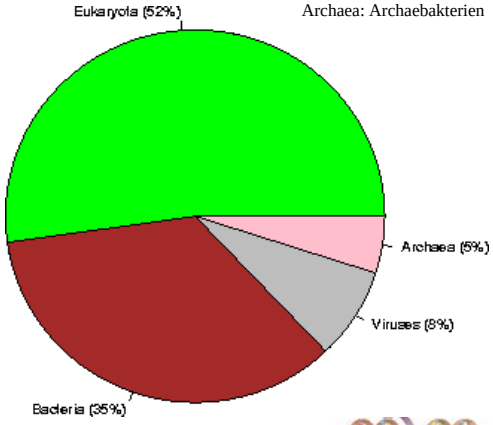
Protein-Datenbanken

- Sequenz-Datenbanken (Vertreter: Swiss-Prot)
- Domain/Familien-Datenbanken (Vertreter: InterPro)
- Struktur-Datenbanken (Vertreter: PDB)
- Vorsicht: Die Grenzen zwischen diesen Datenbank-Typen sind unscharf!

Swiss-Prot

- http://www.ebi.ac.uk/swissprot/
- Repository aller bekannten Proteinsequenzen
- Basiert auf Submission, Übersetzung und aktiver Suche, intensive (manuelle) Datenpflege
  - > 30 "Scientific Database Curators"
  - Redundanzfreiheit
  - Vierteljährliche Releases
- Tools für Protein-Analyse (z.B. Homologie-Modellierung)

Archaea: Archaebakterien

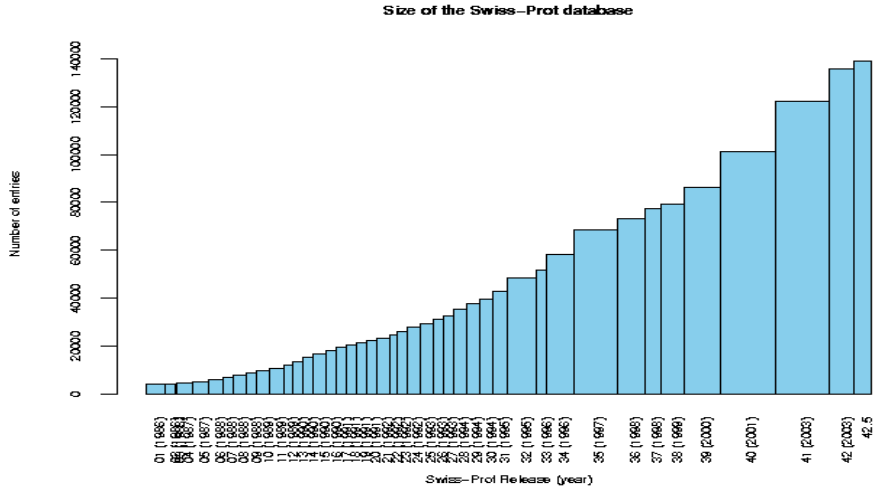


Swiss-Prot: Wachstum

Release 42.5 of 21-Nov-2003 of Swiss-Prot contains 138922 sequence entries,
comprising 51131444 amino acids abstracted from 110725 references.

3077 sequences have been added since release 42, the sequence data of
201 existing entries has been updated and the annotations of
8149 entries have been revised. This represents an increase of 2%.

The growth of the database is summarized below.



Swiss-Prot: Daten

- "Flaches" Datenmodell (Entry-basiertes Modell), sehr ähnlich zu EMBL
  - Autor, Datum, Länge, Methode, letzte Änderung
  - Organismus
  - Proteinsequenz (z.B. im FASTA-Format)
  - Links zu anderen Datenquellen, Literaturreferenzen
- Oracle-Dumps verfügbar (ca. 140 Tabellen)
- XML-Export
- Keine Änderungsübersicht!
- TrEMBL (Translations of EMBL)
  - Supplement zu Swiss-Prot
  - Enthält alle automatisch in AS-Sequenzen übersetzte CDS-Sequenzen aus EMBL
  - Keine Überschneidung mit (manuell) eingebrachten Swiss-Prot-AS-Sequenzen
  - SP-TrEMBL: Geplanter Nachfolger von Swiss-Prot

Swiss-Prot: Beispieleintrag

```
ID   GUNB_CLOTM     STANDARD;      PRT;   563 AA.
AC   P04956;
DT   13-AUG-1987 (Rel. 05, Created)
DT   13-AUG-1987 (Rel. 05, Last sequence update)
DT   01-FEB-1995 (Rel. 31, Last annotation update)
DE   ENDOGLUCANASE B PRECURSOR (EC 3.2.1.4) (EGB) (ENDO
DE   (CELLULASE B).
GN   CELB.
OS   Clostridium thermocellum.
OC   Bacteria; Firmicutes; Bacillus/Clostridium group;
OC   Clostridium.
OX   NCBI_TaxID=1515;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=NCIB 10682;
RX   MEDLINE=86148508; PubMed=3453102;
RA   Grepinet O., Beguin P.;
RT   "Sequence of the cellulase gene of Clostridium the
RT   endoglucanase B.";
RL   Nucleic Acids Res. 14:1791-1799(1986).
CC   -!- FUNCTION: THIS ENZYME CATALYZES THE ENDOHYDROL
CC       GLUCOSIDIC LINKAGES IN CELLULOSE, LICHENIN AND
CC       GLUCANS.
CC   -!- CATALYTIC ACTIVITY: ENDOHYDROLYSIS OF 1,4-BETA
CC       LINKAGES IN CELLULOSE.
CC   -!- DOMAIN: A 24 RESIDUES DOMAIN IS REPEATED TWICE
CC       WELL AS IN OTHER C.THERMOCELLUM CELLULOSOME EN
CC       MAY FUNCTION AS THE BINDING LIGAND FOR THE SL
CC   -!- SIMILARITY: BELONGS TO CELLULASE FAMILY A (FAM
CC       HYDROLASES).
```

```
DR   EMBL; X03592; CAA27266.1; -.
DR   PIR; A23512; CZCLBM.
DR   HSSP; P54583; 1ECE.
DR   InterPro; IPR002105; Dockerin_1.
DR   InterPro; IPR002048; EF-hand.
DR   InterPro; IPR001547; Glyco_hydro_F5.
DR   Pfam; PF00150; cellulase; 1.
DR   Pfam; PF00404; Dockerin_1; 2.
DR   PROSITE; PS00018; EF_HAND; UNKNOWN_1.
DR   PROSITE; PS00448; CLOS_CELLULOSOME_RPT; 2.
DR   PROSITE; PS00659; GLYCOSYL_HYDROL_F5; 1.
KW   Cellulose degradation; Hydrolase; Glycosidase; Repeat; Signal.
FT   SIGNAL        1     27       OR 31.
FT   CHAIN        28    563       ENDOGLUCANASE B.
FT   ACT SITE    204    204       PROTON DONOR (BY SIMILARITY).
FT   ACT SITE    363    363       NUCLEOPHILE (BY SIMILARITY).
FT   DOMAIN      502    557       2 X 24 AA APPROXIMATE REPEATS.
FT   REPEAT      502    526       1.
FT   REPEAT      534    557       2.
SQ   SEQUENCE   563 AA;  63929 MW;  866FE55704A1DE4B CRC64;
     MKKFLVLLIA LIMIATLLVV PGVQTSAEGS YADLAEPDDD WLHVEGTNIV DKYGNKVWIT
     GANWFGFNCR ERMLLDSYHS DIIADIELVA DKGINVVRMP IATDLLYAWS QGIYPPSTDT
     SYNNPALAGL NSYELFNFML ENFKRVGIKV ILDVHSPETD NQGHNYPLWY NTTIITEEIFK
     KAWVWVAERY KNDDTIIGFD LKNEPHTNTG TMKIKAQSAI WDDSNHPNNW KRVAEETALA
     ILEVHPNVLI FVEGVEMYPK DGIWDDETFD TSPWTGNNDY YGNWWGGNLR GVKDYPINLG
     KYQSQLVYSP HDYGPIVYEQ DWFKGDFITA NDEQAKRILY EQCWRDNWAY IMEEGISPLL
     LGEWGGMTEG GHPLLDLNLK YLRCMRDFIL ENKYKLHHTF WCINIDSADT GGLFTRDEGT
     PFPGGRDLKW NDNKYDNYLY PVLWKTEDGK FIGLDHKIPL GRNGISISQL SNYTPSVTPS
     PSATPSPTTI TAPPTDTVTY GDVNGDGRVN SSDVALLKRY LLGLVENINK EAADVNVSGT
     VNSTDLAIMK RYVLRSISEL PYK
//
```

Swiss-Prot: FASTA-Format

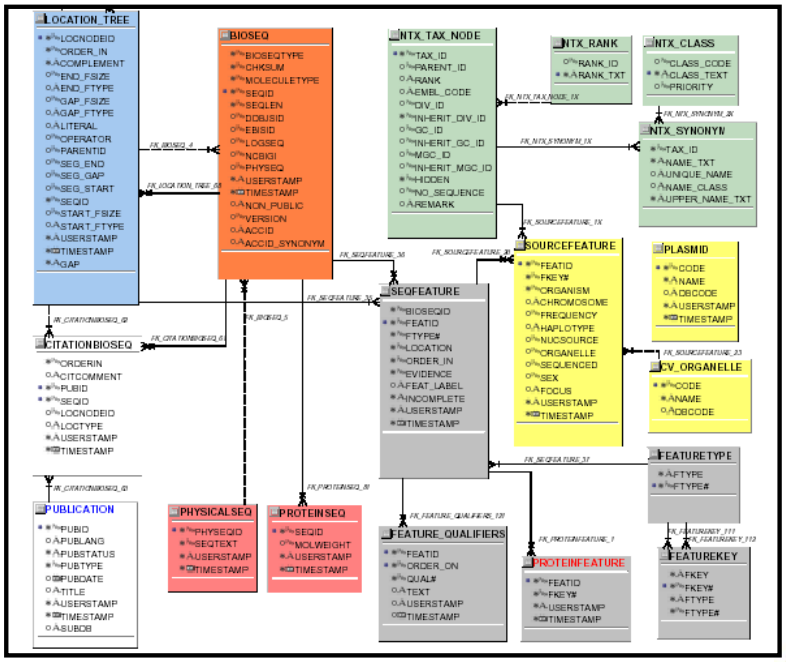
- Alternatives Format für AS-Sequenzen

  - begins with a description line indicated by a “>” sign
  - followed by amino acid seq. in capital letters,
  - no numbers, no blocks
  - line length usually 80 characters

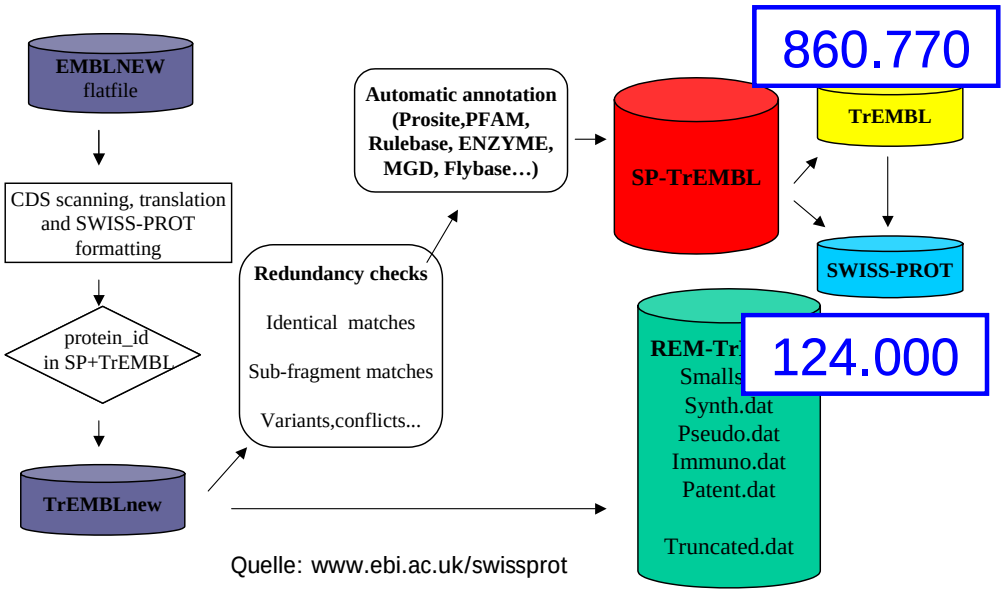
Example:

```
>gi|532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRT
QIWQKHRTSNDSALILLNKHYNLTVTCKRPGNKTVLPVTIMAGLVFHSQKYNLRLRQAWC
HFPSNWKGAWKEVKEEIVNLPKERYRGTNDPKRIFFQRQWGDPETANLWFNCHGEFFYCK
MDWFLNYLNNLTVDADHNECKNTSGTKSGNKRAPGPCVQRTYVACHIRSVIIWLETISKK
TYAPPREGHLECTSTVTGMTVELNYIPKNRTNVTLSPQIESIWAAELDRYKLVEITPIGF
APTEVRRYTGGHERQKRVPFVXXXXXXXXXXXXXXXXXXXXXXVQSQHLLAGILQQQKNL
LAAVEAQQQMLKLTIWGVK
```

Swiss-Prot: Relationales Schema

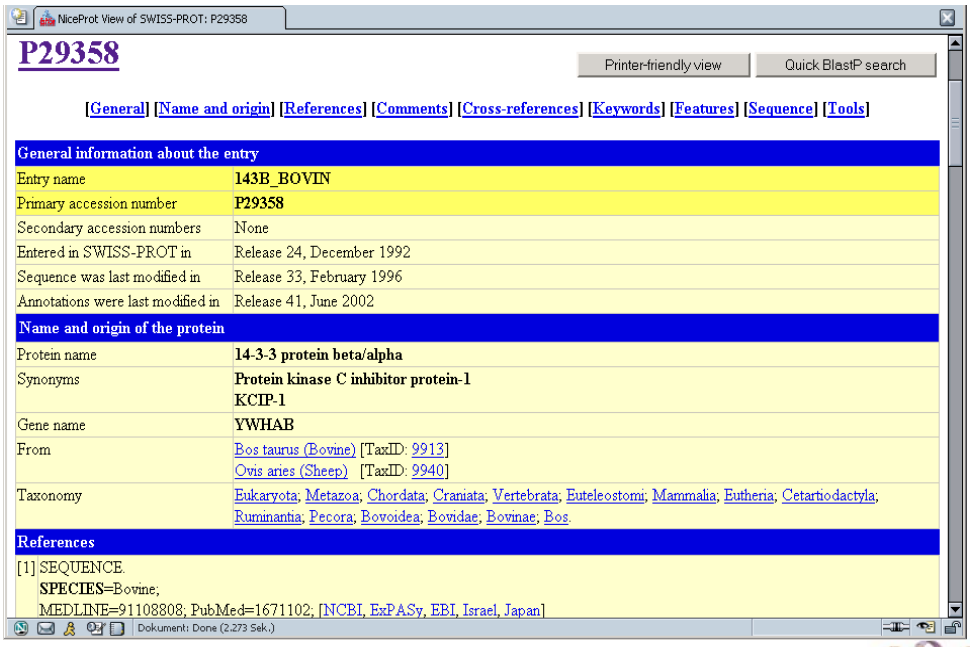


TrEMBL: Datengewinnung



Quelle: www.ebi.ac.uk/swissprot

Swiss-Prot Web Interface



Swiss-Prot: Annotationen

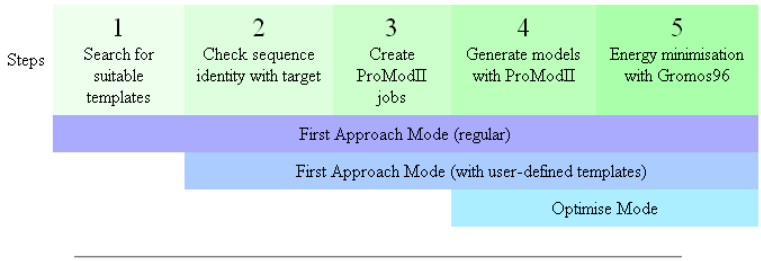
- CC-Felder für Kommentare
  - Unterteilt in Topics
  - Beispiele: Caution, Disease, Function, Regulation, ...
- FT: Feature Table
  - Modifikationen, Sequenzabschnitte, Sekundärstruktur
- KW: Keywords
  - Ca. 800 verschiedene Keywords
- Einträge oft Mischung aus Controlled Vocabularies und Freitext
- Seit kurzem: Evidence Codes für alle Annotationen (Curator, Opinion, By Similarity, Experiment, ...)

Swiss-Prot: Versionierung / Identifikation

- Swiss-Prot Release ca. alle 3 Monate
- ID und AC Line
  - ID: X_Y; X: "Name" des Proteins; Y: "Name" der Spezies
  - Keine Standards für Proteinnamen
  - Spezies mit wissenschaftlichen oder umgangssprachlichen Namen
  - AC: Accession Number
  - Primäre ID
  - Kann mehrere Einträge enthalten (Merged Entries)
- Keine Versionen von Einträgen
  - Last Update
  - Keine Änderungsübersichten

Swiss-Model: 3D-Strukturbestimmung

- Ausgangsproblematik: Nach derzeitigem biochemischem Kenntnisstand ist es nur in Ausnahmefällen möglich, von der AS-Sequenz auf die 3-Struktur zu schließen
- Ausweg: Vergleich mit ähnlichen Sequenzen und deren Struktur (falls bekannt)



Step	Program/Method	Database	Action
1	BLASTP2	ExNRL-3D	Will find all similarities of target sequence with sequences of known structure.
2	SIM	-	Will select all templates with sequence identities above 25% and projected model size larger than 20 residues. Furthermore, this step will detect domains which can be modelled based on unrelated templates
3	-	-	Generate ProModII input files
4	ProModII	ExPDB	Generate all models
5	Gromos96	-	Energy minimsation of all models

BLAST 2.0

- BLAST
  - verschiedene Varianten für AS und Nuk. Sequenzen
  - schneller als dyn. Programmierung
  - aber weniger sensitiv und berücksichtigt keine Lücken
  - findet lokales statt globales Alignment
  - nutzt AS Austauschmatrizen, PAM, BLOSUM
- Ansatz zur Beschleunigung
  - Anfrage-Sequenz wird in Wörter der Länge W (W=3) zerlegt
  - Wortliste wird um ähnliche Wörter erweitert
  - Nur Worte mit Score >= T werden in DB gesucht
  - Wort-Treffer werden nach links und rechts erweitert

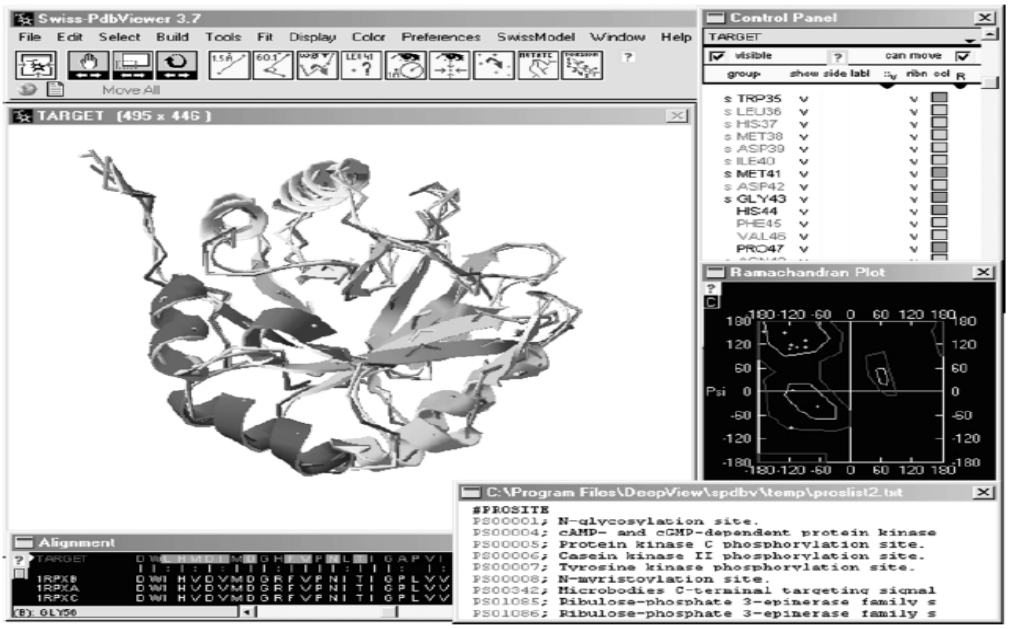
BLAST 2.0

- Beispiel: (W=2, T=8) Anfrage: qlnfsagw

Initiales Wort	Erweiterte Liste
ql	ql, qm, hl, zl
ln	ln, lb
nf	nf, af, ny, df, qf, ef, gf, hf, kf, sf, tf, bf, zf
fs	fs, fa, fn, fd, fg, fp, ft, fb, ys
sa	nothing scores 8 or higher
ag	ag
gw	gw, aw, rw, nw, dw, qw, ew, hw, iw, kw, mw, pw, sw, tw, vw, bw, zw, xw

- Invertierte Liste als Index nutzen
  - bei W=3 und 20AS nur 8000 verschiedene Worte möglich
- Probleme
  - ABCDEGH und ABCDEEFGH

Swiss-Model


Swiss-PdbViewer 3.7
File Edit Select Build Tools Fit Display Color Preferences SwissModel Window Help
Move All
TARGET (495 x 446)
Control Panel
TARGET
visible
can move
group
show side labl
ribn col
s TRP35
s LEU36
s HIS37
s MET38
s ASP39
s ILE40
s MET41
s ASP42
s GLY43
HIS44
PHE45
VAL46
PRO47
Ramachandran Plot
Psi
Alignment
1RPXB
1RPXA
1RPXC
C:\Program Files\DeepView\spdbv\temp\prosist2.txt
#PROSITE
PS00001; N-glycosylation site.
PS00004; cAMP- and cGMP-dependent protein kinase
PS00005; Protein kinase C phosphorylation site.
PS00006; Casein kinase II phosphorylation site.
PS00007; Tyrosine kinase phosphorylation site.
PS00008; N-myristoylation site.
PS00342; Microbodies C-terminal targeting signal
PS01085; Ribulose-phosphate 3-epimerase family s
PS01086; Ribulose-phosphate 3-epimerase family s

InterPro

- Sekundärdatenbank zu Proteinsequenzen (Schwerpunkt: Protein-Domains)
- Motivation
  - Bestimmte Sequenzabschnitte (Motifs) bestimmen Funktion des Proteins
  - Datenbanken zur Beschreibung interessanter Domänen (Proteinfamilien) nötig
  - Untersuchung neuer Sequenzen auf Vorhandensein bekannter Domänen – Rückschlüsse auf Funktion
- InterPro: Integrierte Datenbank von Proteindomänen-Datenbanken

The InterPro consortium:

- Co-ordinated by EBI (R. Apweiler & team)
- PROSITE (A. Bairoch, P. Bucher, N. Hulo, C. Sigrist, L. cerutti. M. Pagni, L. Falquet)
- PRINTS (T. Attwood, P. Bradley)
- PFAM (R. Durbin, A. Bateman, S. Griffiths-Jones)
- PRODOM (D. Kahn, F. Servant)
- SMART (C. Ponting, R. Copley, N. Dickens)
- TIGRFAMs (D. Haft, O. White)

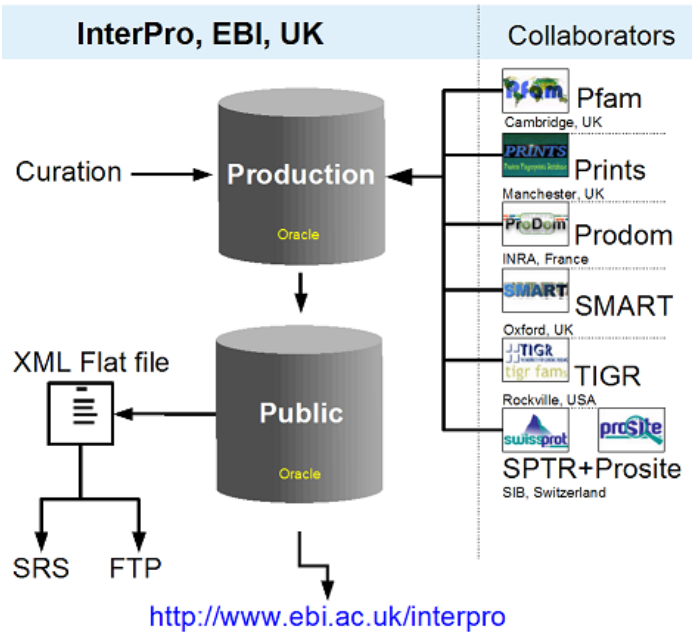





InterPro: Biologischer Fokus

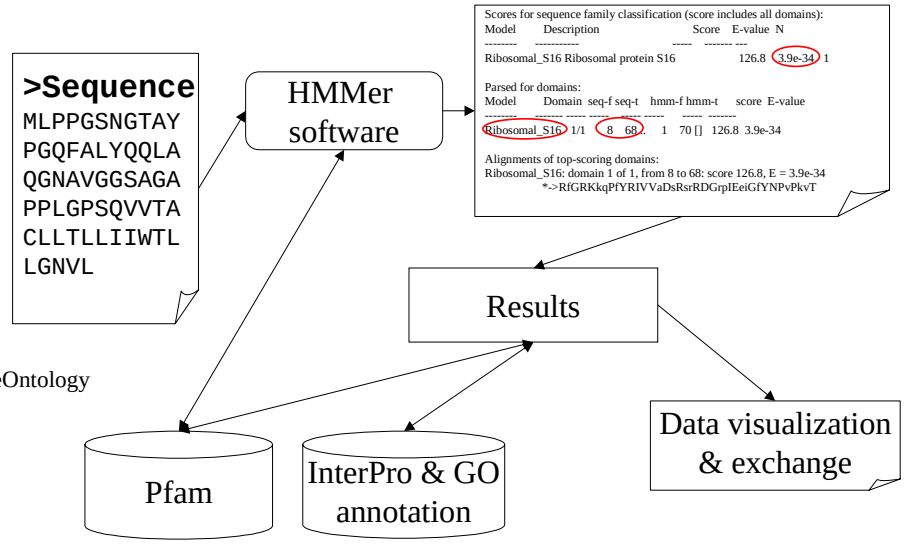
- **Family -** group of evolutionarily related proteins, that share one or more domains/repeats in common.
- **Domain -** independent structural unit which can be found alone or in conjunction with other domains or repeats.
- **Repeat -** region occurring more than once that is not expected to fold into a globular domain on its own.
- **PTM** (post-translational modification) -The sequence motif is defined by the molecular recognition of this region in a cell.

InterPro: Datengewinnung



InterPro: Datengewinnung (2)

- Beispiel: Pfam (Protein families database of alignments and HMMs; Multiple sequence alignments and hidden Markov models of common protein domains)



  - GO: GeneOntology

InterPro: Datengewinnung (3)

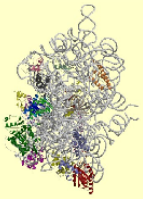


Figure 1: 1pns
Ribosome
Crystal structure of a streptomycin dependent ribosome from e. coli, 30s subunit of 70s ribosome. this file, 1pns, contains the 30s subunit, two trnas, and one mrna molecule. the 50s ribosomal subunit is in file 1pnu.

Key:

Domain	Chain	Start Residue	End Residue
Ribosomal_S2	B	10	226
Ribosomal_S3_C	C	119	202
Ribosomal_S3_N	C	2	62
KH	C	65	112
Ribosomal_S4	D	3	98
S4	D	99	146
Ribosomal_S5	E	5	71
Ribosomal_S5_C	E	80	153
Ribosomal_S6	F	2	93

Accession number: **PF00886**

Ribosomal protein S16

Add Annotation

This family forms **structural complexes** with other Pfam families, to view them click here

INTERPRO description (entry IPR000307)

Ribosomes are the particles that catalyze mRNA-directed protein synthesis in all organisms. The codons of the mRNA are exposed on the ribosome to allow tRNA binding. This leads to the incorporation of amino acids into the growing polypeptide chain in accordance with the genetic information. Incoming amino acid monomers enter the ribosomal A site in the form of aminoacyl-tRNAs complexed with elongation factor Tu (EF-Tu) and GTP. The growing polypeptide chain, situated in the P site as peptidyl-tRNA, is then transferred to aminoacyl-tRNA and the new peptidyl-tRNA, extended by one residue, is translocated to the the P site with the aid the elongation factor G (EF-G) and GTP as the deacylated tRNA is released from the ribosome through one or more exit sites [MEDLINE:21196157], [MEDLINE:21185928]. About 2/3 of the mass of the ribosome consists of RNA and 1/3 of protein. The proteins are named in accordance with the subunit of the ribosome which they belong to - the small (S1 to S31) and the large (L1 to L44). Usually they decorate the rRNA cores of the subunits.

Many of ribosomal proteins, particularly those of the large subunit, are composed of a globular, surfaced-exposed domain with long finger-like projections that extend into the rRNA core to stabilize its structure. Most of the proteins interact with multiple RNA elements, often from different domains. In the large subunit, about 1/3 of the 23S rRNA nucleotides are at least in van der Waal's contact with protein, and L22 interacts with all six domains of the 23S rRNA. Proteins S4 and S7, which initiate assembly of the 16S rRNA, are located at junctions of five and four RNA helices, respectively. In this way proteins serve to organize and stabilize the rRNA tertiary structure. While the crucial activities of decoding and peptide transfer are RNA based, proteins play an active role in functions that may have evolved to streamline the process of protein synthesis. In addition to their function in the ribosome, many ribosomal proteins have some function 'outside' the ribosome [MEDLINE:21185928], [MEDLINE:20566949].

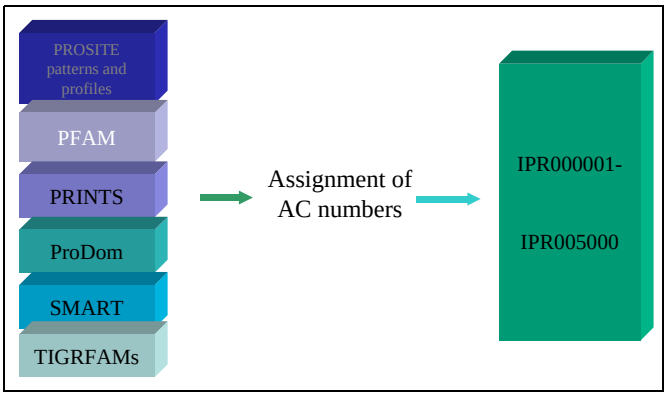
Ribosomal protein S16 is one of the proteins from the small ribosomal subunit. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities PUB00005070, groups:

- Eubacterial S16.
- Algal and plant chloroplast S16.
- Cyanelle S16.
- Neurospora crassa mitochondrial S24 (cyt-21).

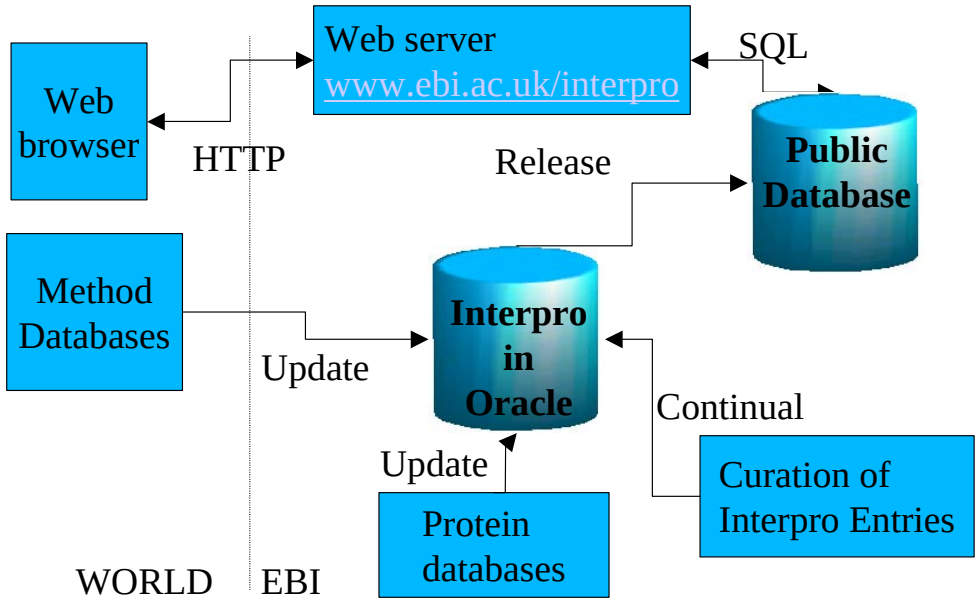
S16 proteins have about 100 amino-acid residues.

InterPro: Datengewinnung (4)

- Quellen bleiben eigenständig
- Regelmäßige Aktualisierungen
- Jeder Entry der Quelle wird Entry in InterPro
  - Aber: Zusammenhänge bleiben erhalten (Verifizierbarkeit!)
- Größtenteils manuelles Verfahren
  - Redundante Einträge
  - Sub/Superdomänen-Relationen zwischen Entries



InterPro: Architektur

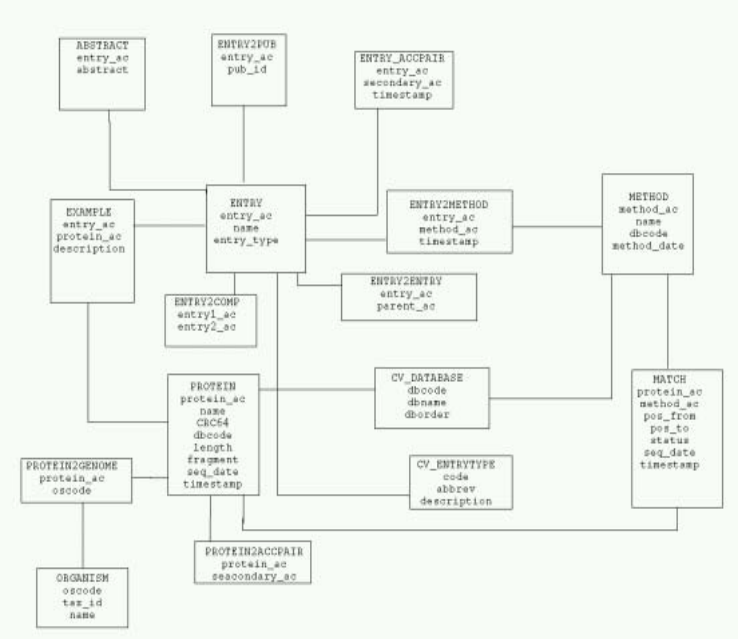


InterPro: Datenarten

- Basic Data
  - InterPro Entries (ENTRY)
  - Proteins (PROTEIN)
  - Methods (METHOD)
- Annotation
  - Abstracts (ABSTRACT)
  - Publications (PUB, AUTHOR, BOOK …)
  - Examples (EXAMPLE)
- Cross References
  - Hierarchical Relationships (ENTRY2ENTRY, ENTRY2COMP)
  - Methods Mapping (ENTRY2METHOD)
  - Matches (MATCH)
- Supporting Data
  - Secondary AC numbers (ENTRY_ACCPAIR)
  - Proteome Analysis Data (PROTEIN2GENOME, ORGANISM)
- Audit Tables

  - Methods - match domains and families
    - Eg: PF00001: 7 transmembrane receptor (rhodopsin family)
  - Proteins
    - Eg: O00155: PROBABLE G PROTEIN-COUPLED RECEPTOR GPR25.
  - Matches – precomputed
    - Eg: PF00001: matches O00155 at amino acids 56-306
  - Entries – logical groupings of Methods
    - Eg: IPR000276: Rhodopsin-like GPCR superfamily

InterPro: Oracle-Schema (Auszug)



- Insgesamt 41 Tabellen (ohne Beziehungstabellen)

InterPro: Tabellen-Attribute

ENTRY

- ENTRY_AC (unique, primary key) IPR000276
- NAME (unique, primary key) Rhodopsin-like GPCR superfamily
- SHORT_NAME (unique, primary key) GPCR_Rhodpsn
- ENTRY_TYPE F
  - D = Domain
  - F = Family
  - P = Pattern
  - R = Repeat

PROTEIN

- PROTEIN_AC (unique, primary key) O00155
- LEN 360
- CRC64 1FAE99650F977E27
- DBCODE S
  - S = SwissProt
  - T = TrEMBL
- NAME GPRP_HUMAN
- FRAGMENT N
  - Y / N

ENTRY2METHOD

- ENTRY_AC IPR000276
  - refers to ENTRY
- METHOD_AC O00155
  - refers to METHOD
- EVIDENCE MAN
  - A controlled vocabulary
  - refers to CV_EVIDENCE
    - Almost all: MAN = maually assigned

METHOD

- METHOD_AC (unique, primary key)PF00001
- NAME 7tm_1
- DBCODE H
  - Code from CV_DATABASE

    D=PRODOM
    P=PROSITE pattern
    Q=PROSITE prereleased profile
    M=PROSITE profile
    H=PFAM
    R=SMART
    t=TIGRFAMS
- METHOD_DATE 2001-08-06 00:00:00.0

MATCH

- PROTEIN_AC: refers to PROTEIN O00155
- METHOD_AC: refers METHOD PF00001
- POS_FROM 56
- POS_TO 306
- STATUS T
  - T = True ? = Unknown F = False N = False negative P = false positive
- DBCODE (D,P,Q,M,R,H,F,R,t) H
- SCORE 0.00000000000000000000000000000000000000000000000000000000000000000000000000000610
- EVIDENCE DEC
  - Contolled from CV_EVIDENCE (DEC=Computed by Timelogic Decypher)

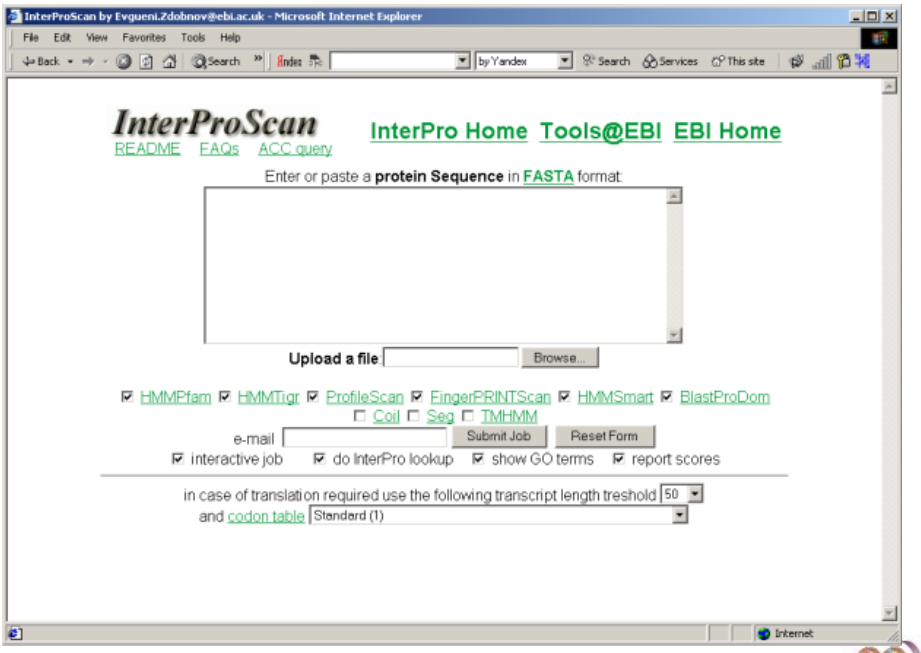
InterPro: Oracle-Anfragen

- How many short (<100 aa) Drosophila proteins have C2H2 zinc fingers (IPR000822) ?
  - select count(p.protein_ac) from protein p, entry2method e, match m
  - where m.protein_ac = p.protein_ac
  - and m.method_ac = e.method_ac
  - and p.len < 100
  - and e.entry_ac = 'IPR000822'
  - and p.protein_ac in (select protein_ac from protein2genome
  - where oscode = 'DROME') ;
- 10

- Which InterPro entries containing only Pfam signatures are common for Human and V. cholerae proteomes (SwissProt proteins only) ?
  - 1 select e.entry_ac, count(e.entry_ac) from entry2method e,protein2genome g, match m, protein p
  - where g.oscode = 'HUMAN' and m.protein_ac = g.protein_ac
  - and g.protein_ac = p.protein_ac and p.dbcode = 'S' and m.method_ac = e.method_ac
  - and m.dbcode = 'H' having count(e.entry_ac) = 1 group by e.entry_ac
  - intersect
  - select e.entry_ac, count(e.entry_ac) from entry2method e,protein2genome g, match m, protein p
  - where g.oscode = 'VIBCH' and m.protein_ac = g.protein_ac
  - and g.protein_ac = p.protein_ac and p.dbcode = 'S' and m.method_ac = e.method_ac
  - and m.dbcode = 'H' having count(e.entry_ac) = 1 group by e.entry_ac

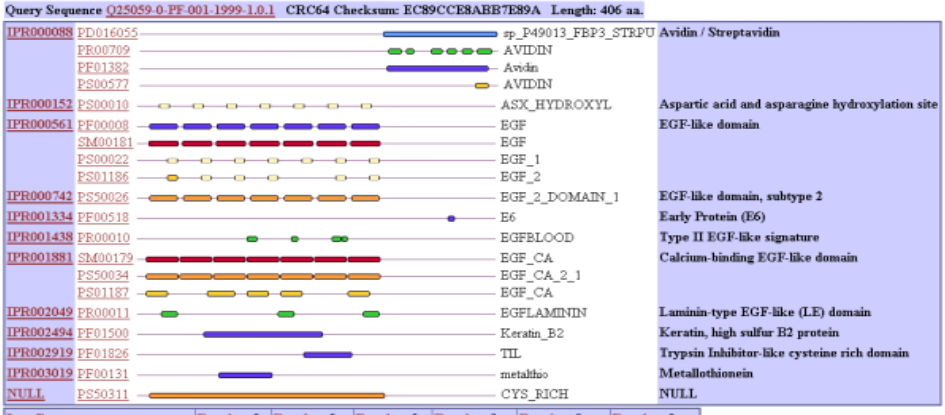
    - IPR000206 1
    - IPR000307 1
    - IPR000398 1
    - …
    - …
    - IPR002930 1
    - IPR003156 1
  - (28 entries)

InterPro: Web Interface



InterPro: Web Interface (2)

Query Sequence Q25059-0-PF-001-1999-1.0.1 **CRC64 Checksum: EC89CCE8ABB7E89A Length: 406 aa.**



InterPro	Results of FPrintScan vs. PRINTS	Results of HMMSmart vs. SMART	Results of HMMPfam vs. PFAM-A	Results of ScanProsite vs. PROSITE	Results of ProfileScan vs. PROFILES	Results of BlastProDom vs. PRODOM
IPR000088 Avidin / Streptavidin	PR00709 [283-297]T [303-311]T [331-343]T [349-359]T [366-377]T [383-399]T		PF01382 [281-395]T	PS00577 [381-395]T		PD016055 [277-405]T
IPR000152 Aspartic acid and asparagine hydroxylation site				PS00010 [27-38]? [65-76]?		

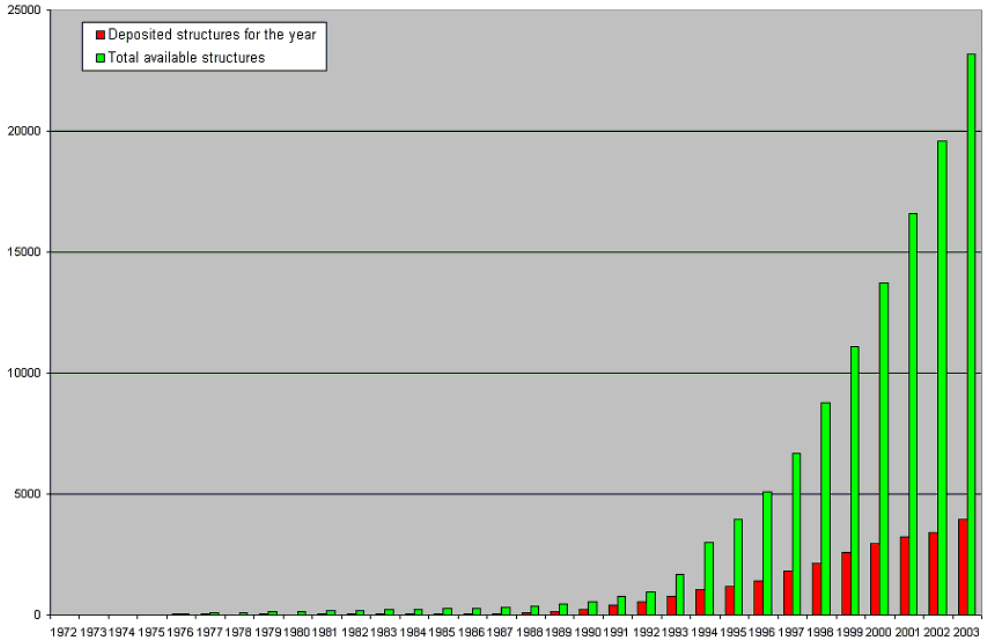
PDB: Protein Data Bank

- Protein-Struktur-Datenbank
- Motivation: Proteine falten sich in komplexe Strukturen, die entscheidend für die Funktion sind

PDB
PROTEIN DATA BANK

- Strukturaufklärung
  - Röntgenkristallographie (seit 50‘er Jahren), Massenspektrometrie, Nuclear Magnetic Resonance (NMR)
- Protein Data Bank
  - Repository aller (bekannten) Protein-3D-Strukturen
  - Seit 1971 in Brookhaven; seit 1999: Rutgers University
- Entry-Based Legacy Format; sehr komplexes 3D-Datenmodell
- Enge Kooperation mit OMG "Specification for Macromolecular Structure, v 1.0" (http://www.omg.org/technology/documents/formal/macro_molecular.htm)

PDB: Wachstum



PDB: Strukturabbildung

Excerpt of ATOM records from a legacy PDB format file

```
...
ATOM      6  CG1 VAL A   1      7.009  20.127   5.418  ...
ATOM      7  CG2 VAL A   1      5.246  18.533   5.681  ...
ATOM      8  N   LEU A   2      9.096  18.040   3.857  ...
ATOM      9  CA  LEU A   2     10.600  17.889   4.283  ...
ATOM     10  C   LEU A   2     11.265  19.184   5.297  ...
ATOM     11  O   LEU A   2     10.813  20.177   4.647  ...
ATOM     12  CB  LEU A   2     11.099  18.007   2.815  ...
ATOM     13  CG  LEU A   2     11.322  16.956   1.934  ...
...
```

A single instance of the AtomSite structure documented in Section 2.3.5.1, "AtomSite," on page 2-15 stores the cartesian position and other information about an atom just as a single ATOM record does in this legacy PDB format. The complete list (an IDL sequence) of all atoms in a macromolecular structure is returned by invoking the **get_atom_site_list** method on an instance of the **Entry** interface object.

As a simple example to illustrate the ease-of-use of the interface definition, the following Java code fragment would print out the atom identifier, atom type, and the cartesian (x,y,z) position for all atoms in the macromolecule 4hhb.

```
Entry e = entryFactory.get_entry_from_id("4hhb");
AtomSite[] a = e.get_atom_site_list();
for (int i = 0; i < a.length; i++)  {
    System.out.println(a[i].id + " " + a[i].type_symbol.id
        + " (" +  a[i].cartn.x + ", " + a[i].cartn.y
        + ", " + a[i].cartn.z + ")");
    }
```

→

```
...
6 C (7.002, 20.127, 5.418)
7 C (5.246, 18.533, 5.681)
8 N (9.096, 18.040, 3.857)
9 C (10.60, 17.889, 4.283)
...
```

PDB: Strukturabbildung (2)

ENTITY_SOURCE_TYPE
ENTITY_SRC_GEN
ENTITY_SRC_NAT
ENTITY_SRC_SYN
ENTRY
ENZYME_CLASS
ENZYME_CLASS_SYNC
ENZYME_STRUCT
EXPERIMENTAL_METHC
EXPTL
EXPTL_CRYSTAL
EXPTL_CRYSTAL_GROV
EXPTL_CRYSTAL_GROV
GEOM_ANGLE
GEOM_BOND
GEOM_CONTACT
GEOM_TORSION
GO_SORT_PDB
GO_TERM
GO_TERM2TERM
GO_TERM_BAYES
GO_TERM_PDB
GO_TERM_SORT
JOURNAL_ABBREVIATI
JOURNAL_NAME

Data items in the GEOM_ANGLE category record details about the molecular and crystal angles, as calculated from the contents of the ATOM, CELL, and SYMMETRY data.

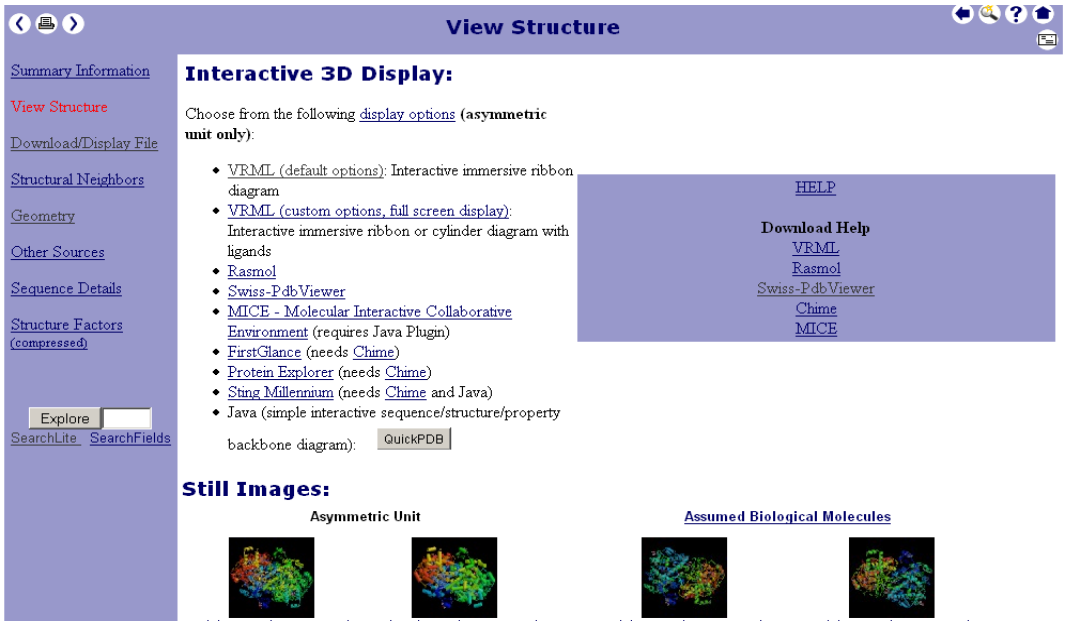
Table Name : geom_angle

Note

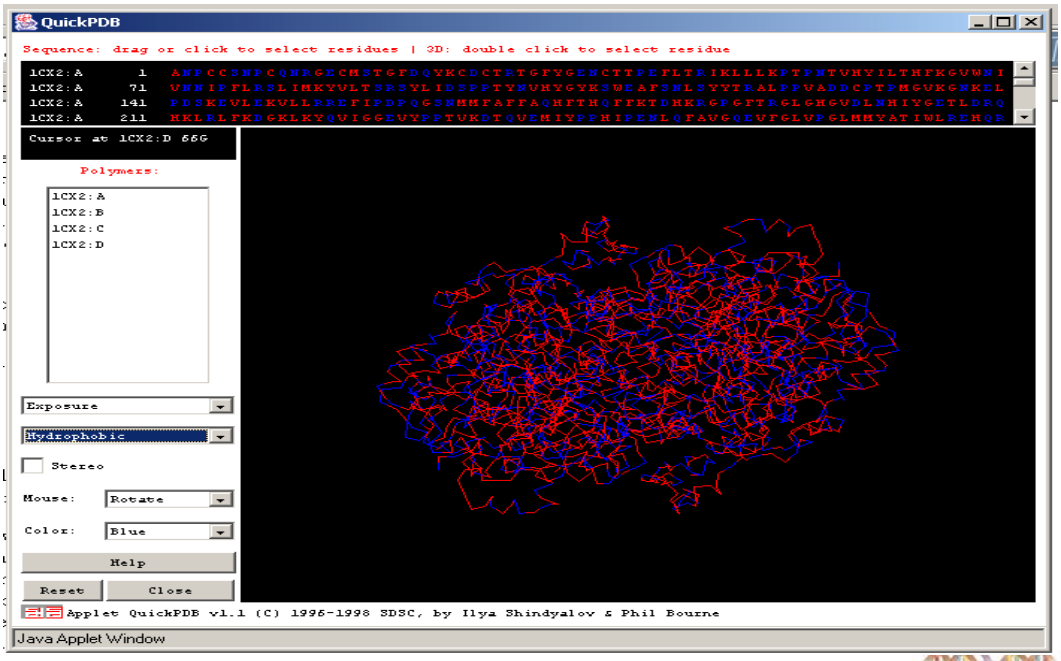
DDL Code

```
CREATE TABLE GEOM_ANGLE(
     PARTITION_RANGE              CHAR(1)              NOT NULL,
     GEOM_ANGLE_ID                NUMBER(38, 0)        NOT NULL,
     ATOM_SITE_AUTH_ASYM_ID_1     VARCHAR2(10),
     ATOM_SITE_AUTH_ASYM_ID_2     VARCHAR2(10),
     ATOM_SITE_AUTH_ASYM_ID_3     VARCHAR2(10),
     ATOM_SITE_AUTH_ATOM_ID_1     VARCHAR2(10),
     ATOM_SITE_AUTH_ATOM_ID_2     VARCHAR2(10),
     ATOM_SITE_AUTH_ATOM_ID_3     VARCHAR2(10),
     ATOM_SITE_AUTH_COMP_ID_1     VARCHAR2(10),
     ATOM_SITE_AUTH_COMP_ID_2     VARCHAR2(10),
     ATOM_SITE_AUTH_COMP_ID_3     VARCHAR2(10),
     ATOM_SITE_AUTH_SEQ_ID_1      VARCHAR2(10),
     ATOM_SITE_AUTH_SEQ_ID_2      VARCHAR2(10),
     ATOM_SITE_AUTH_SEQ_ID_3      VARCHAR2(10),
     ATOM_SITE_ID_1               VARCHAR2(10)         NOT NULL,
     ATOM_SITE_ID_2               VARCHAR2(10)         NOT NULL,
     ATOM_SITE_ID_3               VARCHAR2(10)         NOT NULL,
     ATOM_SITE_LABEL_ALT_ID_1     VARCHAR2(10),
     ATOM_SITE_LABEL_ALT_ID_2     VARCHAR2(10),
     ATOM_SITE_LABEL_ALT_ID_3     VARCHAR2(10),
     ATOM_SITE_LABEL_ASYM_ID_1    VARCHAR2(10),
     ATOM_SITE_LABEL_ASYM_ID_2    VARCHAR2(10),
     ATOM_SITE_LABEL_ASYM_ID_3    VARCHAR2(10),
     ATOM_SITE_LABEL_ATOM_ID_1    VARCHAR2(10),
```

PDB: Web Interface



PDB: Web Interface (2)



Weitere Protein-Datenbanken

pir.georgetown.edu

PIR Protein Information Resource

About PIR | Databases | Search and Retrieval | Download | Support

AN INTEGRATED PUBLIC RESOURCE OF PROTEIN INFORMATICS TO SUPPORT GENOMIC AND PROTEOMIC RESEARCH AND SCIENTIFIC DISCOVERY

PIR produces the **Protein Sequence Database (PSD)** of functionally annotated protein sequences, which grew out of the *Atlas of Protein Sequence and Structure* (1965-1978) edited by Margaret Dayhoff and has been incorporated into an integrated knowledge base system of value-added databases and analytical tools.

ProClass, a central point for exploration of protein information, provides summary descriptions of protein family, function and structure for PIR-PSD, Swiss-Prot, and TrEMBL sequences, with links to over 50 biological databases . Release 2.35, 24-Nov-2003, contains 1,169,177 entries.

PIR-NREF, a comprehensive database for sequence searching and protein identification, contains non-redundant protein sequences from PIR-PSD, Swiss-Prot, TrEMBL, RefSeq, GenPept, and PDB. Release 1.35, 24-Nov-2003, contains 1,397,398 entries.

PIR News Flash
Press Release: Protein Infromation Resource Adds New Tools to Databases

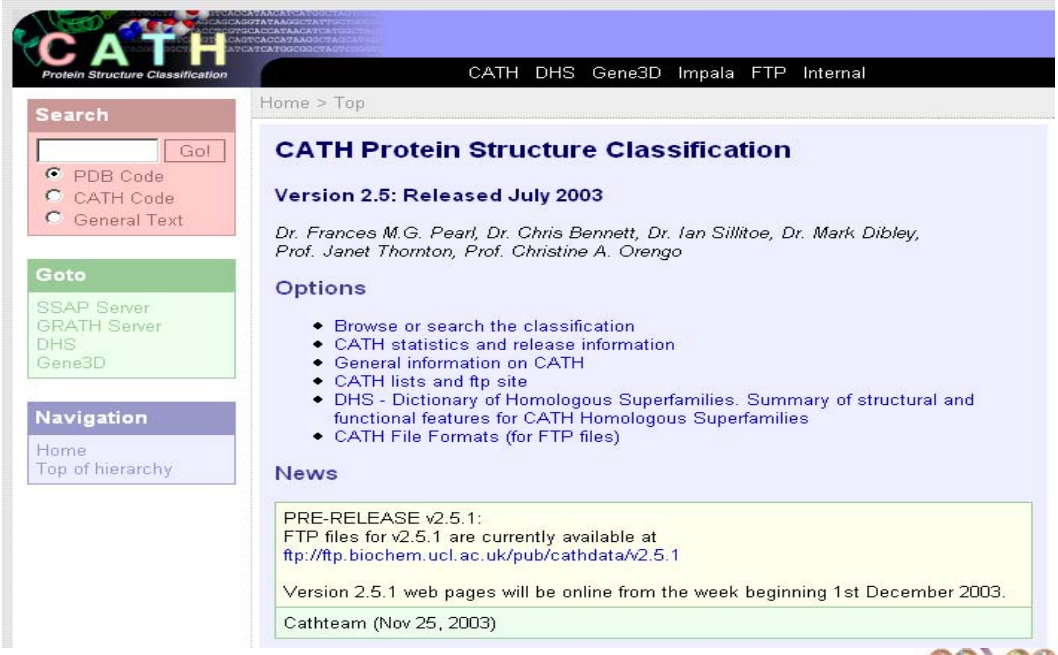
Text Search Protein Databases: GO!

Find an Exact Peptide Match: GO!

Type in a string of single letter amino acid code (at least 3 letters)

PIR has recently joined forces with EBI (European Bioinformatics Institute) and SIB (Swiss Institute of Bioinformatics) to establish the UniProt (United Protein Databases), the central resource of protein sequence and function.

Weitere Protein-Datenbanken (2)



Weitere Protein-Datenbanken (3)

- UniProt
  - Beinhaltet PIR, Swiss-Prot und TrEMBL
  - Ablösung einer langen Parallelentwicklung
  - Erster Release noch nicht verfügbar
- OWL
  - Nicht-redundante Sammlung von Proteinsequenzen
  - Enthält: Swiss-Prot, PIR, GenBank
- ... und viele mehr

Zusammenfassung

- Motivation und historische Entwicklung
- Proteomics
  - Datengewinnung
  - PEDRo-Projekt
- Protein-Datenbanken
  - Sequenz-Datenbanken (Swiss-Prot)
  - Domain/Familien-Datenbanken (InterPro)
  - Struktur-Datenbanken (PDB)
  - Weitere Protein-Datenbanken