Bioinformatics for Proteomics

Ann Loraine
aloraine@uab.edu

What is bioinformatics?

- The science of collecting, processing, organizing, storing, analyzing, and mining biological information, especially data from high-throughput biology, such as genomic sequencing or proteomics.
- Combines aspects of computer science, statistics, biology.
- Different aspects more important at different times, depending on the biological question you want to answer.

Bioinformatics is...

- The computational wing of molecular biology.
- Just another tool in your research repertoire.
- Remember: computers (and computer software programs) are designed by humans for humans. Think about how the tool is designed - be aware of the interface and how it affects what you do.

Useful Texts

- David Mount "Bioinformatics"
- Philip Bourne & Helge Wessig "Structural Bioinformatics"
- Ian Korf, et al. "BLAST"
- Carl Branden & John Tooze "Introduction to Protein Structure"

Bioinformatic data

- ...is information based on bioinformatic analysis of experimental results, such as large sequence databases.
- ...is based on many assumptions and “judgement calls” along the way.
- Should be used with care!

Questions you must answer...

when dealing with bioinformatic data

?

- What is the origin of this information?
  - experimental? computational?

?

- What evidence supports it?
- What are the uncertainties and underlying assumptions?

?

Scenario

Using 2D gel electrophoresis and mass spectrometry, you identify a protein that is differentially expressed in an experimental sample (human tumor) versus a control (normal tissue). MASCOT tells you that the best match is SwissProt accession P31947.

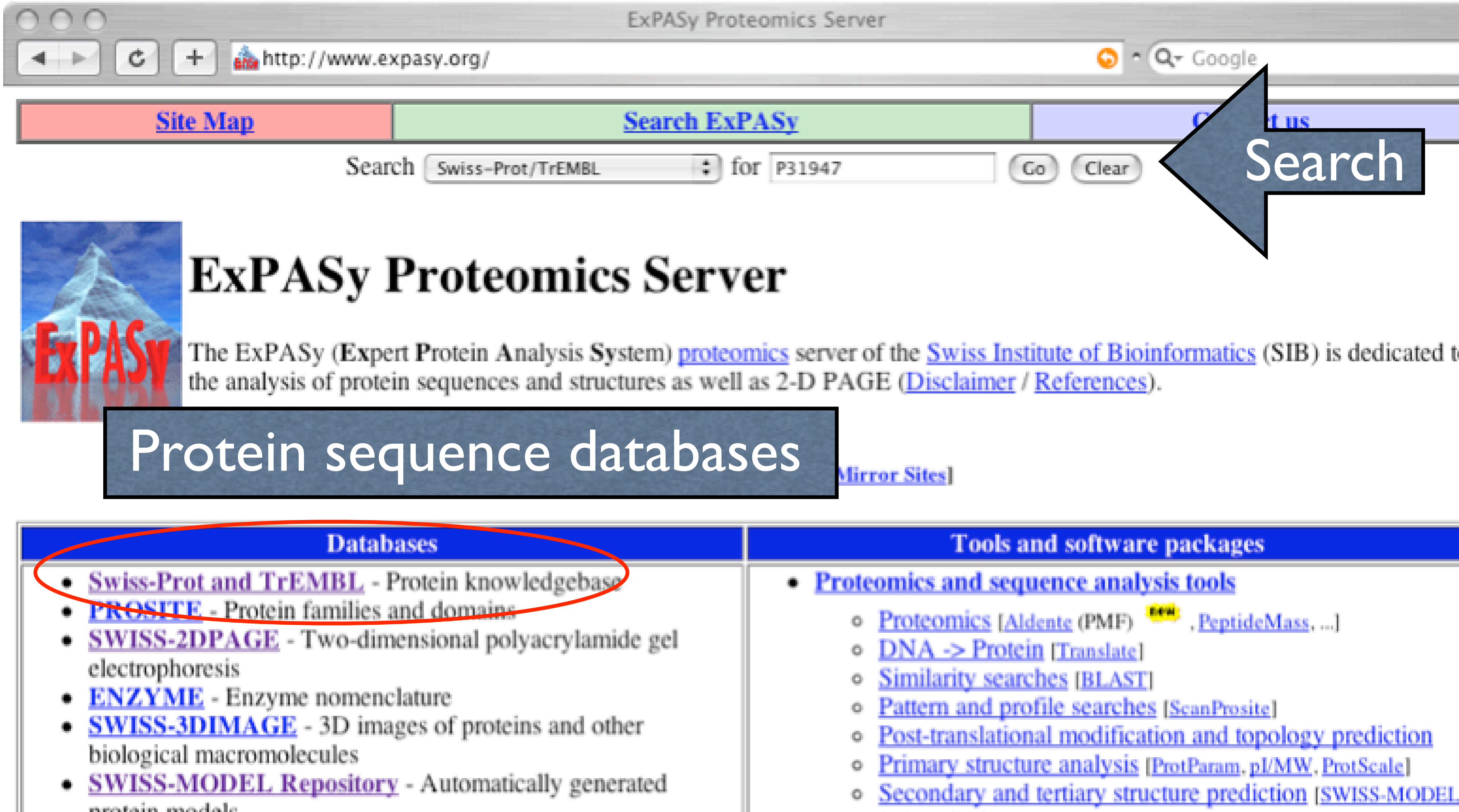
Question: What is it? What is its biological role?

Accession: an id (like a social security number) for an individual record in a sequence (or other type of) database. Each sequence (protein, mRNA, DNA) in a database has a unique "accession."

Questions

- What is it’s biochemical function?
- Where is it localized in the cell?
- What other proteins or pathways does it interact with?
- Others?

Expert Protein Analysis System



Site Map | Search ExPASy | C t us

Search Swiss-Prot/TrEMBL for P31947 Go Clear

Search

ExPASy Proteomics Server

The ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated t the analysis of protein sequences and structures as well as 2-D PAGE (Disclaimer / References).

Protein sequence databases

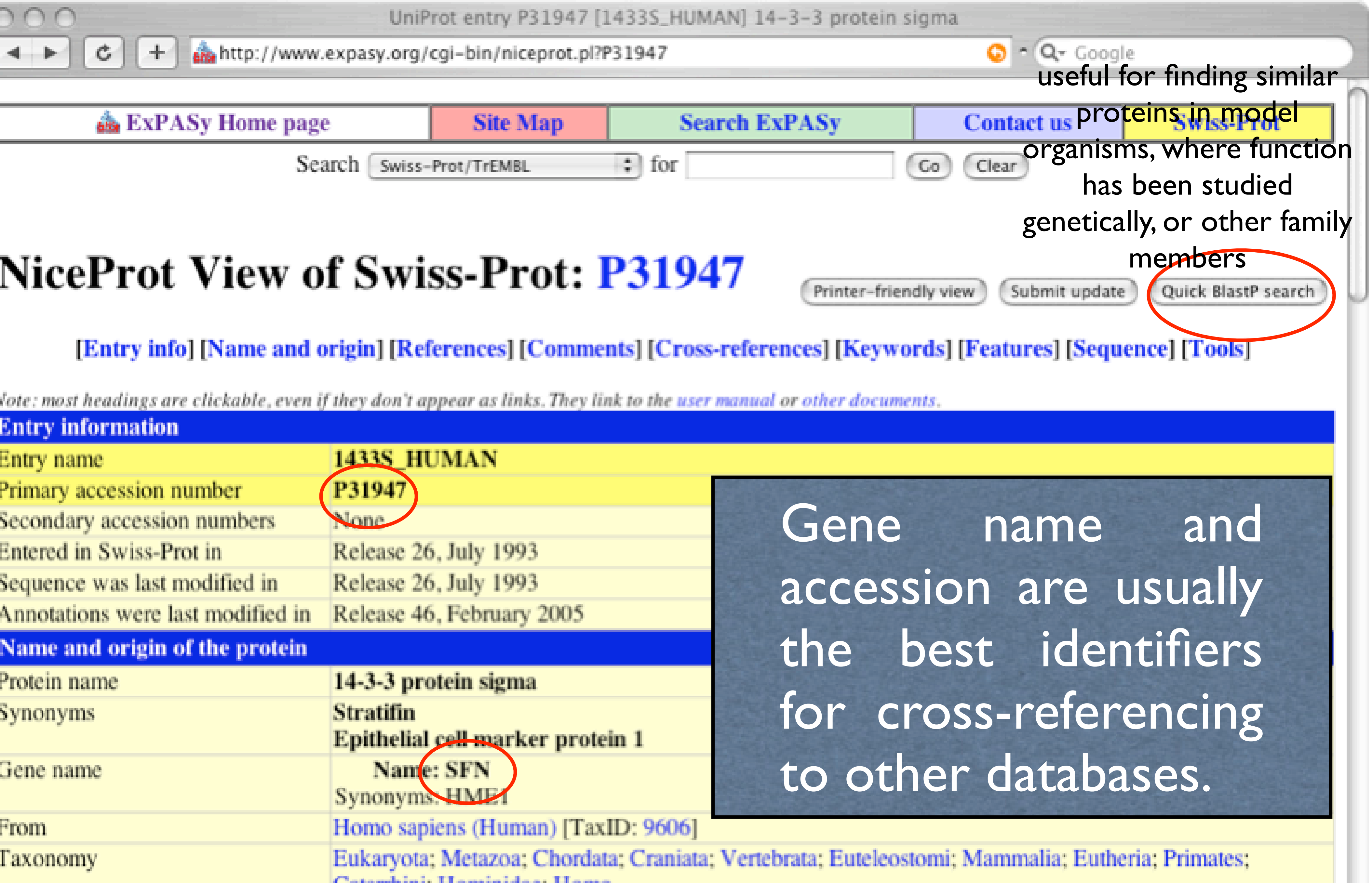
Mirror Sites]

Databases	Tools and software packages
- Swiss-Prot and TrEMBL - Protein knowledgebase - PROSITE - Protein families and domains - SWISS-2DPAGE - Two-dimensional polyacrylamide gel electrophoresis - ENZYME - Enzyme nomenclature - SWISS-3DIMAGE - 3D images of proteins and other biological macromolecules - SWISS-MODEL Repository - Automatically generated protein models	- Proteomics and sequence analysis tools ○ Proteomics [Aldente (PMF) new , PeptideMass, ...] ○ DNA -> Protein [Translate] ○ Similarity searches [BLAST] ○ Pattern and profile searches [ScanProsite] ○ Post-translational modification and topology prediction ○ Primary structure analysis [ProtParam, pI/MW, ProtScale] ○ Secondary and tertiary structure prediction [SWISS-MODEL

SwissProt/trEMBL

- SwissProt, manually-curated protein sequence database; records come from the conceptual translations of full-length cDNAs, usually submitted by individual labs.
- records (one per protein) contain core data (sequence & references) and annotations (bioinformatic analysis results)
- trEMBL: translated DNA sequence records from EMBL (European Molecular Biology Laboratory) and GenBank (US)

SwissProt record



UniProt entry P31947 [1433S_HUMAN] 14-3-3 protein sigma
http://www.expasy.org/cgi-bin/niceprot.pl?P31947
Google
ExPASy Home page
Site Map
Search ExPASy
Contact us
Swiss-Prot
Search
Swiss-Prot/TrEMBL
for
Go
Clear

NiceProt View of Swiss-Prot: P31947

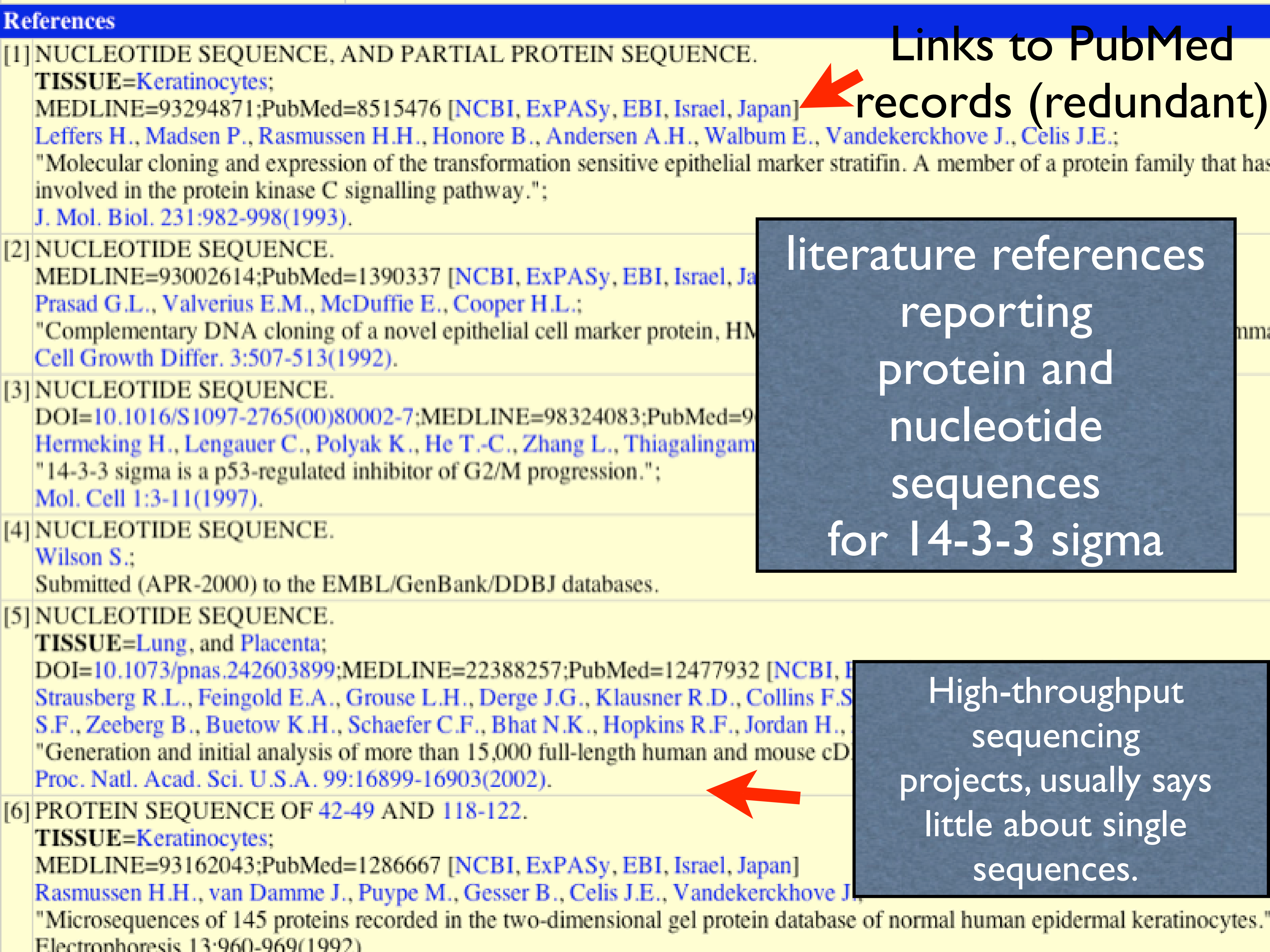
Printer-friendly view
Submit update
Quick BlastP search
[Entry info] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]

Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.

Entry information	
Entry name	1433S_HUMAN
Primary accession number	P31947
Secondary accession numbers	None
Entered in Swiss-Prot in	Release 26, July 1993
Sequence was last modified in	Release 26, July 1993
Annotations were last modified in	Release 46, February 2005
Name and origin of the protein	
Protein name	14-3-3 protein sigma
Synonyms	Stratifin Epithelial cell marker protein 1
Gene name	Name: SFN Synonyms: HME1
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo

useful for finding similar proteins in model organisms, where function has been studied genetically, or other family members

Gene name and accession are usually the best identifiers for cross-referencing to other databases.



References

[1] NUCLEOTIDE SEQUENCE, AND PARTIAL PROTEIN SEQUENCE.
TISSUE=Keratinocytes;
MEDLINE=93294871;PubMed=8515476 [NCBI, ExPASy, EBI, Israel, Japan]
Leffers H., Madsen P., Rasmussen H.H., Honore B., Andersen A.H., Walbum E., Vandekerckhove J., Celis J.E.;
"Molecular cloning and expression of the transformation sensitive epithelial marker stratifin. A member of a protein family that has
involved in the protein kinase C signalling pathway.";
J. Mol. Biol. 231:982-998(1993).

[2] NUCLEOTIDE SEQUENCE.
MEDLINE=93002614;PubMed=1390337 [NCBI, ExPASy, EBI, Israel, Ja
Prasad G.L., Valverius E.M., McDuffie E., Cooper H.L.;
"Complementary DNA cloning of a novel epithelial cell marker protein, HM mma
Cell Growth Differ. 3:507-513(1992).

[3] NUCLEOTIDE SEQUENCE.
DOI=10.1016/S1097-2765(00)80002-7;MEDLINE=98324083;PubMed=9
Hermeking H., Lengauer C., Polyak K., He T.-C., Zhang L., Thiagalingam
"14-3-3 sigma is a p53-regulated inhibitor of G2/M progression.";
Mol. Cell 1:3-11(1997).

[4] NUCLEOTIDE SEQUENCE.
Wilson S.;
Submitted (APR-2000) to the EMBL/GenBank/DDBJ databases.

[5] NUCLEOTIDE SEQUENCE.
TISSUE=Lung, and Placenta;
DOI=10.1073/pnas.242603899;MEDLINE=22388257;PubMed=12477932 [NCBI, E
Strausberg R.L., Feingold E.A., Grouse L.H., Derge J.G., Klausner R.D., Collins F.S
S.F., Zeeberg B., Buetow K.H., Schaefer C.F., Bhat N.K., Hopkins R.F., Jordan H.,
"Generation and initial analysis of more than 15,000 full-length human and mouse cD
Proc. Natl. Acad. Sci. U.S.A. 99:16899-16903(2002).

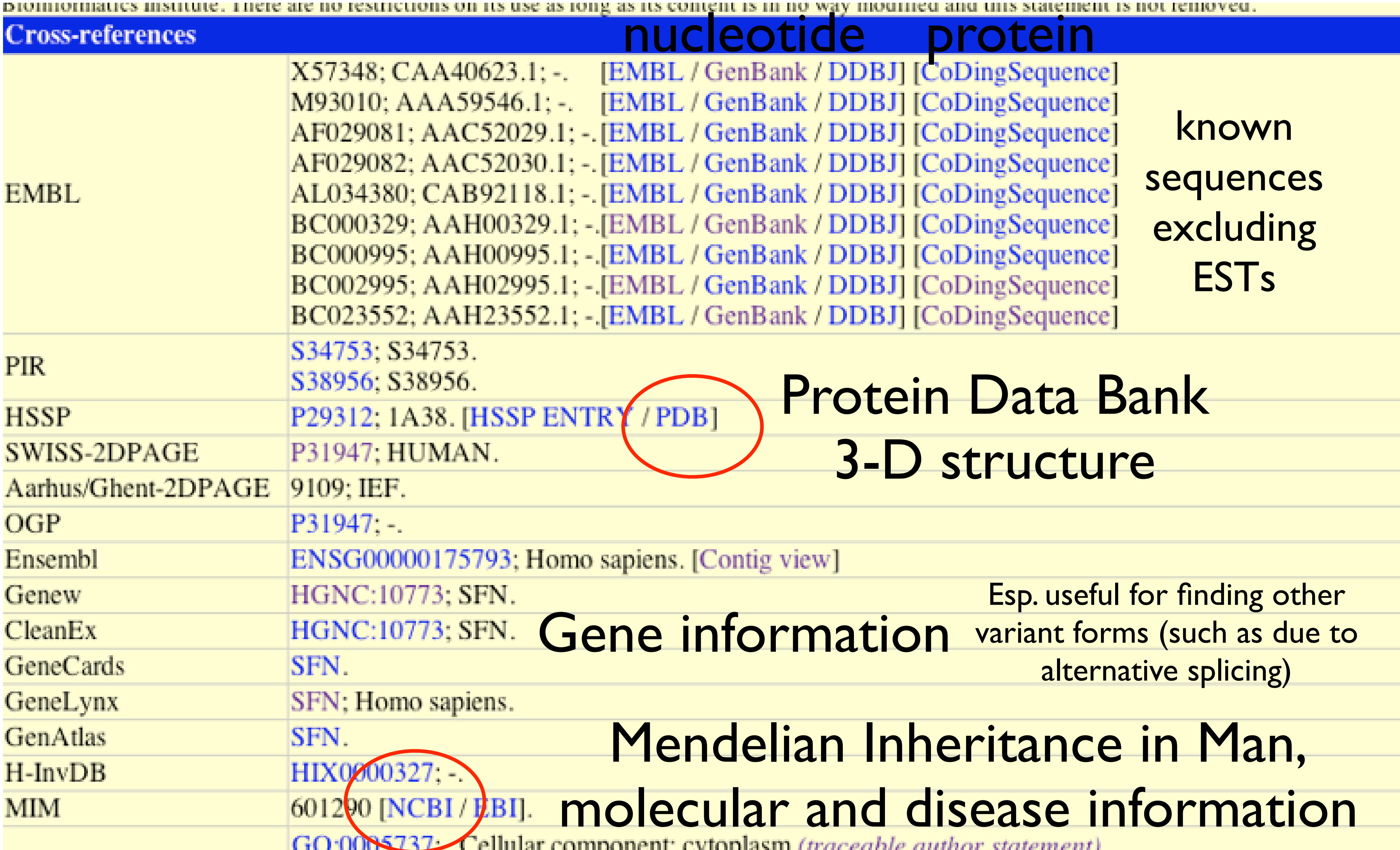
[6] PROTEIN SEQUENCE OF 42-49 AND 118-122.
TISSUE=Keratinocytes;
MEDLINE=93162043;PubMed=1286667 [NCBI, ExPASy, EBI, Israel, Japan]
Rasmussen H.H., van Damme J., Puype M., Gesser B., Celis J.E., Vandekerckhove J.,
"Microsequences of 145 proteins recorded in the two-dimensional gel protein database of normal human epidermal keratinocytes."
Electrophoresis 13:960-969(1992)

Links to PubMed records (redundant)

literature references reporting protein and nucleotide sequences for 14-3-3 sigma

High-throughput sequencing projects, usually says little about single sequences.

Functional Annotations



Bioinformatics Institute. There are no restrictions on its use as long as its content is in no way modified and this statement is not removed.

Cross-references	
EMBL	X57348; CAA40623.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] M93010; AAA59546.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] AF029081; AAC52029.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] AF029082; AAC52030.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] AL034380; CAB92118.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] BC000329; AAH00329.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] BC000995; AAH00995.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] BC002995; AAH02995.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] BC023552; AAH23552.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
PIR	S34753; S34753. S38956; S38956.
HSSP	P29312; 1A38. [HSSP ENTRY / PDB]
SWISS-2DPAGE	P31947; HUMAN.
Aarhus/Ghent-2DPAGE	9109; IEF.
OGP	P31947; -.
Ensembl	ENSG00000175793; Homo sapiens. [Contig view]
Genew	HGNC:10773; SFN.
CleanEx	HGNC:10773; SFN.
GeneCards	SFN.
GeneLynx	SFN; Homo sapiens.
GenAtlas	SFN.
H-InvDB	HIX0000327; -.
MIM	601290 [NCBI / EBI].
	GO:0005737; Cellular component: cytoplasm *(traceable author statement)*

nucleotide

protein

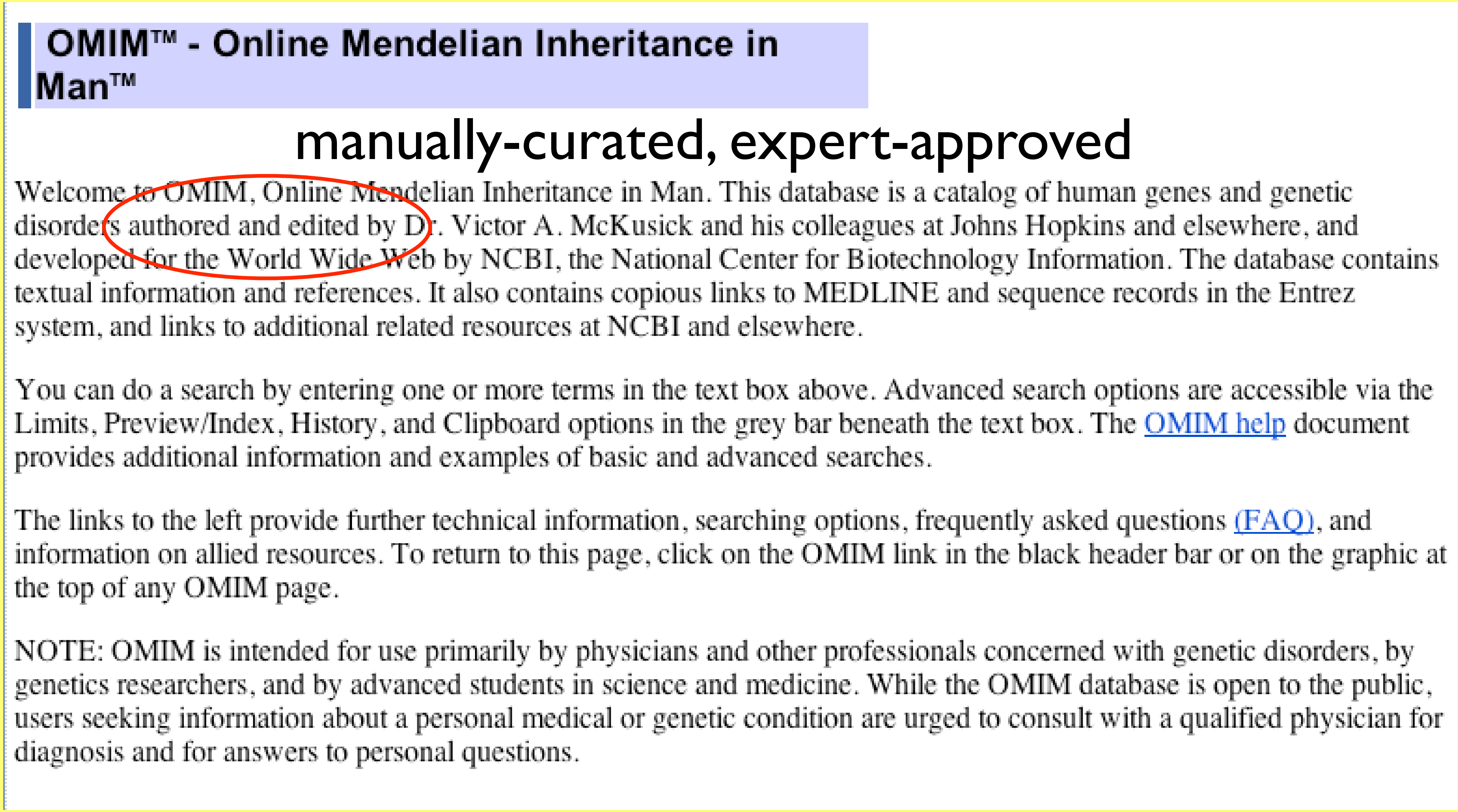
known sequences excluding ESTs

Protein Data Bank
3-D structure

Gene information

Esp. useful for finding other variant forms (such as due to alternative splicing)

Mendelian Inheritance in Man,
molecular and disease information



OMIM™ - Online Mendelian Inheritance in Man™

manually-curated, expert-approved

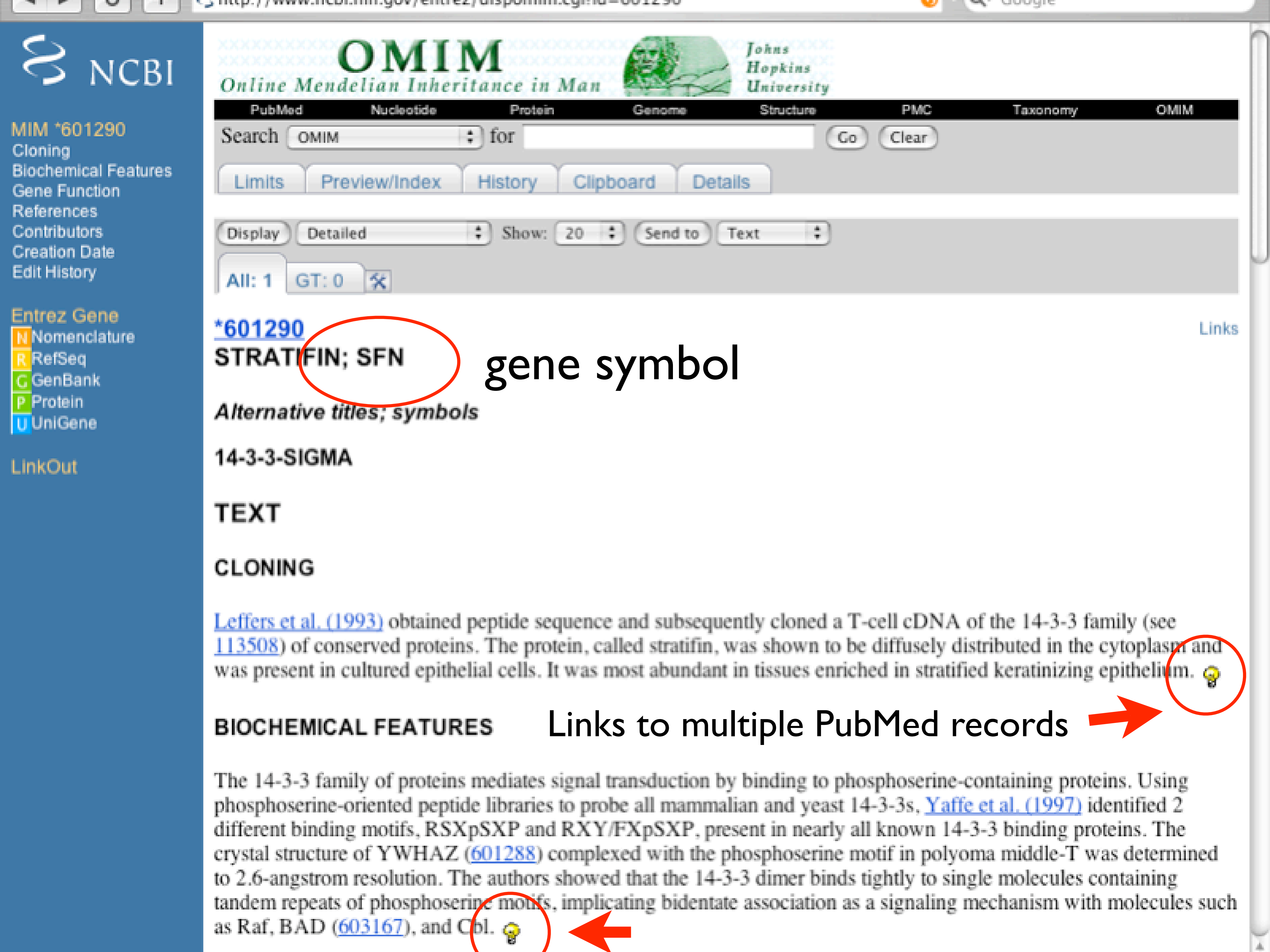
Welcome to OMIM, Online Mendelian Inheritance in Man. This database is a catalog of human genes and genetic disorders authored and edited by Dr. Victor A. McKusick and his colleagues at Johns Hopkins and elsewhere, and developed for the World Wide Web by NCBI, the National Center for Biotechnology Information. The database contains textual information and references. It also contains copious links to MEDLINE and sequence records in the Entrez system, and links to additional related resources at NCBI and elsewhere.

You can do a search by entering one or more terms in the text box above. Advanced search options are accessible via the Limits, Preview/Index, History, and Clipboard options in the grey bar beneath the text box. The OMIM help document provides additional information and examples of basic and advanced searches.

The links to the left provide further technical information, searching options, frequently asked questions (FAQ), and information on allied resources. To return to this page, click on the OMIM link in the black header bar or on the graphic at the top of any OMIM page.

NOTE: OMIM is intended for use primarily by physicians and other professionals concerned with genetic disorders, by genetics researchers, and by advanced students in science and medicine. While the OMIM database is open to the public, users seeking information about a personal medical or genetic condition are urged to consult with a qualified physician for diagnosis and for answers to personal questions.

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM



*601290

STRATIFIN; SFN

gene symbol

Alternative titles; symbols

14-3-3-SIGMA

TEXT

CLONING

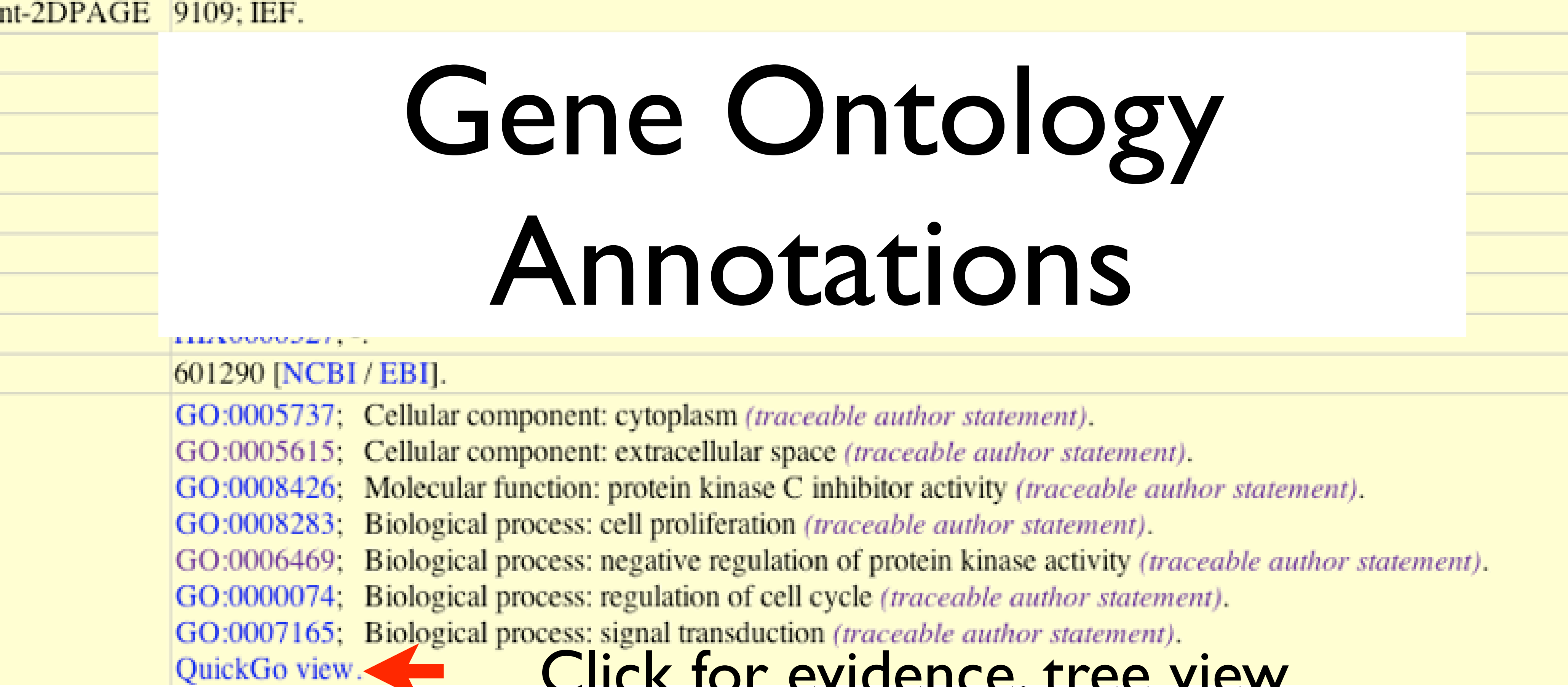
Leffers et al. (1993) obtained peptide sequence and subsequently cloned a T-cell cDNA of the 14-3-3 family (see 113508) of conserved proteins. The protein, called stratifin, was shown to be diffusely distributed in the cytoplasm and was present in cultured epithelial cells. It was most abundant in tissues enriched in stratified keratinizing epithelium.

BIOCHEMICAL FEATURES

Links to multiple PubMed records

The 14-3-3 family of proteins mediates signal transduction by binding to phosphoserine-containing proteins. Using phosphoserine-oriented peptide libraries to probe all mammalian and yeast 14-3-3s, Yaffe et al. (1997) identified 2 different binding motifs, RSXpSXP and RXY/FXpSXP, present in nearly all known 14-3-3 binding proteins. The crystal structure of YWHAZ (601288) complexed with the phosphoserine motif in polyoma middle-T was determined to 2.6-angstrom resolution. The authors showed that the 14-3-3 dimer binds tightly to single molecules containing tandem repeats of phosphoserine motifs, implicating bidentate association as a signaling mechanism with molecules such as Raf, BAD (603167), and Cbl.

Gene Ontology Annotations



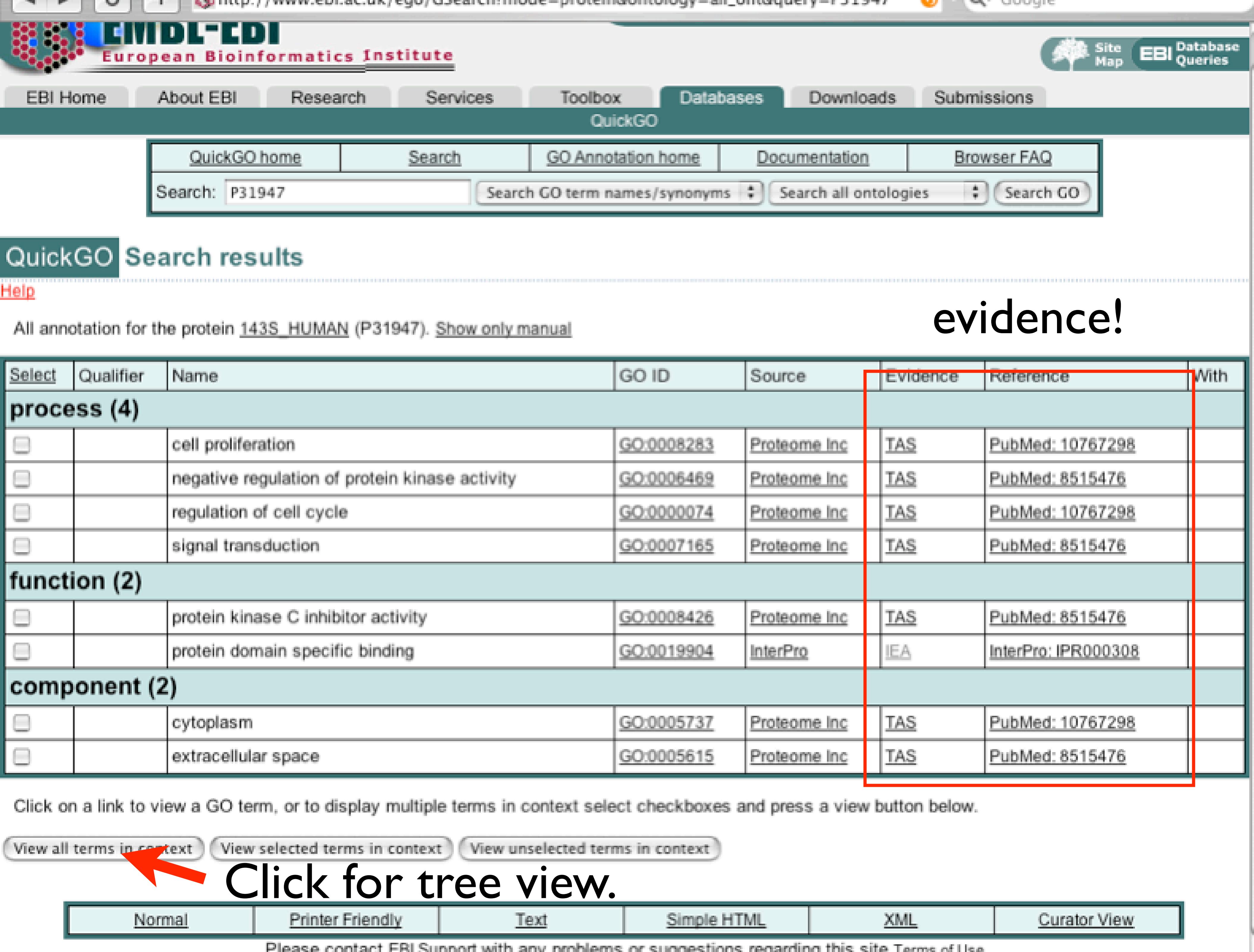
nt-2DPAGE 9109; IEF.

601290 [NCBI / EBI].

GO:0005737; Cellular component: cytoplasm *(traceable author statement)*.
GO:0005615; Cellular component: extracellular space *(traceable author statement)*.
GO:0008426; Molecular function: protein kinase C inhibitor activity *(traceable author statement)*.
GO:0008283; Biological process: cell proliferation *(traceable author statement)*.
GO:0006469; Biological process: negative regulation of protein kinase activity *(traceable author statement)*.
GO:0000074; Biological process: regulation of cell cycle *(traceable author statement)*.
GO:0007165; Biological process: signal transduction *(traceable author statement)*.
QuickGo view.

Click for evidence, tree view

Gene Ontology - a structured, controlled vocabulary describing gene products. There are main branches: biological process, molecular function, and cellular component. The GOA project is annotating human proteins with GO terms.



QuickGO home	Search	GO Annotation home	Documentation	Browser FAQ

Search: P31947 | Search GO term names/synonyms | Search all ontologies | Search GO

QuickGO Search results

Help

All annotation for the protein 143S_HUMAN (P31947). Show only manual

evidence!

Select	Qualifier	Name	GO ID	Source	Evidence	Reference	With
process (4)							
□		cell proliferation	GO:0008283	Proteome Inc	TAS	PubMed: 10767298	
□		negative regulation of protein kinase activity	GO:0006469	Proteome Inc	TAS	PubMed: 8515476	
□		regulation of cell cycle	GO:0000074	Proteome Inc	TAS	PubMed: 10767298	
□		signal transduction	GO:0007165	Proteome Inc	TAS	PubMed: 8515476	
function (2)							
□		protein kinase C inhibitor activity	GO:0008426	Proteome Inc	TAS	PubMed: 8515476	
□		protein domain specific binding	GO:0019904	InterPro	IEA	InterPro: IPR000308	
component (2)							
□		cytoplasm	GO:0005737	Proteome Inc	TAS	PubMed: 10767298	
□		extracellular space	GO:0005615	Proteome Inc	TAS	PubMed: 8515476	

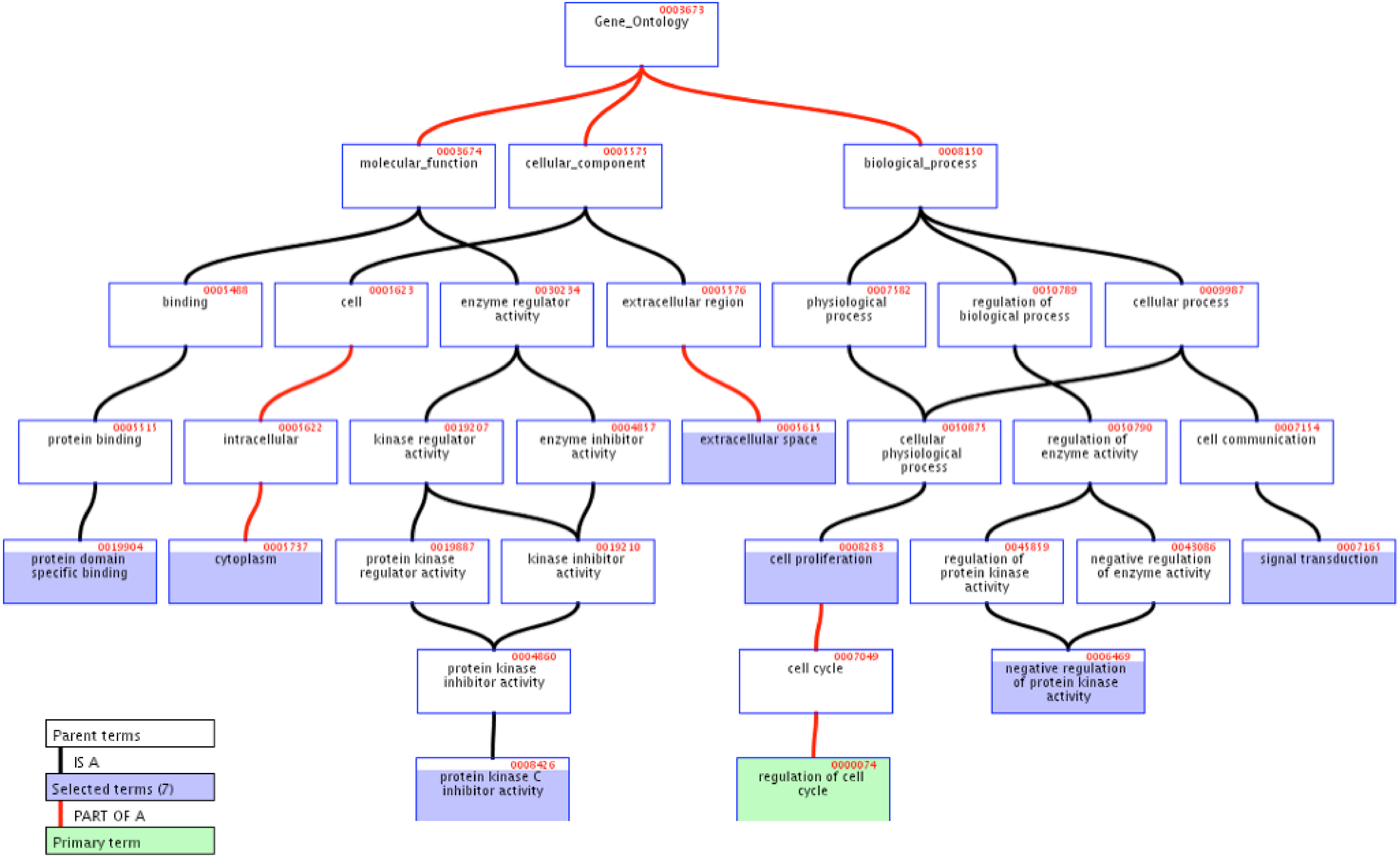
Click on a link to view a GO term, or to display multiple terms in context select checkboxes and press a view button below.

View all terms in context
View selected terms in context
View unselected terms in context

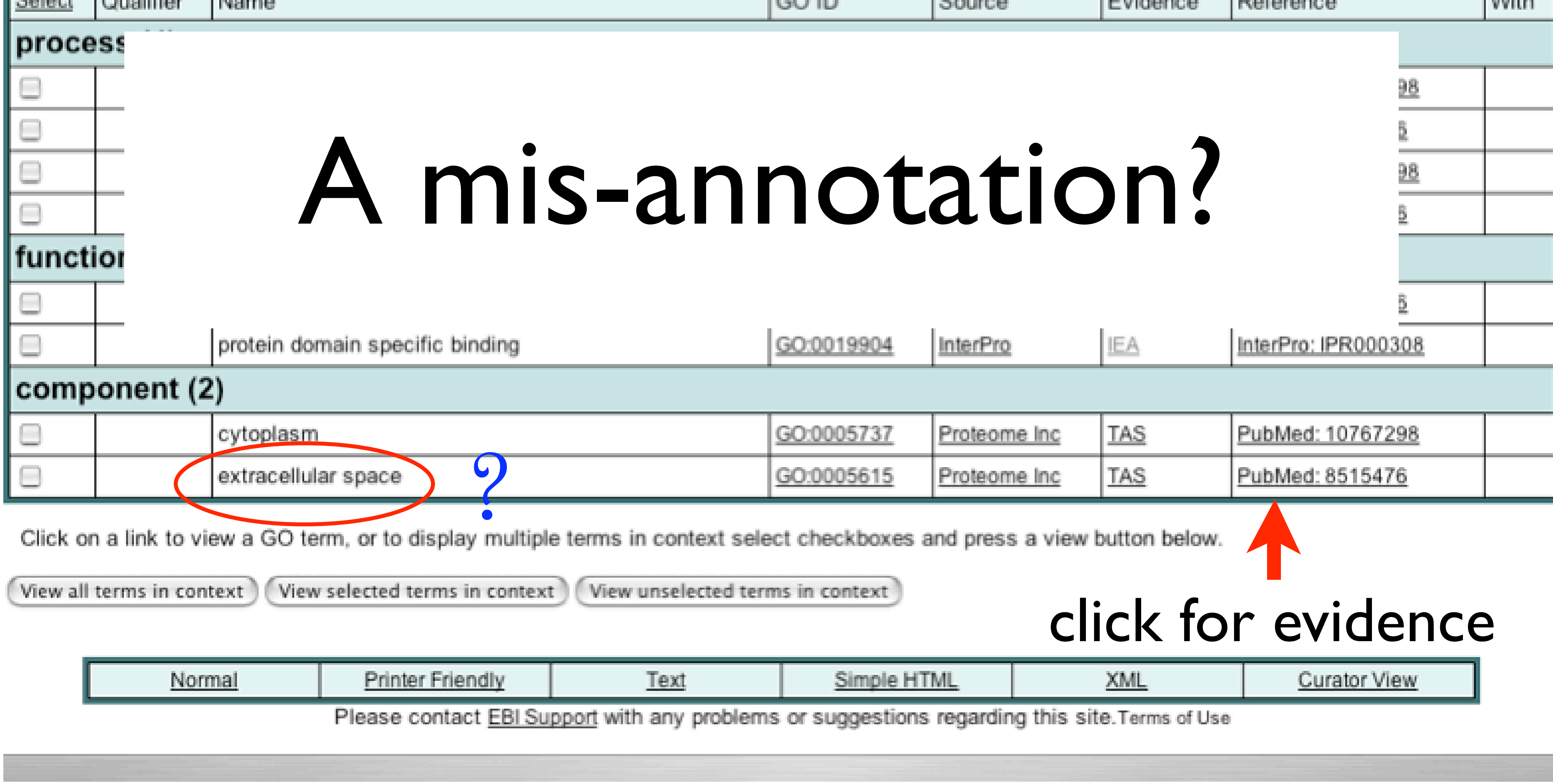
Click for tree view.

Please contact EBI Support with any problems or suggestions regarding this site Terms of Use

GO tree view


0003673
Gene_Ontology
0003674
molecular_function
0005575
cellular_component
0008150
biological_process
0005488
binding
0005623
cell
0030234
enzyme regulator activity
0005576
extracellular region
0007582
physiological process
0050789
regulation of biological process
0009987
cellular process
0005515
protein binding
0005622
intracellular
0019207
kinase regulator activity
0004857
enzyme inhibitor activity
0005615
extracellular space
0050875
cellular physiological process
0050790
regulation of enzyme activity
0007154
cell communication
0019904
protein domain specific binding
0005737
cytoplasm
0019887
protein kinase regulator activity
0019210
kinase inhibitor activity
0008283
cell proliferation
0045859
regulation of protein kinase activity
0043086
negative regulation of enzyme activity
0007165
signal transduction
0004860
protein kinase inhibitor activity
0007049
cell cycle
0006469
negative regulation of protein kinase activity
0008426
protein kinase C inhibitor activity
0000074
regulation of cell cycle
Parent terms
IS A
Selected terms (7)
PART OF A
Primary term

A mis-annotation?



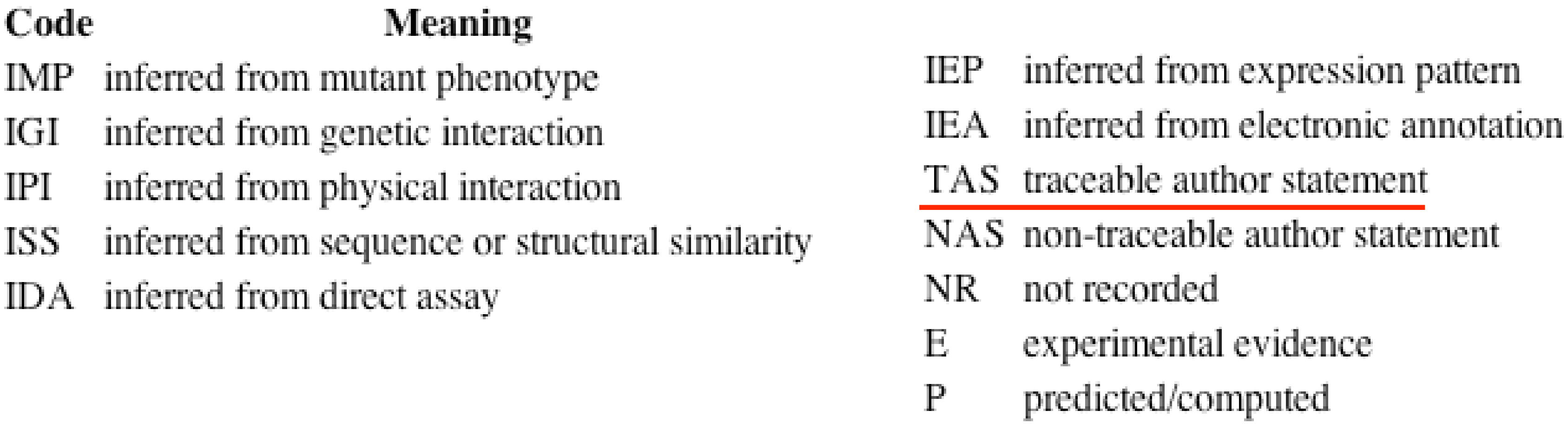
Select	Qualifier	Name	GO ID	Source	Evidence	Reference	With
process							
☐						98	
☐						5	
☐						98	
☐						5	
functio							
☐						5	
☐		protein domain specific binding	GO:0019904	InterPro	IEA	InterPro: IPR000308	
component (2)							
☐		cytoplasm	GO:0005737	Proteome Inc	TAS	PubMed: 10767298	
☐		extracellular space ?	GO:0005615	Proteome Inc	TAS	PubMed: 8515476	

Click on a link to view a GO term, or to display multiple terms in context select checkboxes and press a view button below.

View all terms in context
View selected terms in context
View unselected terms in context

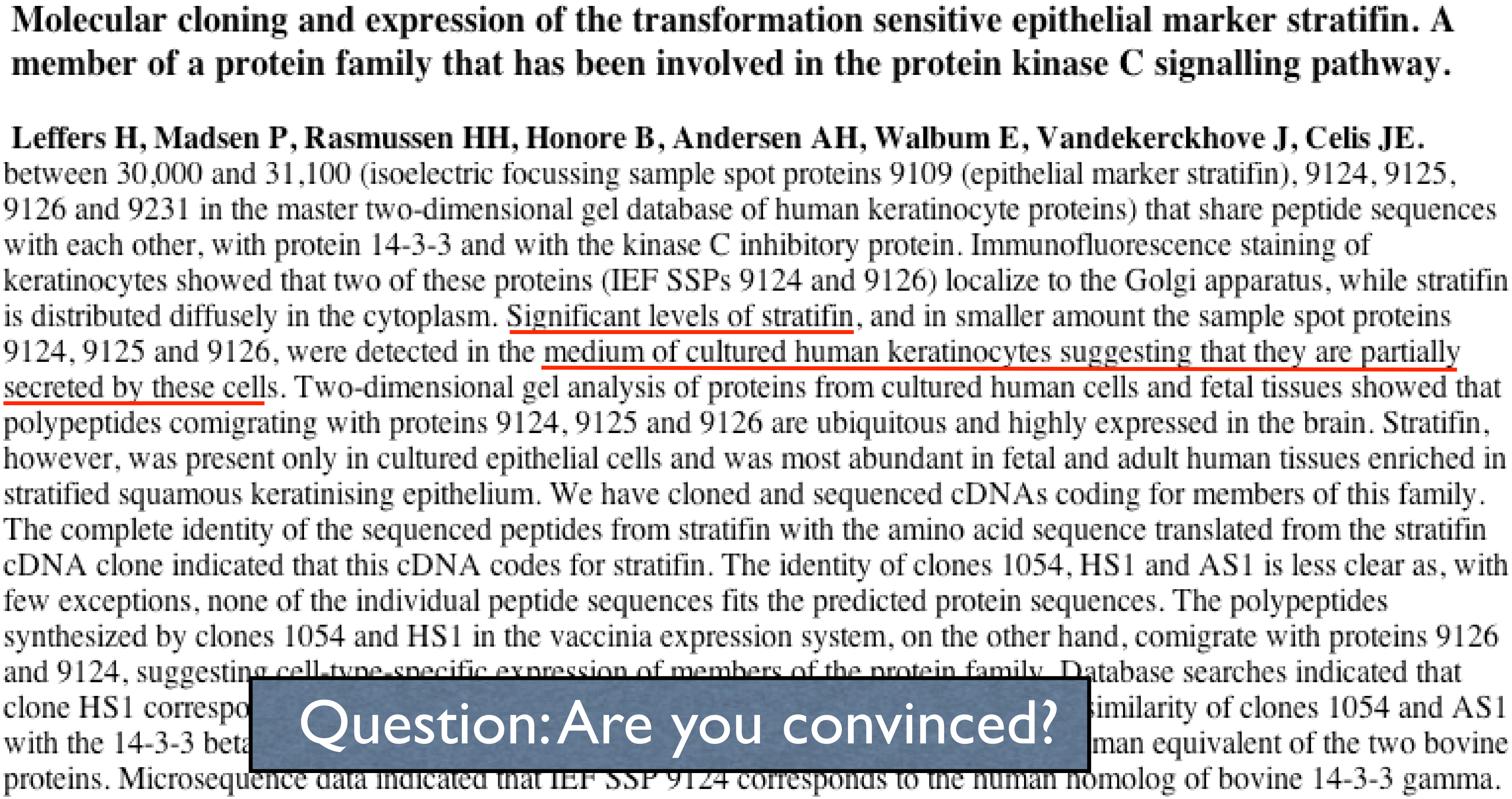
click for evidence

Please contact EBI Support with any problems or suggestions regarding this site. Terms of Use



Code	Meaning
IMP	inferred from mutant phenotype
IGI	inferred from genetic interaction
IPI	inferred from physical interaction
ISS	inferred from sequence or structural similarity
IDA	inferred from direct assay
IEP	inferred from expression pattern
IEA	inferred from electronic annotation
TAS	traceable author statement
NAS	non-traceable author statement
NR	not recorded
E	experimental evidence
P	predicted/computed

Evidence for extracellular stratifin?



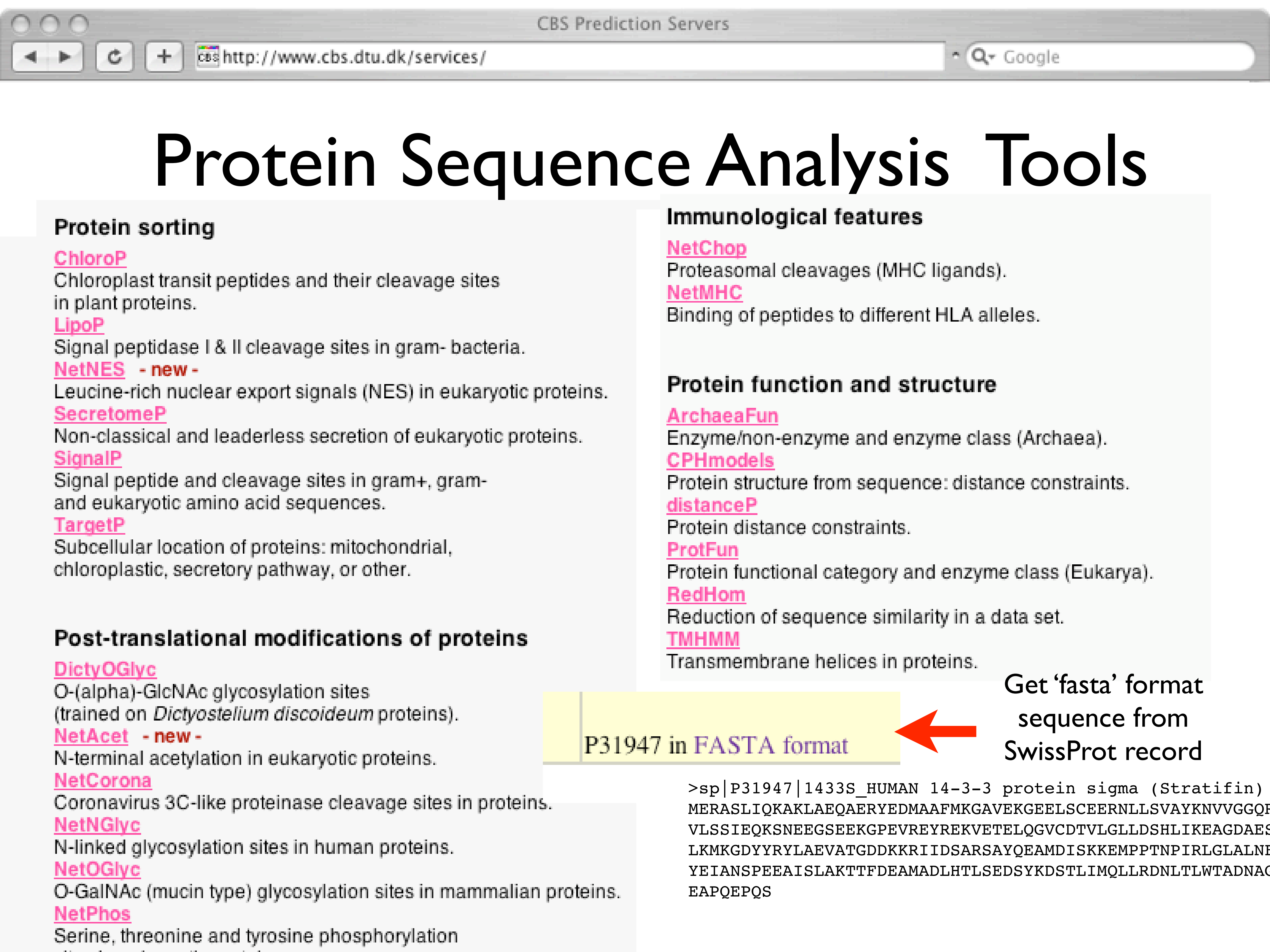
Molecular cloning and expression of the transformation sensitive epithelial marker stratifin. A member of a protein family that has been involved in the protein kinase C signalling pathway.

Leffers H, Madsen P, Rasmussen HH, Honore B, Andersen AH, Walbum E, Vandekerckhove J, Celis JE.

between 30,000 and 31,100 (isoelectric focussing sample spot proteins 9109 (epithelial marker stratifin), 9124, 9125, 9126 and 9231 in the master two-dimensional gel database of human keratinocyte proteins) that share peptide sequences with each other, with protein 14-3-3 and with the kinase C inhibitory protein. Immunofluorescence staining of keratinocytes showed that two of these proteins (IEF SSPs 9124 and 9126) localize to the Golgi apparatus, while stratifin is distributed diffusely in the cytoplasm. Significant levels of stratifin, and in smaller amount the sample spot proteins 9124, 9125 and 9126, were detected in the medium of cultured human keratinocytes suggesting that they are partially secreted by these cells. Two-dimensional gel analysis of proteins from cultured human cells and fetal tissues showed that polypeptides comigrating with proteins 9124, 9125 and 9126 are ubiquitous and highly expressed in the brain. Stratifin, however, was present only in cultured epithelial cells and was most abundant in fetal and adult human tissues enriched in stratified squamous keratinising epithelium. We have cloned and sequenced cDNAs coding for members of this family. The complete identity of the sequenced peptides from stratifin with the amino acid sequence translated from the stratifin cDNA clone indicated that this cDNA codes for stratifin. The identity of clones 1054, HS1 and AS1 is less clear as, with few exceptions, none of the individual peptide sequences fits the predicted protein sequences. The polypeptides synthesized by clones 1054 and HS1 in the vaccinia expression system, on the other hand, comigrate with proteins 9126 and 9124, suggesting cell-type-specific expression of members of the protein family. Database searches indicated that clone HS1 correspo imilarity of clones 1054 and AS1 with the 14-3-3 beta man equivalent of the two bovine proteins. Microsequence data indicated that IEF SSP 9124 corresponds to the human homolog of bovine 14-3-3 gamma.

Question: Are you convinced?

Protein Sequence Analysis Tools



Protein sorting

ChloroP
Chloroplast transit peptides and their cleavage sites in plant proteins.
LipoP
Signal peptidase I & II cleavage sites in gram- bacteria.
NetNES - new -
Leucine-rich nuclear export signals (NES) in eukaryotic proteins.
SecretomeP
Non-classical and leaderless secretion of eukaryotic proteins.
SignalP
Signal peptide and cleavage sites in gram+, gram- and eukaryotic amino acid sequences.
TargetP
Subcellular location of proteins: mitochondrial, chloroplastic, secretory pathway, or other.

Post-translational modifications of proteins

DictyOGlyc
O-(alpha)-GlcNAc glycosylation sites (trained on *Dictyostelium discoideum* proteins).
NetAcet - new -
N-terminal acetylation in eukaryotic proteins.
NetCorona
Coronavirus 3C-like proteinase cleavage sites in proteins.
NetNGlyc
N-linked glycosylation sites in human proteins.
NetOGlyc
O-GalNAc (mucin type) glycosylation sites in mammalian proteins.
NetPhos
Serine, threonine and tyrosine phosphorylation

Immunological features

NetChop
Proteasomal cleavages (MHC ligands).
NetMHC
Binding of peptides to different HLA alleles.

Protein function and structure

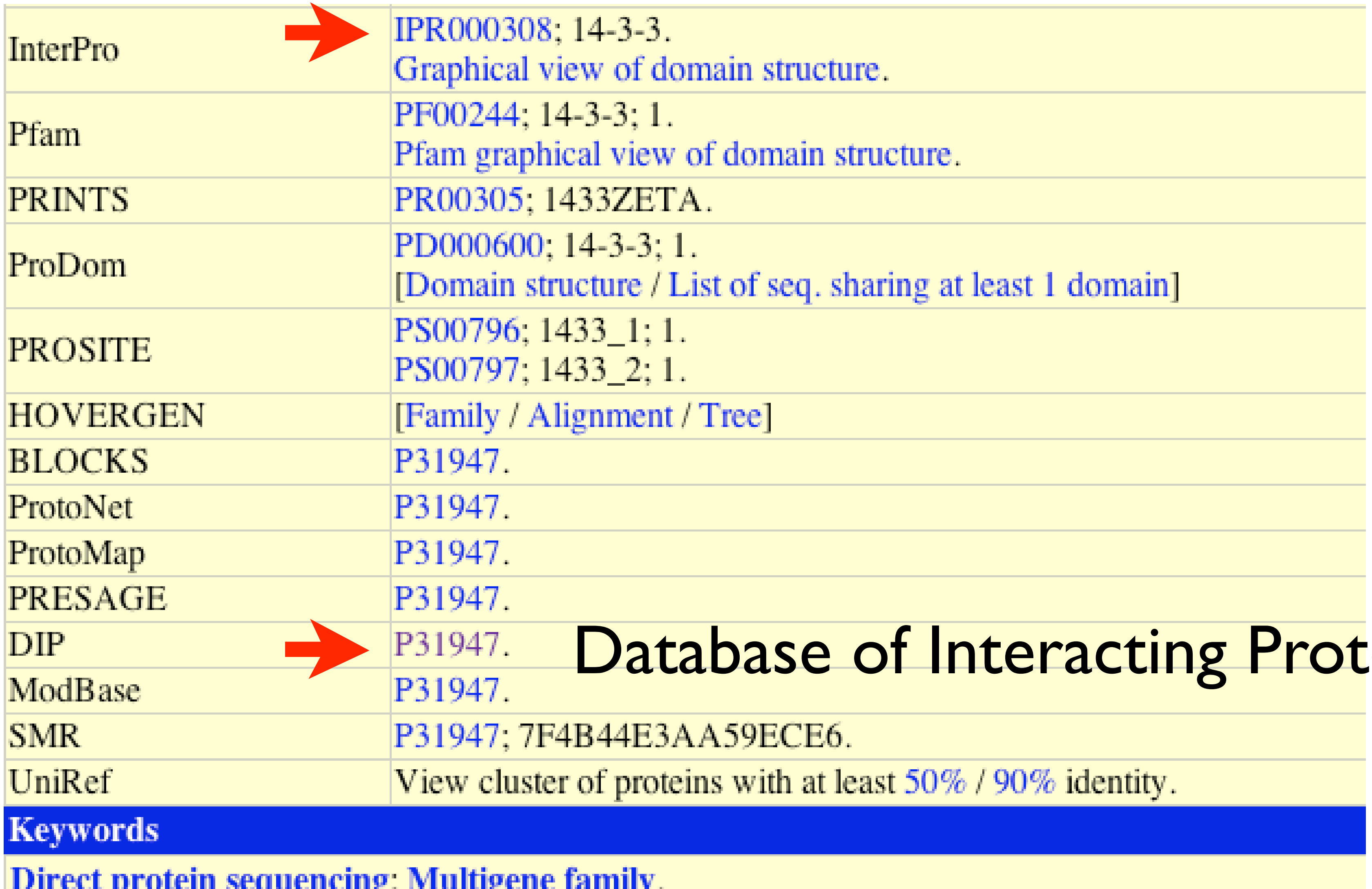
ArchaeaFun
Enzyme/non-enzyme and enzyme class (Archaea).
CPHmodels
Protein structure from sequence: distance constraints.
distanceP
Protein distance constraints.
ProtFun
Protein functional category and enzyme class (Eukarya).
RedHom
Reduction of sequence similarity in a data set.
TMHMM
Transmembrane helices in proteins.

P31947 in FASTA format

Get ‘fasta’ format sequence from SwissProt record

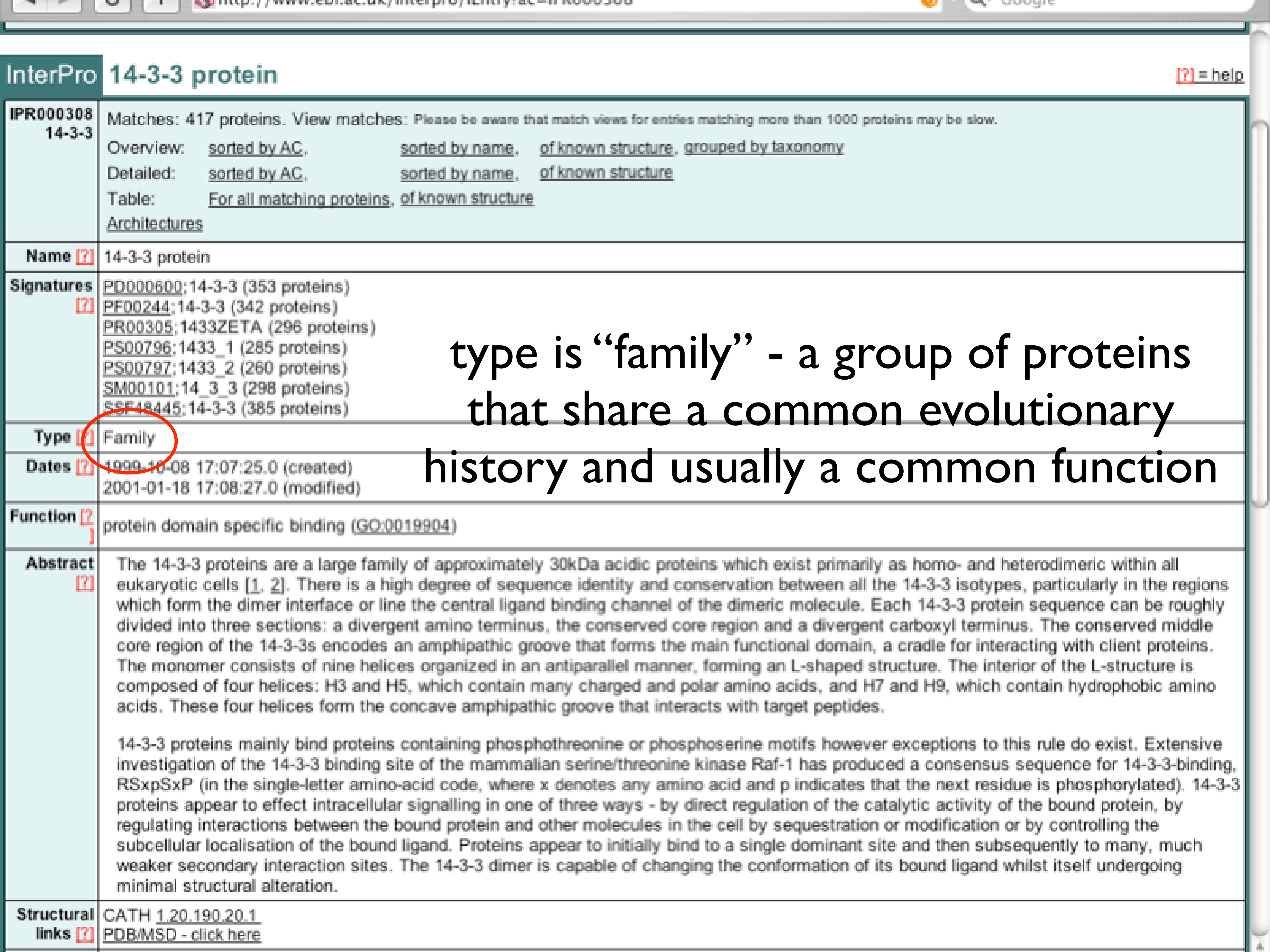
```
>sp|P31947|1433S_HUMAN 14-3-3 protein sigma (Stratifin)
MERASLIQKAKLAEQAERYEDMAAFMKGAVEKGEELSCEERNLLSVAYKNVVGGQR
VLSSIEQKSNEEGSEEKGPEVREYREKVETELQGVCDTVLGLLDSHLIKEAGDAES
LKMKGDYYRYLAEVATGDDKKRIIDSARSAYQEAMDISKKEMPPTNPIRLGLALNF
YEIANSPEEAISLAKTTFDEAMADLHTLSEDSYKDSTLIMQLLRDNLTLWTADNAG
EAPQEPQS
```

InterPro links - more clues about function



InterPro	IPR000308; 14-3-3. Graphical view of domain structure.
Pfam	PF00244; 14-3-3; 1. Pfam graphical view of domain structure.
PRINTS	PR00305; 1433ZETA.
ProDom	PD000600; 14-3-3; 1. [Domain structure / List of seq. sharing at least 1 domain]
PROSITE	PS00796; 1433_1; 1. PS00797; 1433_2; 1.
HOVERGEN	[Family / Alignment / Tree]
BLOCKS	P31947.
ProtoNet	P31947.
ProtoMap	P31947.
PRESAGE	P31947.
DIP	P31947.
ModBase	P31947.
SMR	P31947; 7F4B44E3AA59ECE6.
UniRef	View cluster of proteins with at least 50% / 90% identity.
Keywords	
Direct protein sequencing; Multigene family.	

Database of Interacting Proteins

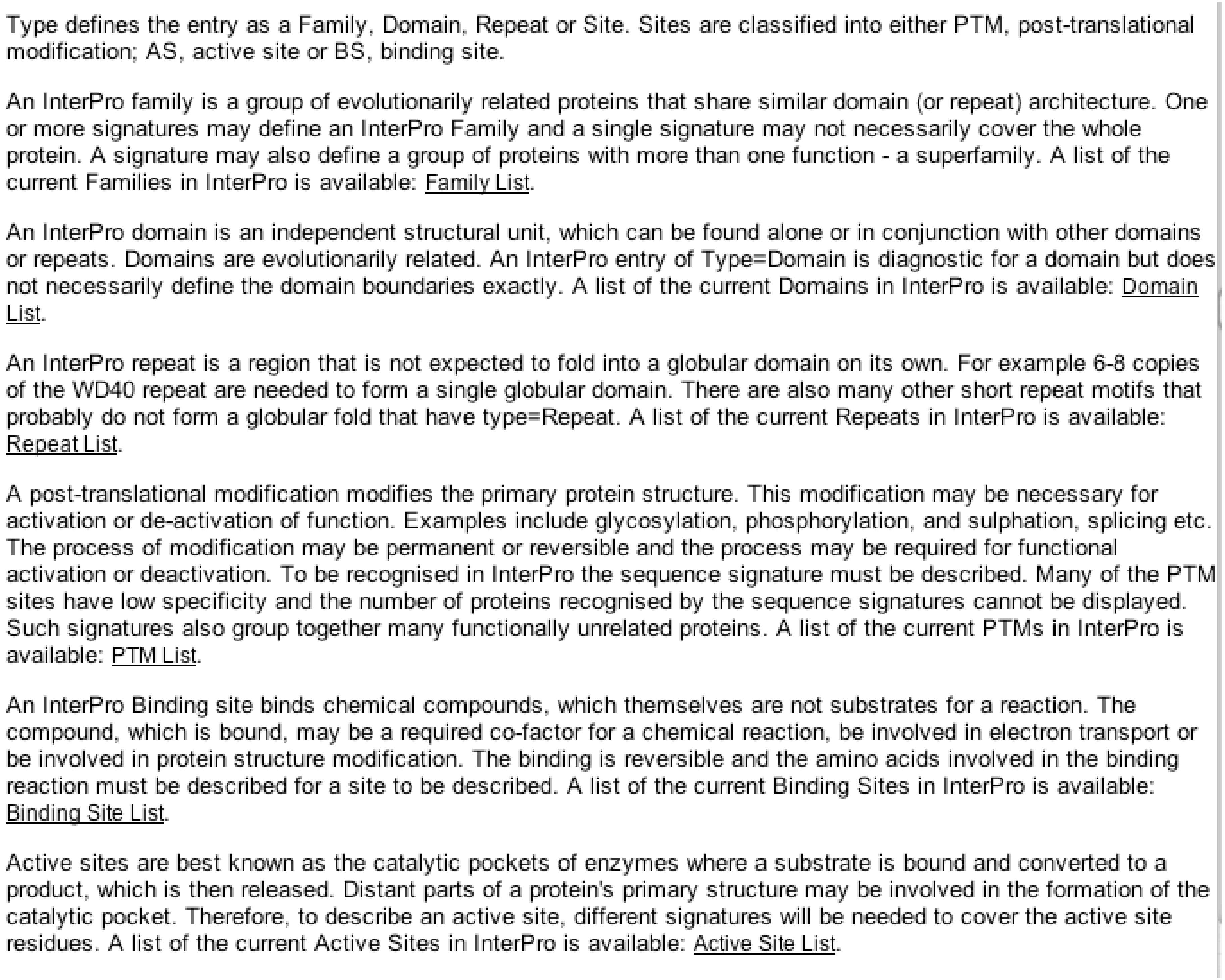


InterPro 14-3-3 protein

[?] = help

IPR000308 14-3-3	Matches: 417 proteins. View matches: Please be aware that match views for entries matching more than 1000 proteins may be slow. Overview: sorted by AC, sorted by name, of known structure, grouped by taxonomy Detailed: sorted by AC, sorted by name, of known structure Table: For all matching proteins, of known structure Architectures
Name [?]	14-3-3 protein
Signatures [?]	PD000600;14-3-3 (353 proteins) PF00244;14-3-3 (342 proteins) PR00305;1433ZETA (296 proteins) PS00796;1433_1 (285 proteins) PS00797;1433_2 (260 proteins) SM00101;14_3_3 (298 proteins) SSF48445;14-3-3 (385 proteins)
Type [?]	Family
Dates [?]	1999-10-08 17:07:25.0 (created) 2001-01-18 17:08:27.0 (modified)
Function [?]	protein domain specific binding (GO:0019904)
Abstract [?]	The 14-3-3 proteins are a large family of approximately 30kDa acidic proteins which exist primarily as homo- and heterodimeric within all eukaryotic cells [1, 2]. There is a high degree of sequence identity and conservation between all the 14-3-3 isotypes, particularly in the regions which form the dimer interface or line the central ligand binding channel of the dimeric molecule. Each 14-3-3 protein sequence can be roughly divided into three sections: a divergent amino terminus, the conserved core region and a divergent carboxyl terminus. The conserved middle core region of the 14-3-3s encodes an amphipathic groove that forms the main functional domain, a cradle for interacting with client proteins. The monomer consists of nine helices organized in an antiparallel manner, forming an L-shaped structure. The interior of the L-structure is composed of four helices: H3 and H5, which contain many charged and polar amino acids, and H7 and H9, which contain hydrophobic amino acids. These four helices form the concave amphipathic groove that interacts with target peptides. 14-3-3 proteins mainly bind proteins containing phosphothreonine or phosphoserine motifs however exceptions to this rule do exist. Extensive investigation of the 14-3-3 binding site of the mammalian serine/threonine kinase Raf-1 has produced a consensus sequence for 14-3-3-binding, RSxpSxP (in the single-letter amino-acid code, where x denotes any amino acid and p indicates that the next residue is phosphorylated). 14-3-3 proteins appear to effect intracellular signalling in one of three ways - by direct regulation of the catalytic activity of the bound protein, by regulating interactions between the bound protein and other molecules in the cell by sequestration or modification or by controlling the subcellular localisation of the bound ligand. Proteins appear to initially bind to a single dominant site and then subsequently to many, much weaker secondary interaction sites. The 14-3-3 dimer is capable of changing the conformation of its bound ligand whilst itself undergoing minimal structural alteration.
Structural links [?]	CATH 1.20.190.20.1 PDB/MSD - click here

type is “family” - a group of proteins that share a common evolutionary history and usually a common function



Type defines the entry as a Family, Domain, Repeat or Site. Sites are classified into either PTM, post-translational modification; AS, active site or BS, binding site.

An InterPro family is a group of evolutionarily related proteins that share similar domain (or repeat) architecture. One or more signatures may define an InterPro Family and a single signature may not necessarily cover the whole protein. A signature may also define a group of proteins with more than one function - a superfamily. A list of the current Families in InterPro is available: Family List.

An InterPro domain is an independent structural unit, which can be found alone or in conjunction with other domains or repeats. Domains are evolutionarily related. An InterPro entry of Type=Domain is diagnostic for a domain but does not necessarily define the domain boundaries exactly. A list of the current Domains in InterPro is available: Domain List.

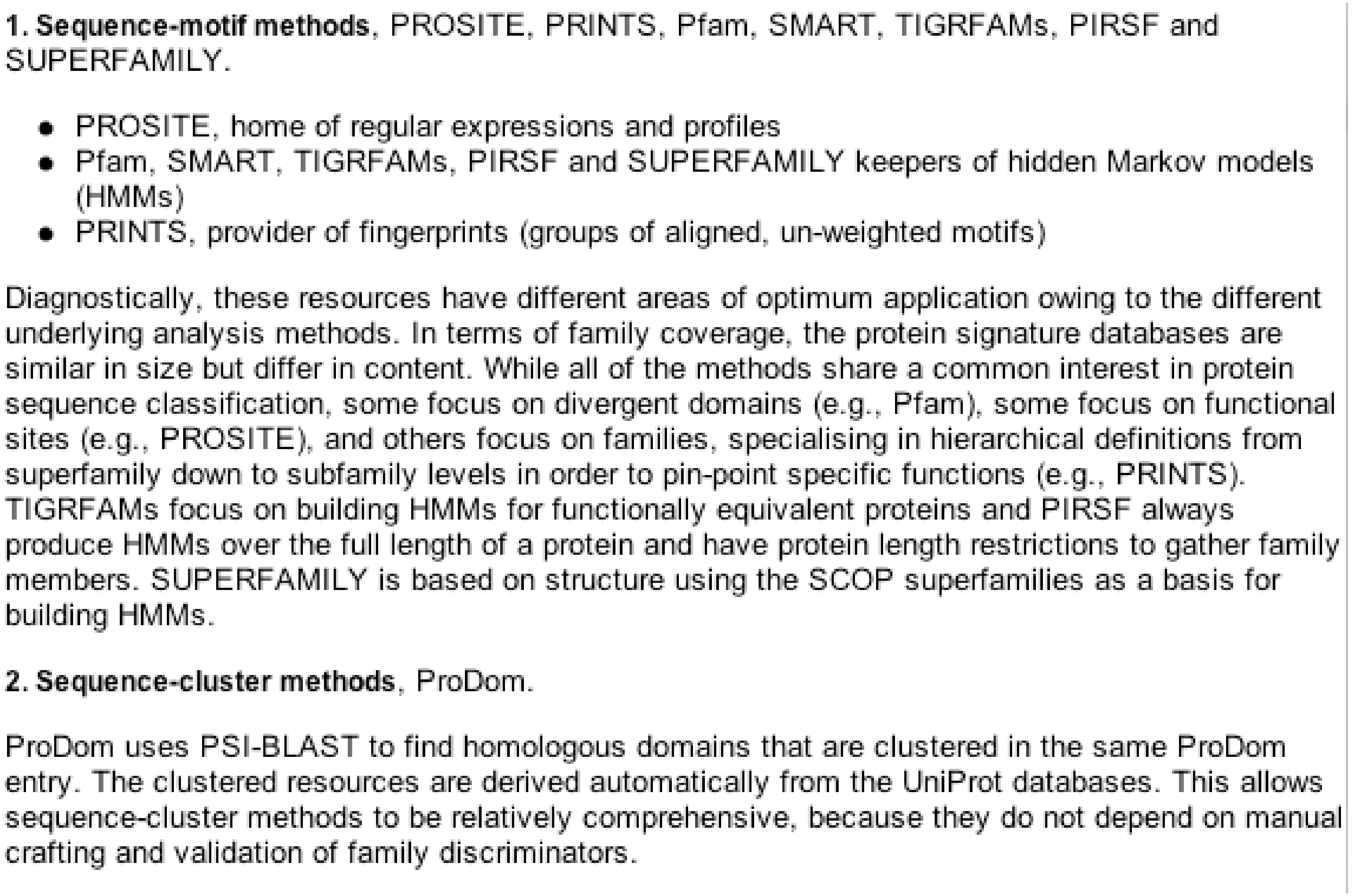
An InterPro repeat is a region that is not expected to fold into a globular domain on its own. For example 6-8 copies of the WD40 repeat are needed to form a single globular domain. There are also many other short repeat motifs that probably do not form a globular fold that have type=Repeat. A list of the current Repeats in InterPro is available: Repeat List.

A post-translational modification modifies the primary protein structure. This modification may be necessary for activation or de-activation of function. Examples include glycosylation, phosphorylation, and sulphation, splicing etc. The process of modification may be permanent or reversible and the process may be required for functional activation or deactivation. To be recognised in InterPro the sequence signature must be described. Many of the PTM sites have low specificity and the number of proteins recognised by the sequence signatures cannot be displayed. Such signatures also group together many functionally unrelated proteins. A list of the current PTMs in InterPro is available: PTM List.

An InterPro Binding site binds chemical compounds, which themselves are not substrates for a reaction. The compound, which is bound, may be a required co-factor for a chemical reaction, be involved in electron transport or be involved in protein structure modification. The binding is reversible and the amino acids involved in the binding reaction must be described for a site to be described. A list of the current Binding Sites in InterPro is available: Binding Site List.

Active sites are best known as the catalytic pockets of enzymes where a substrate is bound and converted to a product, which is then released. Distant parts of a protein's primary structure may be involved in the formation of the catalytic pocket. Therefore, to describe an active site, different signatures will be needed to cover the active site residues. A list of the current Active Sites in InterPro is available: Active Site List.

InterPro member databases



1. Sequence-motif methods, PROSITE, PRINTS, Pfam, SMART, TIGRFAMs, PIRSF and SUPERFAMILY.

- PROSITE, home of regular expressions and profiles
- Pfam, SMART, TIGRFAMs, PIRSF and SUPERFAMILY keepers of hidden Markov models (HMMs)
- PRINTS, provider of fingerprints (groups of aligned, un-weighted motifs)

Diagnostically, these resources have different areas of optimum application owing to the different underlying analysis methods. In terms of family coverage, the protein signature databases are similar in size but differ in content. While all of the methods share a common interest in protein sequence classification, some focus on divergent domains (e.g., Pfam), some focus on functional sites (e.g., PROSITE), and others focus on families, specialising in hierarchical definitions from superfamily down to subfamily levels in order to pin-point specific functions (e.g., PRINTS). TIGRFAMs focus on building HMMs for functionally equivalent proteins and PIRSF always produce HMMs over the full length of a protein and have protein length restrictions to gather family members. SUPERFAMILY is based on structure using the SCOP superfamilies as a basis for building HMMs.

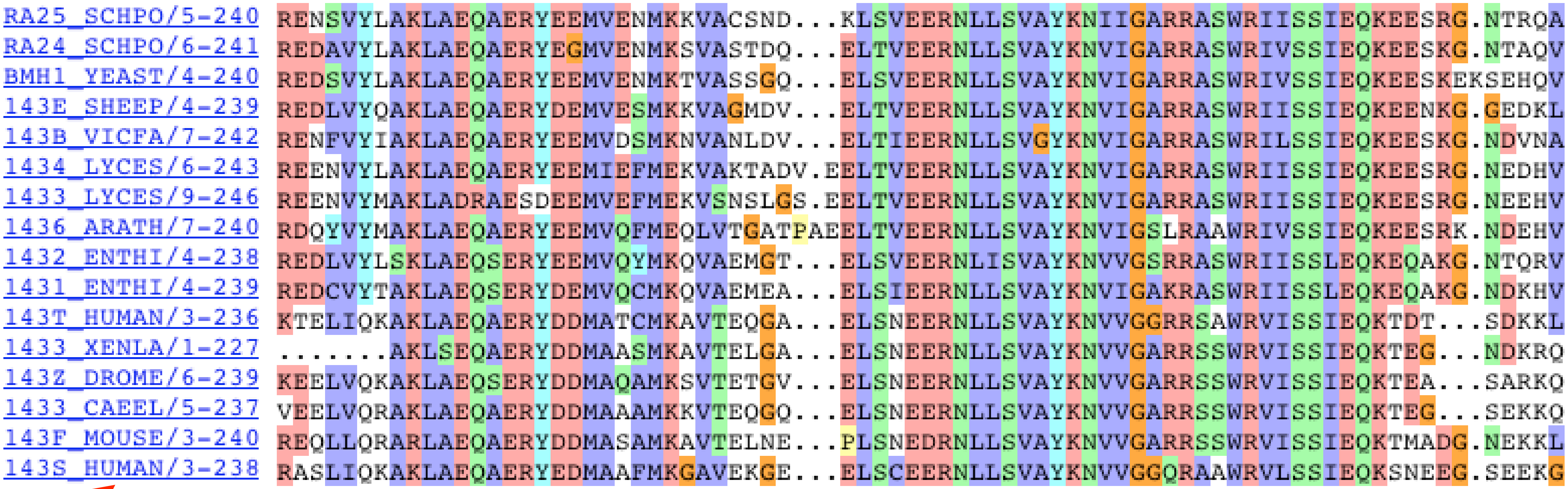
2. Sequence-cluster methods, ProDom.

ProDom uses PSI-BLAST to find homologous domains that are clustered in the same ProDom entry. The clustered resources are derived automatically from the UniProt databases. This allows sequence-cluster methods to be relatively comprehensive, because they do not depend on manual crafting and validation of family discriminators.

Profiles

- Built from multiple alignments involving proteins from many species (usually).
- Capture probability of observing specific amino acids at specific positions.
- Compare a sequence to a profile to get an idea of how well the sequence fits the profile. Is it a true member of the family?
- If yes, this gives you clues about the protein's function. This is a form of transitive annotation. *Use with caution!*

Alignment for 14-3-3



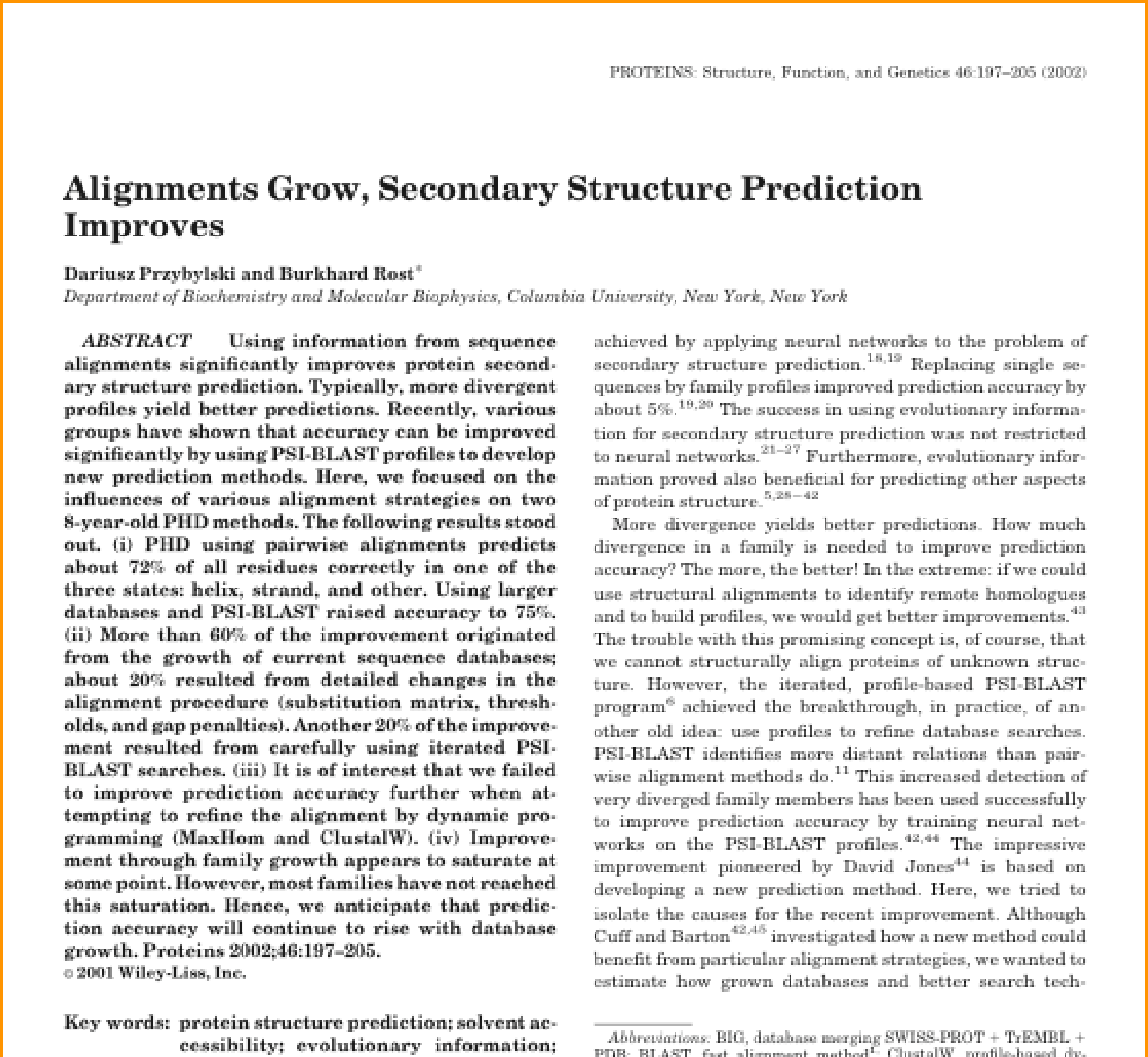
```
RA25_SCHPO/5-240  RENSVYLAKLAEQAERYEEMVENMKKVACSND...KLSVEERNLLSVAYKNIIGARRASWRIISSIEQKEESRG.NTRQA
RA24_SCHPO/6-241  REDAVYLAKLAEQAERYEGMVENMKSVASTDQ...ELTVEERNLLSVAYKNVIGARRASWRIVSSIEQKEESKG.NTAQV
BMH1_YEAST/4-240  REDSVYLAKLAEQAERYEEMVENMKTVASSGQ...ELSVEERNLLSVAYKNVIGARRASWRIVSSIEQKEESKEKSEHQV
143E_SHEEP/4-239  REDLVYQAKLAEQAERYDEMVESMKKVAGMDV...ELTVEERNLLSVAYKNVIGARRASWRIISSIEQKEENKG.GEDKL
143B_VICFA/7-242  RENFVYIAKLAEQAERYEEMVDSMKNVANLDV...ELTIEERNLLSVGYKNVIGARRASWRILSSIEQKEESKG.NDVNA
1434_LYCES/6-243  REENVYLAKLAEQAERYEEMIEFMEKVAKTADV.EELTVEERNLLSVAYKNVIGARRASWRIISSIEQKEESRG.NEDHV
1433_LYCES/9-246  REENVYMAKLADRAESDEEMVEFMEKVSNSLGS.EELTVEERNLLSVAYKNVIGARRASWRIISSIEQKEESRG.NEEHV
1436_ARATH/7-240  RDQYVYMAKLAEQAERYEEMVQFMEQLVTGATPAEELTVEERNLLSVAYKNVIGSLRAAWRIVSSIEQKEESRK.NDEHV
1432_ENTHI/4-238  REDLVYLSKLAEQSERYEEMVQYMKQVAEMGT...ELSVEERNLISVAYKNVVGSRRASWRIISSLEQKEQAKG.NTQRV
1431_ENTHI/4-239  REDCVYTAKLAEQSERYDEMVQCMKQVAEMEA...ELSIEERNLLSVAYKNVIGAKRASWRIISSLEQKEQAKG.NDKHV
143T_HUMAN/3-236  KTELIQKAKLAEQAERYDDMATCMKAVTEQGA...ELSNEERNLLSVAYKNVVGGRRSAWRVISSIEQKTDT...SDKKL
1433_XENLA/1-227  .......AKLSEQAERYDDMAASMKAVTELGA...ELSNEERNLLSVAYKNVVGARRSSWRVISSIEQKTEG...NDKRQ
143Z_DROME/6-239  KEELVQKAKLAEQSERYDDMAQAMKSVTETGV...ELSNEERNLLSVAYKNVVGARRSSWRVISSIEQKTEA...SARKQ
1433_CAEEL/5-237  VEELVQRAKLAEQAERYDDMAAAMKKVTEQGQ...ELSNEERNLLSVAYKNVVGARRSSWRVISSIEQKTEG...SEKKQ
143F_MOUSE/3-240  REQLLQRARLAEQAERYDDMASAMKAVTELNE...PLSNEDRNLLSVAYKNVVGARRSSWRVISSIEQKTMADG.NEKKL
143S_HUMAN/3-238  RASLIQKAKLAEQAERYEDMAAFMKGAVEKGE...ELSCEERNLLSVAYKNVVGGQRAAWRVLSSIEQKSNEEG.SEEKG
```

Note - our original protein sequence

The coloured markup was created by Jalview (Michele Clamp)

Alignments are colored using the ClustalX scheme in Jalview (orange:glycine (G); yellow: Proline (P); blue: small and hydrophobic amino-acids (A, V, L, I, M, F, W); green: hydroxyl and amine amino-acids (S, T, N, Q); red: charged amino-acids (D, E, R, K); cyan: histidine (H) and tyrosine(Y)).

Profiles used in Structure Prediction



PROTEINS: Structure, Function, and Genetics 46:197–205 (2002)

Alignments Grow, Secondary Structure Prediction Improves

Dariusz Przybylski and Burkhard Rost*
Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York

ABSTRACT Using information from sequence alignments significantly improves protein secondary structure prediction. Typically, more divergent profiles yield better predictions. Recently, various groups have shown that accuracy can be improved significantly by using PSI-BLAST profiles to develop new prediction methods. Here, we focused on the influences of various alignment strategies on two 8-year-old PHD methods. The following results stood out. (i) PHD using pairwise alignments predicts about 72% of all residues correctly in one of the three states: helix, strand, and other. Using larger databases and PSI-BLAST raised accuracy to 75%. (ii) More than 60% of the improvement originated from the growth of current sequence databases; about 20% resulted from detailed changes in the alignment procedure (substitution matrix, thresholds, and gap penalties). Another 20% of the improvement resulted from carefully using iterated PSI-BLAST searches. (iii) It is of interest that we failed to improve prediction accuracy further when attempting to refine the alignment by dynamic programming (MaxHom and ClustalW). (iv) Improvement through family growth appears to saturate at some point. However, most families have not reached this saturation. Hence, we anticipate that prediction accuracy will continue to rise with database growth. Proteins 2002;46:197–205.

Key words: protein structure prediction; solvent accessibility; evolutionary information;

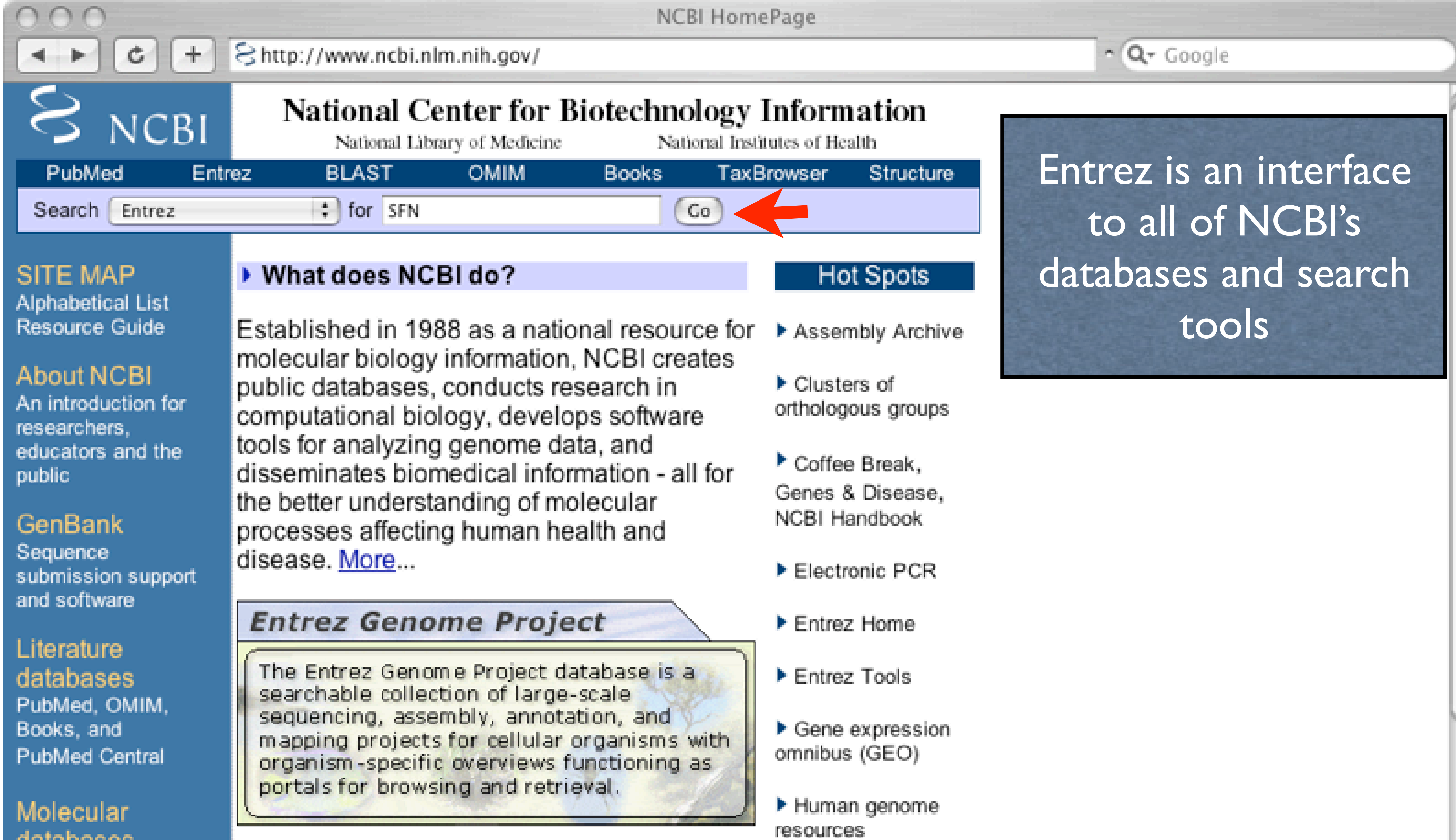
achieved by applying neural networks to the problem of secondary structure prediction.[18,19] Replacing single sequences by family profiles improved prediction accuracy by about 5%.[19,20] The success in using evolutionary information for secondary structure prediction was not restricted to neural networks.[21–27] Furthermore, evolutionary information proved also beneficial for predicting other aspects of protein structure.[5,28–42]

More divergence yields better predictions. How much divergence in a family is needed to improve prediction accuracy? The more, the better! In the extreme: if we could use structural alignments to identify remote homologues and to build profiles, we would get better improvements.[43] The trouble with this promising concept is, of course, that we cannot structurally align proteins of unknown structure. However, the iterated, profile-based PSI-BLAST program[6] achieved the breakthrough, in practice, of another old idea: use profiles to refine database searches. PSI-BLAST identifies more distant relations than pairwise alignment methods do.[11] This increased detection of very diverged family members has been used successfully to improve prediction accuracy by training neural networks on the PSI-BLAST profiles.[42,44] The impressive improvement pioneered by David Jones[44] is based on developing a new prediction method. Here, we tried to isolate the causes for the recent improvement. Although Cuff and Barton[42,45] investigated how a new method could benefit from particular alignment strategies, we wanted to estimate how grown databases and better search tech-

Abbreviations: BIG, database merging SWISS-PROT + TrEMBL + PDB; BLAST, fast alignment method[1]; ClustalW, profile-based dy-

Genomic or gene view

Especially useful for finding alternative forms due to alternative splicing.



Search: Entrez for: SFN

About NCBI: An introduction for researchers, educators and the public

GenBank: Sequence submission support and software

Literature databases: PubMed, OMIM, Books, and PubMed Central

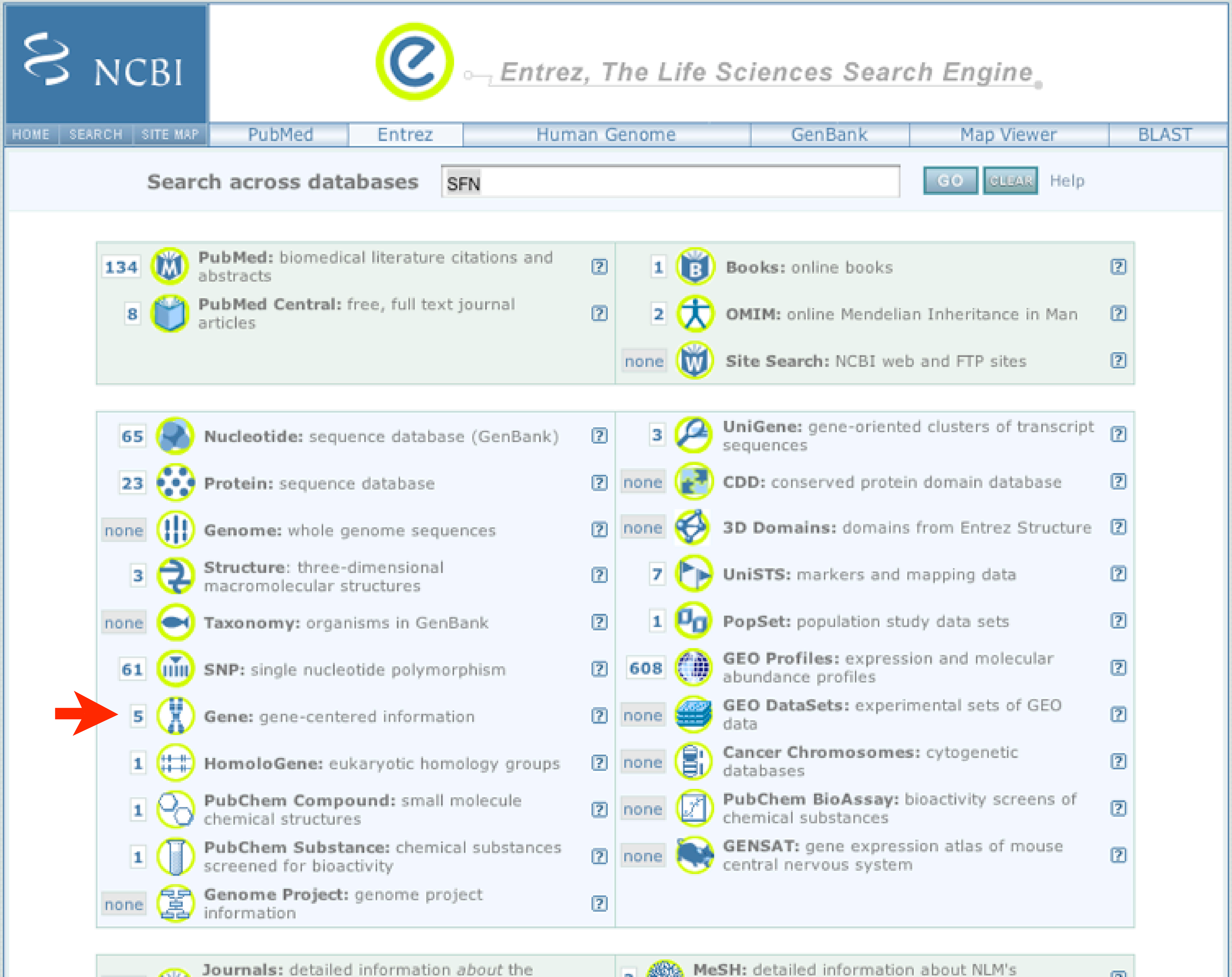
What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...

Entrez Genome Project

The Entrez Genome Project database is a searchable collection of large-scale sequencing, assembly, annotation, and mapping projects for cellular organisms with organism-specific overviews functioning as portals for browsing and retrieval.

Entrez is an interface to all of NCBI's databases and search tools



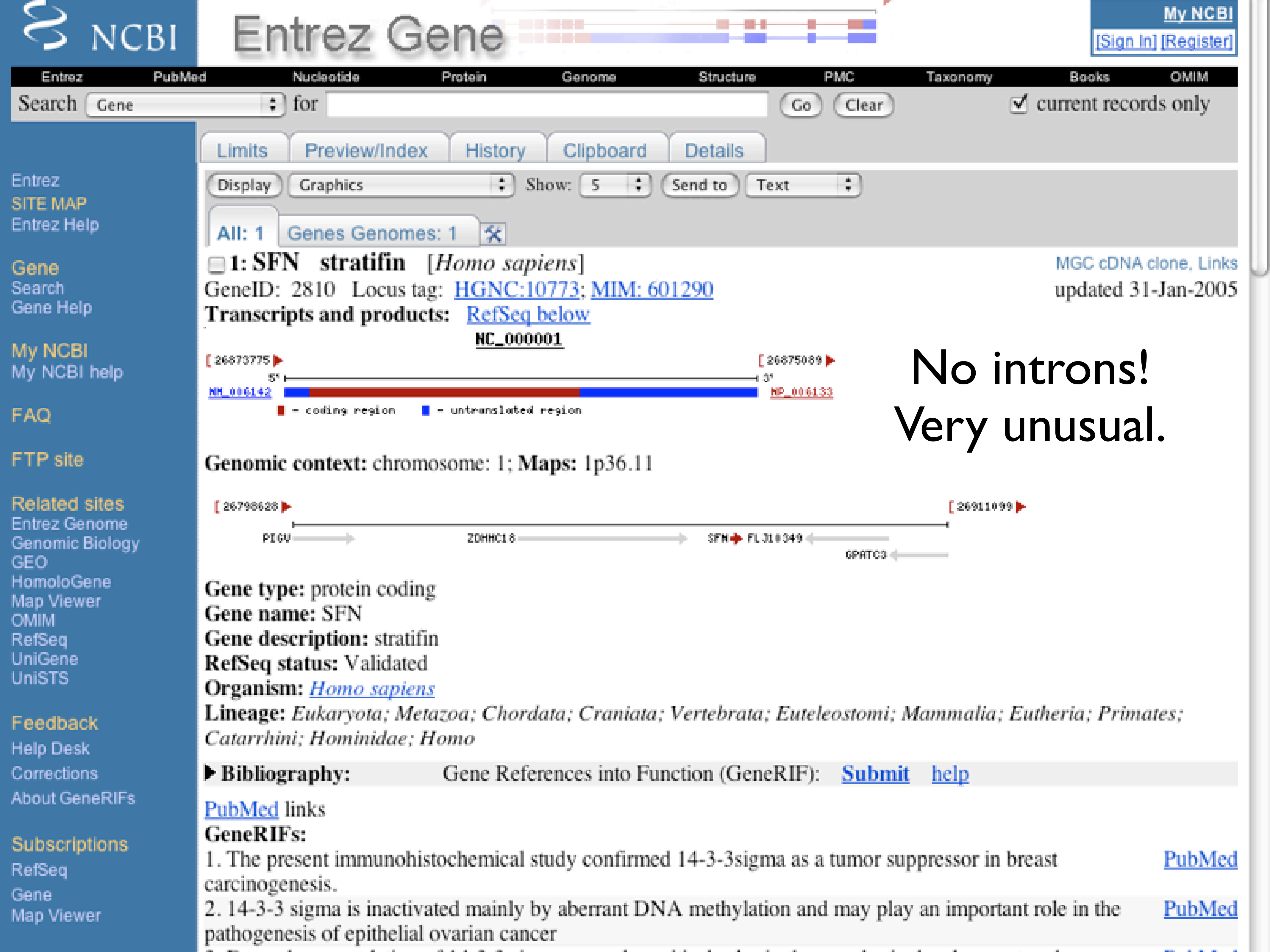
Search across databases SFN

134	**PubMed:** biomedical literature citations and abstracts
8	**PubMed Central:** free, full text journal articles
1	**Books:** online books
2	**OMIM:** online Mendelian Inheritance in Man
none	**Site Search:** NCBI web and FTP sites

65	**Nucleotide:** sequence database (GenBank)
23	**Protein:** sequence database
none	**Genome:** whole genome sequences
3	**Structure**: three-dimensional macromolecular structures
none	**Taxonomy:** organisms in GenBank
61	**SNP:** single nucleotide polymorphism
5	**Gene:** gene-centered information
1	**HomoloGene:** eukaryotic homology groups
1	**PubChem Compound:** small molecule chemical structures
1	**PubChem Substance:** chemical substances screened for bioactivity
none	**Genome Project:** genome project information
3	**UniGene:** gene-oriented clusters of transcript sequences
none	**CDD:** conserved protein domain database
none	**3D Domains:** domains from Entrez Structure
7	**UniSTS:** markers and mapping data
1	**PopSet:** population study data sets
608	**GEO Profiles:** expression and molecular abundance profiles
none	**GEO DataSets:** experimental sets of GEO data
none	**Cancer Chromosomes:** cytogenetic databases
none	**PubChem BioAssay:** bioactivity screens of chemical substances
none	**GENSAT:** gene expression atlas of mouse central nervous system

Journals: detailed information *about* the

MeSH: detailed information about NLM's



□ **1: SFN stratifin** [*Homo sapiens*]

MGC cDNA clone, Links

GeneID: 2810 Locus tag: HGNC:10773; MIM: 601290

updated 31-Jan-2005

Transcripts and products: RefSeq below

No introns!
Very unusual.

Genomic context: chromosome: 1; **Maps:** 1p36.11

Gene type: protein coding
Gene name: SFN
Gene description: stratifin
RefSeq status: Validated
Organism: *Homo sapiens*
Lineage: *Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo*

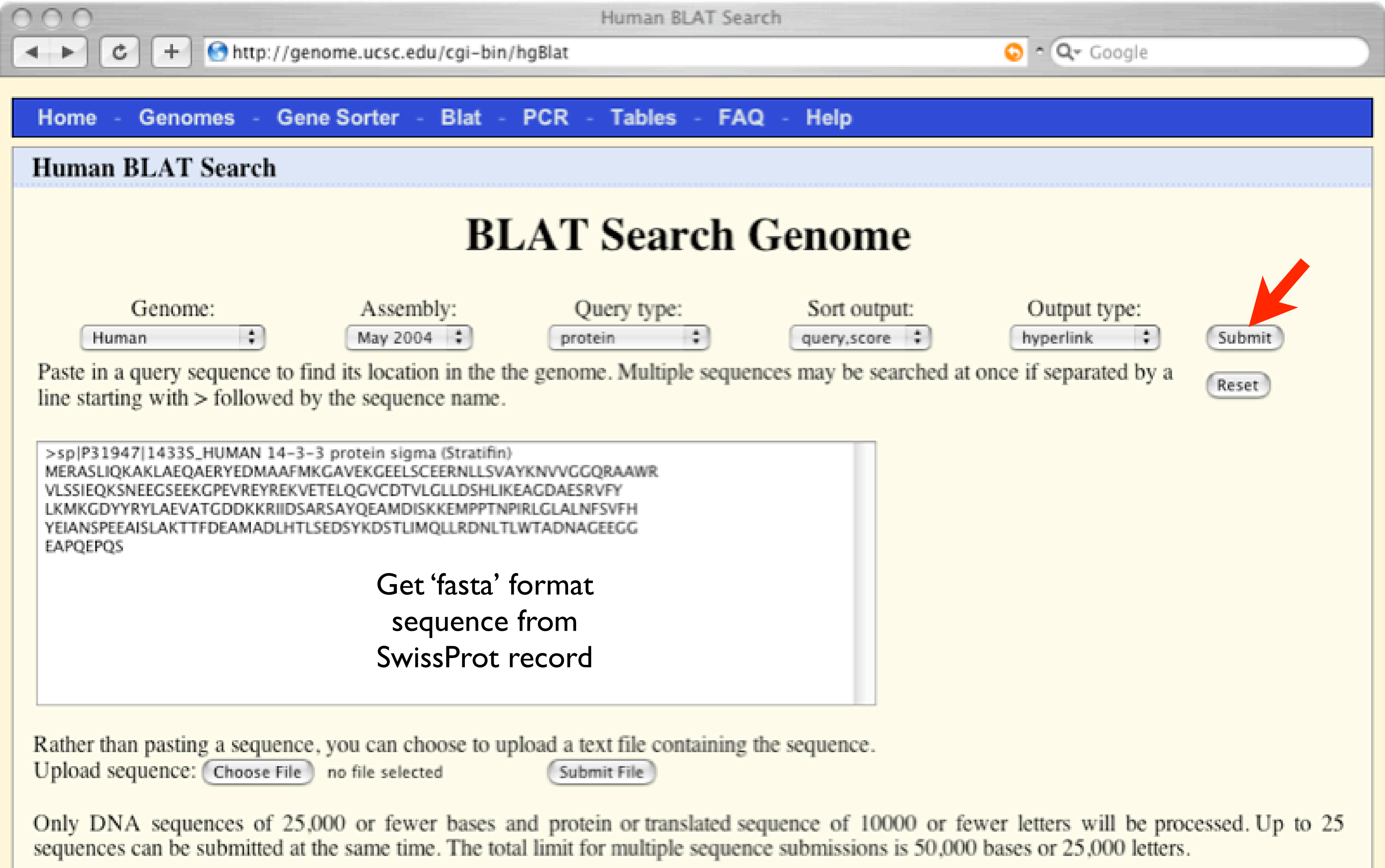
▶ **Bibliography:** Gene References into Function (GeneRIF): **Submit** help

PubMed links

GeneRIFs:

1. The present immunohistochemical study confirmed 14-3-3sigma as a tumor suppressor in breast carcinogenesis. PubMed
2. 14-3-3 sigma is inactivated mainly by aberrant DNA methylation and may play an important role in the pathogenesis of epithelial ovarian cancer PubMed

Search the genome



Human BLAT Search
http://genome.ucsc.edu/cgi-bin/hgBlat
Google
Home - Genomes - Gene Sorter - Blat - PCR - Tables - FAQ - Help

Human BLAT Search

BLAT Search Genome

Genome: Human
Assembly: May 2004
Query type: protein
Sort output: query,score
Output type: hyperlink

Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched at once if separated by a line starting with > followed by the sequence name.

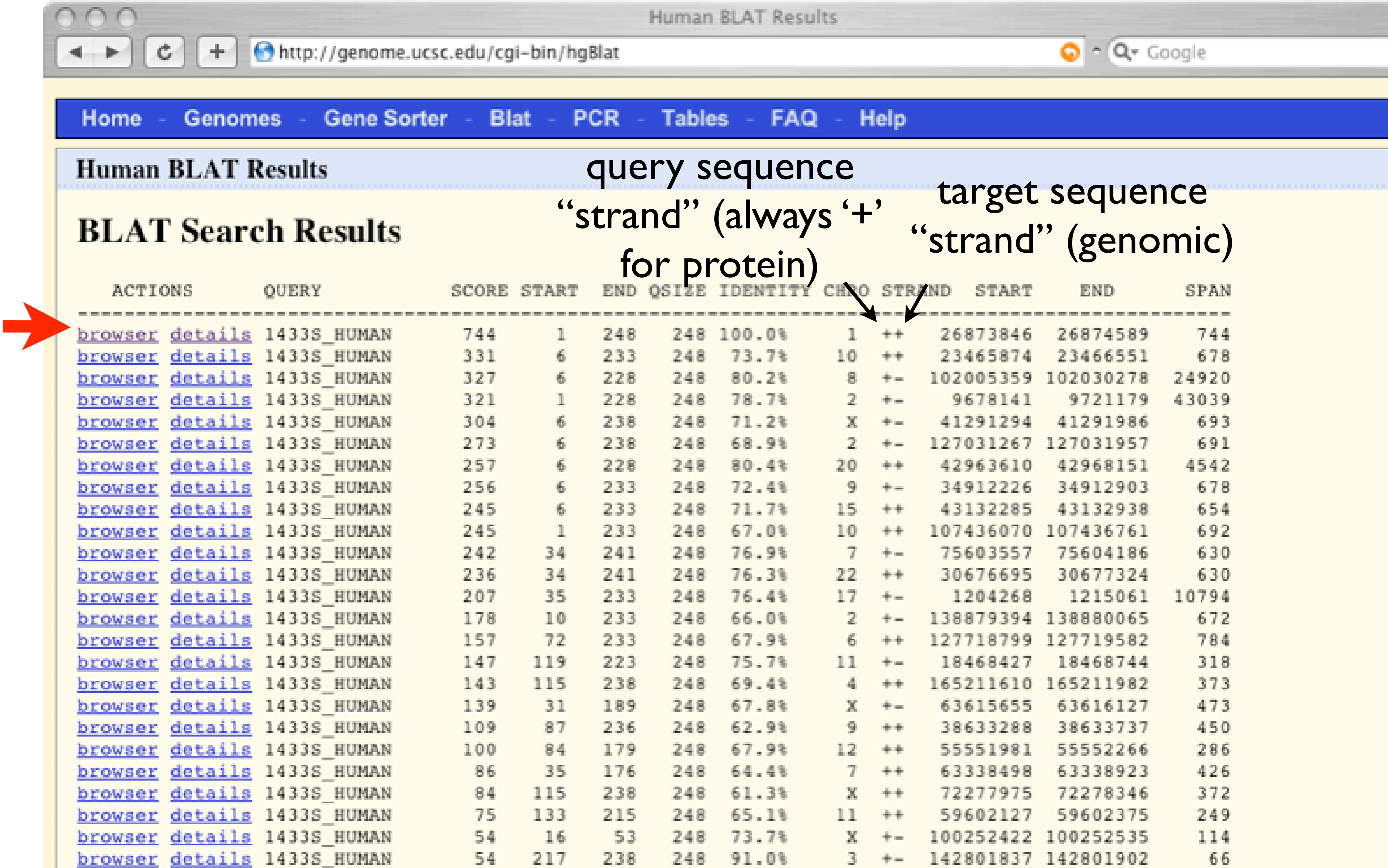
```
>sp|P31947|1433S_HUMAN 14-3-3 protein sigma (Stratifin)
MERASLIQKAKLAEQAERYEDMAAFMKGAVEKGEELSCEERNLLSVAYKNVVGGQRAAWR
VLSSIEQKSNEEGSEEKGPEVREYREKVETELQGVCDTVLGLLDSHLIKEAGDAESRVFY
LKMKGDYYRYLAEVATGDDKKRIIDSARSAYQEAMDISKKEMPPTNPIRLGLALNFSVFH
YEIANSPEEAISLAKTTFDEAMADLHTLSEDSYKDSTLIMQLLRDNLTLWTADNAGEEGG
EAPQEPQS
```

Get ‘fasta’ format sequence from SwissProt record

Rather than pasting a sequence, you can choose to upload a text file containing the sequence.
Upload sequence:

Only DNA sequences of 25,000 or fewer bases and protein or translated sequence of 10000 or fewer letters will be processed. Up to 25 sequences can be submitted at the same time. The total limit for multiple sequence submissions is 50,000 bases or 25,000 letters.

blat results


Human BLAT Results
http://genome.ucsc.edu/cgi-bin/hgBlat
Google
Home - Genomes - Gene Sorter - Blat - PCR - Tables - FAQ - Help

Human BLAT Results

BLAT Search Results

query sequence "strand" (always '+' for protein)

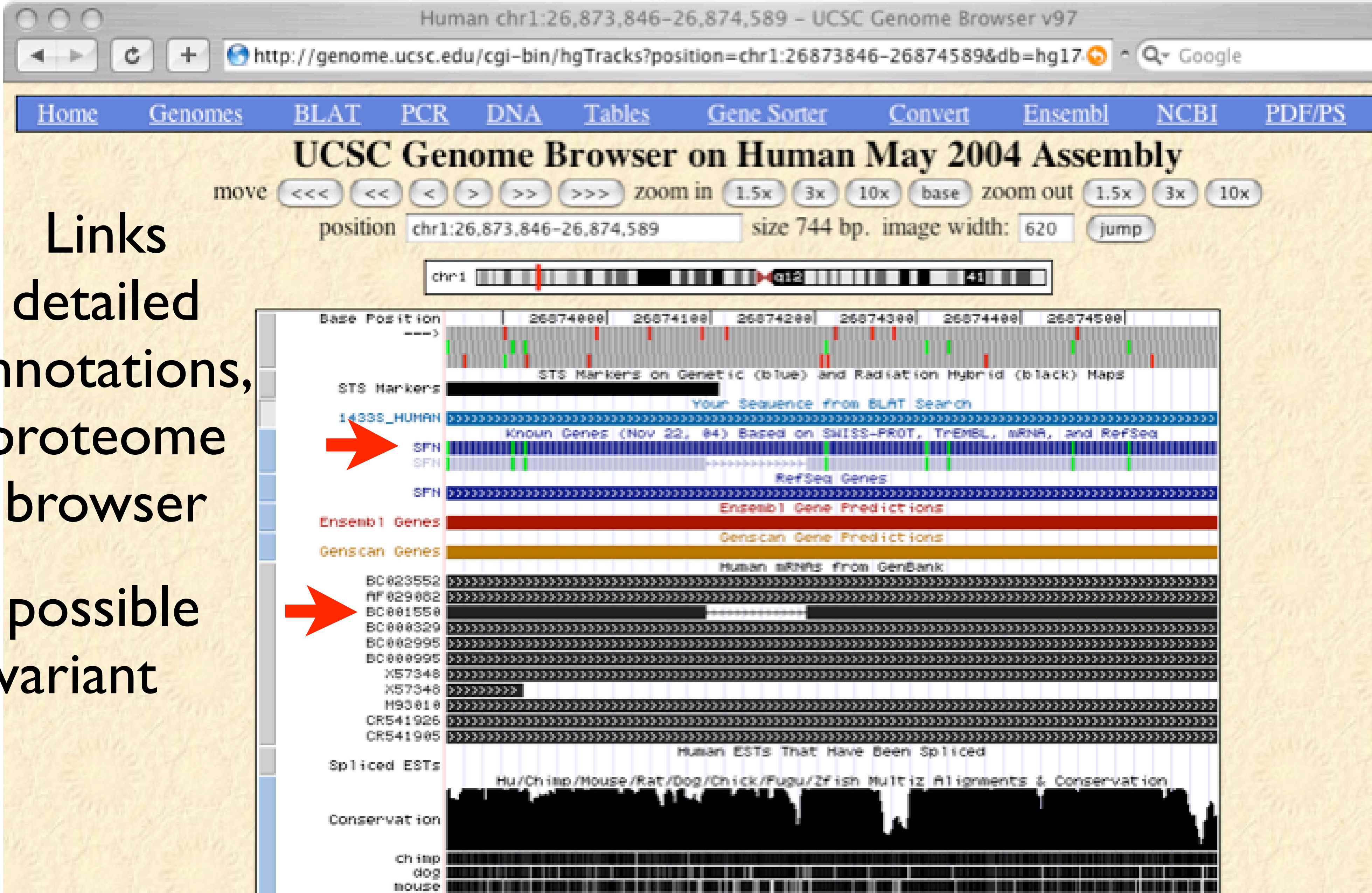
target sequence "strand" (genomic)

```
   ACTIONS       QUERY           SCORE START  END QSIZE IDENTITY CHRO STRAND  START     END       SPAN
---------------------------------------------------------------------------------------------------
browser details 1433S_HUMAN       744     1   248   248 100.0%     1   ++    26873846  26874589    744
browser details 1433S_HUMAN       331     6   233   248  73.7%    10   ++    23465874  23466551    678
browser details 1433S_HUMAN       327     6   228   248  80.2%     8   +-   102005359 102030278  24920
browser details 1433S_HUMAN       321     1   228   248  78.7%     2   +-     9678141   9721179  43039
browser details 1433S_HUMAN       304     6   238   248  71.2%     X   +-    41291294  41291986    693
browser details 1433S_HUMAN       273     6   238   248  68.9%     2   +-   127031267 127031957    691
browser details 1433S_HUMAN       257     6   228   248  80.4%    20   ++    42963610  42968151   4542
browser details 1433S_HUMAN       256     6   233   248  72.4%     9   +-    34912226  34912903    678
browser details 1433S_HUMAN       245     6   233   248  71.7%    15   ++    43132285  43132938    654
browser details 1433S_HUMAN       245     1   233   248  67.0%    10   ++   107436070 107436761    692
browser details 1433S_HUMAN       242    34   241   248  76.9%     7   +-    75603557  75604186    630
browser details 1433S_HUMAN       236    34   241   248  76.3%    22   ++    30676695  30677324    630
browser details 1433S_HUMAN       207    35   233   248  76.4%    17   +-     1204268   1215061  10794
browser details 1433S_HUMAN       178    10   233   248  66.0%     2   +-   138879394 138880065    672
browser details 1433S_HUMAN       157    72   233   248  67.9%     6   ++   127718799 127719582    784
browser details 1433S_HUMAN       147   119   223   248  75.7%    11   +-    18468427  18468744    318
browser details 1433S_HUMAN       143   115   238   248  69.4%     4   ++   165211610 165211982    373
browser details 1433S_HUMAN       139    31   189   248  67.8%     X   +-    63615655  63616127    473
browser details 1433S_HUMAN       109    87   236   248  62.9%     9   ++    38633288  38633737    450
browser details 1433S_HUMAN       100    84   179   248  67.9%    12   ++    55551981  55552266    286
browser details 1433S_HUMAN        86    35   176   248  64.4%     7   ++    63338498  63338923    426
browser details 1433S_HUMAN        84   115   238   248  61.3%     X   ++    72277975  72278346    372
browser details 1433S_HUMAN        75   133   215   248  65.1%    11   ++    59602127  59602375    249
browser details 1433S_HUMAN        54    16    53   248  73.7%     X   +-   100252422 100252535    114
browser details 1433S_HUMAN        54   217   238   248  91.0%     3   +-   142801837 142801902     66
```

A potential variant form?

Links detailed annotations, proteome browser

A possible variant

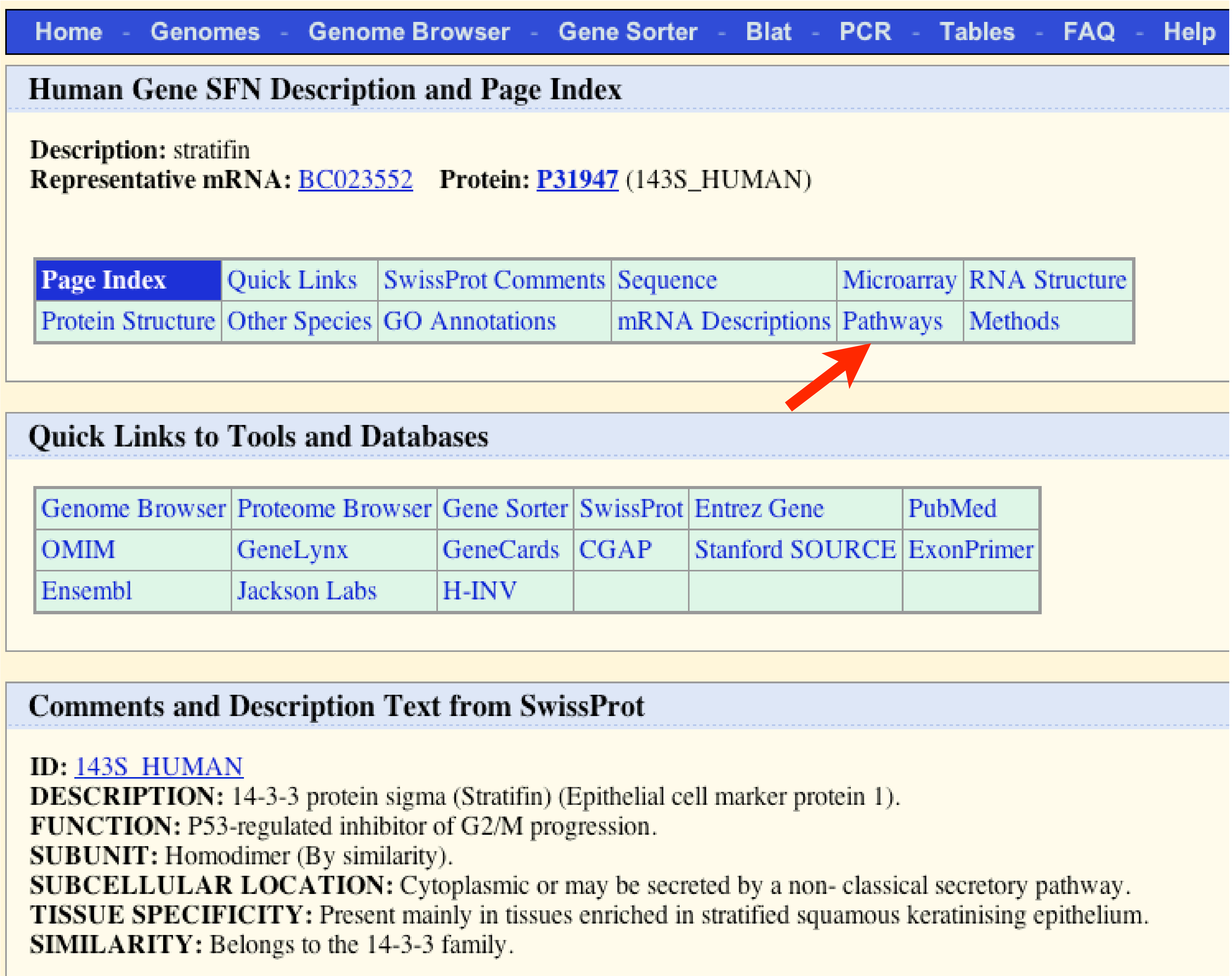


Human chr1:26,873,846-26,874,589 - UCSC Genome Browser v97

http://genome.ucsc.edu/cgi-bin/hgTracks?position=chr1:26873846-26874589&db=hg17

UCSC Genome Browser on Human May 2004 Assembly

position chr1:26,873,846-26,874,589 size 744 bp. image width: 620



Home - Genomes - Genome Browser - Gene Sorter - Blat - PCR - Tables - FAQ - Help

Human Gene SFN Description and Page Index

Description: stratifin
Representative mRNA: BC023552 **Protein:** P31947 (143S_HUMAN)

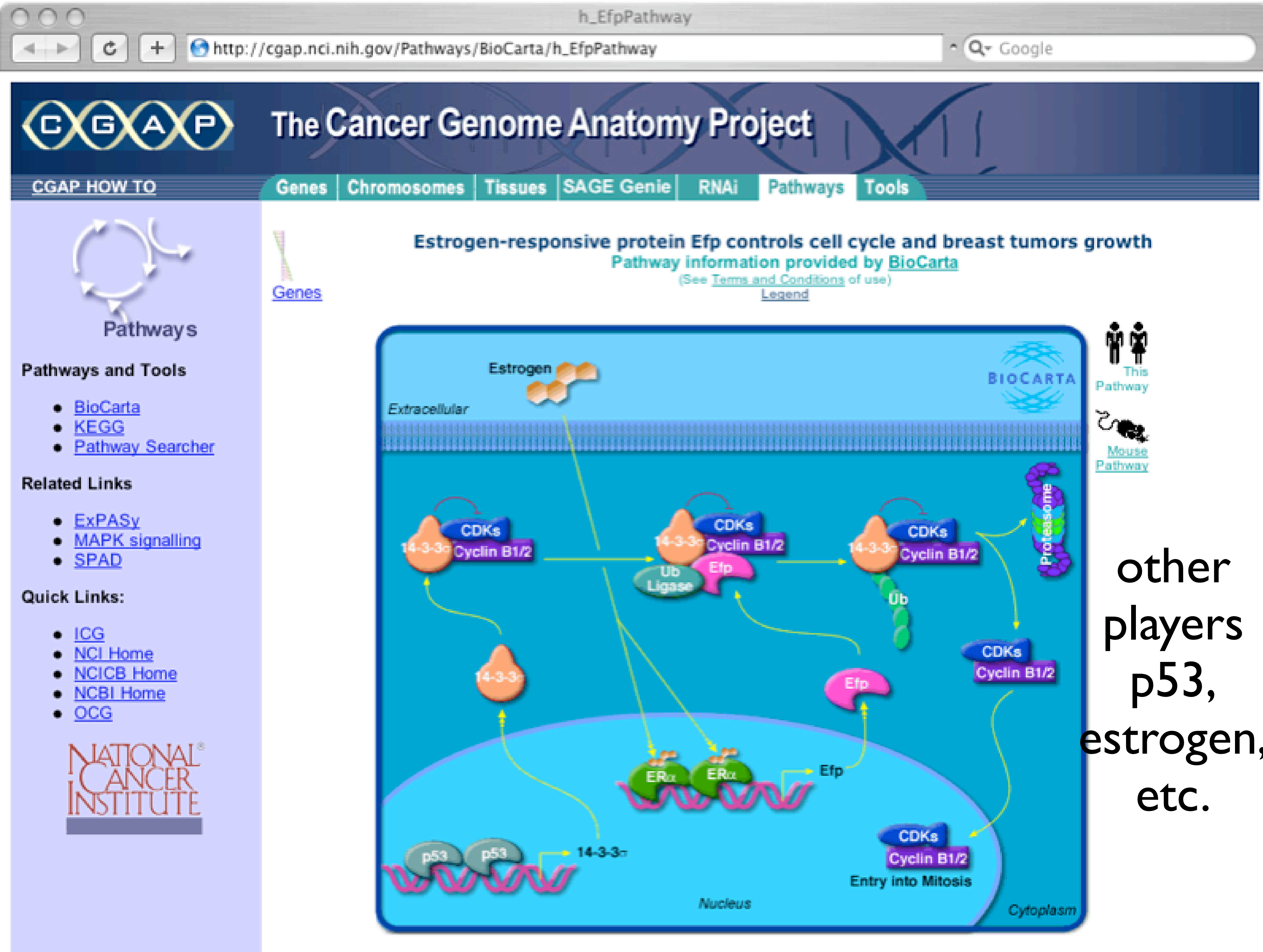
Page Index	Quick Links	SwissProt Comments	Sequence	Microarray	RNA Structure
Protein Structure	Other Species	GO Annotations	mRNA Descriptions	Pathways	Methods

Quick Links to Tools and Databases

Genome Browser	Proteome Browser	Gene Sorter	SwissProt	Entrez Gene	PubMed
OMIM	GeneLynx	GeneCards	CGAP	Stanford SOURCE	ExonPrimer
Ensembl	Jackson Labs	H-INV			

Comments and Description Text from SwissProt

ID: 143S_HUMAN
DESCRIPTION: 14-3-3 protein sigma (Stratifin) (Epithelial cell marker protein 1).
FUNCTION: P53-regulated inhibitor of G2/M progression.
SUBUNIT: Homodimer (By similarity).
SUBCELLULAR LOCATION: Cytoplasmic or may be secreted by a non- classical secretory pathway.
TISSUE SPECIFICITY: Present mainly in tissues enriched in stratified squamous keratinising epithelium.
SIMILARITY: Belongs to the 14-3-3 family.



Pathways and Tools

- BioCarta
- KEGG
- Pathway Searcher

Related Links

- ExPASy
- MAPK signalling
- SPAD

Quick Links:

- ICG
- NCI Home
- NCICB Home
- NCBI Home
- OCG

Estrogen-responsive protein Efp controls cell cycle and breast tumors growth

Pathway information provided by BioCarta

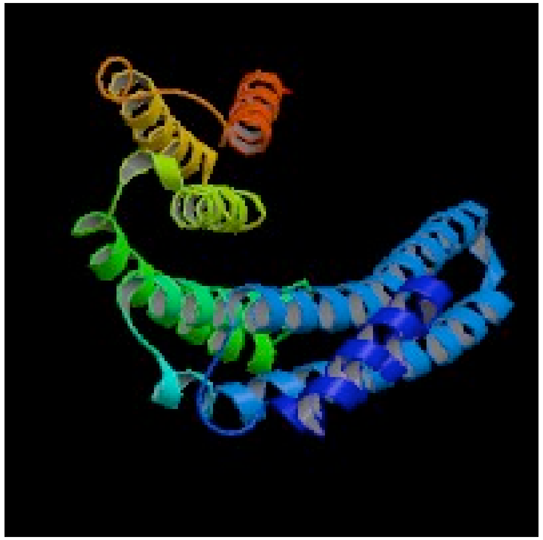
(See Terms and Conditions of use)

Legend

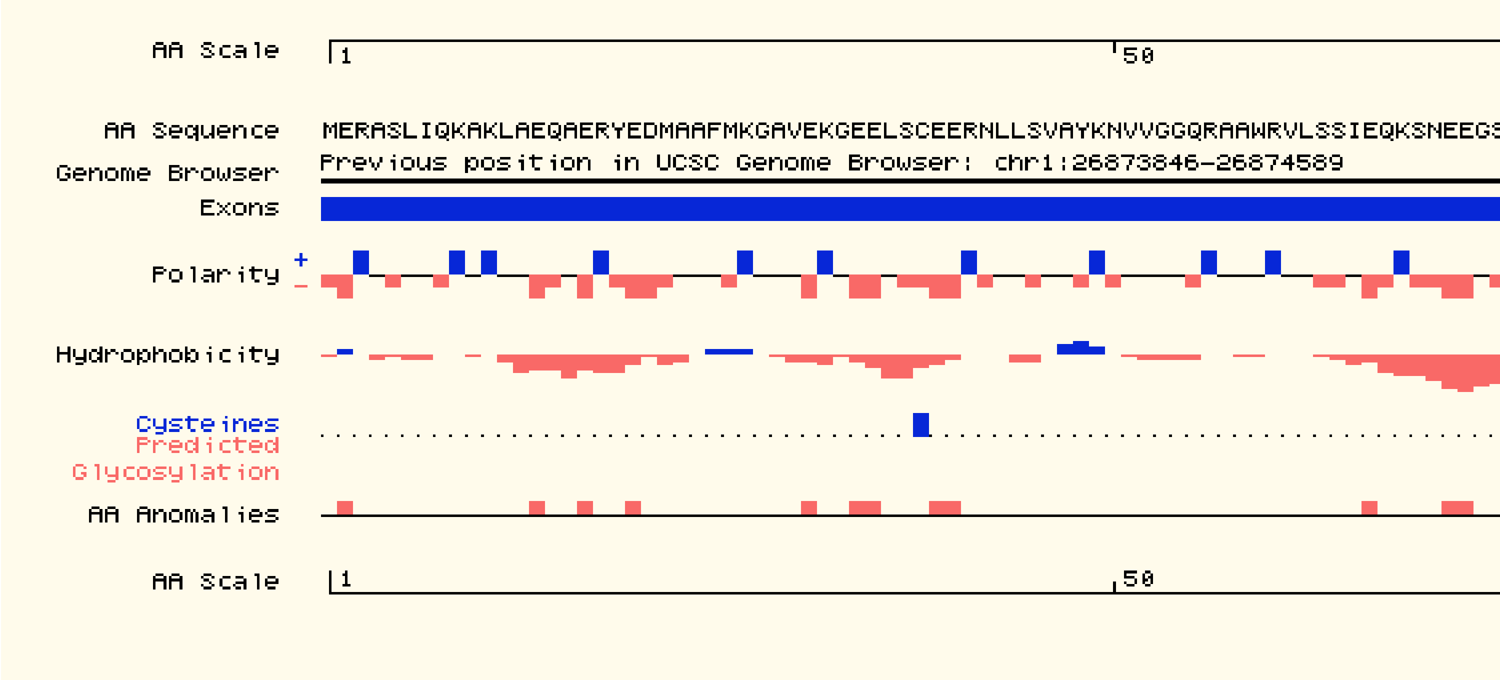
other players p53, estrogen, etc.

Proteome Page

structure prediction



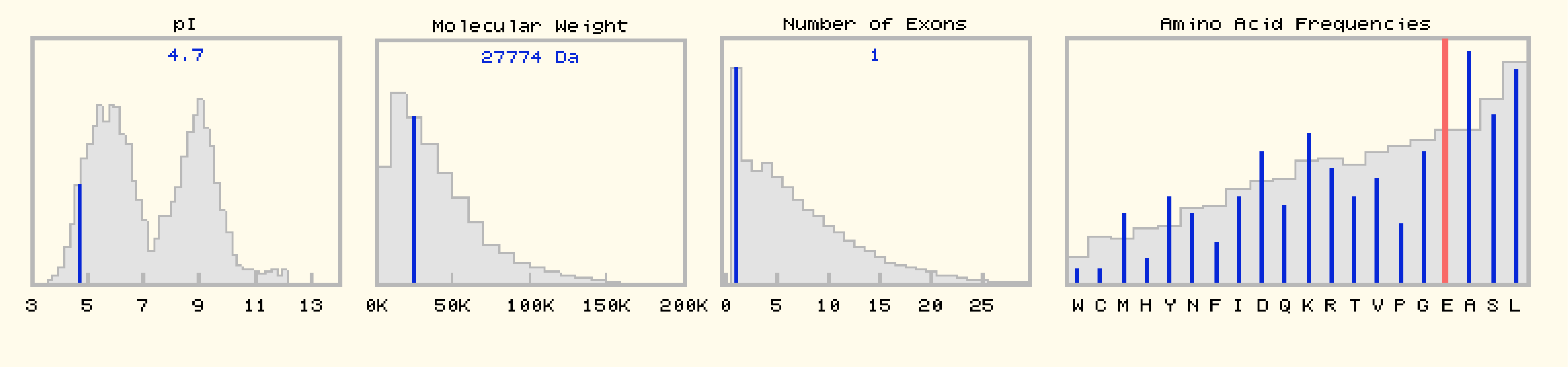
exon structure



MERASLIQKAKLAEQAERYEDMAAFMKGAVEKGEELSCEERNLLSVAYKNVVGGQRAAWRVLSSIEQKSNEEGS

Previous position in UCSC Genome Browser: chr1:26873846-26874589

compare to other proteins/genes



Web pages versus "bulk" download

- Web pages (HTTP) usually for "one at a time" retrieval.
- Bulk download (FTP) usually for fetching entire databases, large files of data needed to answer genome-scale questions.
- NCBI: http://www.ncbi.nlm.nih.gov/Ftp/index.html
- SwissProt: ftp://us.expasy.org/

Build your own?

- Programming toolkits for bioinformatics
  - www.biopython.org, www.bioperl.org
- python, perl (python is easier!)
- most tools have “command-line” versions
- a topic for another lecture
- THANK YOU!