Bioinformatik

Wissenschaftliche Datenbanken in der Genomforschung[1]

JENS HANSEN
INSTITUT FÜR ENTWICKLUNGSGENETIK, GSF, NEUHERBERG

Das mit der Entwicklung der Sequenziertechnologie und ihrer Anwendung einhergehende rapide Anwachsen der Sequenzinformation hat zur Notwendigkeit geführt, die Datenmengen zu strukturieren und durch Sequenzvergleiche auswertbar zu machen.

■ Insbesondere die internationalen Genomsequenzierprojekte haben eine gezielte Sammlung von Sequenzen und ihre Verknüpfung mit Daten zur Annotation dieser Sequenzen notwendig werden lassen.

Nukleotidsequenzdatenbanken

Die klassischen Nukleotidsequenzdatenbanken, wie GenBank[2], die EMBL Nukleotidsequenz Datenbank[3] und die DNA Data Bank of Japan (DDBJ)[4], haben bereits vor dem Beginn der Genomsequenzierprojekte mit der Sammlung weltweit generierter Nukleotidsequenzen begonnen (**Tab. 1**). Inzwischen enthält z.B. GenBank Sequenzen von 205.000 verschiedenen Organismen. Dabei werden entweder einzelne Sequenzen oder große Sequenzpakete übermittelt. Web-basierte Übermittlungswerkzeuge wie BankIt oder Programme wie SeqIn erleichtern die Datenablage und haben zu einem schnellen Anwachsen des Sequenzdatenbestandes in wenigen zentralen Institutionen geführt. Täglicher Abgleich zwischen GenBank, der EMBL Nukleotide Sequence Database und der DNA Database of Japan (DDBJ) stellt die Konsistenz eines gemeinsamen Datenbestandes sicher. Die Übermittlung von Sequenzdaten aus den Forschungslabors ist dabei nach wie vor von großer Bedeutung, um den Datenbestand kontinuierlich zu vergrößern oder bereits existierende Einträge zu aktualisieren.

Nutzen lässt sich aus der Fülle der gesammelten Sequenzinformationen aber erst gewinnen, wenn diese gezielt durchsucht und abgefragt werden können. Dazu hat entscheidend die Entwicklung von Alignmentprogrammen wie BLAST[5] und SSAHA[6] beigetragen, die eine Suche nach ähnlichen Sequenzen in den Datenbanken anhand von Suchsequenzen ermöglichen. Die Bedeutung solcher Programme kann an der enorm weiten Verbreitung von BLAST und der Entwicklung einer ganzen Familie von BLAST-Alignmentprogrammen gesehen werden.

Genomdatenbanken

Die mit der Genomsequenzierung anfallenden großen Mengen genomischer Sequenzen mussten assembliert und annotiert werden. Die annotierte Genomsequenz muss im Gegenzug der wissenschaftlichen Gemeinschaft im Internet zur Verfügung gestellt werden. Zentren, wie das National Center for Biotechnology Information NCBI[7], das EMBL-EBI/Sanger Centre (Ensembl) in Hinxton[8] und die Universität von Kalifornien in Santa Cruz (UCSC)[9] stellen entsprechende Genombrowser zur Verfügung, die dem Nutzer die weitere Analyse der Genomsequenzen und ihrer Annotationsdaten ermöglichen. NCBI hält neben den Genomdaten von 20 Eukaryoten 250 Bakterien- und 2.100 Virengenome bereit. Die zentrale Speicherung und Aufbereitung der Genomdaten so vieler verschiedener Spezies erleichtert die vergleichende Genomanalyse.

Im Laufe der Jahre hat sich gezeigt, dass eine automatische Annotation der genomischen Sequenz nicht ausreicht. Es wurden daher Projekte gegründet, die, aufbauend auf der automatischen Annotation, eine eingehende manuelle Annotation der verschiedenen Genome unter Berücksichtigung der vorhandenen Literatur vornehmen. Entsprechende Projekte sind das Vega- (Ensembl) und das RefSeq-Projekt (NCBI).

Durch die Verlinkung der Sequenzinformation mit verschiedenen genspezifischen, proteinspezifischen, funktionalen sowie biomedizinischen Datenbeständen lassen sich Suchen und Anfragen durchführen. Für diese komplexen, benutzerspezifischen Suchen in den verschiedenen Datenbanken wurden spezielle Werkzeuge entwickelt. Zur Suche in GenBank dient die Entrez-Suchmaschine, die die Daten der Nukleotid- und Proteinsequenzdatenbanken kombiniert mit taxonomischer Information, Genomannotation, Genexpressionsdaten, mit Proteinstruktur- und -domändaten und mit der biomedizinischen Literatur (PubMed). Insgesamt werden dabei etwa 30 verschiedene Datenbanken durchsucht.

Für die komplexe Suche in verschiedenen Datenbanken gleichzeitig wurde SRS (Search and Retrieval System) am EMBL/EBI entwickelt. Für benutzerspezifisches Suchen in Genomannotationsdaten von Ensembl wurde BioMart entwickelt.

Von großer Bedeutung, insbesondere vor dem Hintergrund der Aufklärung genetisch bedingter Erkrankungen, ist eine systematische Erfassung von Mutationen. Eine Vielzahl verschiedener Datenbanken steht hier zur Verfügung, als Beispiel seien The Human Gene Mutation Database (HGMD)[10] und die Human Mutations Database (HmutDB)[11] genannt (**Tab. 1**).

EST-/cDNA-Datenbanken

EST- und cDNA-Sequenzen sind für die Identifizierung von Genen und die Aufklärung der Exon-/Intronstruktur der Transkriptionseinheiten von großem Nutzen. Bereits vor dem Beginn der Genomsequenzierung wurde damit begonnen, überlappende ESTs zu genorientierten Clustern zusammenzusetzen und

1 Der vorliegende Überblick gibt einen kleinen, auf einer subjektiven Auswahl basierenden Querschnitt der im Internet zur Verfügung stehenden Datenressourcen wieder. Es wurden ausschließlich frei zugängliche Datenbanken berücksichtigt. Für die Suche nach Datenbanken, insbesondere Datenbanken mit speziellen Fragestellungen, sei auf die Molecular Biology Database Collection (www.oxfordjournals.org/nar/database/c/)[1] in der Zeitschrift Nucleic Acid Research verwiesen.

so aus nicht-redundanten Sequenzen bestehende Spezies-spezifische Gendatensätze zu gewinnen. Ein Beispiel für einen solchen Ansatz ist das UniGene Projekt (NCBI)[7]. So konnte z.B. die in GenBank vorliegenden 5.3 Mio. humanen ESTs auf eine Zahl von ca. 53.000 Sequenz-Cluster reduziert werden. Seit dem Vorliegen kompletter Genomsequenzen wird inzwischen eine genombasierte Clustering-Methode zur Erstellung der Transkriptsequenzen gewählt, die die distinkten Transkriptionsloci oder annotierten Genen entsprechen. Der Vorteil einer EST-basierten Definition der Genstruktur ergibt sich aus dem Vorliegen von Information über die gewebespezifische Expression, basierend auf der Gewinnung der EST-Sequenzen. Während die umfangreichen Datensätze früher vor allem für die Identifizierung von Transkriptionseinheiten und der Aufklärung ihrer Struktur benutzt wurden, dienen sie heute u. a. zur Erstellung von Mikroarrays für genomweite Untersuchungen der Genexpression.

Proteindatenbanken

Einhergehend mit der fortschreitenden Annotation der verschiedenen Genome und der Aufklärung der biologischen und biochemischen Funktion der einzelnen Genprodukte kommt es zu einem ernormen Anwachsen der Information über Proteine. Eine zentrale Ressource für Proteinsequenzen stellt die Universal Protein Knowledgebase (UniProt)[12] dar. Sie besteht aus drei Komponenten: 1. dem UniProt-Archive, einer nicht-redundanten, stabilen Sammlung aller öffentlich verfügbaren Proteinsequenzen; 2. der UniProt-Knowledgebase, der zentralen Sammlung annotierter und referenzierter Proteinsequenzen, die ihrerseits aus zwei Untereinheiten besteht: UniProt/Swiss-Prot, deren Einträge entweder manuell oder computerbasiert annotiert werden mit anschließend manueller Verifizierung auf der Basis von Literatureinträgen und Sequenzanalysen. Zum anderen UniProt/TrEmbl, die Einträge mit Computer-basierter Annotation und funktionaler Charakterisierung enthält; 3. den UniProt-Referenz-Clustern: Es existieren drei Referenz-Cluster, die eine stufenweise Reduktion der Sequenzdatenmenge mit unterschiedlicher Auflösung zum Ziel haben. Ähnliche Sequenzen und Subsequenzen, über Organismengrenzen hinweg, werden in Abhängigkeit ihrer Ähnlichkeit zusammengefasst (100 %, ≧ 90 %, ≧ 50 % Sequenzidentität). Die Überführung von einer Stufe

Nukleotidsequenzdatenbanken	
http://www.ncbi.nlm.nih.gov/Genbank/index.html	GenBank Genetic Sequence Database (NCBI)
http://www.ebi.ac.uk/embl	EMBL Nucleotide Sequence Database
http://www.ddbj.nig.ac.jp	DNA Data Bank of Japan DDBJ
Genomdatenbanken	
http://www.ensembl.org/index.html	Ensembl Automatic Annotation of Eukaryotic Genomes
http://genome.ucsc.edu	University of California Santa Cruz Genome Bioinformatics
http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi	Entrez Life Science Search Engine
http://srs.ebi.ac.uk	SRS Data Integration Platform
http://mips.gsf.de	MIPS Munich Information Center for Protein Sequences
http://www.informatics.jax.org	MGI Mouse Genome Informatics
http://www.genecards.org/index.shtml	GeneCards Human Genes Database[21]
http://flybase.bio.indiana.edu	Flybase: A Database of the Drosophila Genome[22]
http://mips.gsf.de/proj/yeast	CYGD Comprehensive Yeast Genome Database[23]
http://www.yeastgenome.org	SGD Saccharomyces Database[24]
http://www.wormbase.org	Wormbase: Biology & Genome of C. elegans[25]
http://www.informatics.jax.org	MGD Mouse Genome Database[26]
http://mips.gsf.de/genre/proj/mfungd	MFunGD Mouse Functional Genome Database[27]
Mausressourcen	
http://www.genetrap.de	German Gene Trap Consortium
http://www.genetrap.org	International Gene Trap Consortium
http://www.emma.rm.cnr.it	European Mouse Mutant Archive EMMA
http://jaxmice.jax.org	JaxMice
http://www.mbl.org	The Mouse Brain Library
http://eulep.anat.cam.ac.uk	Pathbase European Mutant Mouse Pathology Database[28]
http://www.mmrrc.org	MMRRC Mutant Mouse Regional Resource Center
Genexpression	
http://genex.hgu.mrc.ac.uk	Emap Edinburgh Mouse Atlas Project[21]
http://www.genepaint.org	Atlas of Gene Expression Patterns in the Mouse[29]
http://www.informatics.jax.org/mgihome/GXD aboutGXD.shtml	The Gene Expression Database[30]
Proteinsequenzdatenbanken	
http://www.expasy.uniprot.org	The Universal Protein Resource UniProt
http://www.ebi.ac.uk/interpro	InterPro Database of Protein Families, Domains and Functional Sites
Proteinstrukturdatenbanken	
http://cathwww.biochem.ucl.ac.uk/latest/index.html	CATH Protein Structure Classification Database
http://scop.berkeley.edu	SCOP Structural Classification of Proteins
http://www.rcsb.org	PDB Protein Data Bank
http://swissmodel.expasy.org/repository	Swiss Model Repository
Protein-Protein Interaktionen	
http://mips.gsf.de/proj/ppi	The MIPS Mammalian Protein-Protein Interaction Database
http://www.bind.ca	BIND Biomolecular Interaction Netwok Database
http://dip.doe-mbi.ucla.edu	DIP Database of Interacting Proteins
Biomedizinische Information	
http://www.ebi.ac.uk/mutations/central	Human Mutations Database (HmutDB)
http://www.hgmd.cf.ac.uk	The Human Gene Mutation Database (HGMD)
http://www.infobiogen.fr/services/chromcancer/index.html	Atlas of Genetics and Cytogenetics in Onkology and Haematology[31]
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM	OMIM[32]
http://www.geneclinics.org	GeneTests Medical Genetics Information Resource[39]

Tab. 1: Eine Auswahl wissenschaftlicher Datenbanken

in die nächste führt zu einer Reduktion der Datenmenge von 40 % bzw. 65 % und damit zu einer Beschleunigung von Ähnlichkeitssuchen im immer weiter anwachsenden Datenbestand.

Proteindomänen und Klassifizierung in funktionale Proteinfamilien und -superfamilien

Eine zentrale Ressource für die Klassifizierung in Proteinfamilien, Domänen und funktionalen Gruppen ist InterPro[13], die den Datenbestand folgender Datenbanken integriert: Protein Families Database (Pfam), Database of Protein Families and Domains PROSITE, PRINTS, ProDom, SMART (Simple Modular Architecture Research Tool), Protein Information Resource PIRSF, Superfamily and TIGRFAMs (The Institute for Genomic Research Protein Families). Das Klassifizierungssystem von InterPro enthält momentan 9.055 Proteinfamilien, 3.585 Domänen, 238 Repeats, 32 aktive Zentren, 22 Bindungsstellen, sowie 21 posttranslationale Modifikationsprozesse und deckt ca. 78 % der Einträge in UniProt-Knowledgebase ab. Jeder Eintrag in InterPro wird beschrieben durch ein oder mehrere Proteinsequenzmuster, die einer biologisch relevanten Proteinfamilie, einer Domäne, einem repetitiven Sequenzmuster oder einer posttranslationalen Modifikation entsprechen. Zwei Arten von Beziehungen zwischen verschiedenen Einträgen können dabei bestehen: eine evolutionäre Verwandtschaft aufgrund eines gemeinsamen Ursprungs oder eine einfache Präsenzbeziehung, wenn es sich um genetisch mobile Domänen handelt.

Proteinstrukturdatenbanken

Ein entscheidender Faktor für ein Verständnis der molekularen Basis der Proteinfunktion ist die dreidimensionale Proteinstruktur. Die Weiterentwicklung von Techniken für die experimentelle Aufklärung der Proteinstruktur, wie Röntgenstrukturanalyse und NMR-Spektroskopie, haben dazu geführt, dass inzwischen weit über 36.344 experimentell aufgeklärte Proteinstrukturen in der Protein Data Bank (PDB)[14] zur Verfügung stehen. PDB umfasst neben Strukturdaten von Protein- und DNA/RNA Molekülen auch Strukturen von Protein-DNA- und Protein-Ligand-Komplexen. Die Proteine in einen evolutionären Zusammenhang aufgrund ihrer konservierten strukturellen Eigenschaften zu bringen, ist Ziel der Datenbank SCOP (Structural Classification of Proteins)[15]. Der SCOP-Datenbank liegt dabei eine phylogenetischer Ansatz zugrunde, der die Proteindomäne als Basiseinheit betrachtet und in ein hierarchisches System bestehend aus Familien und Superfamilien, sowie Faltungsmustern und Klassen einteilt. Die letzte Ausgabe von SCOP umfasst sieben Hauptklassen mit 70.859 Proteindomänen, die in 2.845 Familien, 1.539 Superfamilien und 945 Faltungsmustern klassifiziert wurden. Die erfassten Proteindomänen entsprechen dabei 25.975 Einträgen in PDB. Weitere wichtige Proteinstrukturdatenbanken sind CATH Protein Family Database[16] und MSD Macromolecular Structure Database. Eine Sammlung von 865.445 annotierten Proteinstrukturmodellen, die automatisch durch Homologiemodellierung anhand der in UniProt vorhandenen Proteinsequenzen erzeugt wurden, repräsentiert das Swiss-Modell Repository[17]. Die Sammlung wird laufend aktualisiert und durch neue Strukturmodelle ergänzt.

Vor dem Hintergrund der Frage nach dem Zusammenspiel der Genprodukte bei der Regulation des Metabolismus ist die Erfassung von Protein-Interaktionsdaten in entsprechenden Datenbanken von zunehmender Bedeutung. Zu nennen sind hier Datenbanken wie BIND Biomolecular Interaction Network Database[18], DIP Database of Interacting Proteins[19] und MIPS Mammalian Protein-Protein Interaction Database[20]. Die Interaktionsdatenbanken dokumentieren experimentell nachgewiesene Protein-Protein-Interaktionen zwischen Reaktionspartnern und in Protein-Komplexen. Wichtigste experimentelle Methoden der Interaktionsbestimmung sind die Yeast-Two-Hybrid-Methode und die Immunpräzipitation. Viele Interaktionen sind durch mehr als eine Methode bestimmt worden. Bei der Auswertung der umfangreichen wissenschaftlichen Literatur kommen dabei, je nach Datenbank, manuelle Auswertung durch Annotatoren als auch automatisierte Methoden durch Datenanalyse Programme zum Einsatz.

Fazit

Momentan werden Anstrengungen unternommen mit dem Ziel einer Zusammenfassung und Bündelung biologischer Information im Hinblick auf eine stärkere Integration der Datenbanken. Dieser Trend wird sich in den nächsten Jahren sicherlich verstärken. Für ein Verständnis des Zusammenspiels der verschiedenen Gene in komplexen zellulären Prozessen ist es erforderlich, den funktionalen Kontext der einzelnen Gene verstehen zu lernen. Dies allerdings wird nur möglich sein, wenn es gelingt verschiedene Arten biologischer Information zusammenzuführen. Die verschiedenen Datenbanken müssen dabei nicht notwendigerweise zusammengeführt werden, inzwischen stehen die technischen Möglichkeiten zur Verfügung auch dezentrale Datenbestände zu explorieren. Der Nutzen einer stärkeren Integration der verschiedenen Datenbestände liegt klar auf der Hand: dem Nutzer in der Forschung wird eine leichtere und schnellere Informationsbeschaffung bei gleichzeitig umfassender und tieferer Datenauswertung vor dem Hintergrund von Fragestellungen nach der Regulation des Gesamtorganismus ermöglicht. ■

Literatur

[1] **Galperin, M.** (2006): *Nucl. Acid Res.* 34: D3–5.
[2] **Benson, D.A.,** et al. (2006): *Nucl. Acids Res.* 34: D16–20.
[3] **Cochrane, G.,** et al. (2006): *Nucl. Acids Res.* 34: D10–15.
[4] **Okubo, K.,** et al. (2006): *Nucl. Acids Res.* 34: D6–9.
[5] **Altschul, S.F.,** et al. (1990): *J. Mol. Biol.* 215: 403–410.
[6] **Ning, Z.,** et al. (2001): *Genome Res.* 11: 1725–1729.
[7] **Wheeler, D.L.,** et al. (2006): *Nucl. Acids Res.* 34: D173–180.
[8] **Birney, E.,** et al. (2006): *Nucl. Acids Res.* 34: D556–561.
[9] **Hinrichs, A.S.,** et al. (2006): *Nucl. Acids Res.* 34: D590–598.
[10] **Stenson, P.D.,** et al. (2003): *Hum. Mutat.* 21: 577–581.
[11] **Lehväslaiho, H.,** et al. (1998): *Trends Genet.* 14: 205–206.
[12] **Wu, C.H.,** et al. (2006): *Nucl. Acids Res.* 34: D187–191.
[13] **Mulder, N.J.,** et al. (2005): *Nucl. Acids Res.* 33: D201–205.
[14] **Ivanisenko, V.A.,** et al. (2005): *Nucl. Acids Res.* 33: D183–187.
[15] **Andreeva, A.,** et al. (2004): *Nucl. Acids Res.* 32: D226–229.
[16] **Pearl, F.,** et al. (2005): *Nucl. Acids Res.* 33: D247–251.
[17] **Schwede, T.,** et al. (2003): *Nucl. Acids Res.* 31: 3381–3385.
[18] **Alfarano, C.,** et al. (2005): *Nucl. Acids Res.* 33: D418–424.
[19] **Salwinski, L.,** et al. (2004): *Nucl. Acids Res.* 32: D449–451.
[20] **Pagel, P.,** et al. (2005): *Bioinformatics* 21: 832–834.
[21] **Rebhan, M.,** et al. (1997): GeneCards: encyclopedia for genes, proteins and diseases. Weizmann Institute of Science, Bioinformatics Unit and Genome Center (Rehovot, Israel).
[22] **Drysdale, R.A.,** et al. (2005): *Nucl. Acids Res.* 33: D390–395.
[23] **Güldener, U.,** et al. (2005): *Nucl. Acids Res.* 33: D364–368.
[24] **Hirschman, J.E.,** et al. (2006): *Nucl. Acids Res.* 34: D442–445.
[25] **Chen, N.,** et al. (2005): *Nucl. Acids Res.* 33: D383–389.
[26] **Blake, J.A.,** et al. (2006): *Nucl. Acids Res.* 34: D562–567.
[27] **Ruepp, A.,** et al. (2006): *Nucl. Acids Res.* 34: D568–571.
[28] European mutant mouse pathology database (Pathbase), Pathbase web site, University of Cambridge, World Wide Web (www.pathbase.net).
[29] **Vise, A.,** et al. (2004): *Nucl. Acids Res.* 32: D552–556.
[30] **Hill, D.P.,** et al. (2004): *Nucl. Acids Res.* 32: D568–571.
[31] **Huret, J.L.,** et al. (2003): *Nucl. Acids Res.* 31: 272–274.
[32] **Hamosh, A.,** et al. (2005): *Nucl. Acids Res.* 33: D514–517.
[33] **GeneTests:** Medical Genetics Information Resource (database online). Copyright, University of Washington, Seattle. 1993–2006. Available at www.genetests.org.

Korrespondenzadresse:
Dr. Jens Hansen
GSF – Forschungszentrum für Umwelt und Gesundheit
Institut für Entwicklungsgenetik
Ingolstädter Landstraße 1
D-85764 Neuherberg
hansen@gsf.de