

3 Datenbanken, Alignments, Software

„The good news about computers is that they do what you tell them to do. The bad news is that they do what you tell them to do.“

Theodor H. Nelson, US-amerikanischer Soziologe und IT-Pionier

Ein Wust von Drei- und Vierbuchstabenabkürzungen gehört zum Jargon der Molekularbiologie und auch der Informatik. Wir wollen ihn durchdringen. Wer heute als Molekularbiologe seine *Midlife Crisis* schon fast hinter sich hat, wird sich noch daran erinnern, wie molekulare Datenbanken in den 1980er Jahren auf Disketten an die Forscher in ihre Institute verschickt wurden. Danach kam natürlich die CD und seit den 90er Jahren sind die Datenbanken auf Wechseldatenträgern durch das WWW und komfortable Anwendungen ersetzt. Die erste Datenbank molekularer Sequenzdaten geht auf 1982 zurück. Bereits während wir an der ersten Auflage dieses Buches schrieben, wurde gerade die Schwelle von 100 Gigabasen an gespeicherten Sequenzdaten überschritten. Mit dem immer rasanteren Zuwachs an Daten durch immer schnellere Hochdurchsatztechnologien der DNA-Sequenzierung wird ein Rekord an Datenmenge immer schneller vom nächsten abgelöst.

Übersicht

3.1	Die Datenbanken für molekulare Sequenzdaten	74
3.1.1	Datenbankeinträge: Textbasiertes Suchen und Dateiformate	75
3.1.2	Suche nach Sequenzähnlichkeiten	80
3.2	Alignments	85
3.2.1	Software zur Sequenzverwaltung und Alignmenteditoren	88
3.2.2	Automatische Alinierung	93
3.3	Integrierte Programmpakete für die molekulare Phylogenetik	98
3.3.1	PHYLIP	98
3.3.2	PAUP*	100
3.3.3	MEGA	103
3.3.4	Die anderen Vielzweckalternativen	104
3.4	Speziellere Anwendungen in phylogenetischen Analysen	104
3.4.1	Modeltest	105
3.4.2	MrBayes und BEAST	105
3.4.3	TREE-PUZZLE und Treefinder, PAML und PHYML	105
3.4.4	SplitsTree	107
3.4.5	Mesquite und MacClade	107
3.4.6	NONA, TNT, WinClada, PRAP etc.	108
3.5	Graphische Darstellung von Bäumen	109
3.6	Attraktive Darstellung von Alignments	110
3.7	Leseempfehlungen	111

3.1 Die Datenbanken für molekulare Sequenzdaten

Die Sammlung, Betreuung, Verwaltung und Nutzbarmachung molekularer Sequenzdaten liegt in den Händen dreier großer international kooperierender Datenbanken (Tab. 3.1 auf Seite 76), die ihre Bestände täglich abgleichen: Die **GenBank** des **NIH** (*National Institute of Health*) in den USA, verwaltet vom **NCBI**, dem *National Center for Biotechnology Information*, die **DDBJ**, die *DNA Data Bank of Japan*, verwaltet vom *National Institute of Genetics*, und die europäische **EMBL**-Datenbank (*European Molecular Biology Laboratory*), verwaltet durch das *European Bioinformatics Institute* (**EBI**). Alle Einrichtungen verwalten die Sequenzdaten öffentlich – sie stehen jedem zur Nutzung und Analyse über die WWW-Seiten der Organisationen zur Verfügung: www.ncbi.nlm.nih.gov, www.ddbj.nig.ac.jp und www.ebi.ac.uk.

Neben diesen drei großen, öffentlichen Datenbanken gibt es diverse weitere WWW-zugängliche Datenbanken (Tab. 3.1 auf Seite 76), die auf Initiativen großer oder kleinerer Institute, einzelner Labore oder auch kommerzieller Einrichtungen zurückgehen. Meist haben sie einen klaren Fokus auf bestimmte taxonomische Gruppen oder auf bestimmte Typen von Molekülen. Das J. Craig Venter Institute (Früher: **TIGR** – *The Institute for Genomic Research*), das Maßstäbe für komplette Genomsequenzierungen gesetzt hat, ist hier mit seinen umfangreichen Datensammlungen unbedingt zu nennen. Aber auch Initiativen, mit denen integrierte Datenbanken für Modellorganismen wie den Kreuzblütler *Arabidopsis*, den Nematoden *Caenorhabditis* oder die Fruchtfliege *Drosophila* geschaffen wurden, sind natürlich insbesondere für diejenigen, die mit diesen Organismen arbeiten, hoch interessant. Solche Datenbanken liefern viele Informationen über Allele, Bilder, Forscher, Klone, Mutanten, Phänotypen, Stämme u.s.w., die der betreffenden *Research Community* dienen oder ganz unmittelbar experimentell nützlich sein können. Viele der kleinen, sehr speziellen Initiativen sind zwar prinzipiell nützlich, gehen aber mangels personeller Kontinuität oder anderweitig fehlender Ressourcen auch schnell wieder ein oder werden nicht gepflegt und aktualisiert. Es macht darum wenig Sinn, solche Initiativen erschöpfend aufzulisten, denn einige WWW-Adressen sind in vielen Fällen schon nicht mehr aktiv, bevor die Liste fertig wird. Ein sehr interessantes Projekt für die Phylogenetik ist allerdings **TreeBase**. Diese WWW-basierte Datenbank speichert Informationen über phylogenetische Studien, die über Taxa oder Autoren suchbar sind, aber auch Merkmalsmatrices und Phylogramme, die interaktiv mit dem Java-Applet **ATV** (*A Tree Viewer*) betrachtet werden können.

Schon frühzeitig haben die meisten wissenschaftlichen Zeitschriften gefordert, dass DNA-Sequenzdaten, die in neue Publikationen eingehen sollen, zeitgleich in den zentralen öffentlichen Datenbanken deponiert werden müssen. Alle Daten werden öffentlich gemacht, die Autoren haben lediglich die Möglichkeit, sie bis zur Publikation ihrer Arbeit zurückhalten zu lassen. Die Autoren erhalten dabei von den Datenbanken eine alphanumerische Chiffre für ihren neuen Sequenzeintrag, die so genannte **Accession Number**. Diese **Akzessionsnummer** ist einmalig und eindeutig mit einer Nukleotidsequenz verknüpft. Selbst ein neuer **Datenbankeintrag** (eine *accession*) mit einer identischen Nukleotidsequenz wie ein bereits existierender erhält eine andere Akzessionsnummer. Daneben haben die Datenbankeinträge „Namen“, die in der Frühzeit eine gewisse mnemotechnische Bedeutung hatten, aber heute in der Datenflut kaum noch hilfreich sind und keine nennenswerte Rolle mehr spielen. Bezugnahme und Referenzie-

rung sollte darum immer nur über die eindeutige Akzessionsnummer erfolgen. All dies gilt natürlich für Proteinsequenzen ganz entsprechend, allerdings werden diese praktisch gar nicht mehr direkt ermittelt, sondern nur noch aus den Nukleotidsequenzen abgeleitet. So existieren neben den Nukleotidsequenzdatenbanken Proteinsequenzdatenbanken, die noch schneller wachsen, denn bereits zu einem einzigen komplett sequenzierten neuen Organellengenom gehören dann beispielsweise Dutzende von neuen Proteinsequenzen – zu einem neuen Bakteriengenom schon tausende neuer Proteinsequenzen. Hier wird es kritisch, denn eine Proteinsequenz ist zunächst nur eine *Vorhersage* auf der Grundlage von Genmodellen. In einem Bakteriengenom mag ein ATG als Startcodon, ein durchgehendes Leseraster und ein Stopcodon am Ende noch gut ausreichen, um einen **ORF** (*Open Reading Frame*), ein **offenes Leseraster**, zu definieren. Wenn so ein ORF dann signifikante Ähnlichkeit mit ORFs in anderen Organismen hat oder sogar mit einem funktional charakterisierten Protein, ist die Wahrscheinlichkeit groß, dass es sich um ein echtes Gen und eine vernünftige Proteinsequenz handelt. Bei Eukaryonten mit ihren in aller Regel viel größeren Genomen und komplexeren Genstrukturen ist die Angelegenheit viel schwieriger – durch Introns alleine entstehen schon zahllose Möglichkeiten für Genvorhersagen. In den Organellen wiederum kann je nach Organismengruppe z.B. das RNA-Editing (Abschnitt 1.6.4 auf Seite 30) das Leben schwer machen. Eine handverlesene Datenbank inhaltlich kontrollierter Proteinsequenzen ist die **SWISSPROT**-Datenbank. Sie muss natürlich ergänzt werden durch Datenbanken, in denen sich noch nicht kontrollierte oder verifizierte Proteinübersetzungen nach dem einen oder anderen Modell tummeln. Dies sind beispielsweise **TrEMBL** oder **PIR**. Insbesondere für hypothetische Proteine aus den Kerngenomsequenzen der Eukaryonten ist die Gefahr, in der Datenbank auf eine falsche Proteinübersetzung zu stoßen, gar nicht gering. Mysteriöse Proteinsequenzen sollten immer auf alternatives Spleißen, idealerweise natürlich auf eine verfügbare cDNA-Sequenz, kontrolliert werden.

3.1.1 Datenbankeinträge: Textbasiertes Suchen und Dateiformate

Die Datenbanken machen Vorgaben zur möglichst informativen Beschreibung eines neuen Sequenzeintrages, aber die Verantwortung dafür liegt letztendlich beim einzelnen Wissenschaftler. Hier können sich Irrtümer oder Fehlinformationen einschleichen. Zu den häufigen Fehlern gehört, dass Sequenzen des Vektors (Abschnitt 1.7.1), in den die neue Nukleinsäure kloniert worden ist, auch Teil des Datenbankeintrags geworden sind oder dass falsche Angaben zu Beginn und Ende codierender Regionen (**CDS**, **Codierende Sequenz**) oder Introns in der Nukleotidsequenz gemacht werden. Solche Dinge fallen noch recht schnell auf, viel schwieriger sind natürlich taxonomische Verwechslungen und solche Fälle können den Nutzer durchaus manchmal einige Zeit der Recherche kosten. Zumindest im Prinzip aber sind die Datenbanken textbasiert nach Taxonomie und Informationen zur Nukleotidsequenz durchsuchbar, insbesondere natürlich nach den auf dem Sequenzabschnitt codierten Genen. Bei solchen Suchen können aber die Schwierigkeiten in Details stecken, denn Gene haben leider noch immer kein universell und verbindlich gültiges Benennungsmuster und manches **Gen** ist mit vielen verschiedenen Namen in die Datenbank eingegangen. Im Bereich der **Taxonomie** ist die Situation deutlich besser, denn zumindest die gültige binomiale lateinische Speziesbezeichnung (s. Abschnitt 2.2 auf Seite 51) sollte im Datenbankeintrag stehen. Trivialnamen von

Tabelle 3.1 Liste der großen öffentlichen molekularen Datenbanken (oben) sowie ausgewählter Beispiele für taxonomisch oder molekular spezialisierte Datenbanken (unten).

Datenbanken	WWW - Adresse	Inhalte
DDBJ	www.ddbj.nig.ac.jp	Internationale, öffentliche Datenbanken, die Nukleotid- und Proteinsequenzen verwalten. Integriert sind Literatur-, Genom-, Struktur-, taxonomische und diverse weitere Datenbanken.
EBI / EMBL	www.ebi.ac.uk	
NCBI / Genbank	www.ncbi.nlm.nih.gov	
Beispiele für molekulare Spezialdatenbanken mit taxonomischem Fokus		
CyanoBase	bacteria.kazusa.or.jp/cyanobase	Cyanobakteriengenome
Flybase	flybase.bio.indiana.edu	<i>Drosophila</i>
Gramene	www.gramene.org	Genome in Gräsern
HGMD - Human Genome Mutations Database	www.hgmd.org	Mutationen im menschlichen Genom
HIV Database	www.hiv.lanl.gov	HIV
J. Craig Venter Institute	www.tigr.org	Genomprojekte des TIGR (The Institute for Genomic Research)
TAIR – The Arabidopsis Information Resource	www.arabidopsis.org	<i>Arabidopsis</i>
WormBase	www.wormbase.org	<i>Caenorhabditis</i>
Beispiele für Spezialdatenbanken mit molekular-funktionalem Fokus		
Aramemnon	aramemnon.botanik.uni-koeln.de	Pflanzliche Membranproteine
Cluster of Orthologous Groups	www.ncbi.nlm.nih.gov/COG	Orthologe Proteine in verschiedenen Genomen
Expert Protein Analysis System	www.expasy.ch	Proteomik
GOBASE	gobase.bcm.umontreal.ca	Organellengenome
Kyoto Encyclopedia of Genes and Genomes	www.genome.jp/kegg	Verknüpfung von Genomdaten mit metabolischen Pfaden
Pfam	pfam.sanger.ac.uk	Protein Families Database
Beispiele für Spezialdatenbanken mit taxonomisch-systematischem oder phylogenetischem Fokus		
Angiosperm Phylogeny Group	www.mobot.org/MOBOT/research/APweb	Phylogenie der Angiospermen
Animal Diversity Web	animaldiversity.ummz.umich.edu	Systematik der Metazoa
International Plant Names Index	www.ipni.org/index.html	Gültige Pflanzennamen
TreeBASE	www.treebase.org	Phylogenetische Studien und Stammbäume
Tree of Life Web Project	www.tolweb.org	Phylogenie aller Lebensformen

Arten dürfen dort höchstens ergänzend auftreten. Auf eine textbasierte Suche in den Datenbanken allerdings darf man sich nie verlassen. Die Suche nach Sequenzhomologen, genau genommen zunächst einmal nach signifikant *ähnlichen* Sequenzen, ist hier der verlässlichere Weg, insbesondere um Sequenzeinträge aufzuspüren, die man ansonsten übersehen würde.

Den Zugang zu den Sequenzdatenbanken findet man mit jedem WWW-Browser. Die Startseiten der drei großen Datenbanken im WWW (Tab. 3.1) bieten einen einfachen, sofortigen Einstieg in textbasierte Suchen nach Datenbankeinträgen. Da die Datenbanken inhaltlich abgeglichen sind, ist es praktisch dem Geschmack des Nutzers überlassen, welche Suchformulare ihm ansprechend und übersichtlich erscheinen und welche Aus-

gabeformate er besonders übersichtlich findet. Hinter den Datenbanken am NCBI steht das *Entrez-System*, hinter denen des EBI das **SRS**, das *Sequence Retrieval System*. Die Datenbanken melden sich bereits einfachstmöglich auf den Einstiegsseiten mit einem kleinen Suchfenster, in das der Suchende seine Stichwörter eingeben kann, die dann mit einer logischen UND Verknüpfung zur Durchmusterung der Datenbank eingesetzt werden. Auf der Startseite des NCBI beispielsweise (Abb. 3.1) erlaubt ein Ausklappmenü, die Suche auf einzelne spezielle Datenbanken einzuschränken, also beispielsweise auf Nukleotidsequenzen, Proteinsequenzen, auf die Literaturdatenbank **PubMed**, auf komplettierte Genomprojekte, Proteinstrukturdatenbanken oder auf eine der anderen aus der wachsenden Zahl neuer Spezialdatenbanken. Die textbasierte Suche erlaubt, alle Bereiche eines Datenbankeintrages zu durchsuchen. Sie können also durchaus auch einen Forschernamen, vielleicht ergänzt um den ersten Buchstaben seines Vornamens, oder ein Jahresdatum oder ein Wort im Titel einer Publikation eingeben.

Wer also wissen will, ob schon eine Sequenz der Proteinuntereinheit A des Photosystems II in den Agaven bekannt ist, tippt einfach „Agavaceae psba“ ein (Abb. 3.1). Die Reihenfolge ist unwichtig, Groß- oder Kleinschreibung werden auch nicht berücksichtigt. Natürlich muss der Nutzer hoffen, dass die Taxonomie stimmt und dass sein Wunschgen hoffentlich richtig als *psbA* in den Datenbankeinträgen bezeichnet ist. Unser

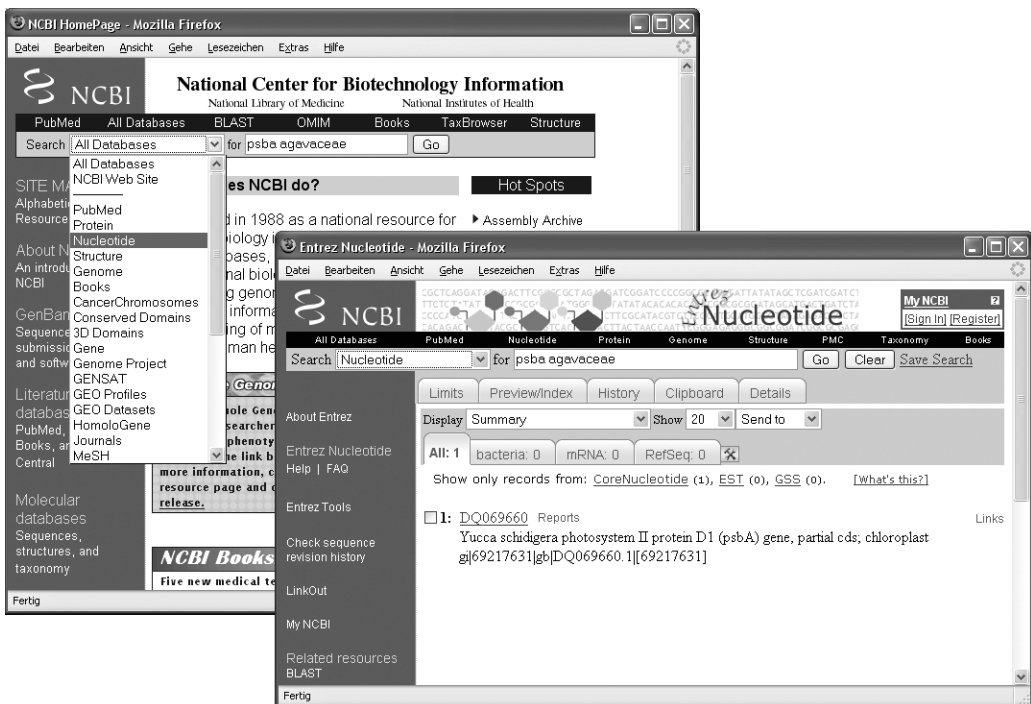


Abbildung 3.1 Oben: Die Startseite des NCBI mit einem einfachen Suchfenster, in das Suchbegriffe eingeegeben werden können, hier im Beispiel einmal „psba agavaceae“. Die Suche wird hier im Ausklappmenü auf Nukleotidsequenzen beschränkt. Unten: Die Suche liefert einen Datenbankeintrag mit der Akzessionsnummer DQ069660, die als Hyperlink direkt zum Datenbankeintrag weiterleitet.

Suchbeispiel liefert im Juni 2008 nur einen Datenbankeintrag: die *psbA*-Sequenz von *Yucca schidigera*. In der Ausgabe funktioniert die Akzessionsnummer DQ069660 als aktiver Querverweis und mit einem Mausklick kann der Datenbankeintrag im **GenBank-Format** aufgerufen werden. Den Aufbau des Datenbankeintrages zeigt Abbildung 3.2 auf der Seite gegenüber. Die Nukleotidsequenz wird idealerweise begleitet von einer Beschreibung der Sequenzeigenschaften, vor allem also der Ausdehnung codierender Regionen, der CDS. In unserem Beispiel beginnt der Sequenzeintrag in Position 1 mit dem Startcodon. Die codierende Region ist aber nicht vollständig: dies wird durch das Zeichen ‘>’ (oder ‘<’) bei den Sequenzkoordinaten angedeutet. Dies ist typischerweise der Fall, wenn Nukleotidsequenzen für phylogenetische Studien gewonnen wurden, weil die Oligonukleotide für die PCR (Abschnitt 1.7.2 auf Seite 36) in konservierten, meist codierenden Regionen ansetzen müssen.

Im Kopfbereich enthält der Datenbankeintrag eine Referenz zu der zugehörigen Publikation, ebenfalls sehr komfortabel mit einem Hyperlink, der Sie per Mausklick zu der Literaturdatenbank **PubMed** bringt, in der Sie zumindest die Zusammenfassung (das *abstract*) der Publikation finden können. Auch die Proteinübersetzung ist per Mausklick auf den Querverweis zum Datenbankeintrag mit der Proteinsequenz abrufbar und ebenso ist der Speziesname mit einer Verknüpfung in den *Taxonomy Browser* der Taxonomiedatenbank versehen. Der *Taxonomy Browser* ist eine fabelhafte Einrichtung, mit der Sie sich sehr schnell über weitere Querverweise in der taxonomischen Hierarchie hinauf- und hinunterbewegen können. In der Abbildung 3.3 auf Seite 80 ist einmal ein Bild für die Situation in der Gattung *Yucca* dargestellt. Das NCBI versteht sich zwar nicht als autoritative Quelle für taxonomische Information, setzt allerdings häufig aktuelle taxonomische Erkenntnisse sehr schnell um. Hilfreich ist der *Taxonomy Browser* in jedem Fall, um Taxa in der Datenbank zu identifizieren – selbst wenn man nicht mit der verwendeten Systematik einverstanden sein sollte. Das nächst höhere taxonomische Niveau über den Agavaceae, die Ordnung Asparagales (Abb. 3.3), sollte natürlich für die Suche nach einem *psbA*-Sequenzeintrag mindestens den schon für *Yucca* gefundenen, idealerweise noch weitere aus anderen Familien liefern. In der Tat identifiziert die Suche „asparagales psba“ sehr viel mehr Datenbankeinträge (Abb. 3.4 auf Seite 81), denn neben dem schon bekannten *Yucca*-Eintrag tauchen nun viele Sequenzen aus den Orchideen (Orchidaceae) auf. Beim näheren Betrachten zeigt sich, dass die meisten Einträge aber nur einen sehr kleinen Bereich des *psbA*-Gens abdecken. Für unsere Datensammlung könnten wir durch Auswahlkästchen in der Ergebnisausgabe (Abb. 3.4 auf Seite 81) einzelne Einträge sehr einfach auswählen.

Das **GenBank-Dateiformat** ist für den Transfer zwischen den phylogenetischen Analyseprogrammen wenig geeignet. Ein Dateiformat, das sowohl für Einzelsequenzen wie auch für multiple Sequenzen funktioniert, ist das **FASTA-Format**. Es ist zwar nicht sehr leistungsfähig im Bezug auf eine datenbankfähige Annotierung, aber dafür mit zahlreichen Programmen kompatibel, von denen es sowohl gelesen als auch geschrieben werden kann. Es gilt für Nukleotid- und Proteinsequenzen gleichermaßen. Das FASTA-Dateiformat sieht jeweils in der ersten Zeile hinter dem Zeichen ‘>’ den Namen der Sequenz vor. Die Sequenz selbst muss nach einem Zeilenumbruch in der nächsten Zeile beginnen und darf sich über beliebige viele Zeilen erstrecken. Im Ausklappmenü können wir das FASTA-Dateiformat unter vielen anderen auswählen (Abb. 3.4). Die Ausgabe für unsere ausgewählten *psbA*-Sequenzen der Asparagales sieht dann z.B. wie in der Abbil-

```

LOCUS      DQ069660                1059 bp    DNA        linear    PLN 15-SEP-2005
DEFINITION Yucca schidigera photosystem II protein D1 (psbA) gene, partial
            cds; chloroplast.
ACCESSION  DQ069660
VERSION    DQ069660.1  GI:69217631
KEYWORDS   .
SOURCE     chloroplast Yucca schidigera
ORGANISM   Yucca schidigera
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; Liliopsida; Asparagales; Agavaceae;
            Yucca.
REFERENCE  1 (bases 1 to 1059)
AUTHORS    Leebens-Mack, J., Raubeson, L.A., Cui, L., Kuehl, J.V., Fourcade, M.H.,
            Chumley, T.W., Boore, J.L., Jansen, R.K. and dePamphilis, C.W.
TITLE      Identifying the Basal Angiosperm Node in Chloroplast Genome
            Phylogenies: Sampling One's Way Out of the Felsenstein Zone
JOURNAL    Mol. Biol. Evol. 22 (10), 1948-1963 (2005)
PUBMED     15944438
REFERENCE  2 (bases 1 to 1059)
AUTHORS    Leebens-Mack, J.H., Raubeson, L.A., Cui, L., Kuehl, J.V.,
            Fourcade, M.H., Chumley, T.W., Boore, J.L., Jansen, R.K. and
            dePamphilis, C.W.
TITLE      Direct Submission
JOURNAL    Submitted (20-MAY-2005) Department of Biology, Institute of
            Molecular Evolutionary Genetics, and The Huck Institutes of Life
            Sciences, The Pennsylvania State University, 201 Life Sciences
            Building, University Park, PA 16802, USA
FEATURES   Location/Qualifiers
            source                1..1059
                                   /organism="Yucca schidigera"
                                   /organelle="plastid:chloroplast"
                                   /mol_type="genomic DNA"
                                   /db_xref="taxon:334597"
            gene                 1..>1059
                                   /gene="psbA"
            CDS                 1..>1059
                                   /gene="psbA"
                                   /codon_start=1
                                   /transl_table=11
                                   /product="photosystem II protein D1"
                                   /protein_id="AA204092.1"
                                   /db_xref="GI:69217632"
                                   /translation="MTAILERRESTSLWGRFCNWITSTENRLYIGWFGVLMIPTLLTA
            TSVFIIAFIAAPPVDIDGIREPVSGSLLYGNIIISGAIPTSAAGLHFYPIWEAASV
            DEWLYNGGPYELIVLHFLLVGACVYMGREWELSFRLGMRPWIAVAYSAPVAAATAVFLI
            YPIQGGSFSDGMPLGISGTNFNMFVQAEHNILMHFFHMLGVAGVFGGSLFSAMHGS
            VTSSLIRETTENESANEGYRFGQEETYNIVAAGHYGRLIFQYASFNNSRSLHFFLA
            AWPVVGWIFTALGISTMAFNLNGFNFNQSVVDSQGRVINTWADIINRANLGMEVMHER
            NAHNFPLDLAAVEVPSTNG"
ORIGIN
1 atgactgcaa ttttagagag acgcgaaagt acaagcctgt ggggtcgctt ctgtaactgg
61 ataaccagca ccgaaaaccg tctttacatt ggatggtttg gtgttttgat gatccctacc
121 ttattgacgc caacttctgt atttattatc gccttcattg ctgctccctc agtagatatt
181 gatggtatct gtgaacctgt ttctgggtct ttactttatg gaacacaat tatttctggt
241 gccattatct ctacttctgc agctataggt ttgcattttt acccgatatg ggaagcagca
301 tctgttgtag agtggttata caacggcggt ccttatgagc taattgttct acacttctta
361 cttggtgtag cttgctacat gggtcgtgaa tgggaactta gtttccgtct gggtatgcgt
421 ccttggaattg cttgtgcata ttcagctcct gttgcagcag ctactgctgt tttcttgatc
481 tatcctatcg gtcaaggaag ttctctgatg ggtatgcctt taggaatatt tggtacttct
541 aacttcatga ttgtattcca ggcggagcac aacatcctta tgcattccatt tccatgttta
601 ggcgtagctg gtgtattcgg cggctcccta ttagtgctta tgcattgttc cttggttaacc
661 tctagtttta tcagggaaac cactgaaaac gagtctgcta atgaaggtta cagattccgtg
721 caagaggaag aaacttataa tatcgtagct gctcatgggt attttggcgg atgtgatcttc
781 caatagcgga gtttcaacaa ttctcgttcc ctacatttct tcttggctgc ttggcctggt
841 gtaggatatct ggttcaactgc tttaggtatt agtactatgg ctttcaacct aaatgggttc
901 aatttcaacc aatctgtagt ttagtagtaa ggcctgtgta ttaacacatg ggctgatatt
961 atcaaccgtg ctcaaccttg tatggaagta atgcattgac gtaattgtca caacttccct
1021 ctagacctag ctgctgttga agttccattc acaaatgga
//

```

Abbildung 3.2 Ein typischer Eintrag in den Nukleotiddatenbanken im **GenBank**-Format: das (partiale) *psbA*-Gen in *Yucca schidigera*. Die Akzessionsnummer (engl. *accession number*), in diesem Fall DQ069660, dient der eindeutigen Identifizierung und sollte als Referenz *verlinkt* werden. Der Artnamen, die Literaturreferenz und die abgeleitete Proteinübersetzung sind *Links* zu den entsprechenden Datenbankeinträgen der *Taxonomy Database*, der PubMed und in die Proteinsequenzdatenbank.

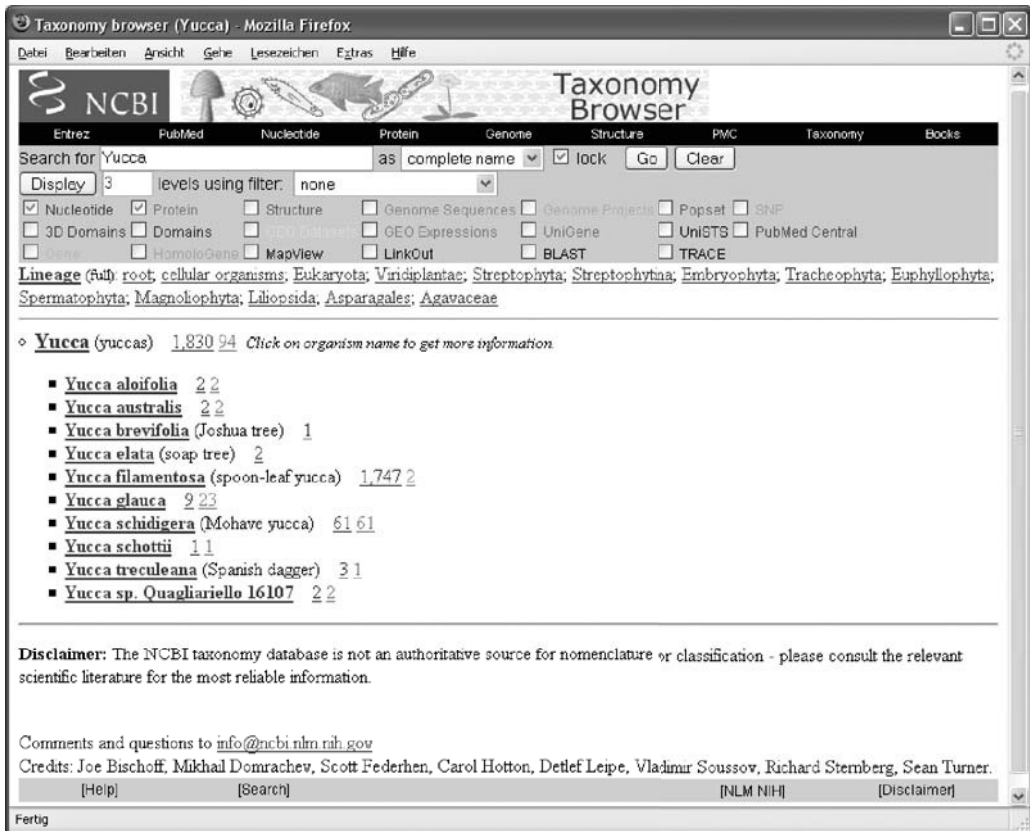


Abbildung 3.3 Der **Taxonomy Browser** des NCBI am Beispiel der Gattung *Yucca*. Für einige der Arten ist außer dem verbindlichen Speziesnamen in Klammern auch der englische Trivialname angegeben. Jedes Niveau der taxonomischen Hierarchie ist direkt per Mausklick aufrufbar. Die aktuelle Anzahl der verfügbaren Einträge in den Nukleotid-, Protein-, Struktur-, Genom- und vielen anderen Datenbanken am NCBI ist nach Auswahl in der Kopfzeile direkt über die Displayfunktion darstellbar. Hier im Beispiel sind die jeweiligen Anzahlen verfügbarer Nukleotid- und Proteinsequenzen für die *Yucca*-Arten angezeigt, die ebenfalls direkt abrufbar sind. Auffällig ist die hohe Zahl von Nukleotideinträgen bei nur zwei Proteinsequenzen für *Y. filamentosa*, die sich durch ein laufendes EST-Sequenzierungsprojekt für diese Art erklärt.

dung 3.5 aus. Im zweiten Ausklappfenster können Sie wählen, ob die Ausgabe direkt in den Webbrowser, ein temporäres *Clipboard* (Zwischenablage) oder in eine neue Datei auf Ihrem Rechner erfolgen soll.

3.1.2 Suche nach Sequenzähnlichkeiten

Nun ist zwar ein Anfang gemacht, aber wir wollen natürlich sicher gehen, dass auch wirklich alle Sequenzen korrekt annotiert sind und wir nicht irgendwo eine homologe Sequenz übersehen haben. Der inzwischen beliebteste Algorithmus zur Suche nach Sequenzähnlichkeiten ist der **BLAST**-Algorithmus. Dieses *Basic Local Alignment Search Tool* geht auf bioinformatische Arbeiten am NCBI zurück (Altschul et al. 1990). Seine

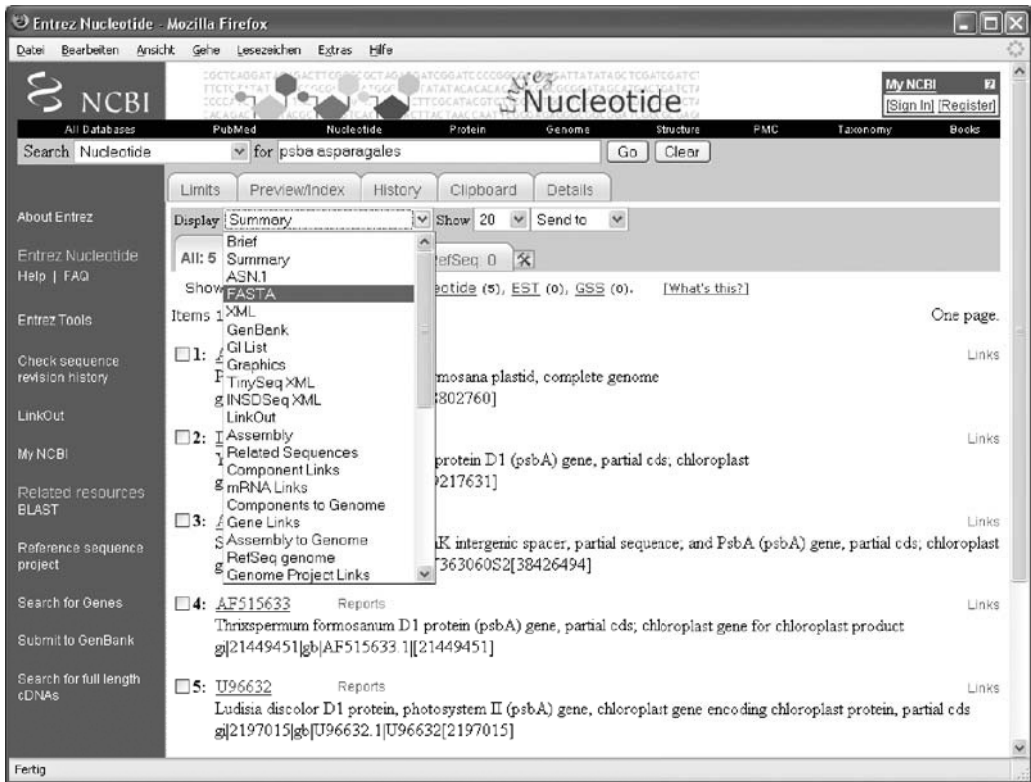


Abbildung 3.4 Identifizierte Datenbankeinträge, hier nach einer Suche für „psba asparagales“, können im Ausgabefenster mit Wahlkästchen ausgewählt werden. Das „Display“-Menü lässt verschiedene Darstellungsformen bzw. Dateiformate zu, hier wird das **FASTA-Format** ausgewählt. Die Ausgabe kann direkt gespeichert werden.

Beliebtheit ist unter anderem in der großen Geschwindigkeit begründet, mit der die riesigen, weiter wachsenden Datenmengen durchsucht werden können. Direkt aus der Kopfzeile des NCBI-Startfensters ist ein Verweis auf die Familie der BLAST-Programme wählbar (Abb. 3.1 auf Seite 77 oben). Möglich sind die Suche nach ähnlichen Nukleotidsequenzen mit einer Nukleotidsequenz als *Query* (**BLASTN**), nach ähnlichen Proteinsequenzen mit einer Proteinsequenz (**BLASTP**), mit einer Nukleotidsequenz unter Übersetzung in alle sechs **Leseraster** gegen eine Proteindatenbank (**BLASTX**) und *vice versa* auch mit einer Proteinsequenz gegen eine übersetzte Nukleotidsequenzdatenbank (**TBLASTN**). Schließlich könnten Sie sogar auch mit den sechs Übersetzungen Ihrer Nukleotidsequenz gegen alle Übersetzungen der Nukleotidsequenzen suchen (**TBLASTX**). Es öffnet sich ein Suchformular (Abb. 3.6 auf Seite 83), in dem Sie Ihre Suchsequenz (die *Query*) im FASTA-Format aus dem Zwischenspeicher einfügen können. Wollen Sie mit einer bereits in der Datenbank vorhandenen Sequenz suchen, reicht es auch, einfach deren Akzessionsnummer einzugeben. Mit verschiedenen Optionen kann die Suche eingegrenzt werden: Sie können nur einen Teil Ihrer Sequenz zur Suche einsetzen oder Sie können statt der kompletten *non-redundant* Datenbank (‘nr’ in der Voreinstellung) nur gegen (andere) Teile der Datenbank (insbesondere ESTs) suchen. Vor allem aber können

```
>gi|69217631|gb|DQ069660.1| Yucca schidigera photosystem II protein D1 (psbA) gene, partial...
ATGACTGCAATTTTAGAGAGACGCGAAAGTACAAGCCTGTGGGGTCGCTTCTGTAAGTGGATAACCAGCA
CCGAAAACCGCTCTTTACATTGGATGGTTTGGTGTGTTTGATGATCCCTACCTTATTGACCGCAACTCTGTCT
ATTTATTATCGCCTTCATGCTGCTCCTCCAGTAGATATTGATGGTATTCGTGAACCTGTTTCTGGGTCT
TTACTTTATGGAACAATATTATTCTGGTGCCATTATTCTACTTCTGCAGCTATAGGTTTGCATTTT
ACCGGATATGGGAAGCAGCATCTGTTGATGAGTGGTTATACAACGGCGGTCTTATGAGCTAATTGTTCT
ACACTTCTTACTTGGTGTAGCTTGCTACATGGGTCGTGAATGGGAACCTAGTTTCCGCTCGGGTATGCGT
CCTTGGATTGCTGTTGCATATTCAGTCTCCTGTTGCAGCAGCTACTGCTGTTTTCTTGATCTATCCTATCG
GTCAAGGAAGTTTCTCTGATGGTATGCCTTTAGGAATATCTGGTACTTTCAACTTCATGATTGTATTCCA
GGCGGAGCACAAACATCCTTATGCATCCATTCACATGTTAGGCGTAGCTGGTGTATTCGGCGGCTCCCTA
TTTAGTGCTATGCATGGTTCCCTTGGTAACCTCTAGTTTAAATCAGGGAACCACTGAAAACGAGTCTGCTA
ATGAAGGTTACAGATTTCGGTCAAGAGGAAGAACTTATAATATCGTAGCTGCTCATGGTTATTTTGGCCG
ATTGATCTTCCAATACGCGAGTTTCAACAATTCTCGTTCCTACATTTCTTCTGGCTGCTTGGCCTGTT
TGAGGTATCTGGTTCACTGCTTTAGGTATTAGTACTATGGCTTTCACCTAAATGGTTTCAATTTCAACC
AATCTGTAGTTGATAGTCAAGCCGTGTGATTAACACATGGGCTGATATCATCAACCGTGCTAACCTTGG
TATGGAAGTAATGCATGAACGTAATGCTCACAACTTCCCTCTAGACCTAGCTGCTGTTGAAGTTCATCT
ACAAATGGA

>gi|21449451|gb|AF515633.1| Thrixspermum formosanum D1 protein (psbA) gene, partial...
GAAAGTACAAGCCTATGGGGTCGCTTCTGCAACTGGATTACCAGTACTGAAAACCGTCTTTACATCGGAT
GGTTTGGTGTGTTTATGATCCCTACTTTATGACCGCAACTTCTGTATTATCATGTGCTTCTATTGCTGC
CCCTCCAGTCGATATTGATGGTATTCGTGAACCTGTTTCTGGGTCTCTACTTTATGGAACAATATTATA
TCAGGTGCCATTATTTCCATTTCCGCGAGCTATAGGTTTGCATTTTACCCAATATGGGAAGCAGCATCTG
TGGATGAGTGGTTATACAATGGCGGTCTTATGAACCTATTGTTCTACACTTTTACTTGGTGTAGCTTG
TTACATGGGTCGTGAGTGGGAACCTAGTTTCCGCTCTGGGTATGCGCCCTTGGATTGCTGTTGCATATTCA
GCTCCTGTTTGGCGGTGCTACGGCTGTTTTCTTGTACTATCCTATCGGTCAAGGAAGTTTTTCTGATGGTA
TGCCTTTAGGAATATCTGGTACTTTCAACTTCATGATTGTATTCCAGGCAGAGCACAAACATTTATGCA
TCCATTTCCACATGTTAGCGGTAGCTGGTGTATTTCGGCGGCTCCCTATTAGTGTATGCATGGTCTTTTG
GTAACCTCTAGTTTAAATCAGGGAACCACTGAAAATGAGTCTGCTAATGAAGGTTACAGATTTGGTCAAG
AAGGAAGAACTTATAATATTGAGCGCTCATGGTTATTTTGGCCGATTGATCTTCCAATATGCTAGTTT
CAACAATCTCGTTCTTGGCATTTCTTCTTGGCTGCTTGGCCTGTAGTGGGTATCTGGTCTACTGCTTTG
GGTATTAGTACTATGGCGTTCAACTTGAACGGTTTTAATTTTAAACCAATCCGTAGTTGATAGCCAAGGTC
GTGTATTAAACACTTGGGCTGATATCATAAATCGTGCTAATCTTGGTATGGAAGTAATGCATGAGCGTAA
TGCACACAACCTCCCTCTAGATTTAGTCTTCTGTA

>gi|2197015|gb|U96632.1|U96632 Ludisia discolor D1 protein ...
TNCCTTATNNCNAACCTCTGTATTATTATCNCNCTCATCNCNCTCCCTCCAGTCGATATTGATGGT
ATTCGTGAACCTGTTTCTGGGTCTCTACTTTATGGAACAATATTATCTCCGGTGCCA...
```

Abbildung 3.5 Beginn einer Beispieldatei mit den ausgewählten *psbA*-Datenbankeinträgen im FASTA-Dateiformat.

Sie taxonomisch begrenzen, was insbesondere sehr nützlich ist, wenn Sie viele Homologe in Gruppen zu erwarten haben, die Sie eigentlich nicht interessieren. Inzwischen genügt die Eingabe der ersten Buchstaben ins Formular und Sie erhalten die taxonomischen Bezeichnungen, die zur Auswahl stehen – Asparagales in unserem Beispiel.

In unserem Beispiel nutzen wir also die *Yucca-psbA*-Sequenz, um in den Asparagales zu suchen. Das Suchfenster akzeptiert auch Mengenoperatoren wie z.B.: 'Insecta OR Vertebrata' oder auch 'Bacteria NOT Gammaproteobacteria'. Ein wichtiger, geschwindigkeitsbestimmender Schritt der Datenbankdurchmusterung ist die Festlegung der Sensitivität. Dafür bietet das Suchformular drei Voreinstellungen an. Wenn wir die Details zu den Parametern betrachten, ist hier die so genannte *Word Size* besonders wichtig – die Anzahl von Nukleotiden, die zumindest an einer Stelle zwischen der Suchsequenz und einem Datenbankeintrag exakt übereinstimmen muss, bevor dieser überhaupt weiter betrachtet wird. Hier kann es im Einzelfall sinnvoll sein, die Voreinstellungen bis auf die Optionen '7' für Nukleotide oder '2' für Proteine für eine sensitivere Suche herunterzustellen. Ein weiterer Punkt betrifft die *Low Complexity*-Filter, die dafür sorgen, Regionen unausgewogener oder repetitiver Basenzusammensetzung nicht in die Sequenzvergleiche mit einzubeziehen. Ein typisches Beispiel sind die Poly-A-Schwänze von mRNAs, deren monotone Entsprechung nun in der Tat zu sinnlosen Scheintreffern in der Daten-

Gene und Stammbäume

Ein Handbuch zur molekularen Phylogenetik

Knoop, V.; Müller, K.

2009, XI, 386 S. 130 Abb., Softcover

ISBN: 978-3-8274-1983-5