahrscheinliche Struktur durch bioinformatische

Methoden vorhergesagt werden konnte, deren biochemische Rolle, die sie in dem Organismus spielen, aber völlig unbekannt ist. Dies sind jedoch die interessantesten Kandidaten für pflanzenspezifische Funktionen und daher Ausgangspunkte für weitergehende molekularbiologische Untersuchungen. Charakteristische Eckdaten dieser ersten Pflanzenchromosomen, wie u. a. Gendichte, Ausmass der Clusterbildung von verwandten Genen, Anteil von repetitiven Strukturelementen und die Veränderung der Verteilung dieser Elemente über das gesamte Chromosom etwa im Bereich von Centromer und Telomer, erlauben interessante Spekulationen über die Evolution von eukaryotischen Genomen.

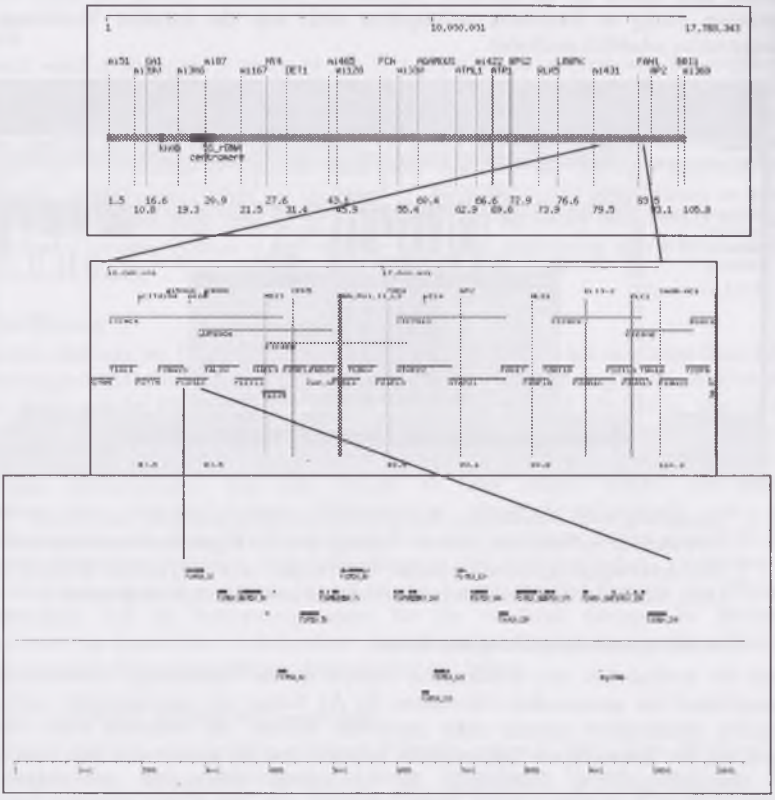

Abb. 2: Chromosom 4 von *Arabidopsis thaliana* beginnend mit der Karte des gesamten Chromosoms, über die Darstellung der für die Sequenzierung verwendeten Subklone bis hin zu den einzelnen Genen, die wiederum mit einer Datenbank mit allen verfügbaren Informationen zu diesen Genen verknüpft sind.

6 Literatur

BEVAN, M.; BANCROFT, I.; MEWES, H.W.; MARTIENSSEN, R.; MCCOMBIE, R. (1999): Clearing a path through the jungle: progress in Arabidopsis genomics. Bioessays 21(2):110-20.

BEVAN, M.; BANCROFT, I.; ...SCHUELLER, C.; AND CHALWATZIS, N. (1998): Analysis of 1,9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana*. Nature 391: 485-488.

BURSET, M.; GUIGO, R. (1996): Evaluation of gene structure prediction programs. Genomics. 34(3):353-367.

THE EUROPEAN UNION ARABIDOPSIS GENOME SEQUENCING CONSORTIUM: K. MAYER, C. SCHUELLER , ...,C. BIELKE, D. FRISHMAN, D. HAASE, K. LEMCKE, H.W. MEWES, S. STOCKER, P. ZACCARIA, AND M. BEVAN AND THE COLD SPRING HARBOR, WASHINGTON UNIVERSITY IN ST LOUIS AND PE BIOSYSTEMS ARABIDOPSIS SEQUENCING CONSORTIUM: R.K. WILSON, ..., AND W.R MCCOMBIE (1999) Sequence analysis of chromosome 4 of the plant Arabidopsis thaliana. Nature , vol. 402, 769-777.

FRISHMAN, D.; MEWES, H.W. (1997): PEDANTic genome analysis. Trends Genet. 13:415-6.

MEWES, H.W.; FRISHMAN, D.; GRUBER, C.; GEIER, B.; HAASE, D.; KAPS, A.; LEMCKE, K.; MANNHAUPT, G ; PFEIFFER, F.; SCHUELLER, C.; STOCKER, S.; WEIL, B. (2000): MIPS: a database for genomes and protein sequences. Nucleic Acids Research. 28(1):37-40.

PAVY, N.; ROMBAUTS, S.; DEHAIS, P.; MATHE, C.; RAMANA, DVV.; LEROY, P.; ROUZE, P. (1999): Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. [Article] Bioinformatics. 15(11) 887-899.

WAMBUTT, R ; ... MAYER, K.; SCHUELLER, C.; BEVAN, M. (2000): Progress in Arabidopsis genome sequencing and functional genomics. Journal of Biotechnology 78: 281 - 292.

TERRYN, N.; HEIJNEN, L ; ... SCHUELLER, C ; ... VOS, P. (1999): Evidence for an ancient chromosomal duplication in *Arabidopsis thaliana* by sequencing and analyzing a 400-kb contig at the APETALA2 locus on chromosome 4. FEBS Letters 445: 237-245.

ZACCHARIA, P.; MEWES, H.W. (1999): Homology based gene prediction in *Arabidopsis thaliana* Proceedings of the German Conference on Bioinformatics.