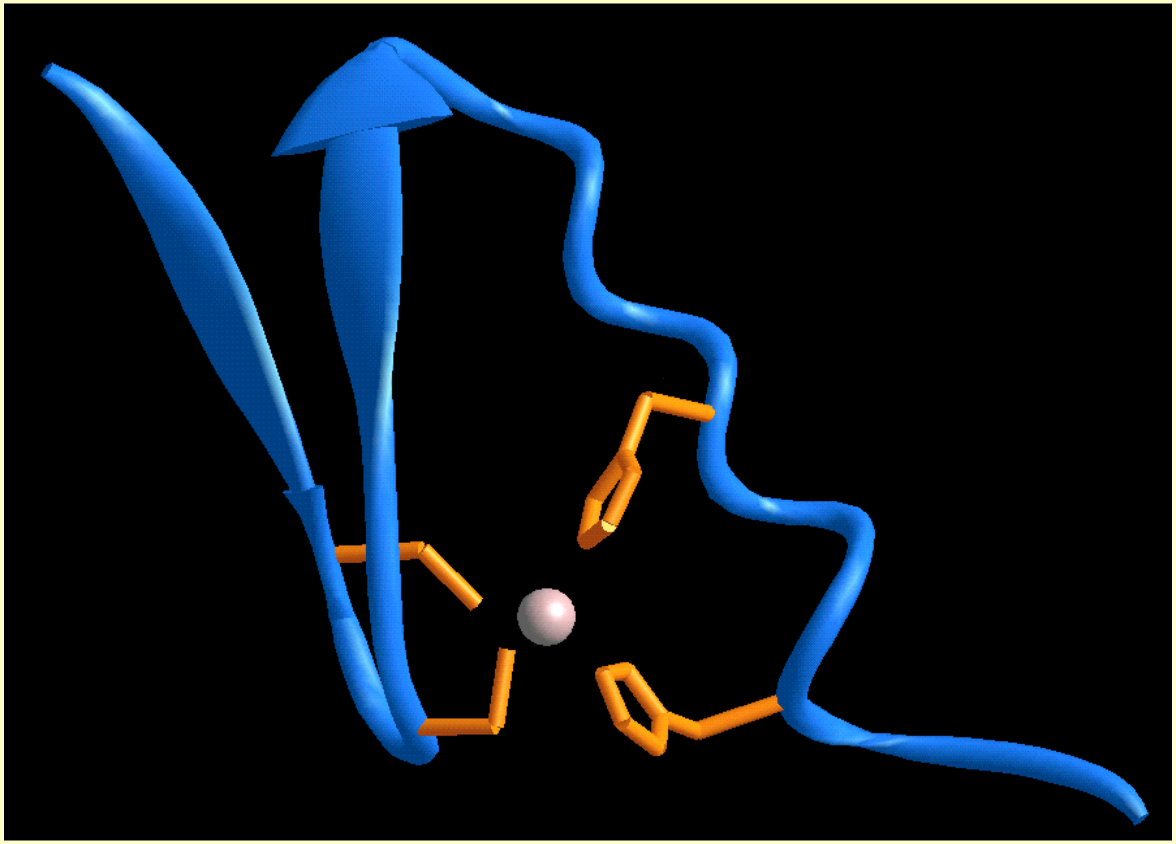
Bioinformatics

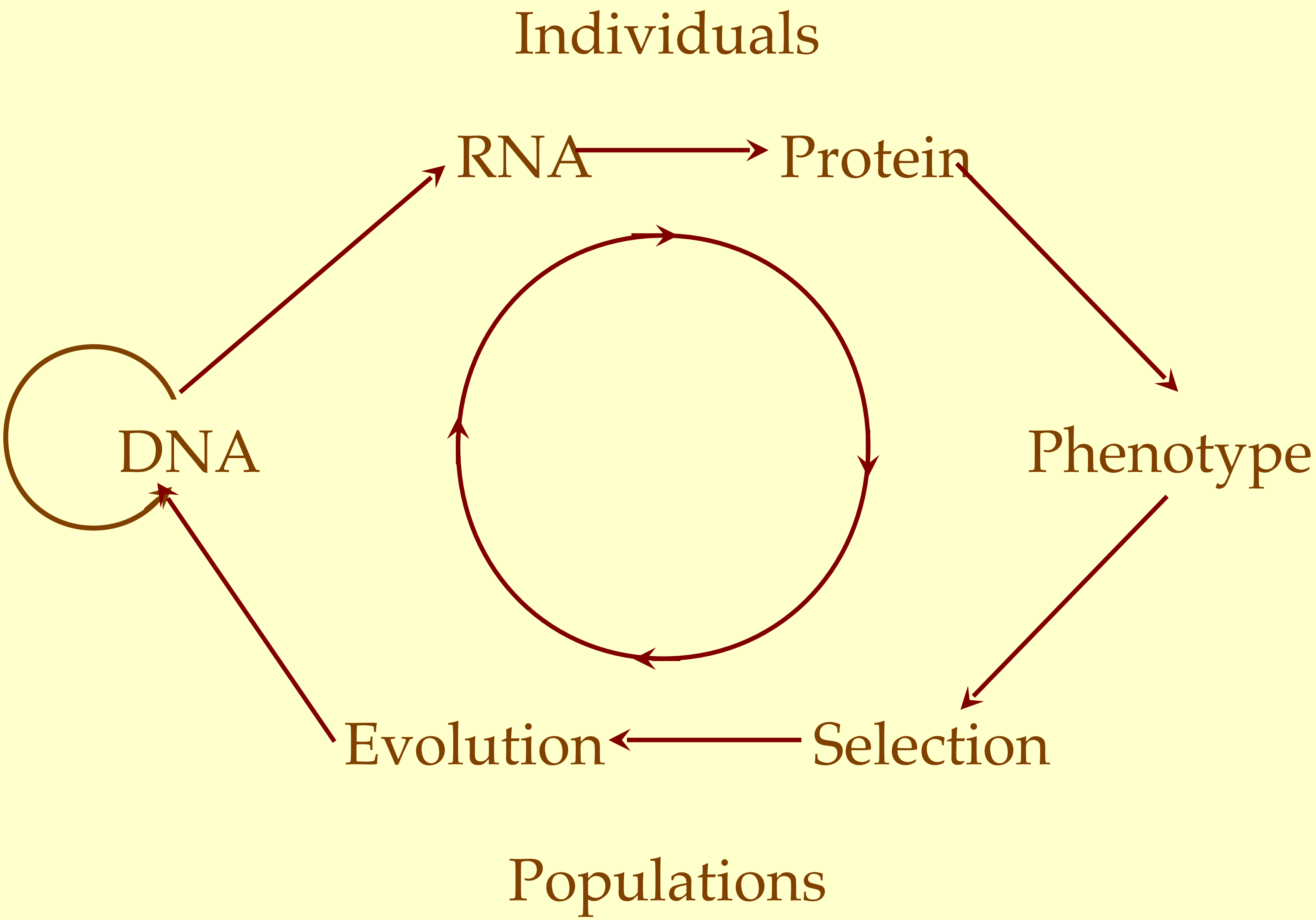
Genomics& Medicine

http://biochem118.stanford.edu/



Doug Brutlag, Professor Emeritus
of Biochemistry & Medicine (by courtesy)
Stanford University School of Medicine

What is Bioinformatics?



Individuals

RNA

Protein

DNA

Phenotype

Evolution

Selection

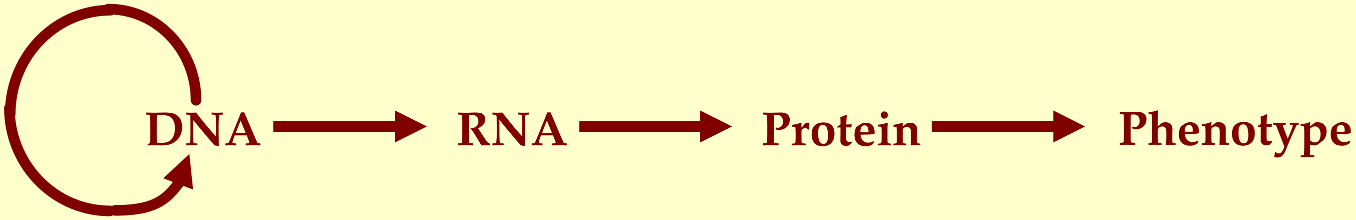
Populations

Biological Information

Computational Goals of Bioinformatics

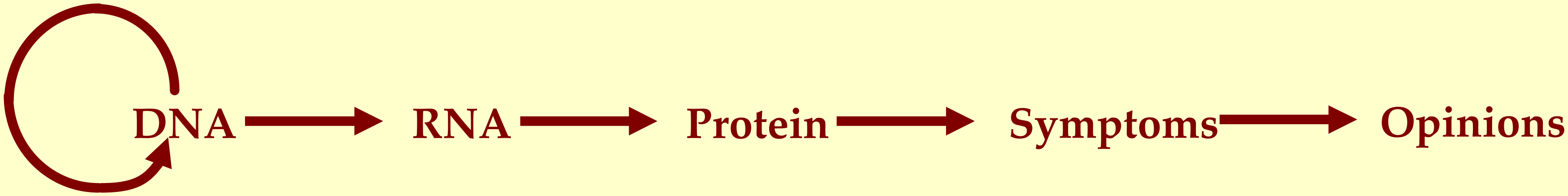
- Learn & Generalize: Discover conserved patterns (models) of sequences, structures, metabolism & chemistries from well-studied examples.
- Prediction: Infer function or structure of newly sequenced genes, genomes, proteomes or proteins from these generalizations.
- Organize & Integrate: Develop a systematic and genomic approach to molecular interactions, metabolism, cell signaling, gene expression… Basis of systems biology
- Simulate: Model gene expression, gene regulation, protein folding, protein-protein interaction, protein-ligand binding, catalytic function, metabolism… Goal of systems biology.
- Engineer: Construct novel organisms or novel functions or novel regulation of genes and proteins. Basis of synthetic biology.
- Target: Mutations, RNAi to specific genes and transcripts or drugs to specific protein targets. Practical biological and medical use of bioinformatics.

Central Paradigm of Molecular Biology



DNA → **RNA** → **Protein** → **Phenotype**

Central Paradigm of Medicine

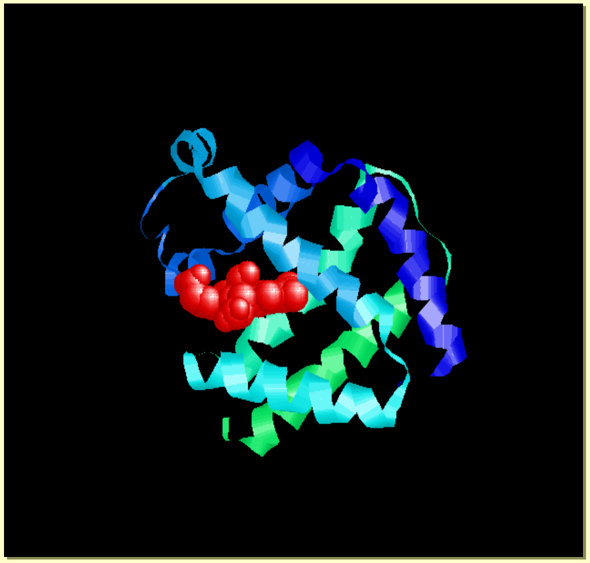


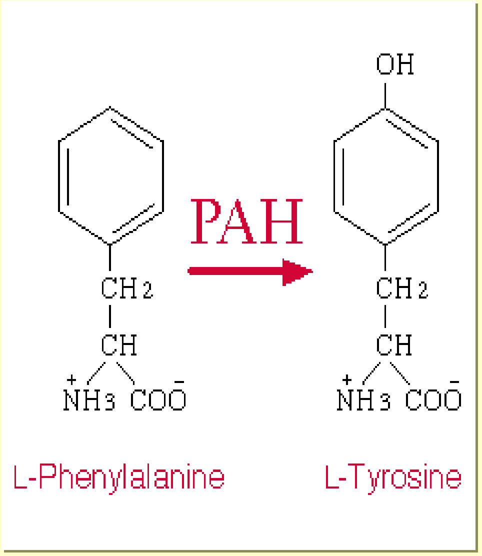
DNA → **RNA** → **Protein** → **Symptoms** → **Opinions**

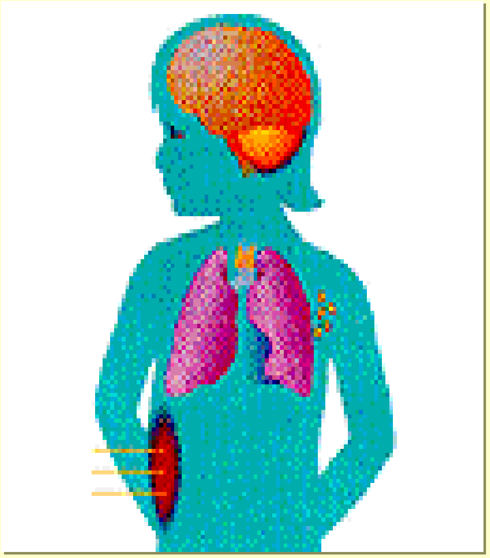
Central Paradigm of Bioinformatics

Genetic Information → Molecular Structure → Biochemical Function → Phenotype (Symptoms)

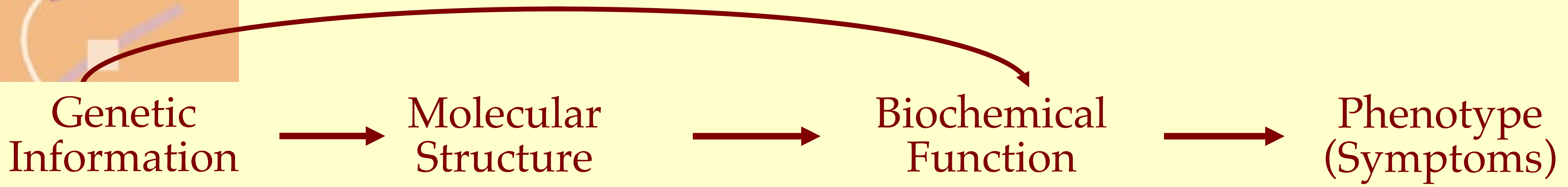
MVHLTPEEKT
AVNALWGKVN
VDAVGGEALG
RLLVVYPWTQ
RFFESFGDLS
SPDAVMGNPK
VKAHGKKVLG
AFSDGLAHLD
NLKGTFSQLS
ELHCDKLHVD
PENFRLLGNV
LVCVLARNFG
KEFTPQMQAA
YQKVVAGVAN
ALAHKYH





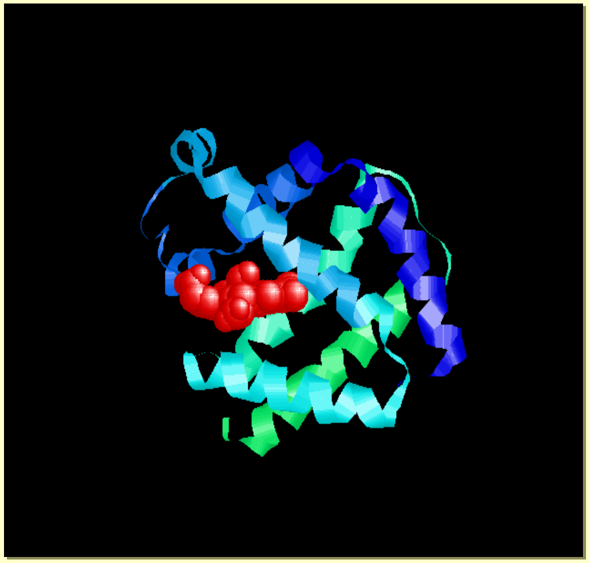


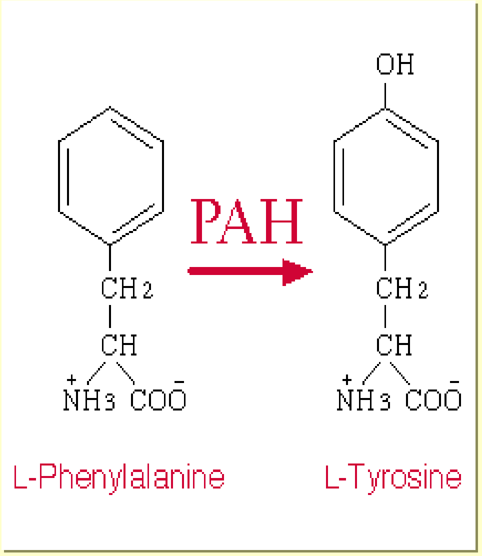
Central Paradigm of Bioinformatics

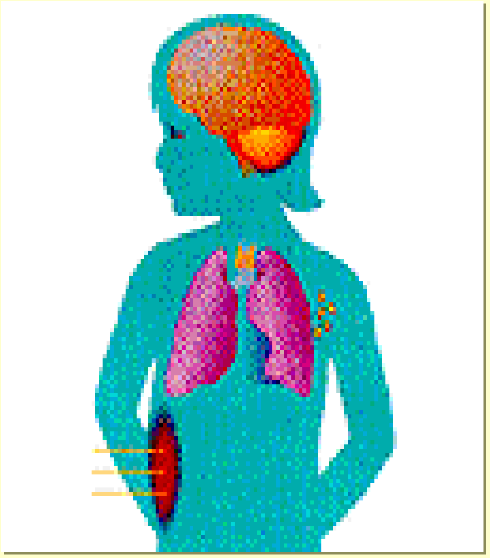


Genetic Information → Molecular Structure → Biochemical Function → Phenotype (Symptoms)

MVHLTPEEKT
AVNALWGKVN
VDAVGGEALG
RLLVVYPWTQ
RFFESFGDLS
SPDAVMGNPK
VKAHGKKVLG
AFSDGLAHLD
NLKGTFSQLS
ELHCDKLHVD
PENFRLLGNV
LVCVLARNFG
KEFTPQMQAA
YQKVVAGVAN
ALAHKYH





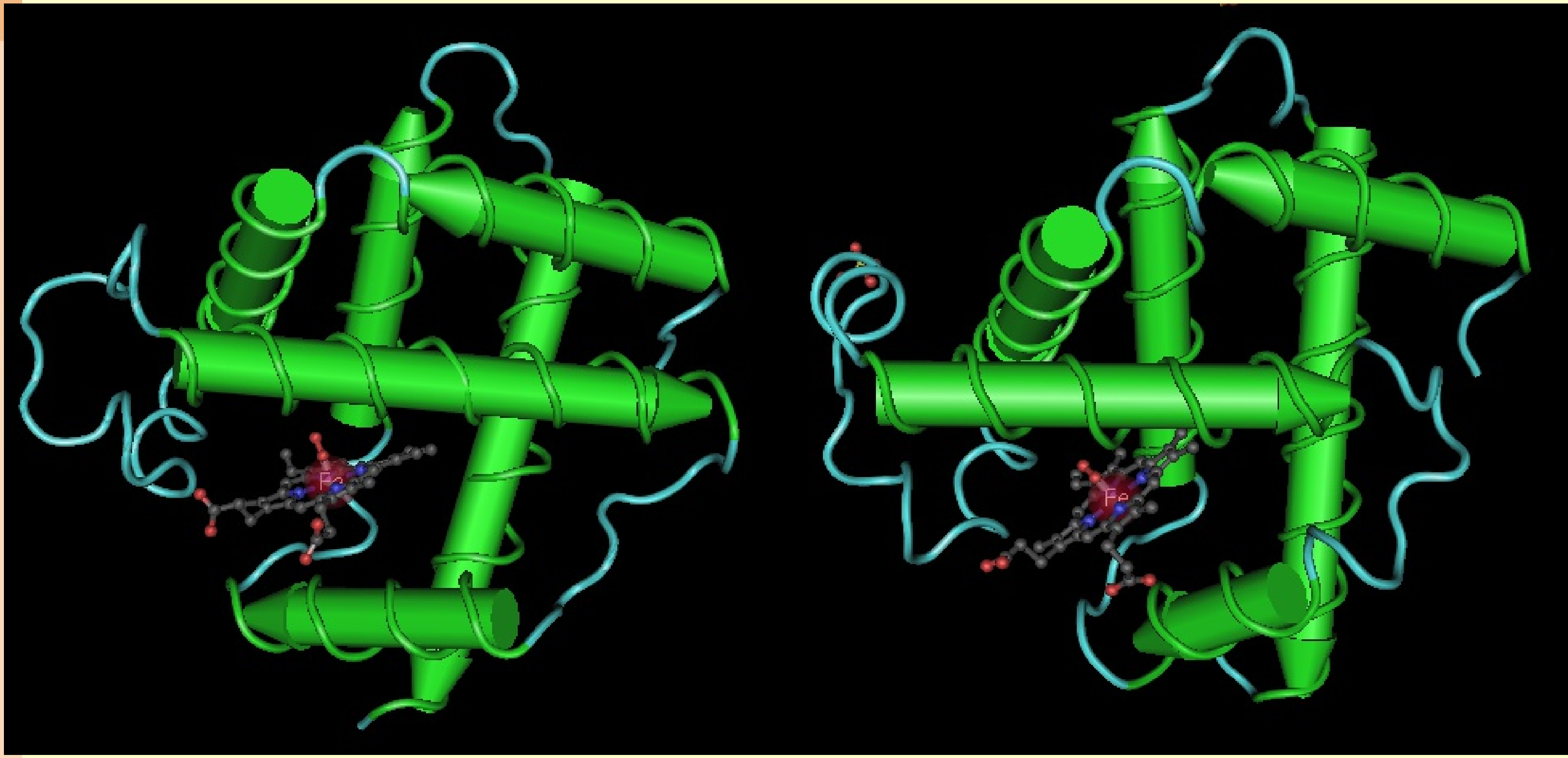


Challenges Understanding Genetic Information

Genetic Information → Molecular Structure → Biochemical Function → Phenotype

- Genetic information is redundant
- Structural information is redundant

Soybean Leghemoglobin and Sperm Whale Myoglobin



Soybean Leghemoglobin

Sperm Whale Myoglobin

Challenges Understanding Genetic Information

Genetic Information → Molecular Structure → Biochemical Function → Phenotype

- Genetic information is redundant
- Structural information is redundant
- Genes and proteins are one dimensional but their function depends on three-dimensional structure

Challenges Understanding Genetic Information

Genetic Information → Molecular Structure → Biochemical Function → Phenotype

- Genetic information is redundant
- Structural information is redundant
- Genes and proteins are one dimensional but their function depends on three-dimensional structure
- Genes and proteins are meta-stable

Discovering Function from Protein Sequence

Sequences of Common Structure or Function

Sequence Similarity

```
                     10        20        30        40              50
Query         HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
               |:| :|: | |:||||     | |:||| |: : :|:|  :|  |      |:  |
Database      VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
                      10          20        30        40        50
```

Dayhoff's PAM Matrix
Amino Acid Replacement Matrix (1978)

Cys	C	12																			
Ser	S	0	2																		
Thr	T	-2	1	3																	
Pro	P	-3	1	0	6																
Ala	A	-2	1	1	1	2															
Gly	G	-3	1	0	-1	1	5														
Asn	N	-4	1	0	-1	0	0	2													
Asp	D	-5	0	0	-1	0	1	2	4												
Glu	E	-5	0	0	-1	0	0	1	3	4											
Gln	Q	-5	-1	-1	0	0	-1	1	2	2	4										
His	H	-3	-1	-1	0	-1	-2	2	1	1	3	6									
Arg	R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
Lys	K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
Met	M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
Ile	I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
Leu	L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
Val	V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
Phe	F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Tyr	Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
Trp	W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

```
                 10        20        30        40              50
Query     VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
           |:| :|: | |:||||     | |:||| |: : :|:|  :|  |      |:  |
Database  HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
                  10          20        30        40        50
```

$$Score = \varsigma_{Region\_Start}^{Region\_End} (Similarity\_Weights)\, 2 \varsigma_{Region\_start}^{Region\_End} (Gap\_Penalties)$$

Discovering Function from Protein Sequence

Sequences of Common Structure or Function

↓

Sequence Similarity

```
                     10        20        30        40              50
Query         HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
               |:| :|: | |:||||     | |:||| |: : :|:|  :|  |      |:  | 
Database      VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
                      10          20        30        40        50
```

Discovering Function from Protein Sequence

Consensus Sequences or Sequence Motifs

Zinc Finger (C2H2 type)

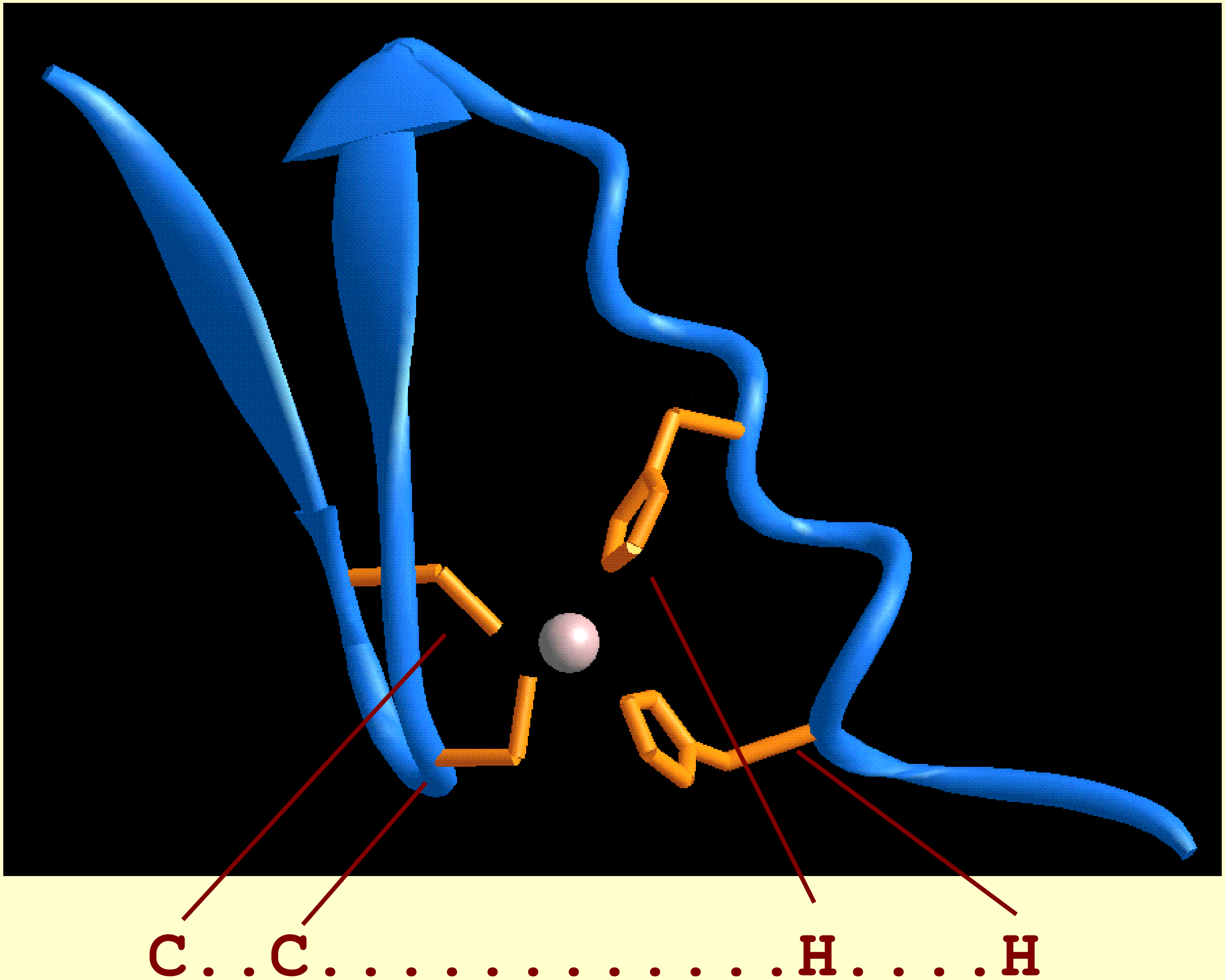
C X{2,4} C X{12} H X{3,5} H

Sequences of Common Structure or Function

Sequence Similarity

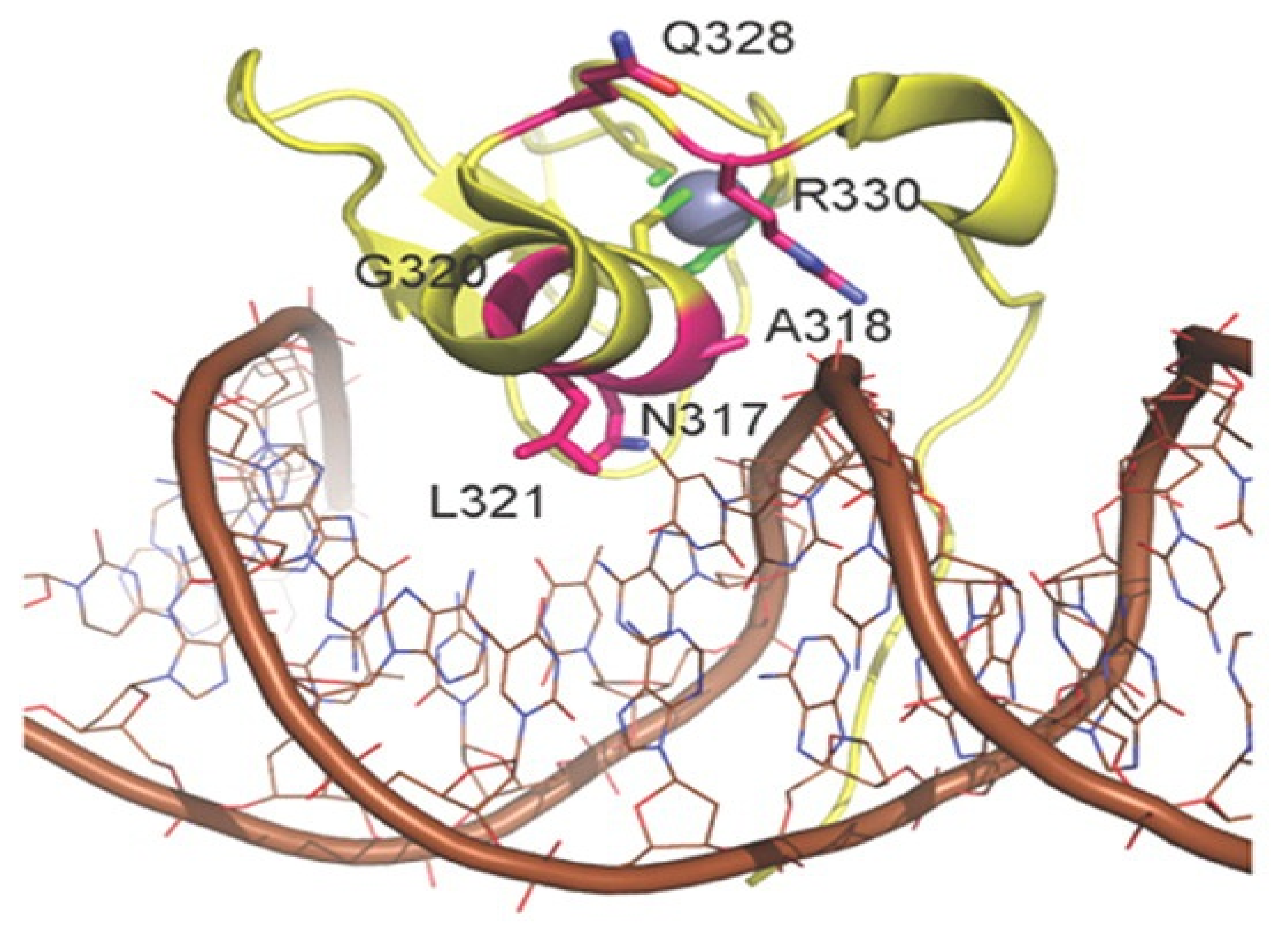
```
                 10        20        30        40              50
Query     HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
           |:| :|: | |:||||     | |:||| |: : :|:|  :|  |      |:  |
Database  VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
                  10          20        30        40        50
```

A Typical Motif:
Zinc Finger DNA Binding Motif



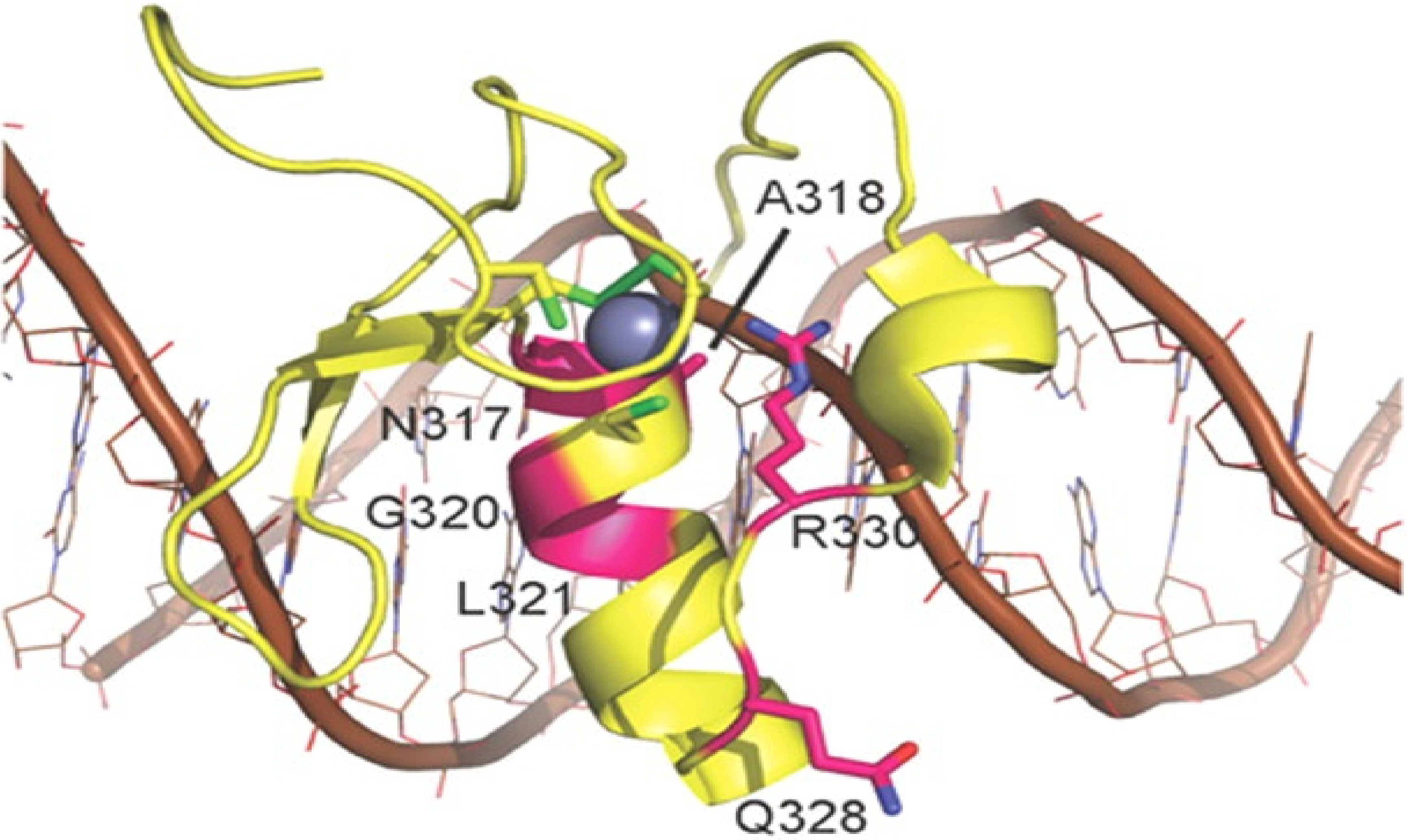
© Doug Brutlag 2015

Zinc Finger DNA Binding Motif



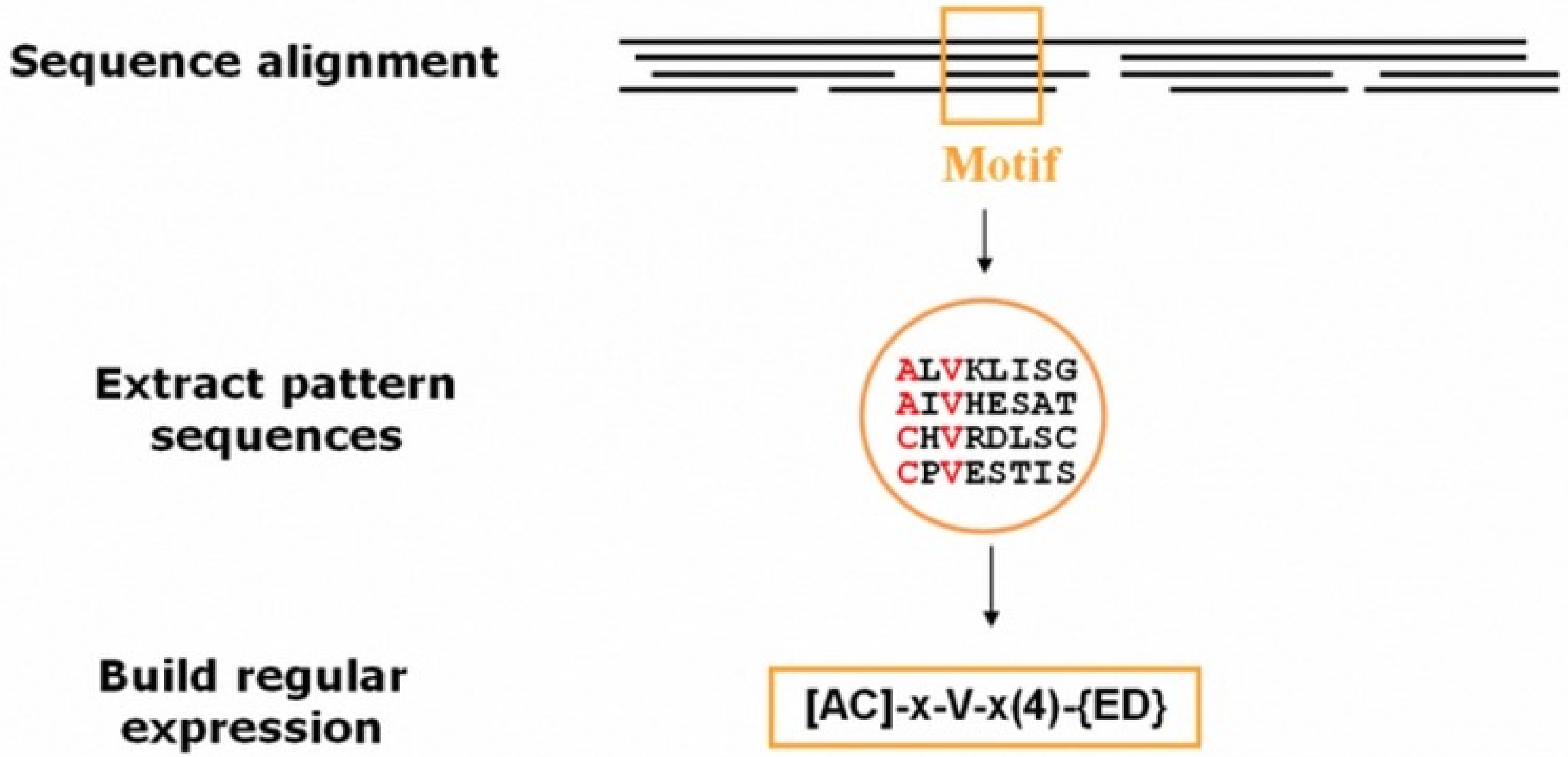
Greif et al, Blood (July 12, 2012) 120

Zinc Finger DNA Binding Motif



Greif et al, Blood (July 12, 2012) 120

Protein Motifs from Multiple Sequence Alignments



Sequence alignment

Motif

Extract pattern sequences

ALVKLISG
AIVHESAT
CHVRDLSC
CPVESTIS

Build regular expression

[AC]-x-V-x(4)-{ED}

EBI Course on Protein Motifs/Signatures
http://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi

PROSITE Patterns

http://expasy.org/prosite/

- Active site of trypsin-like serine proteases

G D S G G

- Zinc Finger (C$_2$H$_2$ type)

C-X(2,4)-C-X(12)-H-X(3,5)-H

- N-Glycosylation Site

N-[^P]-[S T]-[^P]

- Homeobox Domain Signature

[LIVMF]-X(5)-[LIVM]-X(4)-[IV]-[RKQ]-X-W-X(8)-[RK]

Discovering Function from Protein Sequence

Consensus Sequences or Sequence Motifs

Zinc Finger (C2H2 type)
C X{2,4} C X{12} H X{3,5} H

Sequences of Common Structure or Function

Sequence Similarity

```
                  10          20        30        40        50
Query     HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
           |:| :|: | |:||||     | |:||| |: : :|:|  :|  |      |:  |
Database  VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
                  10        20        30        40              50
```

Discovering Function from Protein Sequence

BLOCKs, PRINTs, PSSMS or Weight Matrices

Position

	1	2	3	4	5	6	7	8	9	10	11	12
A	2	1	3	13	10	12	67	4	13	9	1	2
R	7	5	8	9	4	0	1	16	7	0	1	0
N	0	8	0	1	0	0	0	2	1	1	10	0
D	0	1	0	1	13	0	0	12	1	0	4	0
C	0	0	1	0	0	0	0	0	0	2	2	1
Q	1	1	21	8	10	0	0	7	6	0	0	2
E	2	0	0	9	21	0	0	15	7	3	3	0
G	9	7	1	4	0	0	8	0	0	0	46	0
H	4	3	1	1	2	0	0	2	2	0	5	0
I	10	0	11	1	2	10	0	4	9	3	0	16
L	16	1	17	0	1	31	0	3	11	24	0	14
K	3	4	5	10	11	1	1	13	10	0	5	2
M	7	1	1	0	0	0	0	0	5	7	1	8
F	4	0	3	0	0	4	0	0	0	10	0	0
P	0	6	0	1	0	0	0	0	0	0	0	0
S	1	17	0	8	3	1	3	0	2	2	2	0
T	5	22	3	11	1	5	0	2	2	2	0	5
W	2	0	0	0	0	0	0	0	0	1	0	1
Y	1	0	4	2	0	1	0	0	2	4	0	1
V	6	3	1	1	2	15	0	0	2	12	0	28

Consensus Sequences or Sequence Motifs

Zinc Finger (C2H2 type)

C X{2,4} C X{12} H X{3,5} H

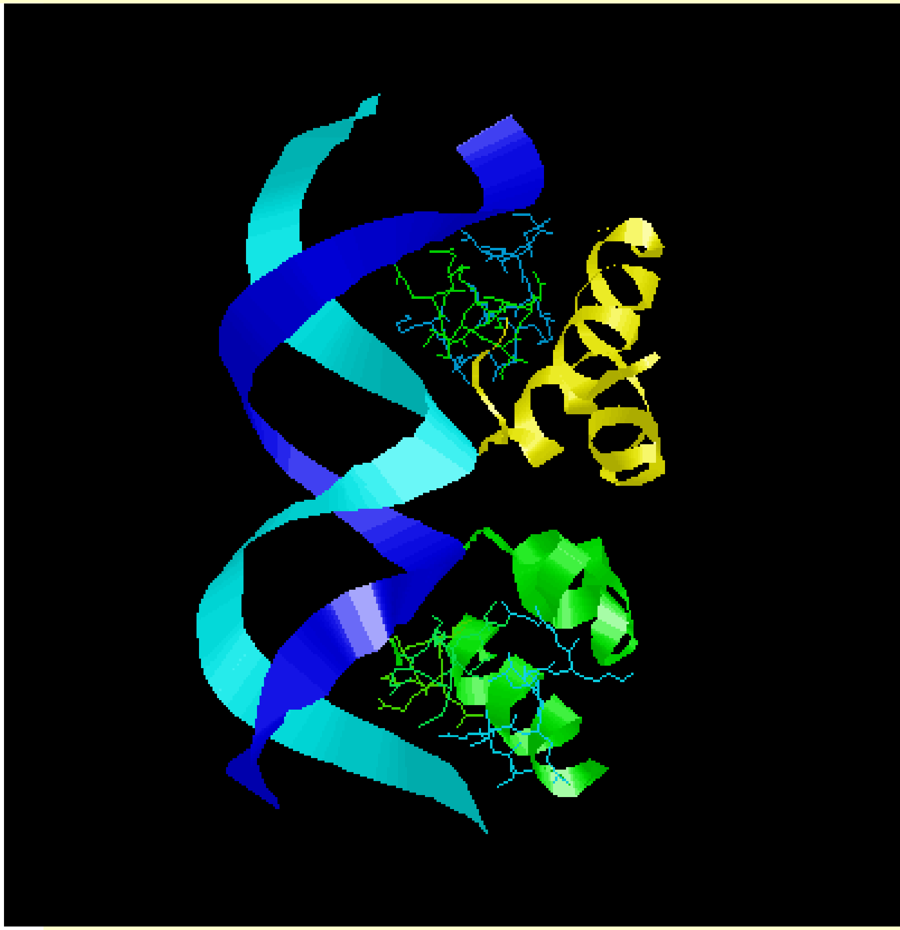
Sequences of Common Structure or Function

Sequence Similarity

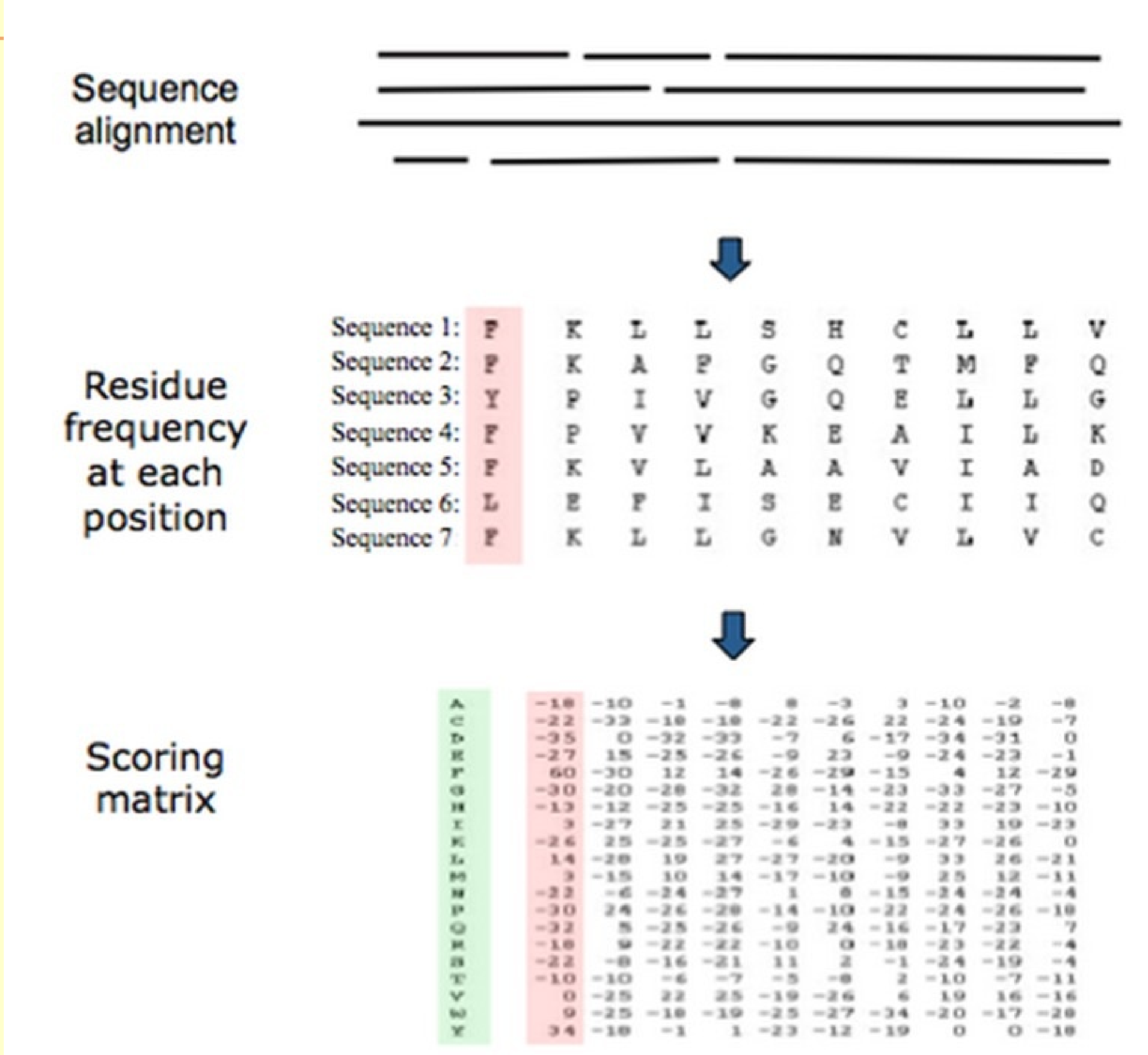
```
                      10        20        30        40             50
Query         HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
               |:| :|: | |:||||     | |:||| |: : :|:|  :|  |      |:  |
Database      VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
                       10          20        30        40        50
```

Position-Specific Scoring Matrix for Prokaryotic Helix-Turn-Helix Motifs

Sequence	Helix											Turn			Helix							
RCRO_LAMBD	F	G	Q	T	K	T	**A**	K	D	L	**G**	V	Y	Q	S	A	**I**	N	K	A	I	H
RCRO_BP434	M	T	Q	T	E	L	**A**	T	K	A	**G**	V	K	Q	Q	S	**I**	Q	L	I	E	A
RCRO_BPP22	G	T	Q	R	A	V	**A**	K	A	L	**G**	I	S	D	A	A	**V**	S	Q	W	K	E
RPC1_LAMBD	L	S	Q	E	S	V	**A**	D	K	M	**G**	M	G	Q	S	G	**V**	G	A	L	F	N
RPC1_BP434	L	N	Q	A	E	L	**A**	Q	K	V	**G**	T	T	Q	Q	S	**I**	E	Q	L	E	N
RPC1_BPP22	I	R	Q	A	A	L	**G**	K	M	V	**G**	V	S	N	V	A	**I**	S	Q	W	E	R
RPC2_LAMBD	L	G	T	E	K	T	**A**	E	A	V	**G**	V	D	K	S	Q	**I**	S	R	W	K	R
LACR_ECOLI	V	T	L	Y	D	V	**A**	E	Y	A	**G**	V	S	Y	Q	T	**V**	S	R	V	V	N
CRP_ECOLI	I	T	Q	Q	E	I	**G**	Q	I	V	**G**	C	S	R	E	T	**V**	G	R	I	L	K
TRPR_ECOLI	M	S	Q	R	E	L	**K**	N	E	L	**G**	A	G	I	A	T	**I**	T	R	G	S	N
RPC1_CPP22	R	G	Q	R	K	V	**A**	D	A	L	**G**	I	N	E	S	Q	**I**	S	R	W	K	G
GALR_ECOLI	A	T	I	K	D	V	**A**	R	L	A	**G**	V	S	V	A	T	**V**	S	R	V	I	N
Y77_BPT7	L	S	H	R	S	L	**G**	E	L	Y	**G**	V	S	Q	S	T	**I**	T	R	I	L	Q
TER3_ECOLI	L	T	T	R	K	L	**A**	Q	K	L	**G**	V	E	Q	P	T	**L**	Y	W	H	V	K
VIVB_BPT7	D	Y	Q	A	I	F	**A**	Q	Q	L	**G**	G	T	Q	S	A	**A**	S	Q	I	D	E
DEOR_ECOLI	L	H	L	K	D	A	**A**	A	L	L	**G**	V	S	E	M	T	**I**	R	R	D	L	N
RP32_BACSU	R	T	L	E	E	V	**G**	K	V	F	**G**	V	T	R	E	R	**I**	R	Q	I	E	A
Y28_BPT7	E	S	N	V	S	L	**A**	R	T	Y	**G**	V	S	Q	Q	T	**I**	C	D	I	R	K
IMMRE_BPPH	S	T	L	E	A	V	**A**	G	A	L	**G**	I	Q	V	S	A	**I**	V	G	E	E	T



Blocks or Finger Prints from Multiple Sequence Alignments



Sequence alignment

Residue frequency at each position

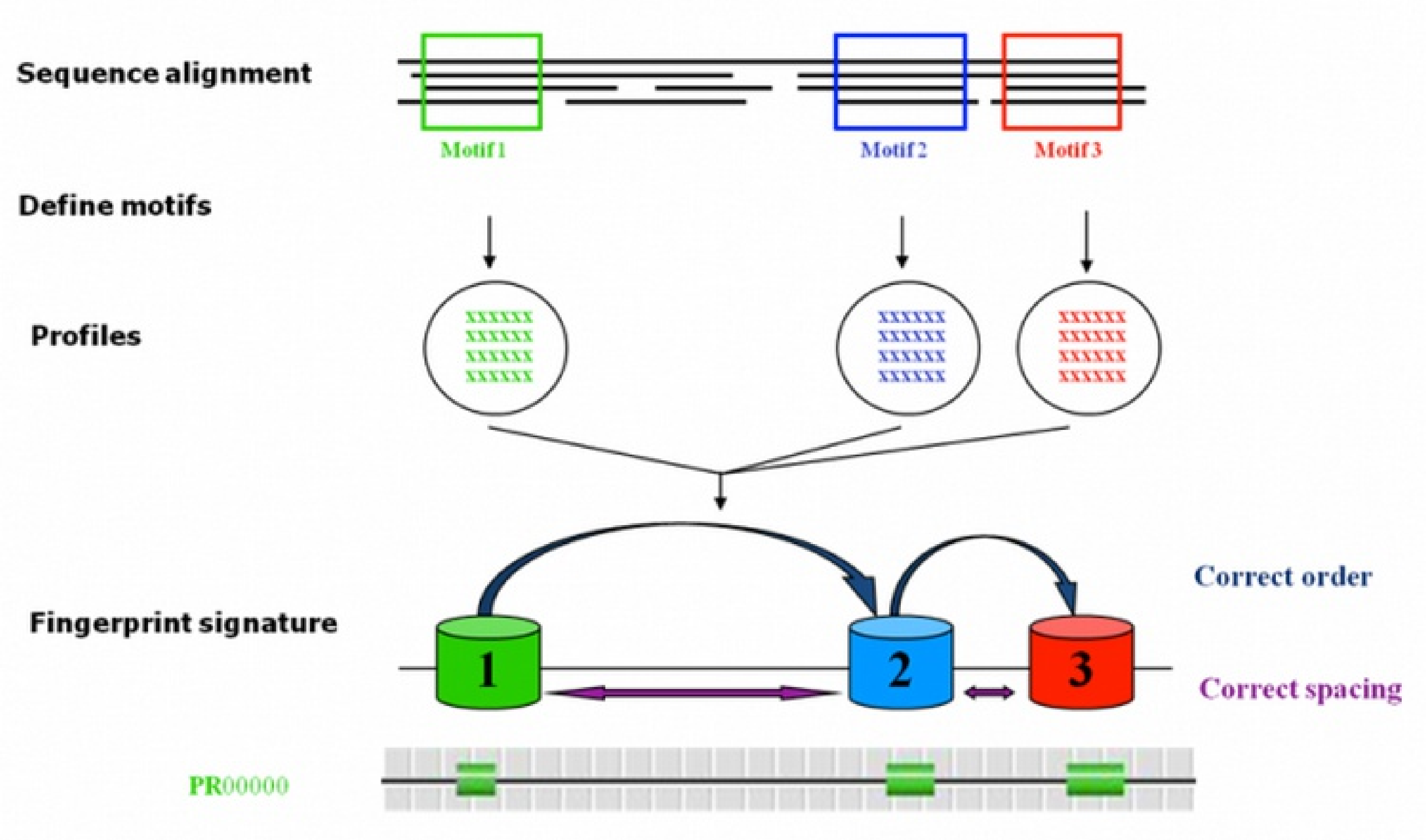
Sequence 1:	F	K	L	L	S	H	C	L	L	V
Sequence 2:	F	K	A	F	G	Q	T	M	F	Q
Sequence 3:	Y	P	I	V	G	Q	E	L	L	G
Sequence 4:	F	P	V	V	K	E	A	I	L	K
Sequence 5:	F	K	V	L	A	A	V	I	A	D
Sequence 6:	L	E	F	I	S	E	C	I	I	Q
Sequence 7	F	K	L	L	G	N	V	L	V	C

Scoring matrix

A	-18	-10	-1	-8	8	-3	3	-10	-2	-8
C	-22	-33	-18	-18	-22	-26	22	-24	-19	-7
D	-35	0	-32	-33	-7	6	-17	-34	-31	0
E	-27	15	-25	-26	-9	23	-9	-24	-23	-1
F	60	-30	12	14	-26	-29	-15	4	12	-29
G	-30	-20	-28	-32	28	-14	-23	-33	-27	-5
H	-13	-12	-25	-25	-16	14	-22	-22	-23	-10
I	3	-27	21	25	-29	-23	-8	33	19	-23
K	-26	25	-25	-27	-6	4	-15	-27	-26	0
L	14	-28	19	27	-27	-20	-9	33	26	-21
M	3	-15	10	14	-17	-10	-9	25	12	-11
N	-22	-6	-24	-27	1	8	-15	-24	-24	-4
P	-30	24	-26	-28	-14	-10	-22	-24	-26	-18
Q	-32	5	-25	-26	-9	24	-16	-17	-23	7
R	-18	9	-22	-22	-10	0	-18	-23	-22	-4
S	-22	-8	-16	-21	11	2	-1	-24	-19	-4
T	-10	-10	-6	-7	-5	-8	2	-10	-7	-11
V	0	-25	22	25	-19	-26	6	19	16	-16
W	9	-25	-18	-19	-25	-27	-34	-20	-17	-28
Y	34	-18	-1	1	-23	-12	-19	0	0	-18

EBI Course on Protein Motifs/Signatures
http://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi

Finger Prints from Multiple Sequence Alignments



EBI Course on Protein Motifs/Signatures
http://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi

Discovering Function from Protein Sequence

BLOCKs, PRINTs, PSSMS or Weight Matrices

Position

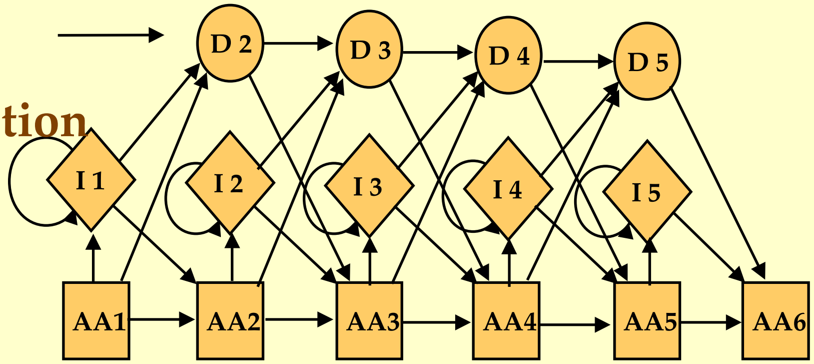
	1	2	3	4	5	6	7	8	9	10	11	12
A	2	1	3	13	10	12	67	4	13	9	1	2
R	7	5	8	9	4	0	1	16	7	0	1	0
N	0	8	0	1	0	0	0	2	1	1	10	0
D	0	1	0	1	13	0	0	12	1	0	4	0
C	0	0	1	0	0	0	0	0	0	2	2	1
Q	1	1	21	8	10	0	0	7	6	0	0	2
E	2	0	0	9	21	0	0	15	7	3	3	0
G	9	7	1	4	0	0	8	0	0	0	46	0
H	4	3	1	1	2	0	0	2	2	0	5	0
I	10	0	11	1	2	10	0	4	9	3	0	16
L	16	1	17	0	1	31	0	3	11	24	0	14
K	3	4	5	10	11	1	1	13	10	0	5	2
M	7	1	1	0	0	0	0	0	5	7	1	8
F	4	0	3	0	0	4	0	0	0	10	0	0
P	0	6	0	1	0	0	0	0	0	0	0	0
S	1	17	0	8	3	1	3	0	2	2	2	0
T	5	22	3	11	1	5	0	2	2	2	0	5
W	2	0	0	0	0	0	0	0	0	1	0	1
Y	1	0	4	2	0	1	0	0	2	4	0	1
V	6	3	1	1	2	15	0	0	2	12	0	28

Consensus Sequences or Sequence Motifs

Zinc Finger (C2H2 type)
C X{2,4} C X{12} H X{3,5} H

Sequences of Common Structure or Function

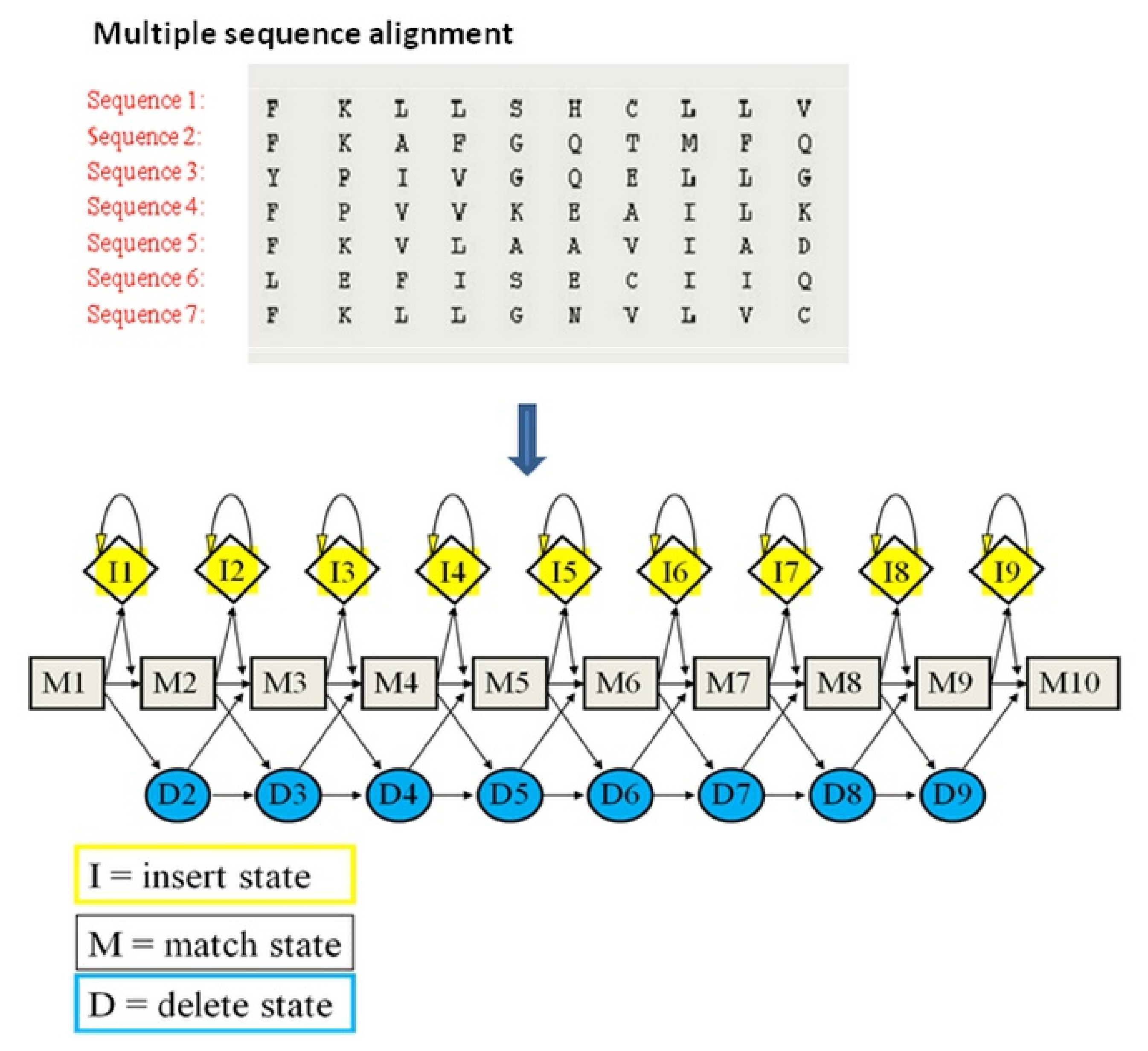
Profiles, PSI-BLAST Hidden Markov Models



Sequence Similarity

```
                   10       20         30       40              50
Query       HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
             |:| :|: | |:||||     | |:||| |: : :|:|  :|  |      |:  |
Database    VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
                    10         20         30       40        50
```

Hidden Markov Models from Multiple Sequence Alignments



Multiple sequence alignment

Sequence 1:	F	K	L	L	S	H	C	L	L	V
Sequence 2:	F	K	A	F	G	Q	T	M	F	Q
Sequence 3:	Y	P	I	V	G	Q	E	L	L	G
Sequence 4:	F	P	V	V	K	E	A	I	L	K
Sequence 5:	F	K	V	L	A	A	V	I	A	D
Sequence 6:	L	E	F	I	S	E	C	I	I	Q
Sequence 7:	F	K	L	L	G	N	V	L	V	C

I = insert state

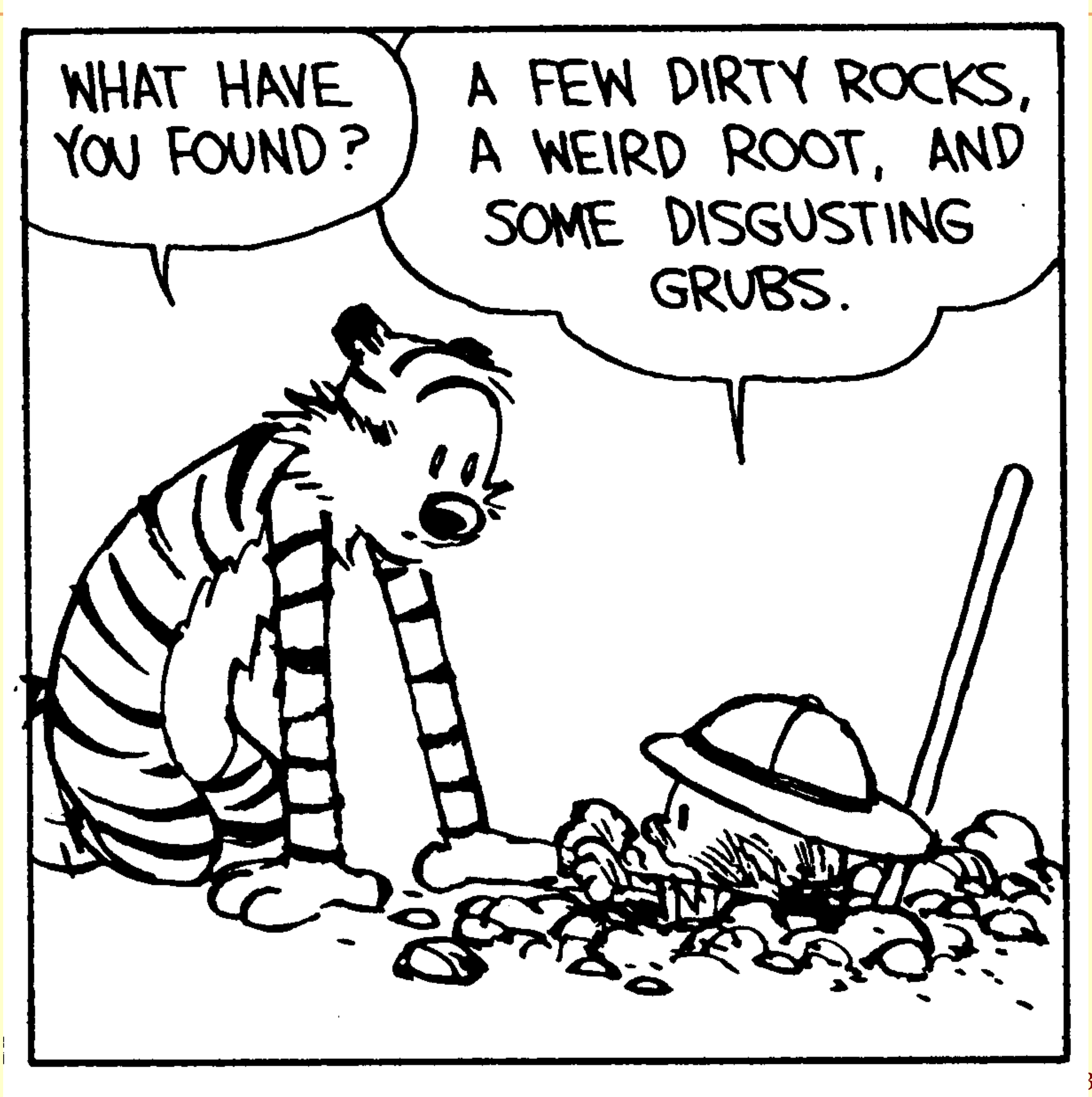
M = match state

D = delete state

EBI Course on Protein Motifs/Signatures
http://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi

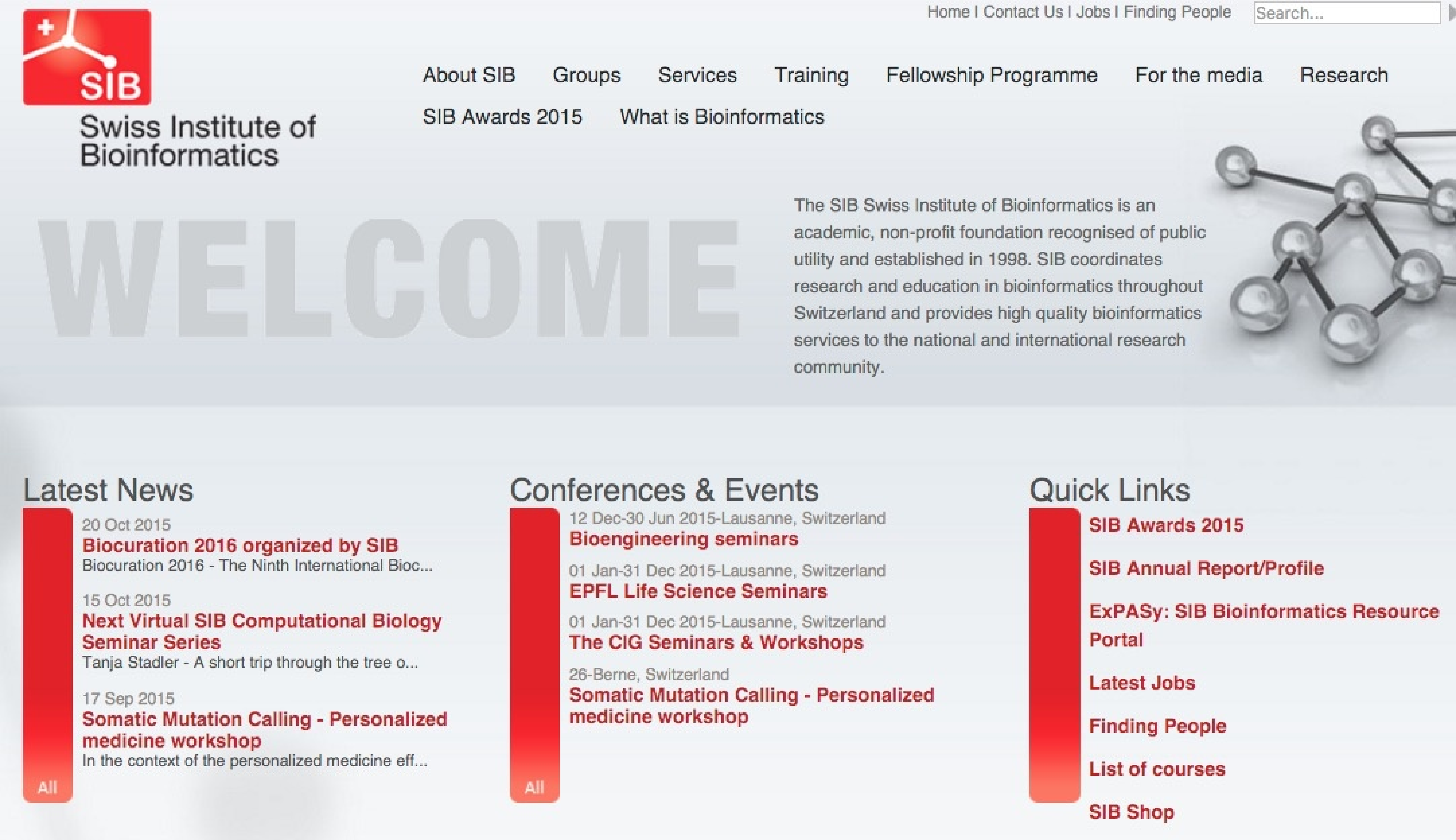
Data Mining: The Search for Buried Treasure

Data Mining:
The Search for Buried Treasure



Data Mining:
The Search for Buried Treasure

Swiss Institute of Bioinformatics
http://www.isb-sib.ch/



The SIB Swiss Institute of Bioinformatics is an academic, non-profit foundation recognised of public utility and established in 1998. SIB coordinates research and education in bioinformatics throughout Switzerland and provides high quality bioinformatics services to the national and international research community.

Latest News

20 Oct 2015
Biocuration 2016 organized by SIB
Biocuration 2016 - The Ninth International Bioc...

15 Oct 2015
Next Virtual SIB Computational Biology Seminar Series
Tanja Stadler - A short trip through the tree o...

17 Sep 2015
Somatic Mutation Calling - Personalized medicine workshop
In the context of the personalized medicine eff...

All

Conferences & Events

12 Dec-30 Jun 2015-Lausanne, Switzerland
Bioengineering seminars

01 Jan-31 Dec 2015-Lausanne, Switzerland
EPFL Life Science Seminars

01 Jan-31 Dec 2015-Lausanne, Switzerland
The CIG Seminars & Workshops

26-Berne, Switzerland
Somatic Mutation Calling - Personalized medicine workshop

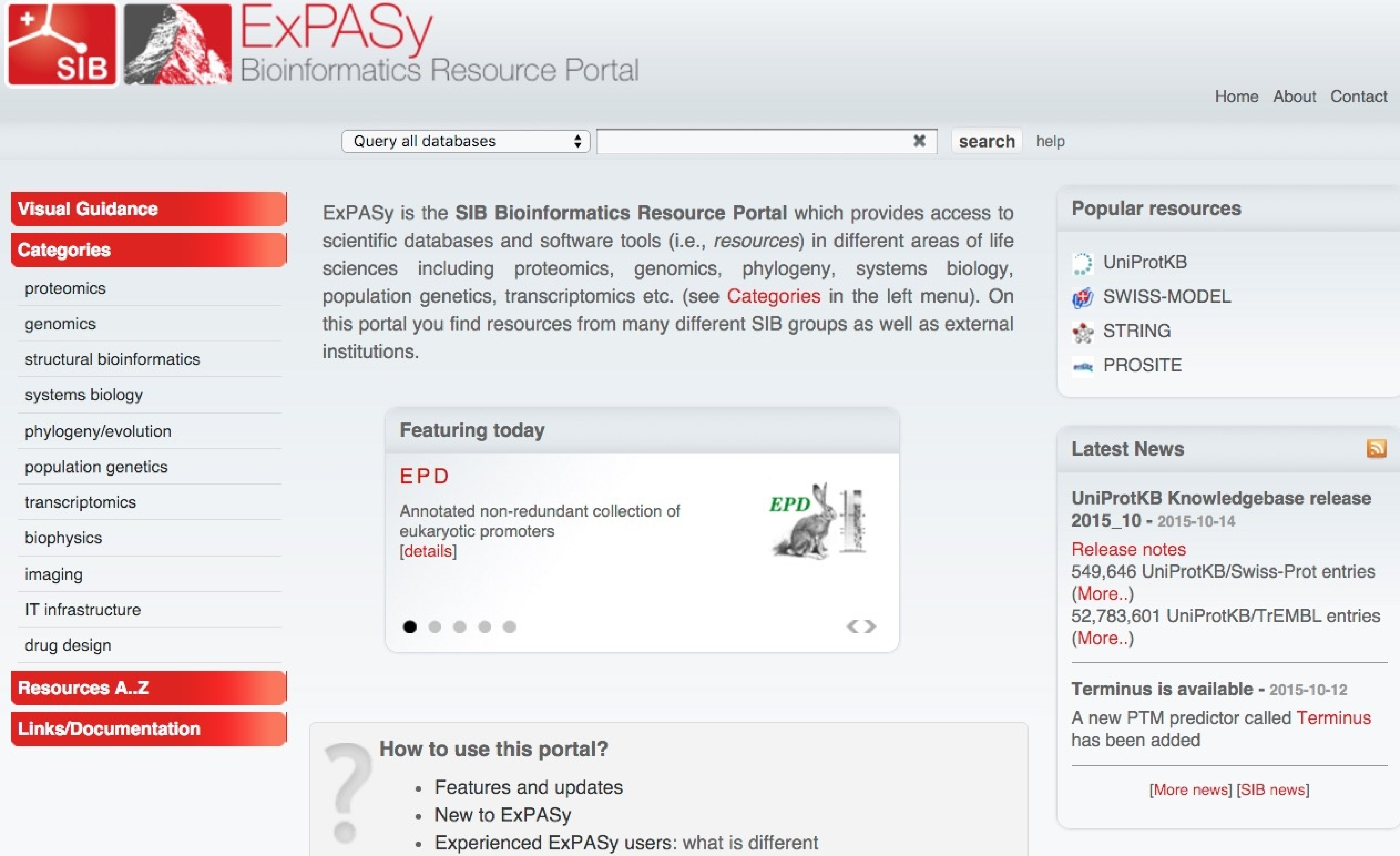
All

Quick Links

- **SIB Awards 2015**
- **SIB Annual Report/Profile**
- **ExPASy: SIB Bioinformatics Resource Portal**
- **Latest Jobs**
- **Finding People**
- **List of courses**
- **SIB Shop**

Expasy Bioinformatics Resource Portal

http://expasy.org/



ExPASy is the **SIB Bioinformatics Resource Portal** which provides access to scientific databases and software tools (i.e., *resources*) in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc. (see Categories in the left menu). On this portal you find resources from many different SIB groups as well as external institutions.

Featuring today

EPD

Annotated non-redundant collection of eukaryotic promoters
[details]

Latest News

UniProtKB Knowledgebase release 2015_10 - 2015-10-14

Release notes
549,646 UniProtKB/Swiss-Prot entries (More..)
52,783,601 UniProtKB/TrEMBL entries (More..)

Terminus is available - 2015-10-12

A new PTM predictor called Terminus has been added

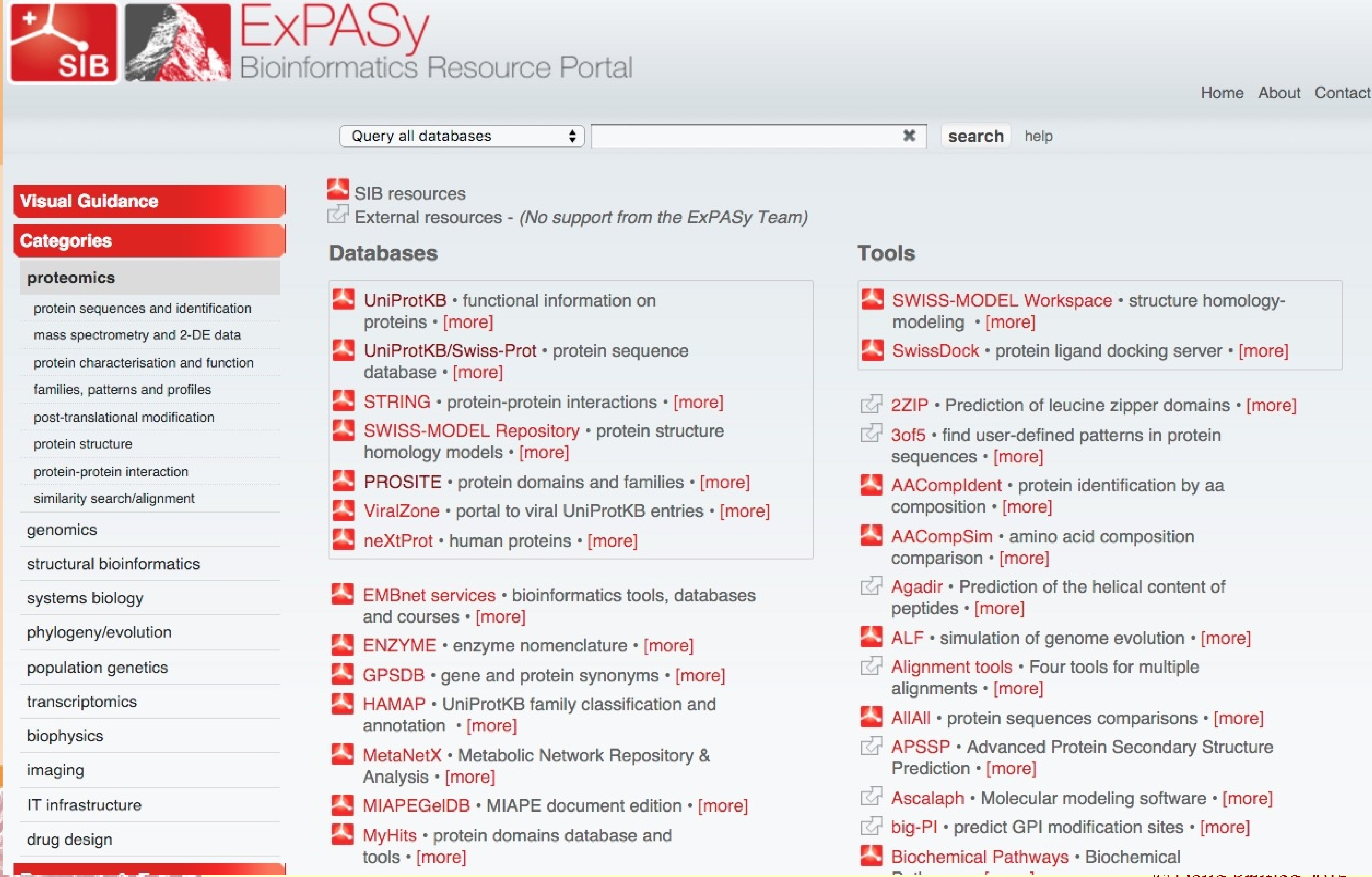
[More news] [SIB news]

How to use this portal?

- Features and updates
- New to ExPASy
- Experienced ExPASy users: what is different

Expasy Proteomics Resource Portal

http://expasy.org/proteomics



proteomics

- protein sequences and identification
- mass spectrometry and 2-DE data
- protein characterisation and function
- families, patterns and profiles
- post-translational modification
- protein structure
- protein-protein interaction
- similarity search/alignment

- genomics
- structural bioinformatics
- systems biology
- phylogeny/evolution
- population genetics
- transcriptomics
- biophysics
- imaging
- IT infrastructure
- drug design

🟥 SIB resources
↗ External resources - *(No support from the ExPASy Team)*

Databases

- 🟥 UniProtKB • functional information on proteins • [more]
- 🟥 UniProtKB/Swiss-Prot • protein sequence database • [more]
- 🟥 STRING • protein-protein interactions • [more]
- 🟥 SWISS-MODEL Repository • protein structure homology models • [more]
- 🟥 PROSITE • protein domains and families • [more]
- 🟥 ViralZone • portal to viral UniProtKB entries • [more]
- 🟥 neXtProt • human proteins • [more]

- 🟥 EMBnet services • bioinformatics tools, databases and courses • [more]
- 🟥 ENZYME • enzyme nomenclature • [more]
- 🟥 GPSDB • gene and protein synonyms • [more]
- 🟥 HAMAP • UniProtKB family classification and annotation • [more]
- 🟥 MetaNetX • Metabolic Network Repository & Analysis • [more]
- 🟥 MIAPEGelDB • MIAPE document edition • [more]
- 🟥 MyHits • protein domains database and tools • [more]

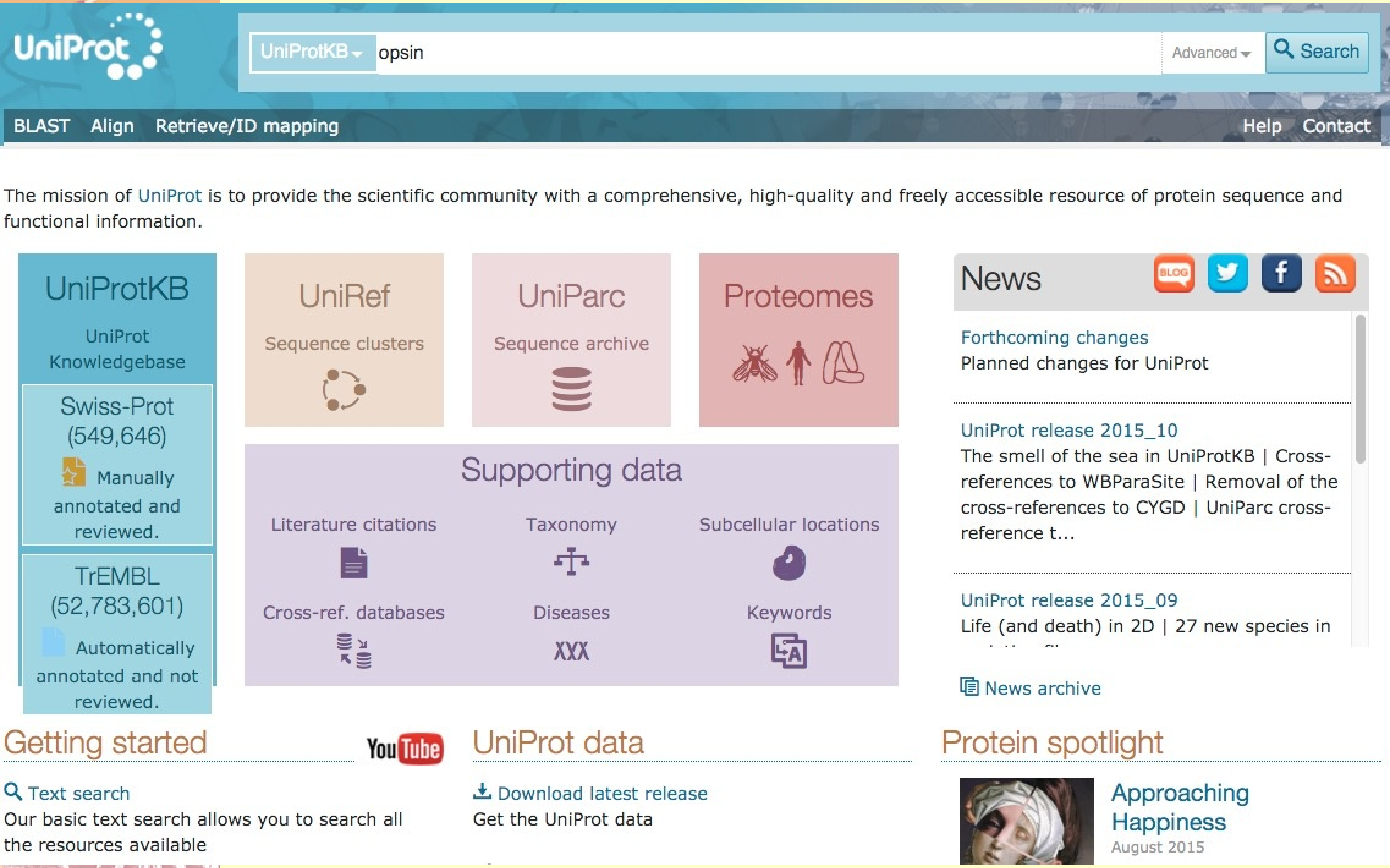
Tools

- 🟥 SWISS-MODEL Workspace • structure homology-modeling • [more]
- 🟥 SwissDock • protein ligand docking server • [more]

- ↗ 2ZIP • Prediction of leucine zipper domains • [more]
- ↗ 3of5 • find user-defined patterns in protein sequences • [more]
- 🟥 AACompIdent • protein identification by aa composition • [more]
- 🟥 AACompSim • amino acid composition comparison • [more]
- ↗ Agadir • Prediction of the helical content of peptides • [more]
- 🟥 ALF • simulation of genome evolution • [more]
- ↗ Alignment tools • Four tools for multiple alignments • [more]
- 🟥 AllAll • protein sequences comparisons • [more]
- ↗ APSSP • Advanced Protein Secondary Structure Prediction • [more]
- ↗ Ascalaph • Molecular modeling software • [more]
- ↗ big-PI • predict GPI modification sites • [more]
- 🟥 Biochemical Pathways • Biochemical

UniProt Knowledge Base

http://www.uniprot.org/



The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

Forthcoming changes

Planned changes for UniProt

UniProt release 2015_10

The smell of the sea in UniProtKB | Cross-references to WBParaSite | Removal of the cross-references to CYGD | UniParc cross-reference t...

UniProt release 2015_09

Life (and death) in 2D | 27 new species in

Text search

Our basic text search allows you to search all the resources available

Download latest release

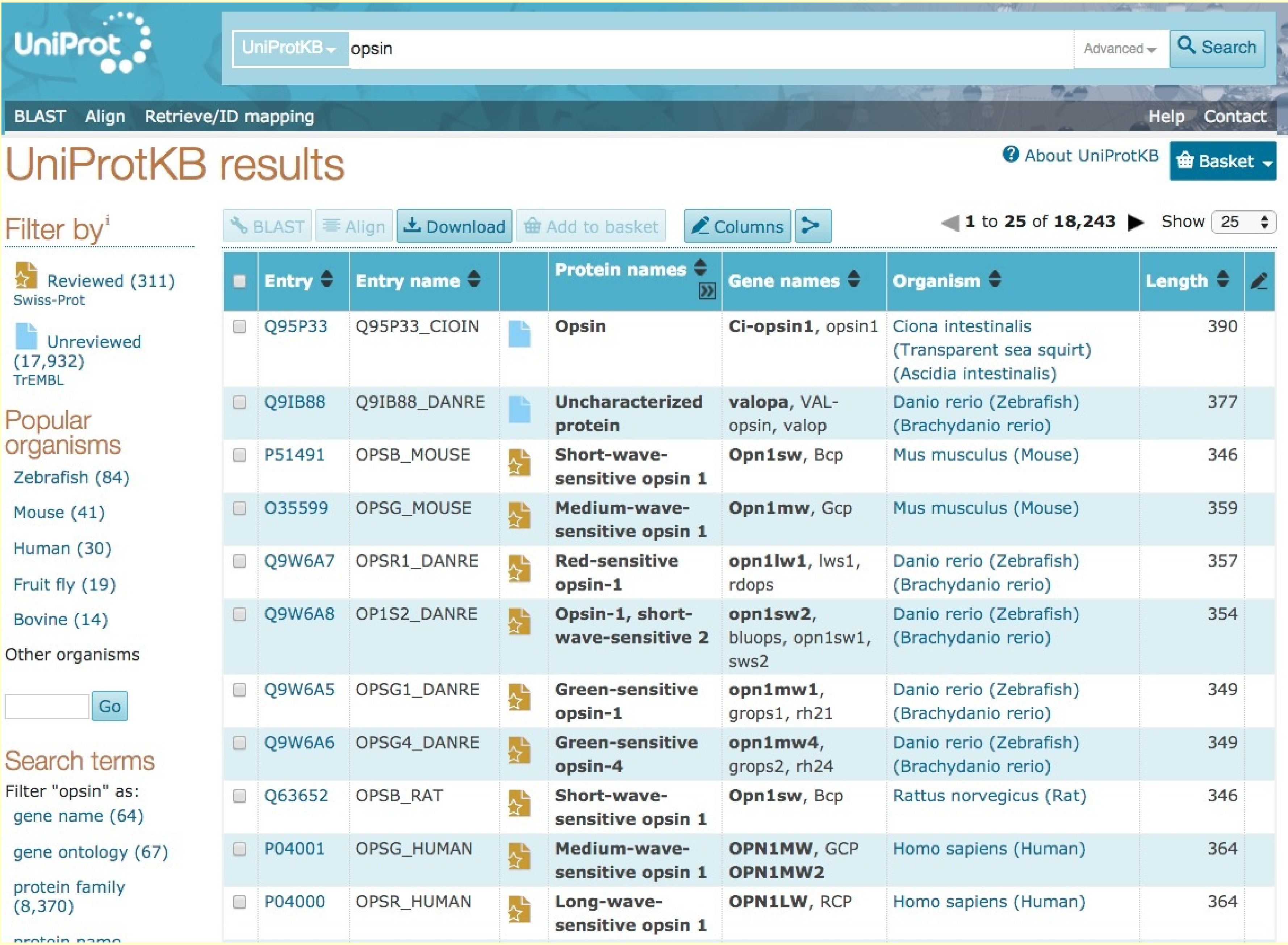
Get the UniProt data

Approaching Happiness

August 2015

UniProt Opsin Seach Results

http://www.uniprot.org/uniprot/?query=opsin&sort=score



UniProtKB results

Filter by

Reviewed (311)
Swiss-Prot

Unreviewed (17,932)
TrEMBL

Popular organisms

Zebrafish (84)
Mouse (41)
Human (30)
Fruit fly (19)
Bovine (14)
Other organisms

Search terms

Filter "opsin" as:

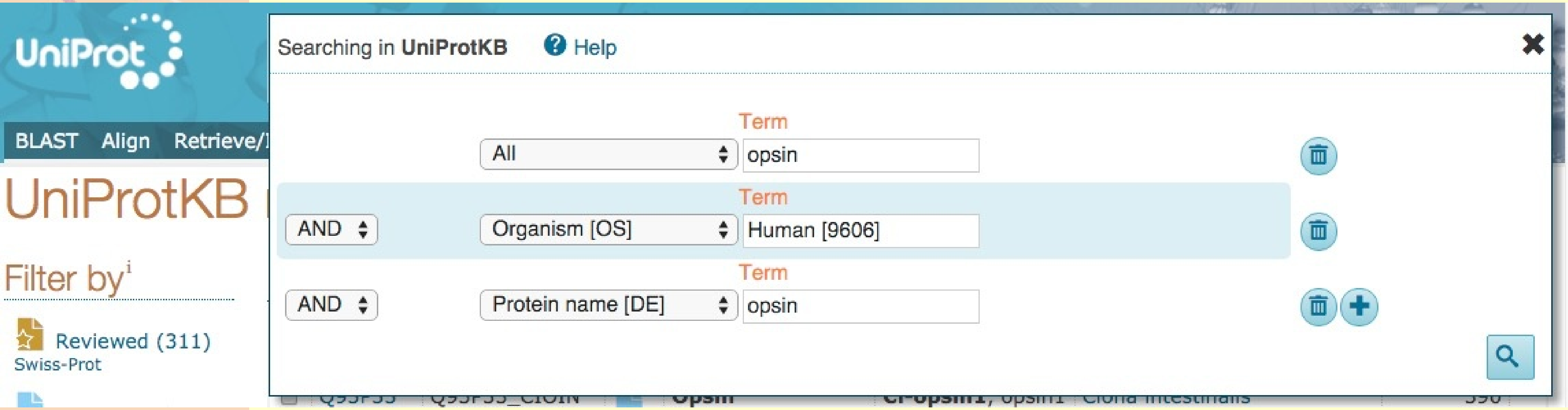
gene name (64)
gene ontology (67)
protein family (8,370)
protein name

1 to 25 of 18,243 Show 25

☐	Entry	Entry name		Protein names	Gene names	Organism	Length
☐	Q95P33	Q95P33_CIOIN	📄	**Opsin**	**Ci-opsin1**, opsin1	Ciona intestinalis (Transparent sea squirt) (Ascidia intestinalis)	390
☐	Q9IB88	Q9IB88_DANRE	📄	**Uncharacterized protein**	**valopa**, VAL-opsin, valop	Danio rerio (Zebrafish) (Brachydanio rerio)	377
☐	P51491	OPSB_MOUSE	★	**Short-wave-sensitive opsin 1**	**Opn1sw**, Bcp	Mus musculus (Mouse)	346
☐	O35599	OPSG_MOUSE	★	**Medium-wave-sensitive opsin 1**	**Opn1mw**, Gcp	Mus musculus (Mouse)	359
☐	Q9W6A7	OPSR1_DANRE	★	**Red-sensitive opsin-1**	**opn1lw1**, lws1, rdops	Danio rerio (Zebrafish) (Brachydanio rerio)	357
☐	Q9W6A8	OP1S2_DANRE	★	**Opsin-1, short-wave-sensitive 2**	**opn1sw2**, bluops, opn1sw1, sws2	Danio rerio (Zebrafish) (Brachydanio rerio)	354
☐	Q9W6A5	OPSG1_DANRE	★	**Green-sensitive opsin-1**	**opn1mw1**, grops1, rh21	Danio rerio (Zebrafish) (Brachydanio rerio)	349
☐	Q9W6A6	OPSG4_DANRE	★	**Green-sensitive opsin-4**	**opn1mw4**, grops2, rh24	Danio rerio (Zebrafish) (Brachydanio rerio)	349
☐	Q63652	OPSB_RAT	★	**Short-wave-sensitive opsin 1**	**Opn1sw**, Bcp	Rattus norvegicus (Rat)	346
☐	P04001	OPSG_HUMAN	★	**Medium-wave-sensitive opsin 1**	**OPN1MW**, GCP **OPN1MW2**	Homo sapiens (Human)	364
☐	P04000	OPSR_HUMAN	★	**Long-wave-sensitive opsin 1**	**OPN1LW**, RCP	Homo sapiens (Human)	364

UniProt Advanced Opsin Search

http://www.uniprot.org/uniprot/?query=opsin+AND+organism%3A%22Human+%5B960

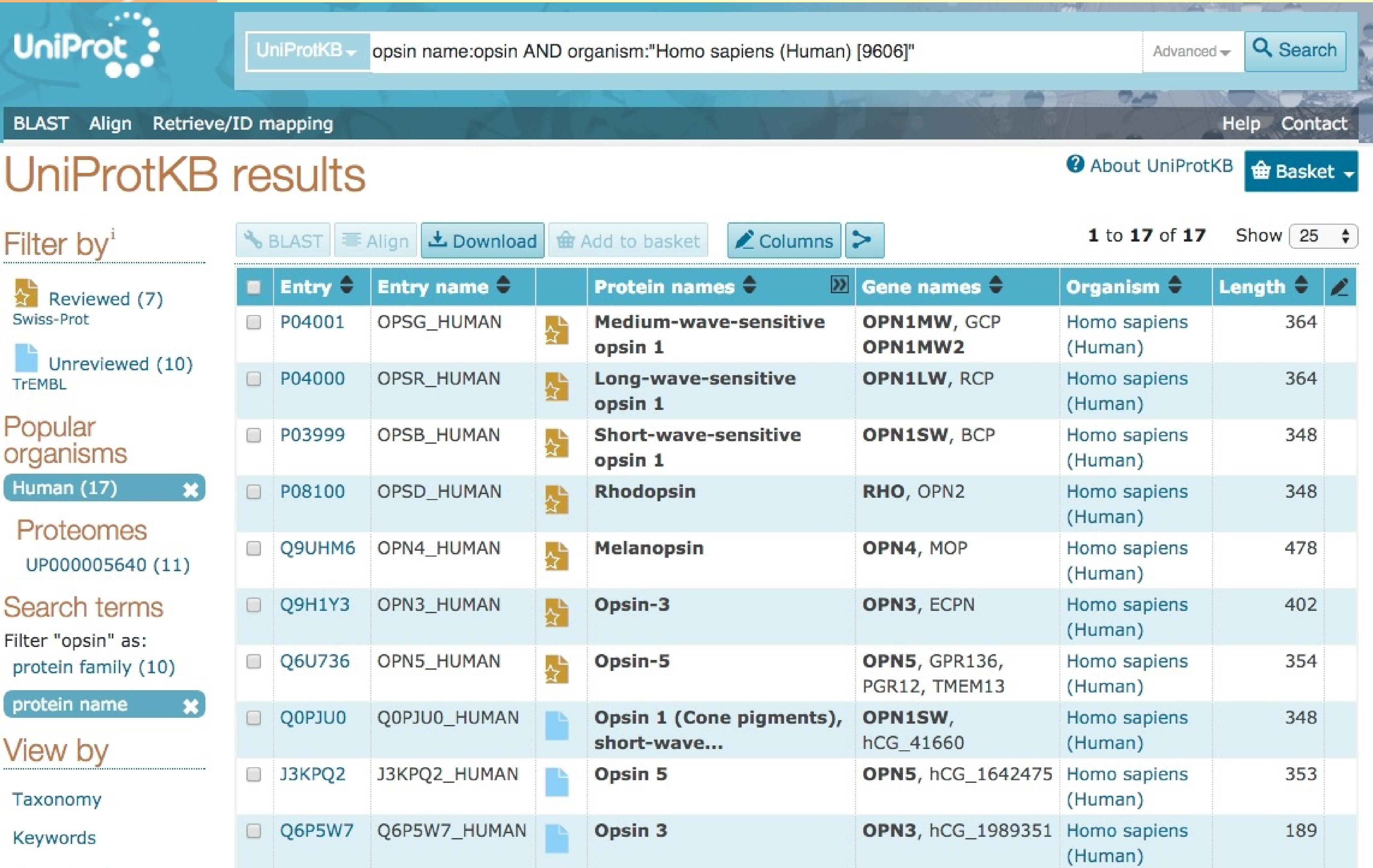


Searching in **UniProtKB**

		Term
	All	opsin
AND	Organism [OS]	Human [9606]
AND	Protein name [DE]	opsin

UniProt Human Opsin Entries

http://www.uniprot.org/



opsin name:opsin AND organism:"Homo sapiens (Human) [9606]"

UniProtKB results

1 to 17 of 17 Show 25

Filter by

Reviewed (7)
Swiss-Prot

Unreviewed (10)
TrEMBL

Popular organisms

Human (17) ✖

Proteomes

UP000005640 (11)

Search terms

Filter "opsin" as:
protein family (10)
protein name ✖

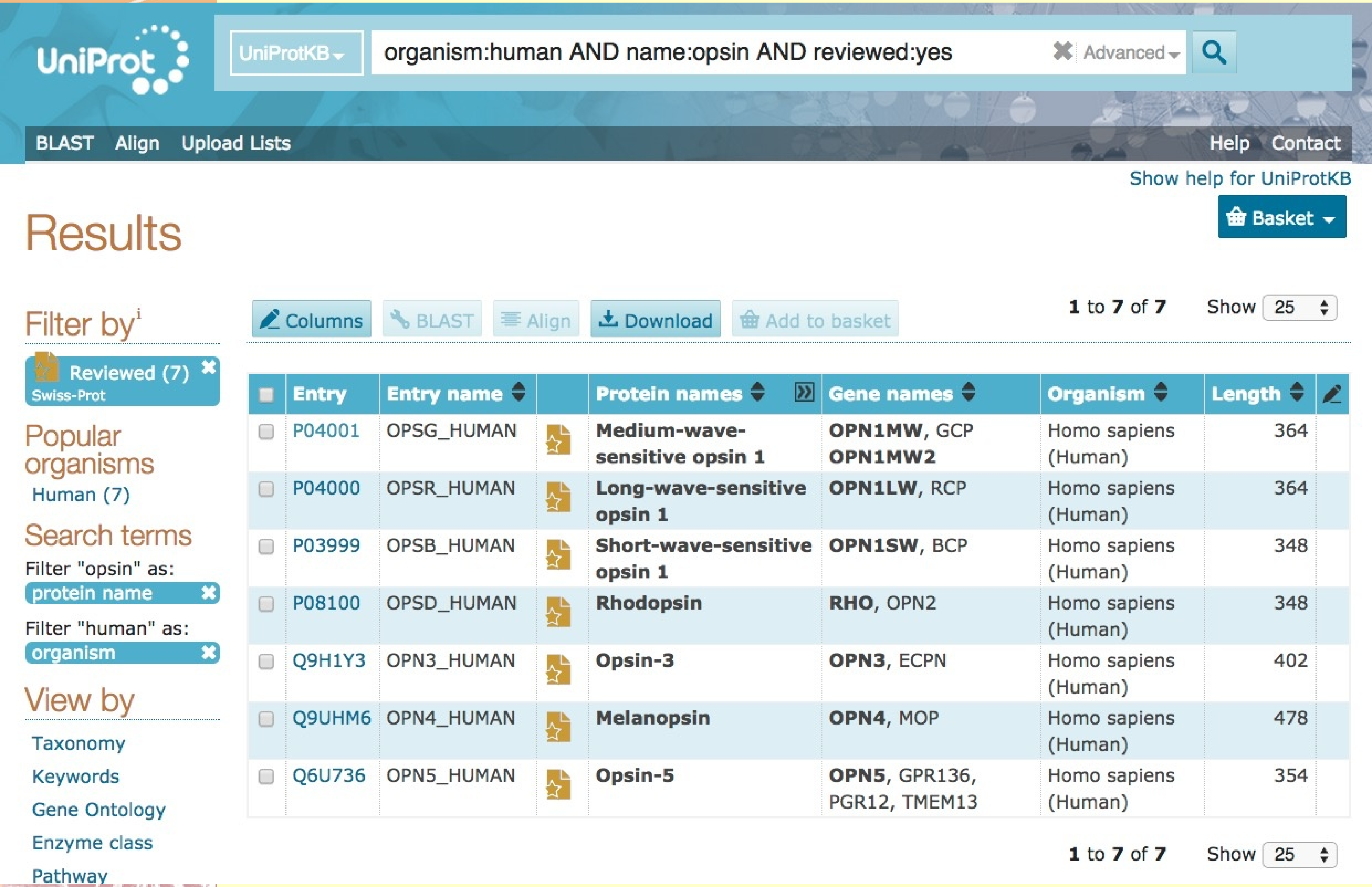
View by

Taxonomy
Keywords

☐	Entry	Entry name		Protein names	Gene names	Organism	Length
☐	P04001	OPSG_HUMAN	★	**Medium-wave-sensitive opsin 1**	**OPN1MW**, GCP **OPN1MW2**	Homo sapiens (Human)	364
☐	P04000	OPSR_HUMAN	★	**Long-wave-sensitive opsin 1**	**OPN1LW**, RCP	Homo sapiens (Human)	364
☐	P03999	OPSB_HUMAN	★	**Short-wave-sensitive opsin 1**	**OPN1SW**, BCP	Homo sapiens (Human)	348
☐	P08100	OPSD_HUMAN	★	**Rhodopsin**	**RHO**, OPN2	Homo sapiens (Human)	348
☐	Q9UHM6	OPN4_HUMAN	★	**Melanopsin**	**OPN4**, MOP	Homo sapiens (Human)	478
☐	Q9H1Y3	OPN3_HUMAN	★	**Opsin-3**	**OPN3**, ECPN	Homo sapiens (Human)	402
☐	Q6U736	OPN5_HUMAN	★	**Opsin-5**	**OPN5**, GPR136, PGR12, TMEM13	Homo sapiens (Human)	354
☐	Q0PJU0	Q0PJU0_HUMAN		**Opsin 1 (Cone pigments), short-wave...**	**OPN1SW**, hCG_41660	Homo sapiens (Human)	348
☐	J3KPQ2	J3KPQ2_HUMAN		**Opsin 5**	**OPN5**, hCG_1642475	Homo sapiens (Human)	353
☐	Q6P5W7	Q6P5W7_HUMAN		**Opsin 3**	**OPN3**, hCG_1989351	Homo sapiens (Human)	189

UniProt Human Opsin Entries Reviewed

http://www.uniprot.org/



organism:human AND name:opsin AND reviewed:yes

Filter "opsin" as: protein name

Filter "human" as: organism

1 to 7 of 7

☐	Entry	Entry name		Protein names	Gene names	Organism	Length
☐	P04001	OPSG_HUMAN	☆	**Medium-wave-sensitive opsin 1**	**OPN1MW**, GCP **OPN1MW2**	Homo sapiens (Human)	364
☐	P04000	OPSR_HUMAN	☆	**Long-wave-sensitive opsin 1**	**OPN1LW**, RCP	Homo sapiens (Human)	364
☐	P03999	OPSB_HUMAN	☆	**Short-wave-sensitive opsin 1**	**OPN1SW**, BCP	Homo sapiens (Human)	348
☐	P08100	OPSD_HUMAN	☆	**Rhodopsin**	**RHO**, OPN2	Homo sapiens (Human)	348
☐	Q9H1Y3	OPN3_HUMAN	☆	**Opsin-3**	**OPN3**, ECPN	Homo sapiens (Human)	402
☐	Q9UHM6	OPN4_HUMAN	☆	**Melanopsin**	**OPN4**, MOP	Homo sapiens (Human)	478
☐	Q6U736	OPN5_HUMAN	☆	**Opsin-5**	**OPN5**, GPR136, PGR12, TMEM13	Homo sapiens (Human)	354

1 to 7 of 7

UniProt Human Opsin OPN1MW Entry

http://www.uniprot.org/uniprot/P04001

P04001 - OPSG_HUMAN

Protein: **Medium-wave-sensitive opsin 1**

Gene: **OPN1MW** more

Organism: *Homo sapiens (Human)*

Status: Reviewed - ●●●●● - Experimental evidence at protein level

Function

Visual pigments are the light-absorbing molecules that mediate vision. They consist of an apoprotein, opsin, covalently linked to cis-retinal.

Absorption
Abs(max)=530 nm

GO - Molecular function
- G-protein coupled receptor activity Source: UniProtKB-KW
- photoreceptor activity Source: ProtInc

GO - Biological process
- G-protein coupled receptor signaling pathway Source: ProtInc
- phototransduction, visible light Source: Reactome
- positive regulation of cytokinesis Source: UniProt
- protein-chromophore linkage Source: UniProtKB-KW
- retinoid metabolic process Source: Reactome
- visual perception Source: ProtInc

Complete GO annotation...

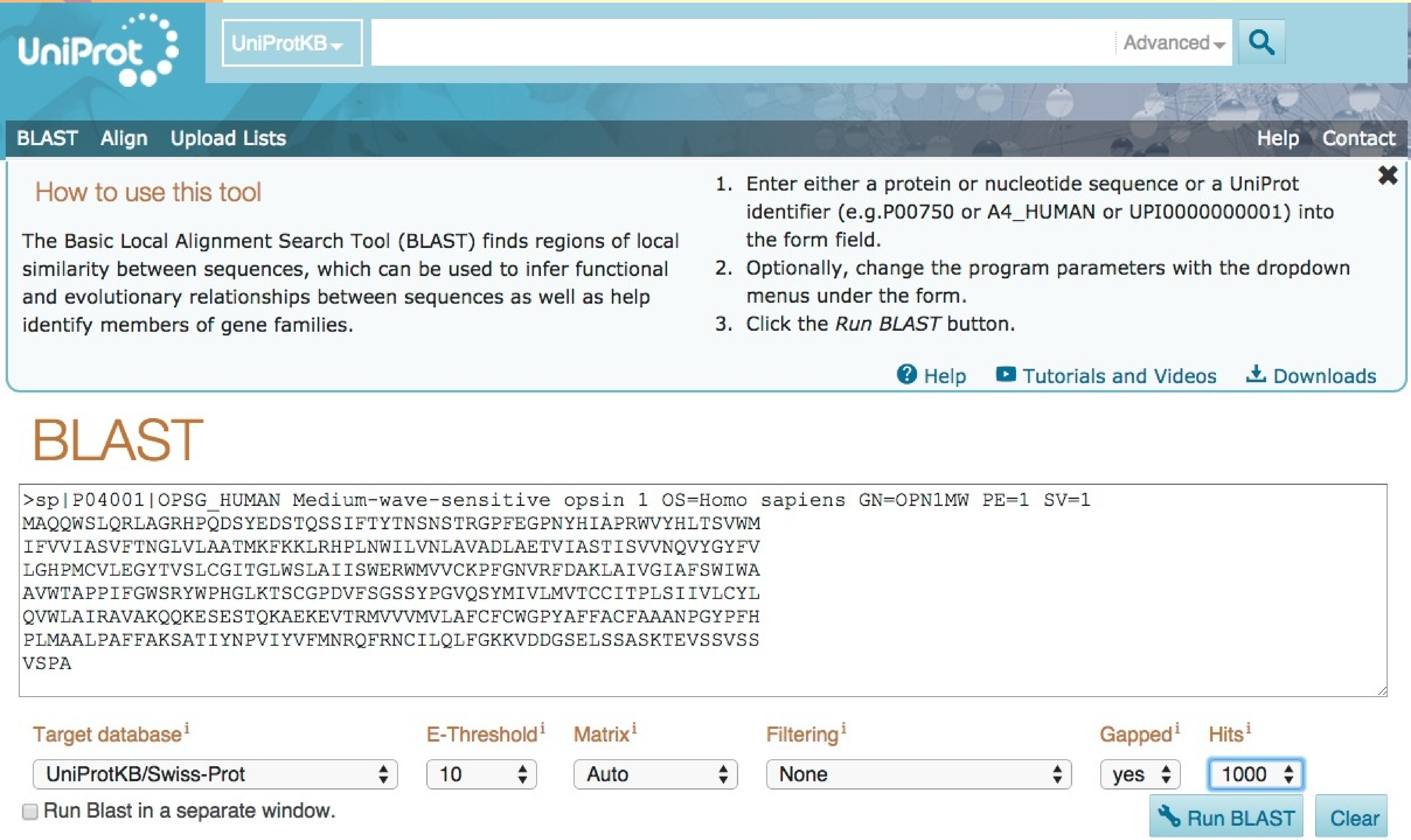
Keywords - Molecular function
G-protein coupled receptor, Photoreceptor protein, Receptor, Retinal protein, Transducer

Keywords - Biological process
Sensory transduction, Vision

Keywords - Ligand
Chromophore

Blast UniProt Human Opsin OPN1MW Entry

http://www.uniprot.org/uniprot/P04001



How to use this tool

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences, which can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

1. Enter either a protein or nucleotide sequence or a UniProt identifier (e.g.P00750 or A4_HUMAN or UPI0000000001) into the form field.
2. Optionally, change the program parameters with the dropdown menus under the form.
3. Click the *Run BLAST* button.

BLAST

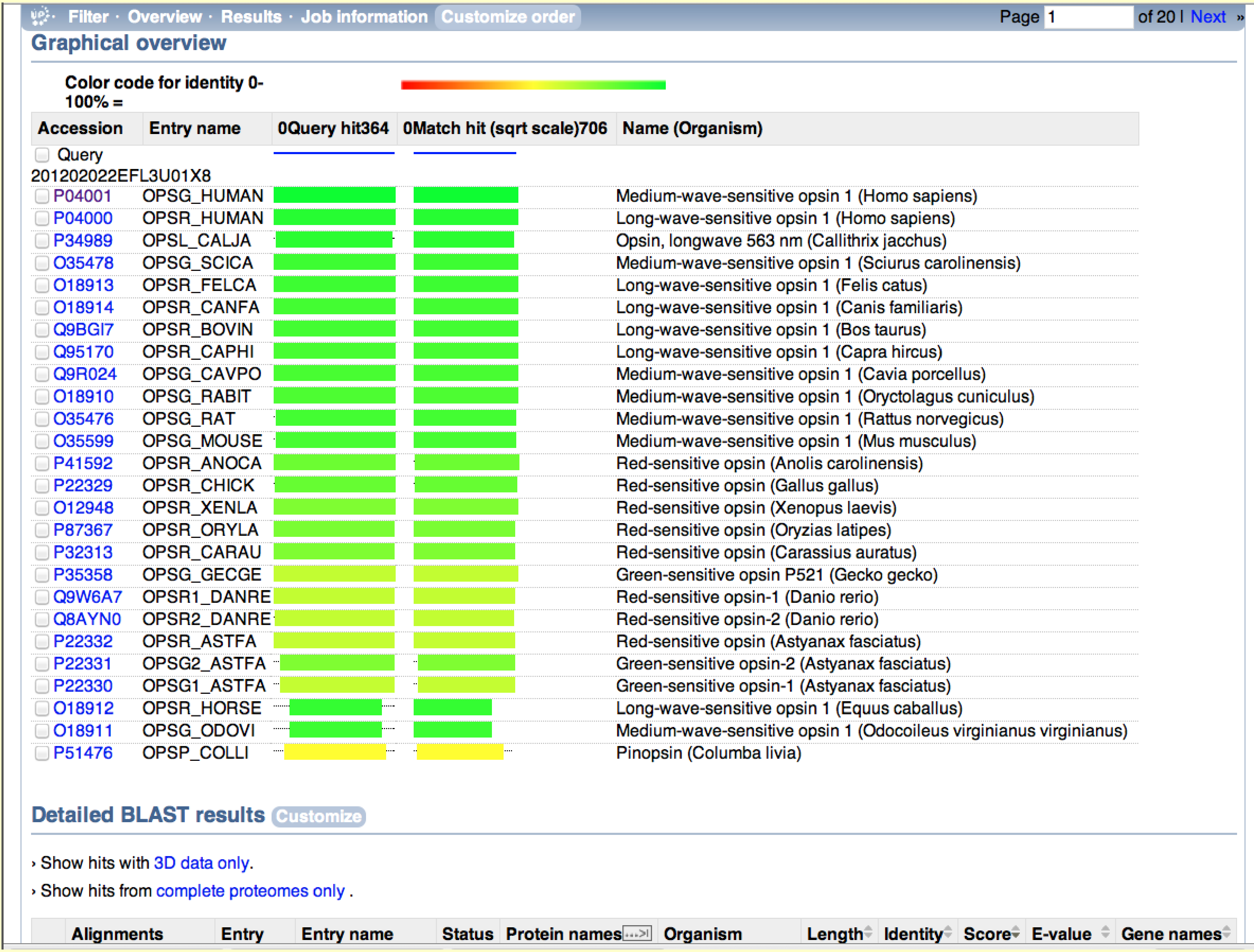
```
>sp|P04001|OPSG_HUMAN Medium-wave-sensitive opsin 1 OS=Homo sapiens GN=OPN1MW PE=1 SV=1
MAQQWSLQRLAGRHPQDSYEDSTQSSIFTYTNSNSTRGPFEGPNYHIAPRWVYHLTSVWM
IFVVIASVFTNGLVLAATMKFKKLRHPLNWILVNLAVADLAETVIASTISVVNQVYGYFV
LGHPMCVLEGYTVSLCGITGLWSLAIISWERWMVVCKPFGNVRFDAKLAIVGIAFSWIWA
AVWTAPPIFGWSRYWPHGLKTSCGPDVFSGSSYPGVQSYMIVLMVTCCITPLSIIVLCYL
QVWLAIRAVAKQQKESESTQKAEKEVTRMVVVMVLAFCFCWGPYAFFACFAAANPGYPFH
PLMAALPAFFAKSATIYNPVIYVFMNRQFRNCILQLFGKKVDDGSELSSASKTEVSSVSS
VSPA
```

Target database	E-Threshold	Matrix	Filtering	Gapped	Hits
UniProtKB/Swiss-Prot	10	Auto	None	yes	1000

☐ Run Blast in a separate window.

Blast UniProt Human OPN1MW Results

http://www.uniprot.org/uniprot/P04001



Graphical overview

Color code for identity 0-100% =

Accession	Entry name	0Query hit364	0Match hit (sqrt scale)706	Name (Organism)
☐ Query 201202022EFL3U01X8				
☐ P04001	OPSG_HUMAN	🟩	🟩	Medium-wave-sensitive opsin 1 (Homo sapiens)
☐ P04000	OPSR_HUMAN	🟩	🟩	Long-wave-sensitive opsin 1 (Homo sapiens)
☐ P34989	OPSL_CALJA	🟩	🟩	Opsin, longwave 563 nm (Callithrix jacchus)
☐ O35478	OPSG_SCICA	🟩	🟩	Medium-wave-sensitive opsin 1 (Sciurus carolinensis)
☐ O18913	OPSR_FELCA	🟩	🟩	Long-wave-sensitive opsin 1 (Felis catus)
☐ O18914	OPSR_CANFA	🟩	🟩	Long-wave-sensitive opsin 1 (Canis familiaris)
☐ Q9BGI7	OPSR_BOVIN	🟩	🟩	Long-wave-sensitive opsin 1 (Bos taurus)
☐ Q95170	OPSR_CAPHI	🟩	🟩	Long-wave-sensitive opsin 1 (Capra hircus)
☐ Q9R024	OPSG_CAVPO	🟩	🟩	Medium-wave-sensitive opsin 1 (Cavia porcellus)
☐ O18910	OPSG_RABIT	🟩	🟩	Medium-wave-sensitive opsin 1 (Oryctolagus cuniculus)
☐ O35476	OPSG_RAT	🟩	🟩	Medium-wave-sensitive opsin 1 (Rattus norvegicus)
☐ O35599	OPSG_MOUSE	🟩	🟩	Medium-wave-sensitive opsin 1 (Mus musculus)
☐ P41592	OPSR_ANOCA	🟩	🟩	Red-sensitive opsin (Anolis carolinensis)
☐ P22329	OPSR_CHICK	🟩	🟩	Red-sensitive opsin (Gallus gallus)
☐ O12948	OPSR_XENLA	🟩	🟩	Red-sensitive opsin (Xenopus laevis)
☐ P87367	OPSR_ORYLA	🟩	🟩	Red-sensitive opsin (Oryzias latipes)
☐ P32313	OPSR_CARAU	🟩	🟩	Red-sensitive opsin (Carassius auratus)
☐ P35358	OPSG_GECGE	🟨	🟨	Green-sensitive opsin P521 (Gecko gecko)
☐ Q9W6A7	OPSR1_DANRE	🟨	🟨	Red-sensitive opsin-1 (Danio rerio)
☐ Q8AYN0	OPSR2_DANRE	🟨	🟨	Red-sensitive opsin-2 (Danio rerio)
☐ P22332	OPSR_ASTFA	🟨	🟨	Red-sensitive opsin (Astyanax fasciatus)
☐ P22331	OPSG2_ASTFA	🟨	🟨	Green-sensitive opsin-2 (Astyanax fasciatus)
☐ P22330	OPSG1_ASTFA	🟨	🟨	Green-sensitive opsin-1 (Astyanax fasciatus)
☐ O18912	OPSR_HORSE	🟩	🟩	Long-wave-sensitive opsin 1 (Equus caballus)
☐ O18911	OPSG_ODOVI	🟩	🟩	Medium-wave-sensitive opsin 1 (Odocoileus virginianus virginianus)
☐ P51476	OPSP_COLLI	🟨	🟨	Pinopsin (Columba livia)

Detailed BLAST results

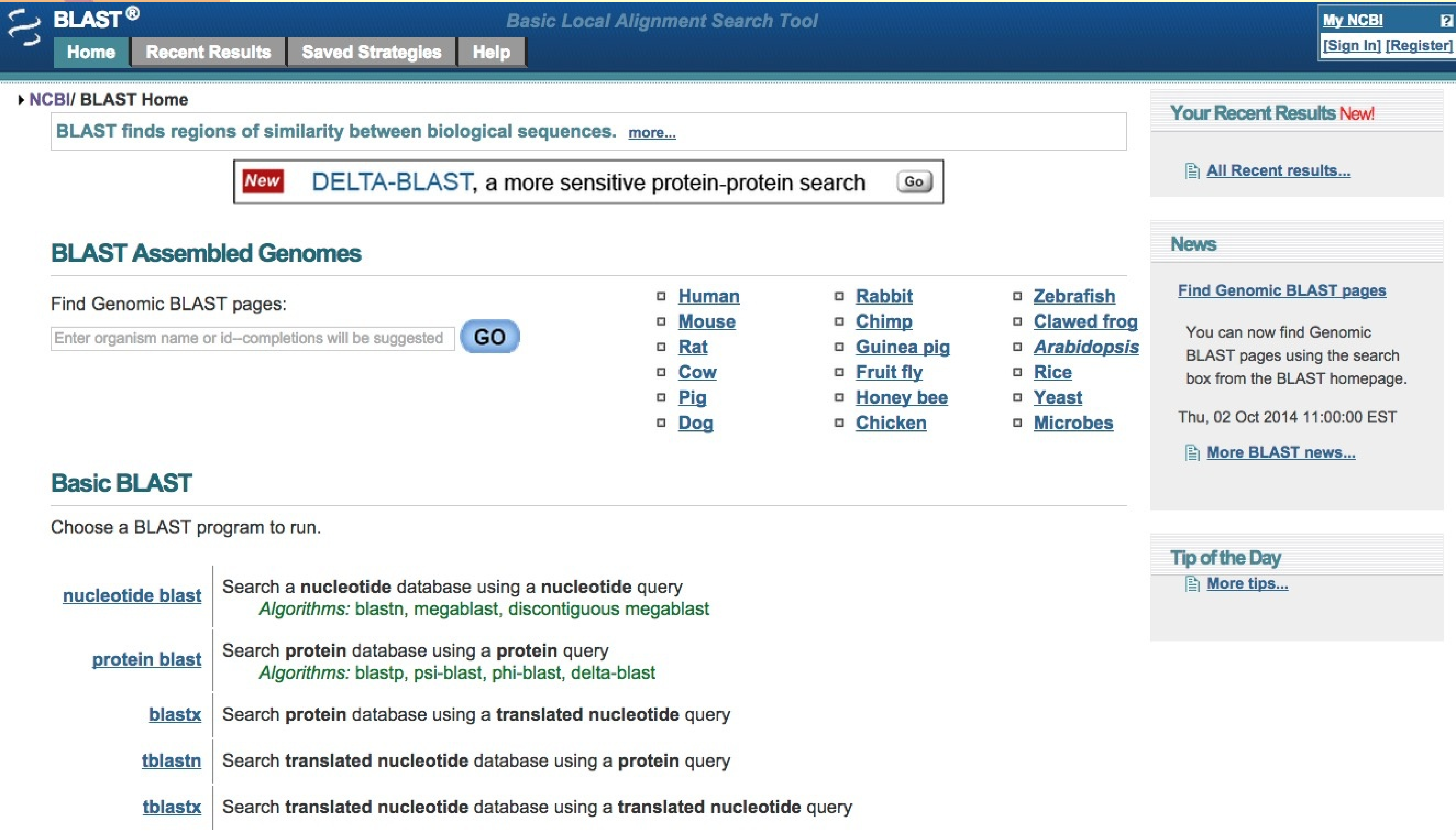
› Show hits with 3D data only.

› Show hits from complete proteomes only .

Alignments	Entry	Entry name	Status	Protein names	Organism	Length	Identity	Score	E-value	Gene names

NCBI BLAST Home Page

http://blast.ncbi.nlm.nih.gov/



BLAST finds regions of similarity between biological sequences. more...

New DELTA-BLAST, a more sensitive protein-protein search

BLAST Assembled Genomes

Find Genomic BLAST pages:

Enter organism name or id--completions will be suggested

- Human
- Mouse
- Rat
- Cow
- Pig
- Dog
- Rabbit
- Chimp
- Guinea pig
- Fruit fly
- Honey bee
- Chicken
- Zebrafish
- Clawed frog
- *Arabidopsis*
- Rice
- Yeast
- Microbes

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Your Recent Results New!

All Recent results...

News

Find Genomic BLAST pages

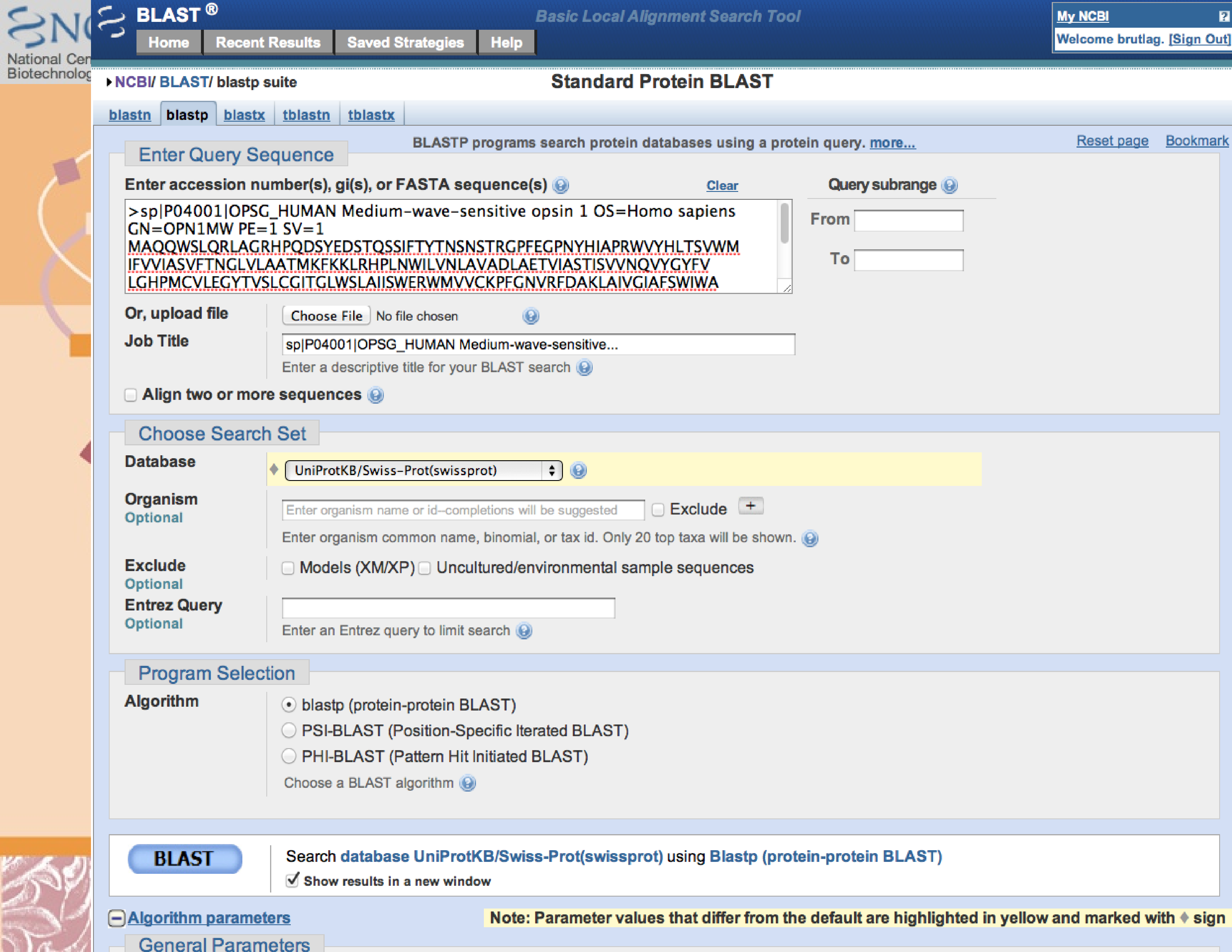
You can now find Genomic BLAST pages using the search box from the BLAST homepage.

Thu, 02 Oct 2014 11:00:00 EST

More BLAST news...

Tip of the Day

More tips...



Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
>sp|P04001|OPSG_HUMAN Medium-wave-sensitive opsin 1 OS=Homo sapiens
GN=OPN1MW PE=1 SV=1
MAQQWSLQRLAGRHPQDSYEDSTQSSIFTYTNSNSTRGPFEGPNYHIAPRWVYHLTSVWM
IFVVIASVFTNGLVLAATMKFKKLRHPLNWILVNLAVADLAETVIASTISVVNQVYGYFV
LGHPMCVLEGYTVSLCGITGLWSLAIISWERWMVVCKPFGNVRFDAKLAIVGIAFSWIWA
```

Query subrange

From

To

Or, upload file No file chosen

Job Title sp|P04001|OPSG_HUMAN Medium-wave-sensitive...

Enter a descriptive title for your BLAST search

☐ **Align two or more sequences**

Choose Search Set

Database ♦ UniProtKB/Swiss-Prot(swissprot)

Organism Optional

Enter organism name or id--completions will be suggested ☐ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query Optional

Enter an Entrez query to limit search

Program Selection

Algorithm

● blastp (protein-protein BLAST)

○ PSI-BLAST (Position-Specific Iterated BLAST)

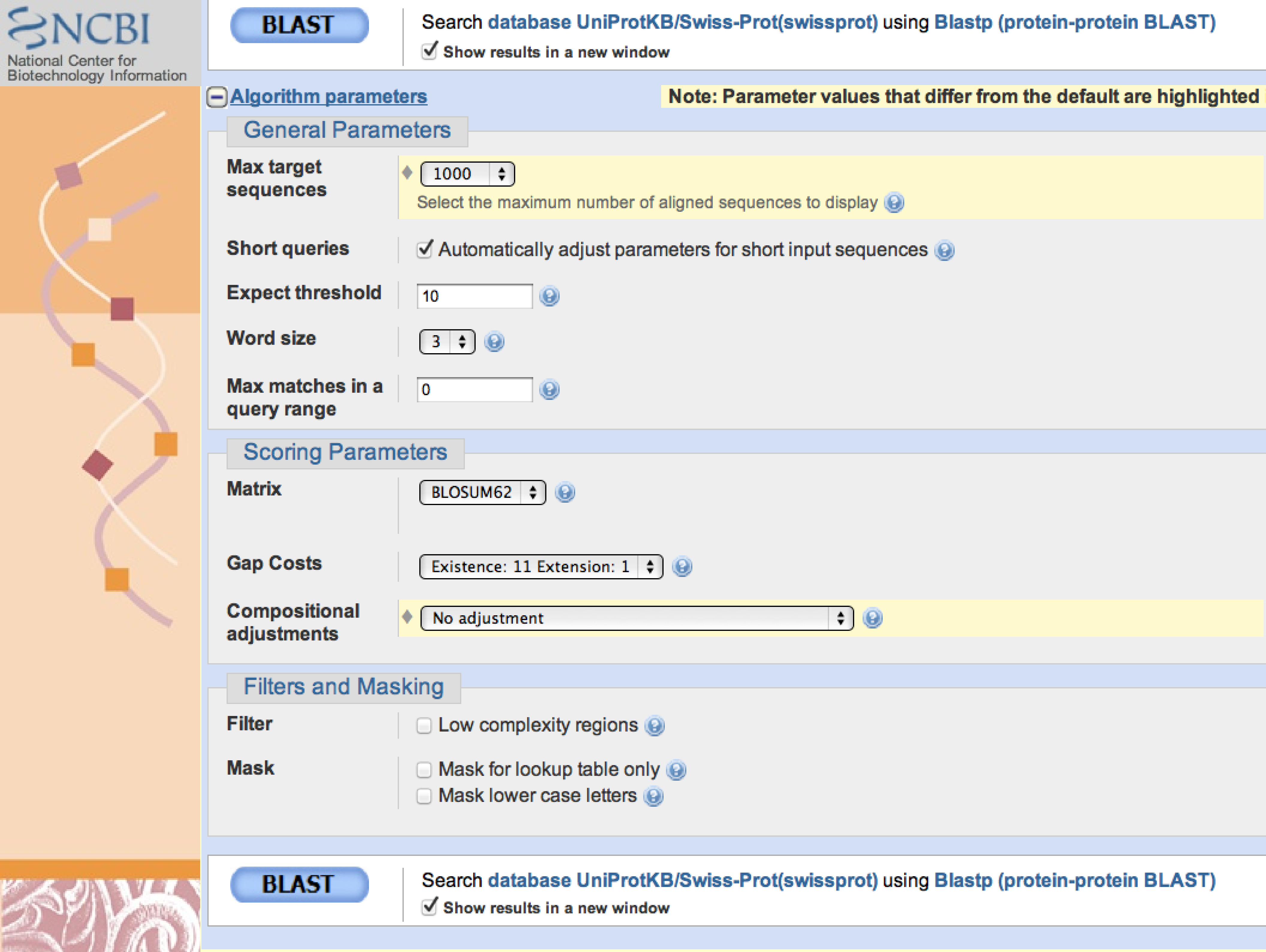
○ PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

Search database UniProtKB/Swiss-Prot(swissprot) using Blastp (protein-protein BLAST)

☑ Show results in a new window

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign



Search database UniProtKB/Swiss-Prot(swissprot) using Blastp (protein-protein BLAST)

☑ Show results in a new window

Algorithm parameters

Note: Parameter values that differ from the default are highlighted

General Parameters

Max target sequences: 1000
Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: No adjustment

Filters and Masking

Filter: ☐ Low complexity regions

Mask: ☐ Mask for lookup table only
☐ Mask lower case letters

Search database UniProtKB/Swiss-Prot(swissprot) using Blastp (protein-protein BLAST)

☑ Show results in a new window

Discovering Function from Protein Sequence

BLOCKs, PRINTs, PSSMS or Weight Matrices

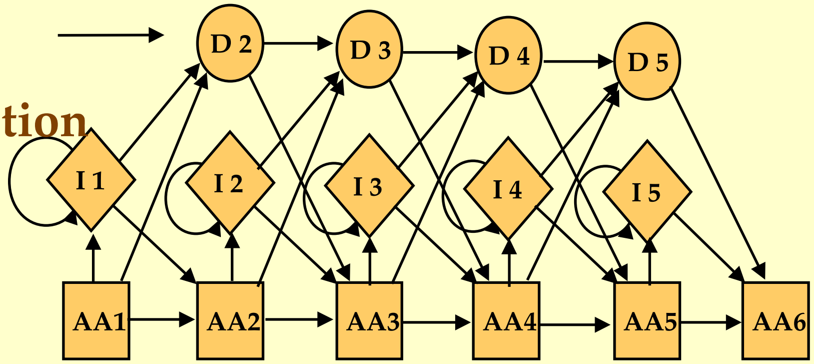
	Position											
	1	2	3	4	5	6	7	8	9	10	11	12
A	2	1	3	13	10	12	67	4	13	9	1	2
R	7	5	8	9	4	0	1	16	7	0	1	0
N	0	8	0	1	0	0	0	2	1	1	10	0
D	0	1	0	1	13	0	0	12	1	0	4	0
C	0	0	1	0	0	0	0	0	0	2	2	1
Q	1	1	21	8	10	0	0	7	6	0	0	2
E	2	0	0	9	21	0	0	15	7	3	3	0
G	9	7	1	4	0	0	8	0	0	0	46	0
H	4	3	1	1	2	0	0	2	2	0	5	0
I	10	0	11	1	2	10	0	4	9	3	0	16
L	16	1	17	0	1	31	0	3	11	24	0	14
K	3	4	5	10	11	1	1	13	10	0	5	2
M	7	1	1	0	0	0	0	0	5	7	1	8
F	4	0	3	0	0	4	0	0	0	10	0	0
P	0	6	0	1	0	0	0	0	0	0	0	0
S	1	17	0	8	3	1	3	0	2	2	2	0
T	5	22	3	11	1	5	0	2	2	2	0	5
W	2	0	0	0	0	0	0	0	0	1	0	1
Y	1	0	4	2	0	1	0	0	2	4	0	1
V	6	3	1	1	2	15	0	0	2	12	0	28

Consensus Sequences or Sequence Motifs

Zinc Finger (C2H2 type)
C X{2,4} C X{12} H X{3,5} H

Sequences of Common Structure or Function

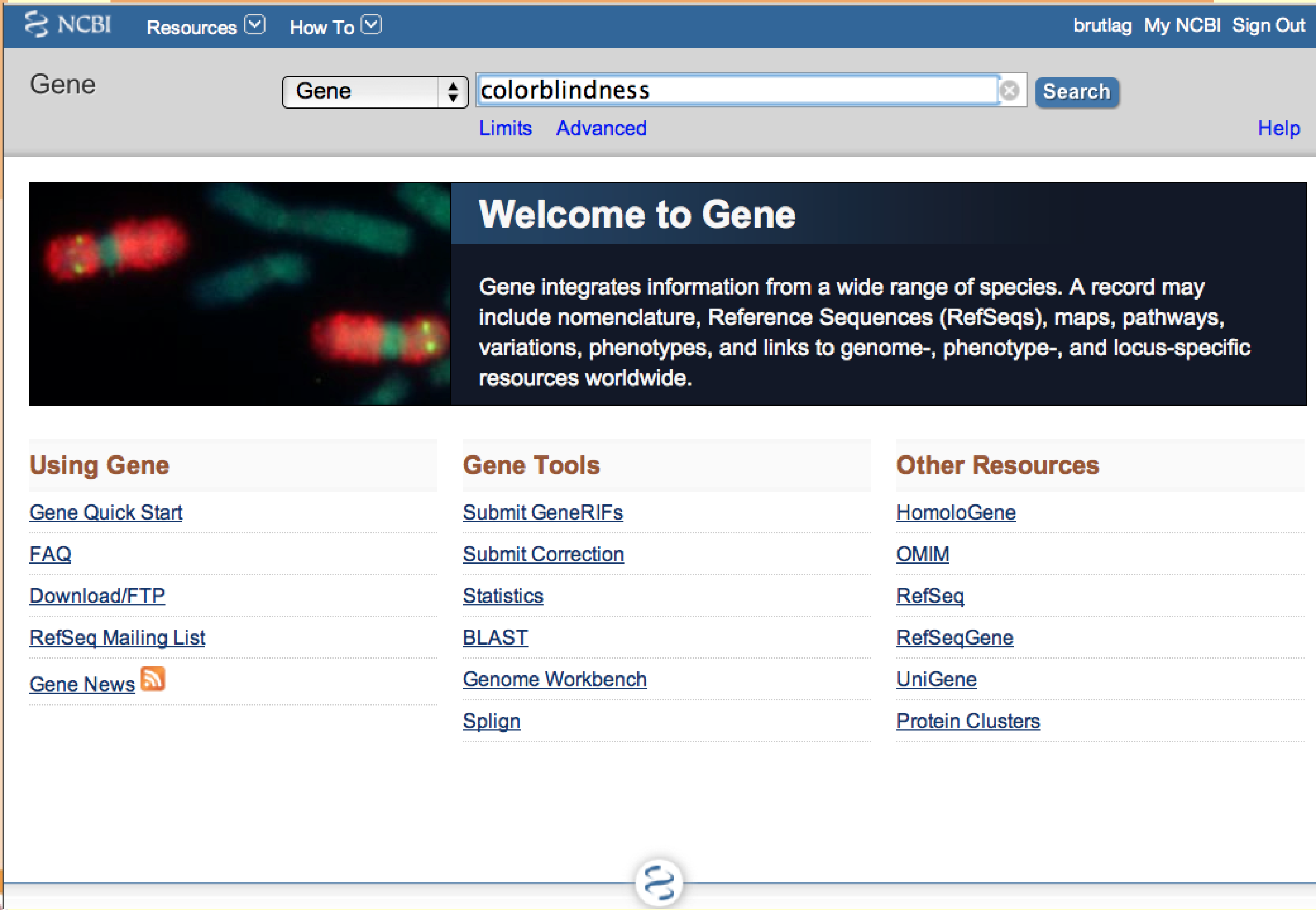
Profiles, PSI-BLAST Hidden Markov Models



Sequence Similarity

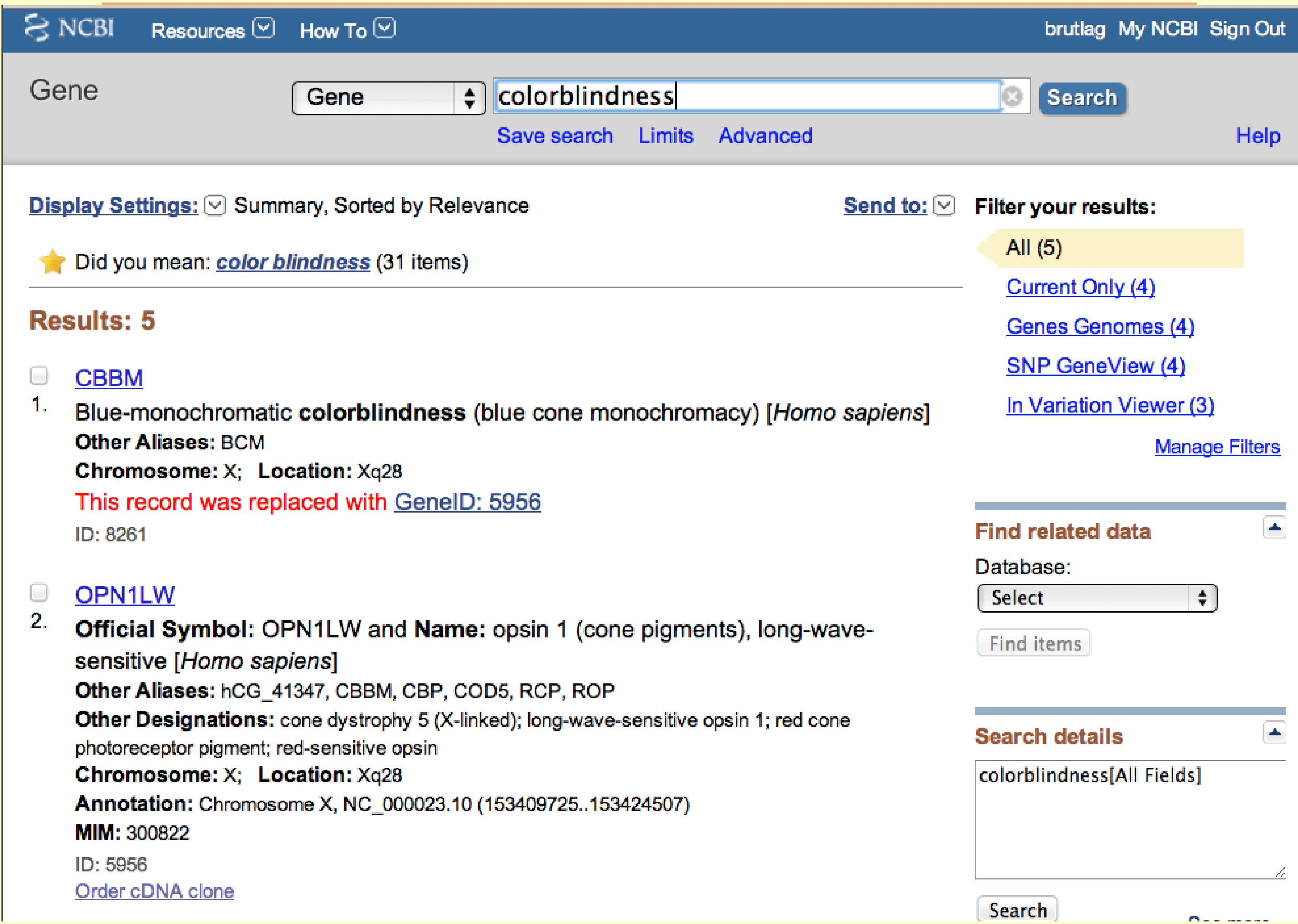
```
                     10        20        30        40              50
Query         HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
               |:| :|: | |:||||     | |:||| |: : :|:|  :|  |      |:  |
Database      VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
                      10          20        30        40        50
```

Entrez Gene search for Colorblindness



Gene integrates information from a wide range of species. A record may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.

Entrez Gene search for Colorblindness



colorblindness

Display Settings: Summary, Sorted by Relevance

Send to:

Did you mean: ***color blindness*** (31 items)

Results: 5

☐ 1. CBBM

Blue-monochromatic **colorblindness** (blue cone monochromacy) [*Homo sapiens*]

Other Aliases: BCM

Chromosome: X; **Location:** Xq28

This record was replaced with GeneID: 5956

ID: 8261

☐ 2. OPN1LW

Official Symbol: OPN1LW and **Name:** opsin 1 (cone pigments), long-wave-sensitive [*Homo sapiens*]

Other Aliases: hCG_41347, CBBM, CBP, COD5, RCP, ROP

Other Designations: cone dystrophy 5 (X-linked); long-wave-sensitive opsin 1; red cone photoreceptor pigment; red-sensitive opsin

Chromosome: X; **Location:** Xq28

Annotation: Chromosome X, NC_000023.10 (153409725..153424507)

MIM: 300822

ID: 5956

Order cDNA clone

Filter your results:

- All (5)
- Current Only (4)
- Genes Genomes (4)
- SNP GeneView (4)
- In Variation Viewer (3)

Manage Filters

Find related data

Database: Select

Find items

Search details

colorblindness[All Fields]

Search

Entrez Gene search for Colorblindness

Display Settings: Full Report

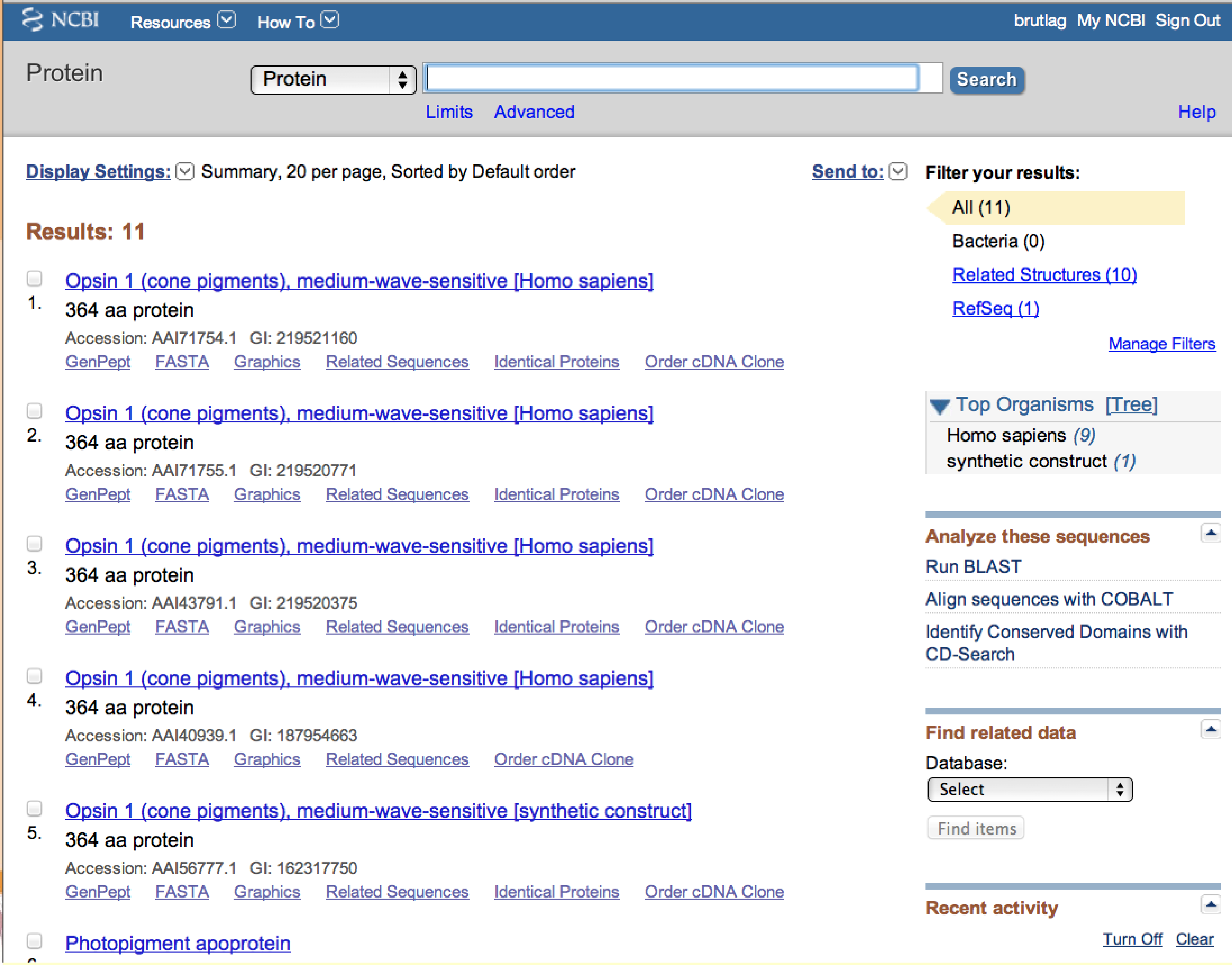
OPN1MW opsin 1 (cone pigments), medium-wave-sensitive [*Homo sapiens*]

Gene ID: 2652, updated on 18-Nov-2011

Summary

Official Symbol	OPN1MW provided by HGNC
Official Full Name	opsin 1 (cone pigments), medium-wave-sensitive provided by HGNC
Primary source	HGNC:4206
See related	Ensembl:ENSG00000147380; HPRD:02365; MIM:300821; Vega:OTTHUMG00000022652
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	CBD; GCP; GOP; CBBM; COD5; OPN1MW1; OPN1MW2
Summary	This gene encodes for a light absorbing visual pigment of the opsin gene family. The encoded protein is called green cone photopigment or medium-wavelength sensitive opsin. Opsins are G-protein coupled receptors with seven transmembrane domains, an N-terminal extracellular domain, and a C-terminal cytoplasmic domain. The long-wavelength opsin gene and multiple copies of the medium-wavelength opsin gene are tandemly arrayed on the X chromosome and frequent unequal recombination and gene conversion may occur between these sequences. X chromosomes may have fusions of the medium- and long-wavelength opsin genes or may have more than one copy of these genes. Defects in this gene are the cause of deutanopic colorblindness. [provided by RefSeq, Mar 2009]

Entrez Gene search for Opsins



Display Settings: Summary, 20 per page, Sorted by Default order

Send to:

Results: 11

☐ 1. Opsin 1 (cone pigments), medium-wave-sensitive [Homo sapiens]
364 aa protein
Accession: AAI71754.1 GI: 219521160
GenPept FASTA Graphics Related Sequences Identical Proteins Order cDNA Clone

☐ 2. Opsin 1 (cone pigments), medium-wave-sensitive [Homo sapiens]
364 aa protein
Accession: AAI71755.1 GI: 219520771
GenPept FASTA Graphics Related Sequences Identical Proteins Order cDNA Clone

☐ 3. Opsin 1 (cone pigments), medium-wave-sensitive [Homo sapiens]
364 aa protein
Accession: AAI43791.1 GI: 219520375
GenPept FASTA Graphics Related Sequences Identical Proteins Order cDNA Clone

☐ 4. Opsin 1 (cone pigments), medium-wave-sensitive [Homo sapiens]
364 aa protein
Accession: AAI40939.1 GI: 187954663
GenPept FASTA Graphics Related Sequences Order cDNA Clone

☐ 5. Opsin 1 (cone pigments), medium-wave-sensitive [synthetic construct]
364 aa protein
Accession: AAI56777.1 GI: 162317750
GenPept FASTA Graphics Related Sequences Identical Proteins Order cDNA Clone

☐ Photopigment apoprotein

Filter your results:

All (11)
Bacteria (0)
Related Structures (10)
RefSeq (1)

Manage Filters

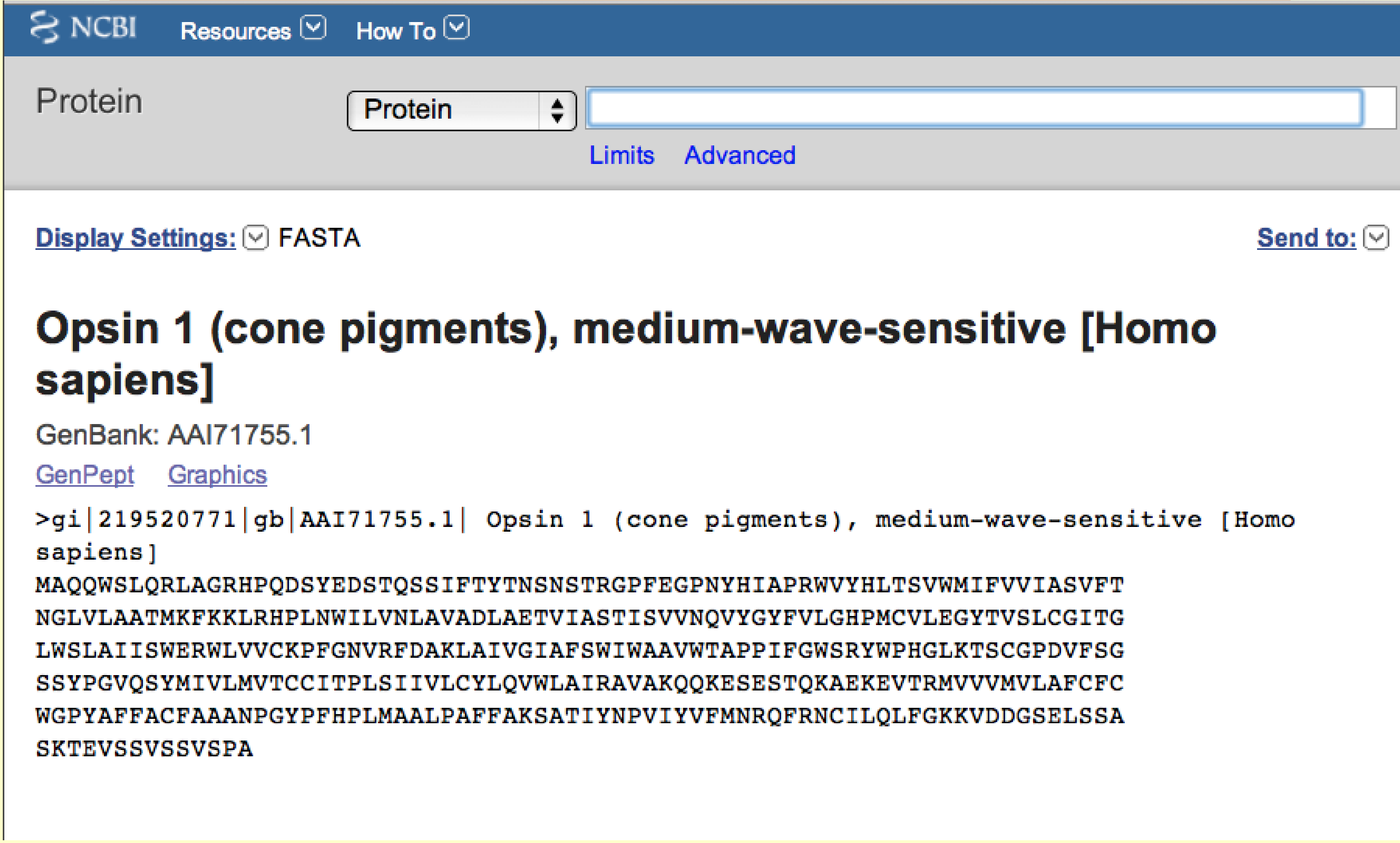
Top Organisms [Tree]
Homo sapiens *(9)*
synthetic construct *(1)*

Analyze these sequences
Run BLAST
Align sequences with COBALT
Identify Conserved Domains with CD-Search

Find related data
Database:
Select
Find items

Recent activity
Turn Off Clear

Entrez Gene search for Opsins



Display Settings: FASTA

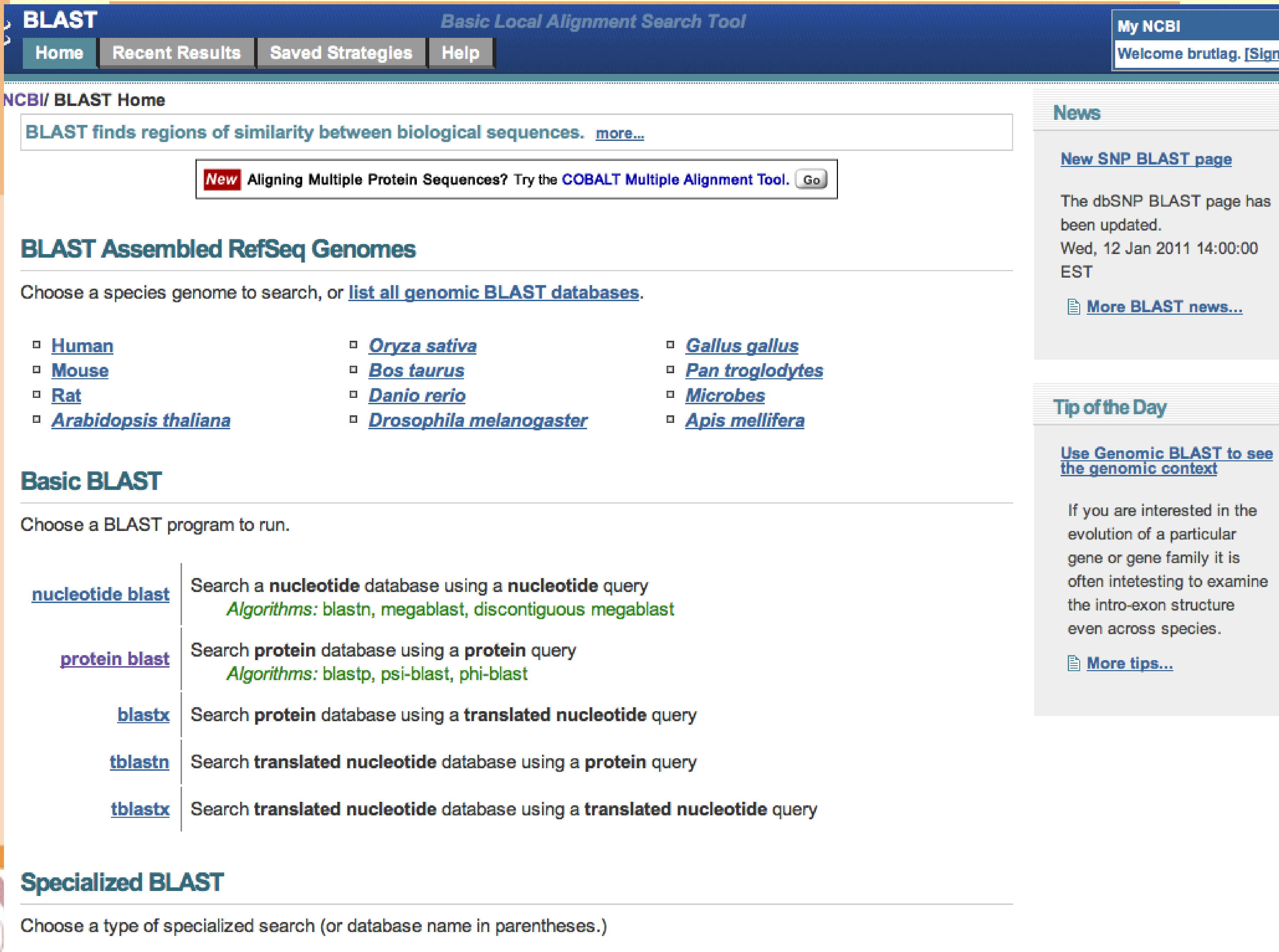
Opsin 1 (cone pigments), medium-wave-sensitive [Homo sapiens]

GenBank: AAI71755.1

```
>gi|219520771|gb|AAI71755.1| Opsin 1 (cone pigments), medium-wave-sensitive [Homo
sapiens]
MAQQWSLQRLAGRHPQDSYEDSTQSSIFTYTNSNSTRGPFEGPNYHIAPRWVYHLTSVWMIFVVIASVFT
NGLVLAATMKFKKLRHPLNWILVNLAVADLAETVIASTISVVNQVYGYFVLGHPMCVLEGYTVSLCGITG
LWSLAIISWERWLVVCKPFGNVRFDAKLAIVGIAFSWIWAAVWTAPPIFGWSRYWPHGLKTSCGPDVFSG
SSYPGVQSYMIVLMVTCCITPLSIIVLCYLQVWLAIRAVAKQQKESESTQKAEKEVTRMVVVMVLAFCFC
WGPYAFFACFAAANPGYPFHPLMAALPAFFAKSATIYNPVIYVFMNRQFRNCILQLFGKKVDDGSELSSA
SKTEVSSVSSVSPA
```

BLAST Similarity Search

http://www.ncbi.nlm.nih.gov/BLAST/



NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

New Aligning Multiple Protein Sequences? Try the COBALT Multiple Alignment Tool. Go

BLAST Assembled RefSeq Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

News

New SNP BLAST page

The dbSNP BLAST page has been updated.
Wed, 12 Jan 2011 14:00:00 EST

More BLAST news...

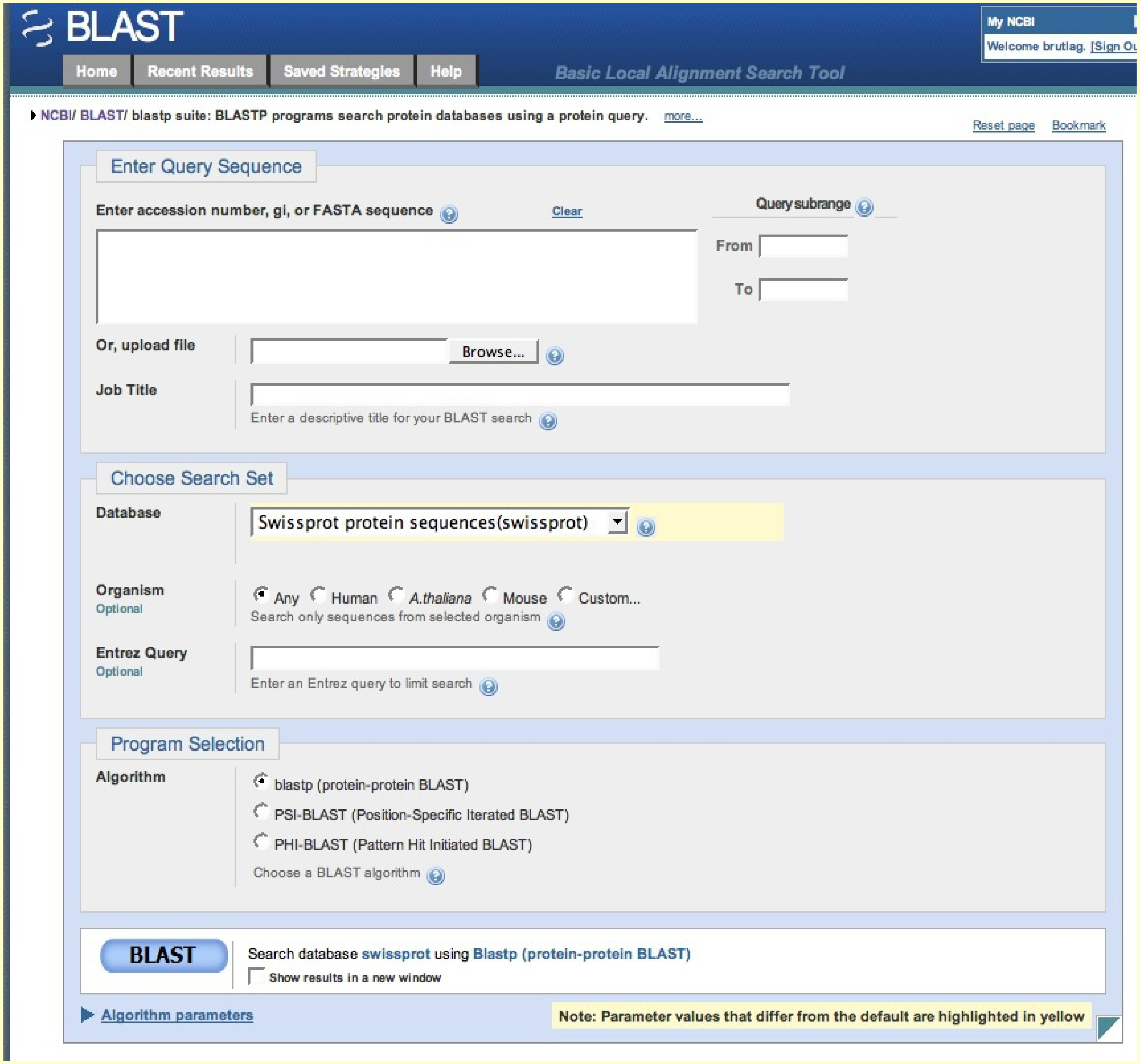
Tip of the Day

Use Genomic BLAST to see the genomic context

If you are interested in the evolution of a particular gene or gene family it is often intetesting to examine the intro-exon structure even across species.

More tips...

Choose Standard Protein-Protein BLAST
http://www.ncbi.nlm.nih.gov/BLAST/



NCBI/ BLAST/ blastp suite: BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

Query subrange

From

To

Or, upload file

Job Title

Enter a descriptive title for your BLAST search

Choose Search Set

Database: Swissprot protein sequences(swissprot)

Organism Optional: ● Any ○ Human ○ A.thaliana ○ Mouse ○ Custom...

Search only sequences from selected organism

Entrez Query Optional

Enter an Entrez query to limit search

Program Selection

Algorithm:
- ● blastp (protein-protein BLAST)
- ○ PSI-BLAST (Position-Specific Iterated BLAST)
- ○ PHI-BLAST (Pattern Hit Initiated BLAST)

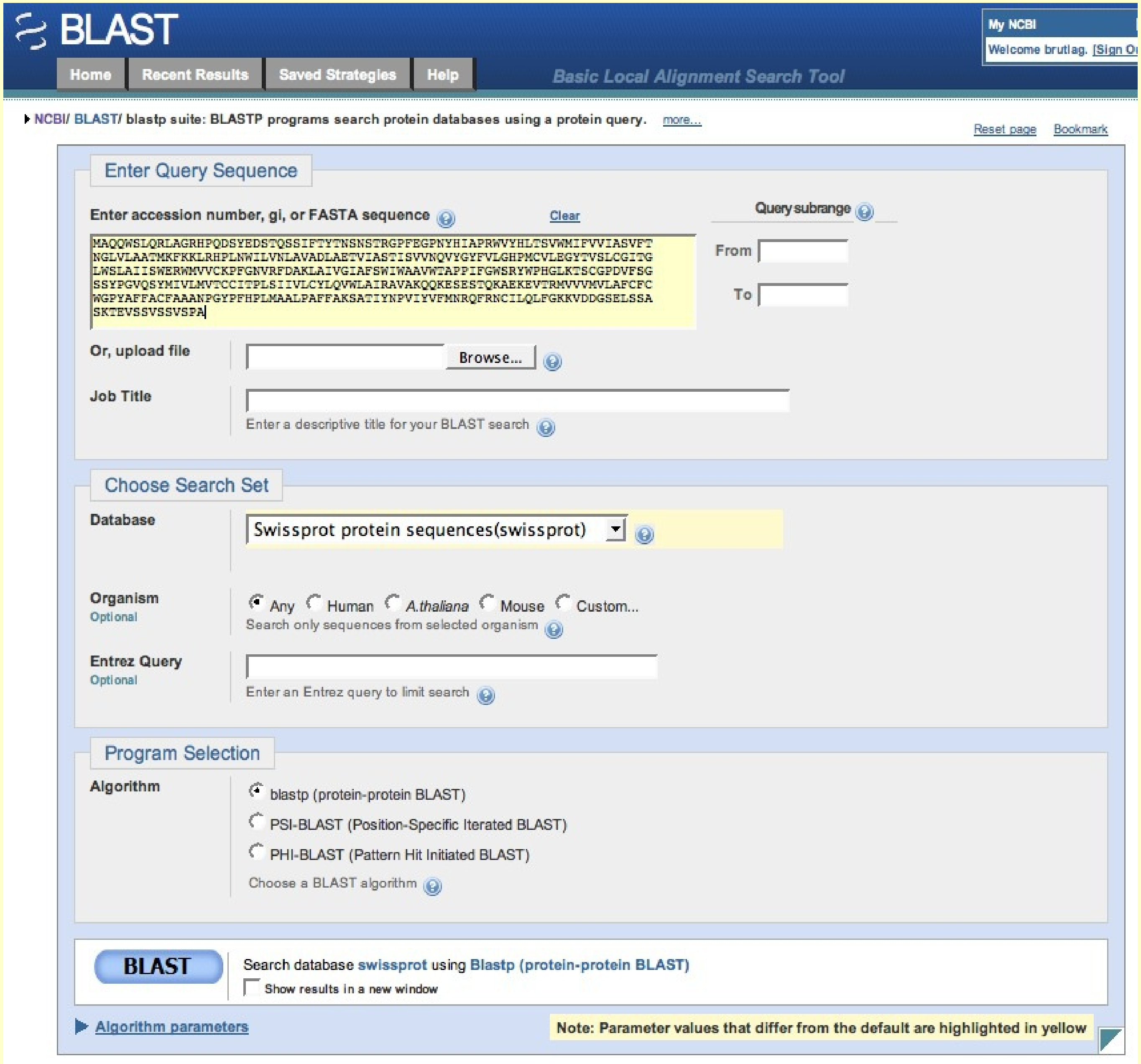
Choose a BLAST algorithm

Search database swissprot using Blastp (protein-protein BLAST)

☐ Show results in a new window

Note: Parameter values that differ from the default are highlighted in yellow

Paste Sequence, Choose SwissProt Database and BLAST!



NCBI/ BLAST/ blastp suite: BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

```
MAQQWSLQRLAGRHPQDSYEDSTQSSIFTYTNSNSTRGPFEGPNYHIAPRWVYHLTSVWMIFVVIASVFT
NGLVLAATMKFKKLRHPLNWILVNLAVADLAETVIASTISVVNQVYGYFVLGHPMCVLEGYTVSLCGITG
LWSLAIISWERWMVVCKPFGNVRFDAKLAIVGIAFSWIWAAVWTAPPIFGWSRYWPHGLKTSCGPDVFSG
SSYPGVQSYMIVLMVTCCITPLSIIVLCYLQVWLAIRAVAKQQKESESTQKAEKEVTRMVVVMVLAFCFC
WGPYAFFACFAAANPGYPFHPLMAALPAFFAKSATIYNPVIYVFMNRQFRNCILQLFGKKVDDGSELSSA
SKTEVSSVSSVSPA
```

Query subrange

From:

To:

Or, upload file:

Job Title:

Enter a descriptive title for your BLAST search

Choose Search Set

Database: Swissprot protein sequences(swissprot)

Organism (Optional): ● Any ○ Human ○ *A.thaliana* ○ Mouse ○ Custom...

Search only sequences from selected organism

Entrez Query (Optional):

Enter an Entrez query to limit search

Program Selection

Algorithm:

● blastp (protein-protein BLAST)

○ PSI-BLAST (Position-Specific Iterated BLAST)

○ PHI-BLAST (Pattern Hit Initiated BLAST)

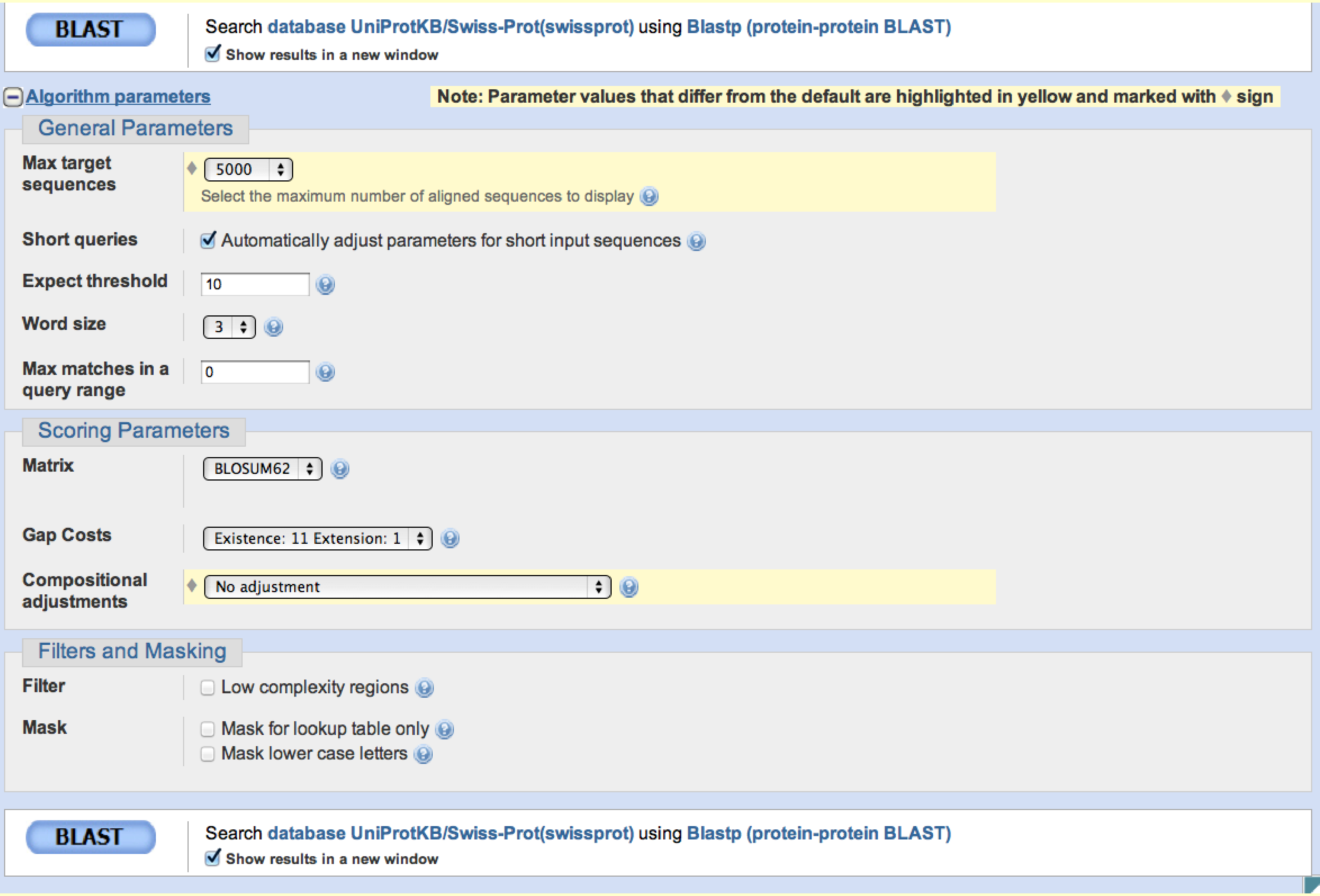
Choose a BLAST algorithm

Search database swissprot using Blastp (protein-protein BLAST)

☐ Show results in a new window

Note: Parameter values that differ from the default are highlighted in yellow

Optional Parameters



Search database UniProtKB/Swiss-Prot(swissprot) using Blastp (protein-protein BLAST)

☑ Show results in a new window

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

General Parameters

Max target sequences ♦ 5000 — Select the maximum number of aligned sequences to display

Short queries ☑ Automatically adjust parameters for short input sequences

Expect threshold 10

Word size 3

Max matches in a query range 0

Scoring Parameters

Matrix BLOSUM62

Gap Costs Existence: 11 Extension: 1

Compositional adjustments ♦ No adjustment

Filters and Masking

Filter ☐ Low complexity regions

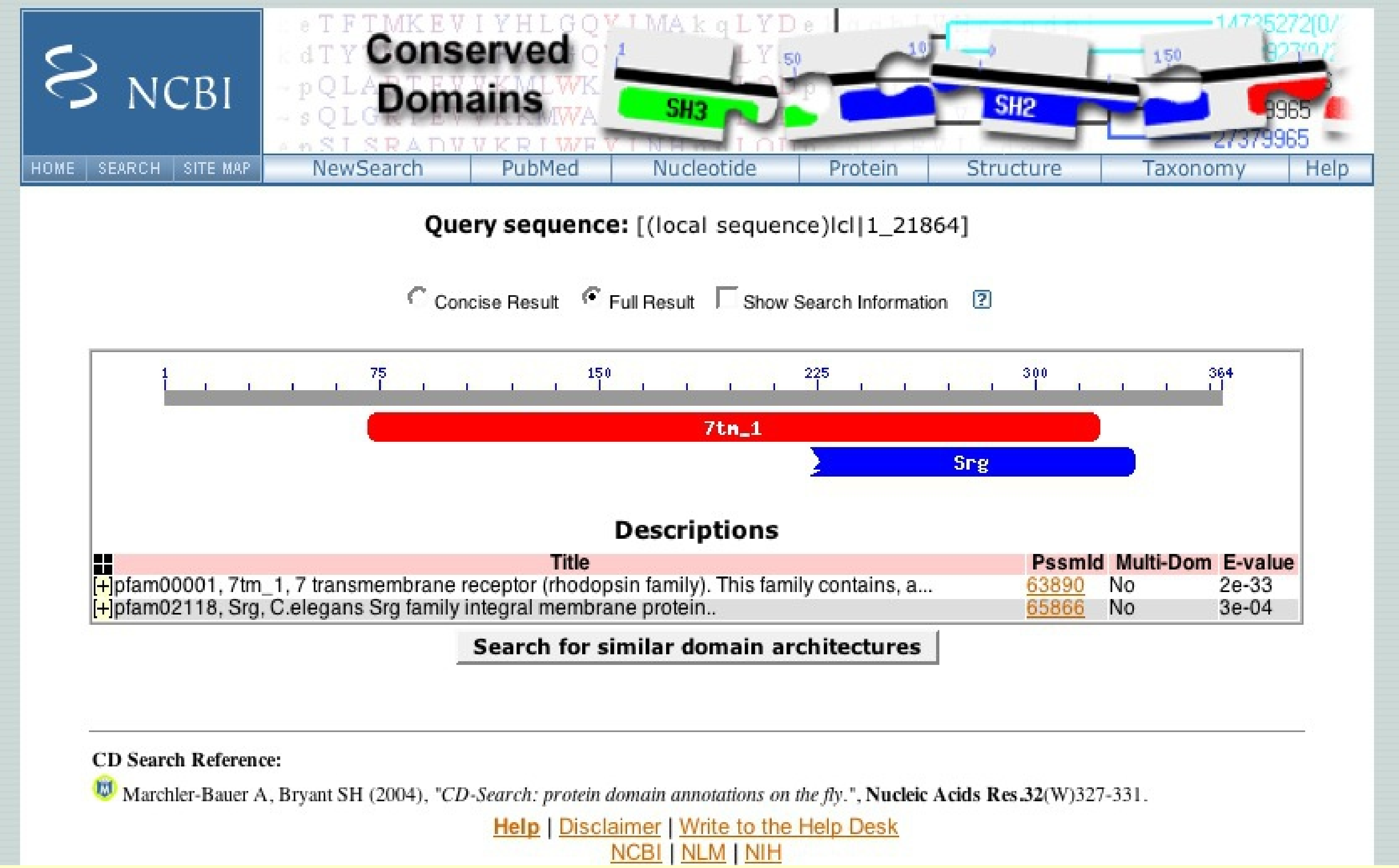
Mask ☐ Mask for lookup table only

☐ Mask lower case letters

Search database UniProtKB/Swiss-Prot(swissprot) using Blastp (protein-protein BLAST)

☑ Show results in a new window

BLAST Conserved Domain Output



Query sequence: [(local sequence)lcl|1_21864]

○ Concise Result ● Full Result ☐ Show Search Information

Descriptions

Title	PssmId	Multi-Dom	E-value
[+]pfam00001, 7tm_1, 7 transmembrane receptor (rhodopsin family). This family contains, a...	63890	No	2e-33
[+]pfam02118, Srg, C.elegans Srg family integral membrane protein..	65866	No	3e-04

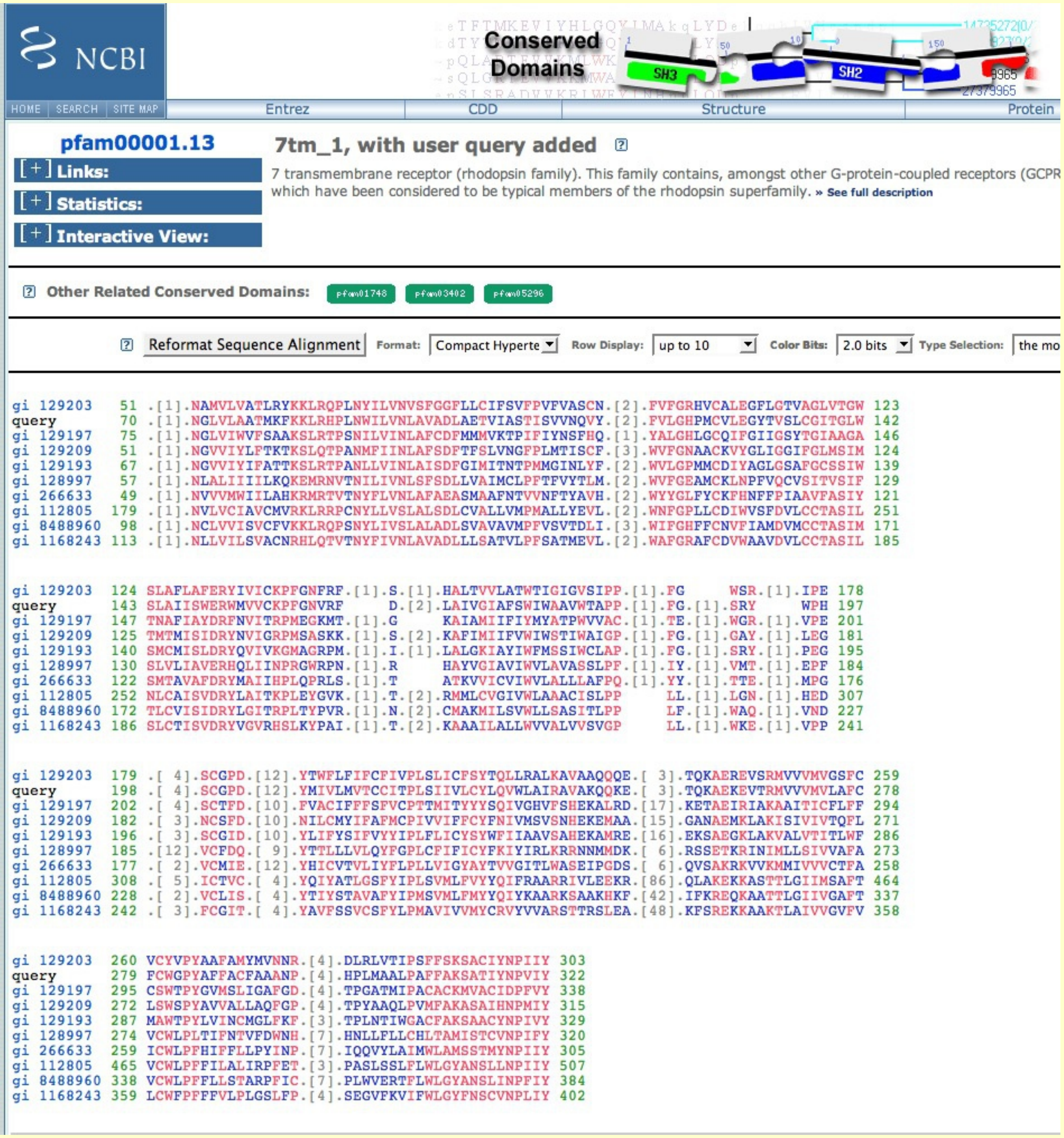
CD Search Reference:

Marchler-Bauer A, Bryant SH (2004), "*CD-Search: protein domain annotations on the fly.*", **Nucleic Acids Res.32**(W)327-331.

Help | Disclaimer | Write to the Help Desk

NCBI | NLM | NIH

Sequence Aligned with Domain



pfam00001.13

7tm_1, with user query added

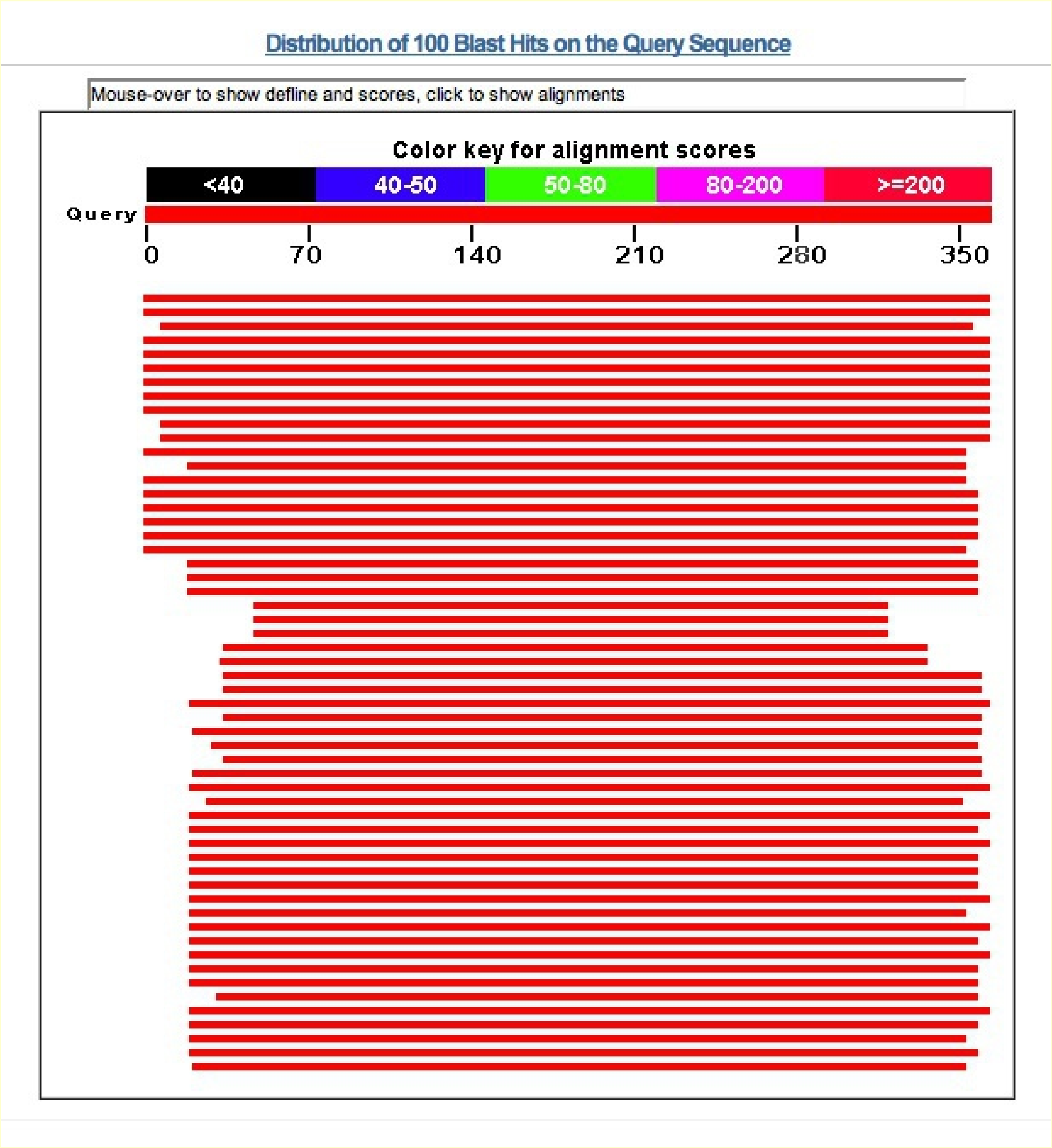
7 transmembrane receptor (rhodopsin family). This family contains, amongst other G-protein-coupled receptors (GCPR which have been considered to be typical members of the rhodopsin superfamily. » **See full description**

Other Related Conserved Domains: pfam01748 pfam03402 pfam05296

Format: Compact Hyperte **Row Display:** up to 10 **Color Bits:** 2.0 bits **Type Selection:** the mo

```
gi 129203   51 .[1].NAMVLVATLRYKKLRQPLNYILVNVSFGGFLLCIFSVFPVFVASCN.[2].FVFGRHVCALEGFLGTVAGLVTGW 123
query       70 .[1].NGLVLAATMKFKKLRHPLNWILVNLAVADLAETVIASTISVVNQVY.[2].FVLGHPMCVLEGYTVSLCGITGLW 142
gi 129197   75 .[1].NGLVIWVFSAAKSLRTPSNILVINLAFCDFMMMVKTPIFIYNSFHQ.[1].YALGHLGCQIFGIIGSYTGIAAGA 146
gi 129209   51 .[1].NGVVIYLFTKTKSLQTPANMFIINLAFSDFTFSLVNGFPLMTISCF.[3].WVFGNAACKVYGLIGGIFGLMSIM 124
gi 129193   67 .[1].NGVVIYIFATTKSLRTPANLLVINLAISDFGIMITNTPMMGINLYF.[2].WVLGPMMCDIYAGLGSAFGCSSIW 139
gi 128997   57 .[1].NLALIIIILKQKEMRNVTNILIVNLSFSDLLVAIMCLPFTFVYTLM.[2].WVFGEAMCKLNPFVQCVSITVSIF 129
gi 266633   49 .[1].NVVVMWIILAHKRMRTVTNYFLVNLAFAEASMAAFNTVVNFTYAVH.[2].WYYGLFYCKFHNFFPIAAVFASIY 121
gi 112805  179 .[1].NVLVCIAVCMVRKLRRPCNYLLVSLALSDLCVALLVMPMALLYEVL.[2].WNFGPLLCDIWVSFDVLCCTASIL 251
gi 8488960  98 .[1].NCLVVISVCFVKKLRQPSNYLIVSLALADLSVAVAVMPFVSVTDLI.[3].WIFGHFFCNVFIAMDVMCCTASIM 171
gi 1168243 113 .[1].NLLVILSVACNRHLQTVTNYFIVNLAVADLLLSATVLPFSATMEVL.[2].WAFGRAFCDVWAAVDVLCCTASIL 185

gi 129203  124 SLAFLAFERYIVICKPFGNFRF.[1].S.[1].HALTVVLATWTIGIGVSIPP.[1].FG     WSR.[1].IPE 178
query      143 SLAIISWERWMVVCKPFGNVRF     D.[2].LAIVGIAFSWIWAAVWTAPP.[1].FG.[1].SRY     WPH 197
gi 129197  147 TNAFIAYDRFNVITRPMEGKMT.[1].G     KAIAMIIFIYMYATPWVVAC.[1].TE.[1].WGR.[1].VPE 201
gi 129209  125 TMTMISIDRYNVIGRPMSASKK.[1].S.[2].KAFIMIIFVWIWSTIWAIGP.[1].FG.[1].GAY.[1].LEG 181
gi 129193  140 SMCMISLDRYQVIVKGMAGRPM.[1].I.[1].LALGKIAYIWFMSSIWCLAP.[1].FG.[1].SRY.[1].PEG 195
gi 128997  130 SLVLIAVERHQLIINPRGWRPN.[1].R     HAYVGIAVIWVLAVASSLPF.[1].IY.[1].VMT.[1].EPF 184
gi 266633  122 SMTAVAFDRYMAIIHPLQPRLS.[1].T     ATKVVICVIWVLALLLAFPQ.[1].YY.[1].TTE.[1].MPG 176
gi 112805  252 NLCAISVDRYLAITKPLEYGVK.[1].T.[2].RMMLCVGIVWLAAACISLPP     LL.[1].LGN.[1].HED 307
gi 8488960 172 TLCVISIDRYLGITRPLTYPVR.[1].N.[2].CMAKMILSVWLLSASITLPP     LF.[1].WAQ.[1].VND 227
gi 1168243 186 SLCTISVDRYVGVRHSLKYPAI.[1].T.[2].KAAAILALLWVVALVVSVGP     LL.[1].WKE.[1].VPP 241

gi 129203  179 .[ 4].SCGPD.[12].YTWFLFIFCFIVPLSLICFSYTQLLRALKAVAAQQQE.[ 3].TQKAEREVSRMVVVMVGSFC 259
query      198 .[ 4].SCGPD.[12].YMIVLMVTCCITPLSIIVLCYLQVWLAIRAVAKQQKE.[ 3].TQKAEKEVTRMVVVMVLAFC 278
gi 129197  202 .[ 4].SCTFD.[10].FVACIFFFSFVCPTTMITYYYSQIVGHVFSHEKALRD.[17].KETAEIRIAKAAITICFLFF 294
gi 129209  182 .[ 3].NCSFD.[10].NILCMYIFAFMCPIVVIFFCYFNIVMSVSNHEKEMAA.[15].GANAEMKLAKISIVIVTQFL 271
gi 129193  196 .[ 3].SCGID.[10].YLIFYSIFVYYIPLFLICYSYWFIIAAVSAHEKAMRE.[16].EKSAEGKLAKVALVTITLWF 286
gi 128997  185 .[12].VCFDQ.[ 9].YTTLLLVLQYFGPLCFIFICYFKIYIRLKRRNNMMDK.[ 6].RSSETKRINIMLLSIVVAFA 273
gi 266633  177 .[ 2].VCMIE.[12].YHICVTVLIYFLPLLVIGYAYTVVGITLWASEIPGDS.[ 6].QVSAKRKVVKMMIVVVCTFA 258
gi 112805  308 .[ 5].ICTVC.[ 4].YQIYATLGSFYIPLSVMLFVYYQIFRAARRIVLEEKR.[86].QLAKEKKASTTLGIIMSAFT 464
gi 8488960 228 .[ 2].VCLIS.[ 4].YTIYSTAVAFYIPMSVMLFMYYQIYKAARKSAAKHKF.[42].IFKREQKAATTLGIIVGAFT 337
gi 1168243 242 .[ 3].FCGIT.[ 4].YAVFSSVCSFYLPMAVIVVMYCRVYVVARSTTRSLEA.[48].KFSREKKAAKTLAIVVGVFV 358

gi 129203  260 VCYVPYAAFAMYMVNNR.[4].DLRLVTIPSFFSKSACIYNPIIY 303
query      279 FCWGPYAFFACFAAANP.[4].HPLMAALPAFFAKSATIYNPVIY 322
gi 129197  295 CSWTPYGVMSLIGAFGD.[4].TPGATMIPACACKMVACIDPFVY 338
gi 129209  272 LSWSPYAVVALLAQFGP.[4].TPYAAQLPVMFAKASAIHNPMIY 315
gi 129193  287 MAWTPYLVINCMGLFKF.[3].TPLNTIWGACFAKSAACYNPIVY 329
gi 128997  274 VCWLPLTIFNTVFDWNH.[7].HNLLFLLCHLTAMISTCVNPIFY 320
gi 266633  259 ICWLPFHIFFLLPYINP.[7].IQQVYLAIMWLAMSSTMYNPIIY 305
gi 112805  465 VCWLPFFILALIRPFET.[3].PASLSSLFLWLGYANSLLNPIIY 507
gi 8488960 338 VCWLPFFLLSTARPFIC.[7].PLWVERTFLWLGYANSLINPFIY 384
gi 1168243 359 LCWFPFFFVLPLGSLFP.[4].SEGVFKVIFWLGYFNSCVNPLIY 402
```

Most Significant Similarity Hits

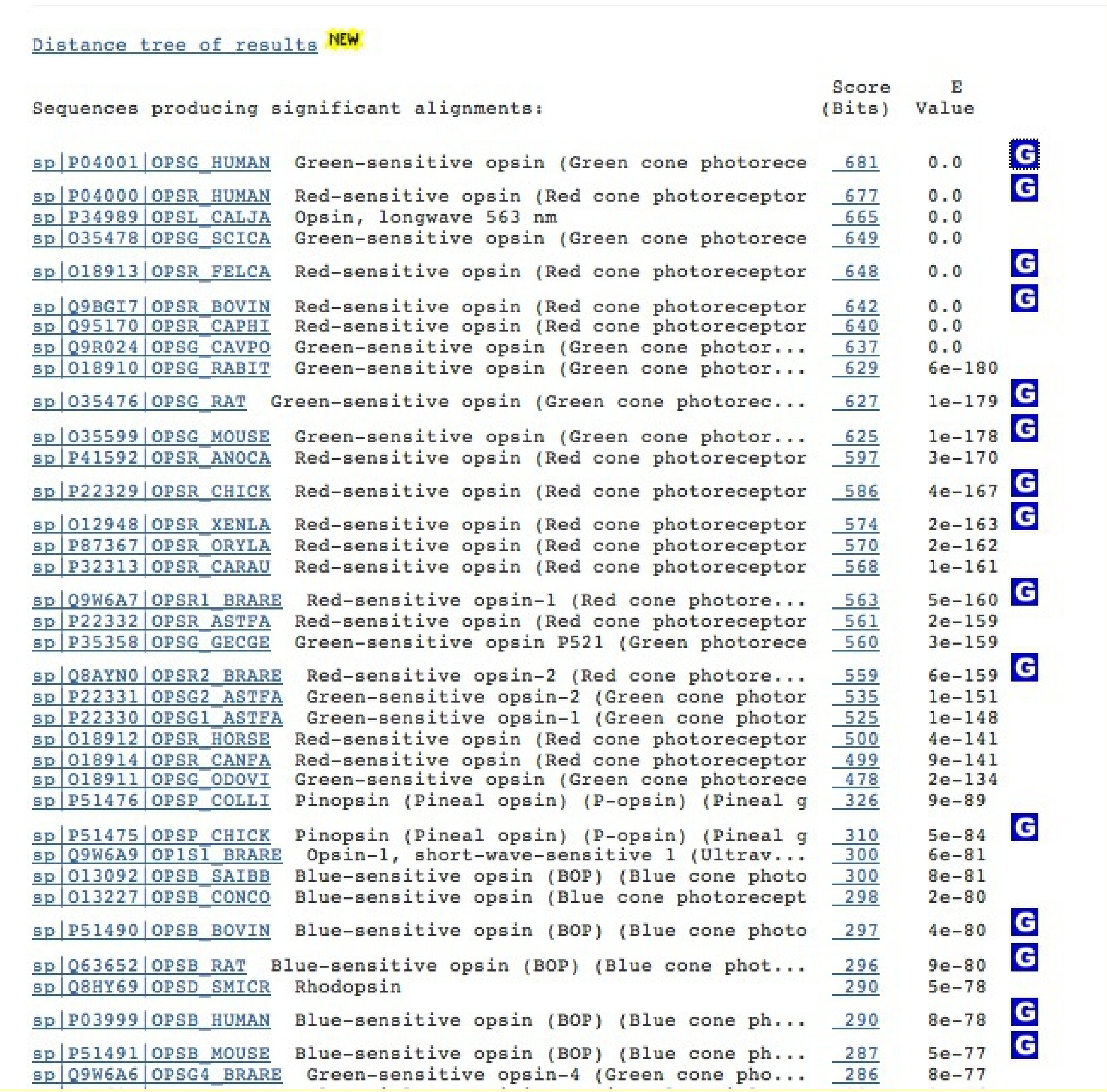


Distribution of 100 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

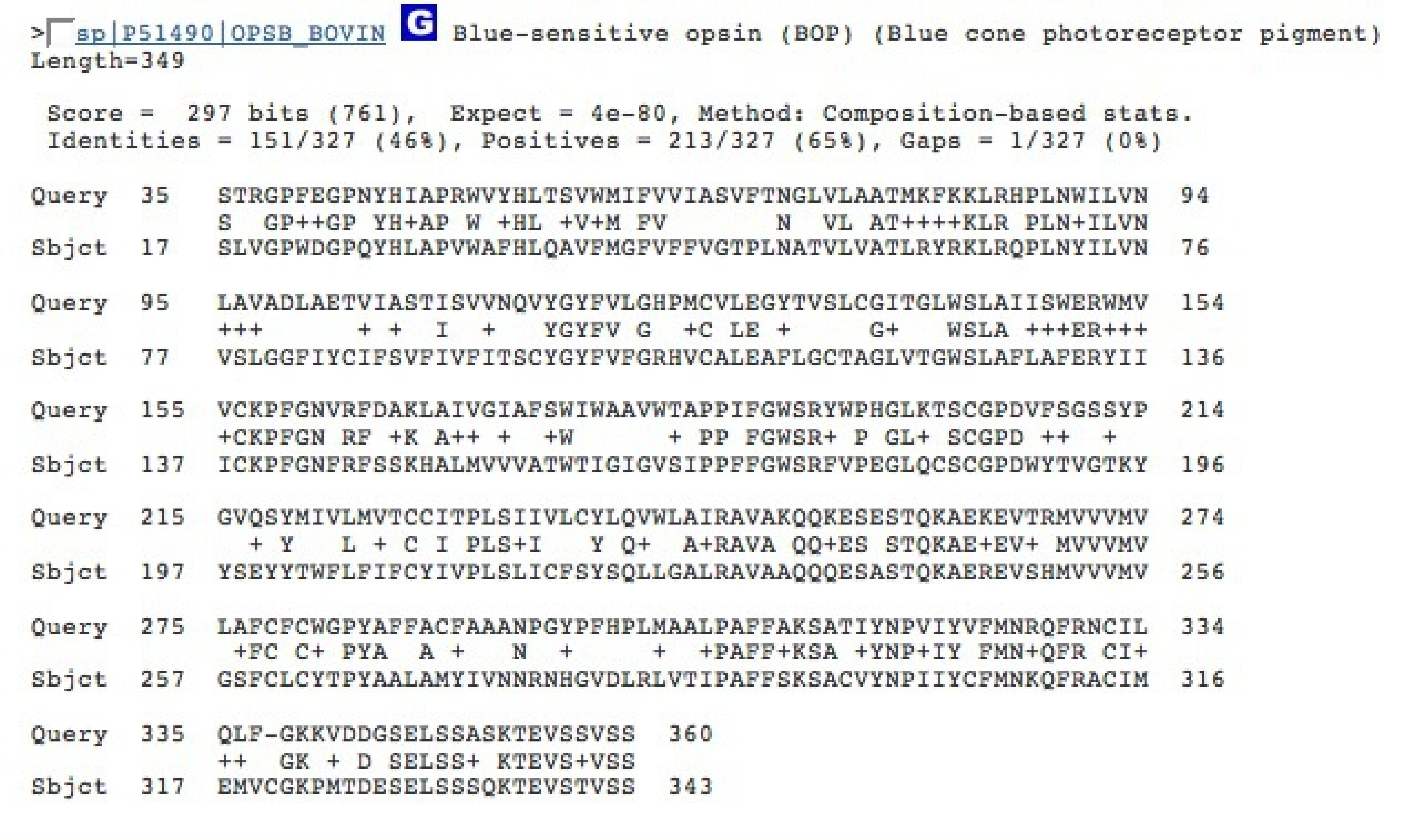
© Doug Brutlag 2015

Most Significant Similarity Hits



```
Distance tree of results NEW

                                                                Score    E
Sequences producing significant alignments:                     (Bits)  Value

sp|P04001|OPSG_HUMAN  Green-sensitive opsin (Green cone photorece   681    0.0    G

sp|P04000|OPSR_HUMAN  Red-sensitive opsin (Red cone photoreceptor   677    0.0    G
sp|P34989|OPSL_CALJA  Opsin, longwave 563 nm                        665    0.0
sp|O35478|OPSG_SCICA  Green-sensitive opsin (Green cone photorece   649    0.0

sp|O18913|OPSR_FELCA  Red-sensitive opsin (Red cone photoreceptor   648    0.0    G

sp|Q9BGI7|OPSR_BOVIN  Red-sensitive opsin (Red cone photoreceptor   642    0.0    G
sp|Q95170|OPSR_CAPHI  Red-sensitive opsin (Red cone photoreceptor   640    0.0
sp|Q9R024|OPSG_CAVPO  Green-sensitive opsin (Green cone photor...   637    0.0
sp|O18910|OPSG_RABIT  Green-sensitive opsin (Green cone photor...   629    6e-180

sp|O35476|OPSG_RAT  Green-sensitive opsin (Green cone photorec...   627    1e-179 G

sp|O35599|OPSG_MOUSE  Green-sensitive opsin (Green cone photor...   625    1e-178 G
sp|P41592|OPSR_ANOCA  Red-sensitive opsin (Red cone photoreceptor   597    3e-170

sp|P22329|OPSR_CHICK  Red-sensitive opsin (Red cone photoreceptor   586    4e-167 G

sp|O12948|OPSR_XENLA  Red-sensitive opsin (Red cone photoreceptor   574    2e-163 G
sp|P87367|OPSR_ORYLA  Red-sensitive opsin (Red cone photoreceptor   570    2e-162
sp|P32313|OPSR_CARAU  Red-sensitive opsin (Red cone photoreceptor   568    1e-161

sp|Q9W6A7|OPSR1_BRARE  Red-sensitive opsin-1 (Red cone photore...   563    5e-160 G
sp|P22332|OPSR_ASTFA  Red-sensitive opsin (Red cone photoreceptor   561    2e-159
sp|P35358|OPSG_GECGE  Green-sensitive opsin P521 (Green photorece   560    3e-159

sp|Q8AYN0|OPSR2_BRARE  Red-sensitive opsin-2 (Red cone photore...   559    6e-159 G
sp|P22331|OPSG2_ASTFA  Green-sensitive opsin-2 (Green cone photor   535    1e-151
sp|P22330|OPSG1_ASTFA  Green-sensitive opsin-1 (Green cone photor   525    1e-148
sp|O18912|OPSR_HORSE  Red-sensitive opsin (Red cone photoreceptor   500    4e-141
sp|O18914|OPSR_CANFA  Red-sensitive opsin (Red cone photoreceptor   499    9e-141
sp|O18911|OPSG_ODOVI  Green-sensitive opsin (Green cone photorece   478    2e-134
sp|P51476|OPSP_COLLI  Pinopsin (Pineal opsin) (P-opsin) (Pineal g   326    9e-89

sp|P51475|OPSP_CHICK  Pinopsin (Pineal opsin) (P-opsin) (Pineal g   310    5e-84  G
sp|Q9W6A9|OP1S1_BRARE  Opsin-1, short-wave-sensitive 1 (Ultrav...   300    6e-81
sp|O13092|OPSB_SAIBB  Blue-sensitive opsin (BOP) (Blue cone photo   300    8e-81
sp|O13227|OPSB_CONCO  Blue-sensitive opsin (Blue cone photorecept   298    2e-80

sp|P51490|OPSB_BOVIN  Blue-sensitive opsin (BOP) (Blue cone photo   297    4e-80  G

sp|Q63652|OPSB_RAT  Blue-sensitive opsin (BOP) (Blue cone phot...   296    9e-80  G
sp|Q8HY69|OPSD_SMICR  Rhodopsin                                     290    5e-78

sp|P03999|OPSB_HUMAN  Blue-sensitive opsin (BOP) (Blue cone ph...   290    8e-78  G

sp|P51491|OPSB_MOUSE  Blue-sensitive opsin (BOP) (Blue cone ph...   287    5e-77  G
sp|Q9W6A6|OPSG4_BRARE  Green-sensitive opsin-4 (Green cone pho...   286    8e-77
```

Bovine Blue Opsin Similarity



```
>☐sp|P51490|OPSB_BOVIN G Blue-sensitive opsin (BOP) (Blue cone photoreceptor pigment)
Length=349

 Score =  297 bits (761),  Expect = 4e-80, Method: Composition-based stats.
 Identities = 151/327 (46%), Positives = 213/327 (65%), Gaps = 1/327 (0%)

Query  35    STRGPFEGPNYHIAPRWVYHLTSVWMIFVVIASVFTNGLVLAATMKFKKLRHPLNWILVN  94
             S  GP++GP YH+AP W +HL +V+M FV       N  VL AT++++KLR PLN+ILVN
Sbjct  17    SLVGPWDGPQYHLAPVWAFHLQAVFMGFVFFVGTPLNATVLVATLRYRKLRQPLNYILVN  76

Query  95    LAVADLAETVIASTISVVNQVYGYFVLGHPMCVLEGYTVSLCGITGLWSLAIISWERWMV  154
             +++      + +  I  +   YGYFV G  +C LE +     G+   WSLA +++ER+++
Sbjct  77    VSLGGFIYCIFSVFIVFITSCYGYFVFGRHVCALEAFLGCTAGLVTGWSLAFLAFERYII  136

Query  155   VCKPFGNVRFDAKLAIVGIAFSWIWAAVWTAPPIFGWSRYWPHGLKTSCGPDVFSGSSYP  214
             +CKPFGN RF +K A++ +  +W      + PP FGWSR+ P GL+ SCGPD ++  +
Sbjct  137   ICKPFGNFRFSSKHALMVVVATWTIGIGVSIPPFFGWSRFVPEGLQCSCGPDWYTVGTKY  196

Query  215   GVQSYMIVLMVTCCITPLSIIVLCYLQVWLAIRAVAKQQKESESTQKAEKEVTRMVVVMV  274
               + Y   L + C I PLS+I   Y Q+  A+RAVA QQ+ES STQKAE+EV+ MVVVMV
Sbjct  197   YSEYYTWFLFIFCYIVPLSLICFSYSQLLGALRAVAAQQQESASTQKAEREVSHMVVVMV  256

Query  275   LAFCFCWGPYAFFACFAAANPGYPFHPLMAALPAFFAKSATIYNPVIYVFMNRQFRNCIL  334
              +FC C+ PYA  A +   N  +     +  +PAFF+KSA +YNP+IY FMN+QFR CI+
Sbjct  257   GSFCLCYTPYAALAMYIVNNRNHGVDLRLVTIPAFFSKSACVYNPIIYCFMNKQFRACIM  316

Query  335   QLF-GKKVDDGSELSSASKTEVSSVSS  360
             ++  GK + D SELSS+ KTEVS+VSS
Sbjct  317   EMVCGKPMTDESELSSSQKTEVSTVSS  343
```

Discovering Function from Protein Sequence

BLOCKs, PRINTs, PSSMS or Weight Matrices

Position

	1	2	3	4	5	6	7	8	9	10	11	12
A	2	1	3	13	10	12	67	4	13	9	1	2
R	7	5	8	9	4	0	1	16	7	0	1	0
N	0	8	0	1	0	0	0	2	1	1	10	0
D	0	1	0	1	13	0	0	12	1	0	4	0
C	0	0	1	0	0	0	0	0	0	2	2	1
Q	1	1	21	8	10	0	0	7	6	0	0	2
E	2	0	0	9	21	0	0	15	7	3	3	0
G	9	7	1	4	0	0	8	0	0	0	46	0
H	4	3	1	1	2	0	0	2	2	0	5	0
I	10	0	11	1	2	10	0	4	9	3	0	16
L	16	1	17	0	1	31	0	3	11	24	0	14
K	3	4	5	10	11	1	1	13	10	0	5	2
M	7	1	1	0	0	0	0	0	5	7	1	8
F	4	0	3	0	0	4	0	0	0	10	0	0
P	0	6	0	1	0	0	0	0	0	0	0	0
S	1	17	0	8	3	1	3	0	2	2	2	0
T	5	22	3	11	1	5	0	2	2	2	0	5
W	2	0	0	0	0	0	0	0	0	1	0	1
Y	1	0	4	2	0	1	0	0	2	4	0	1
V	6	3	1	1	2	15	0	0	2	12	0	28

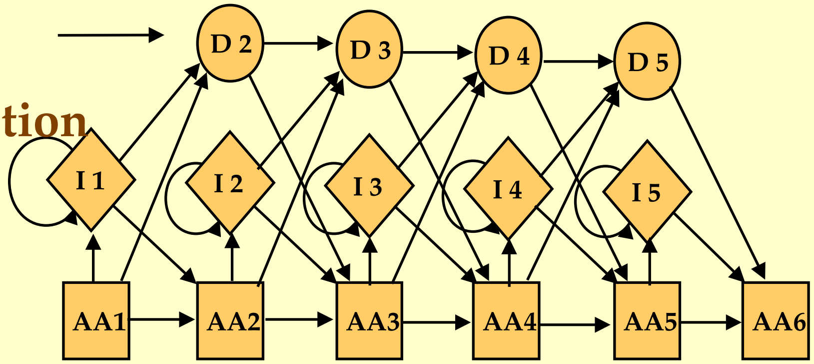
Consensus Sequences or Sequence Motifs

Zinc Finger (C2H2 type)

C X{2,4} C X{12} H X{3,5} H

Sequences of Common Structure or Function

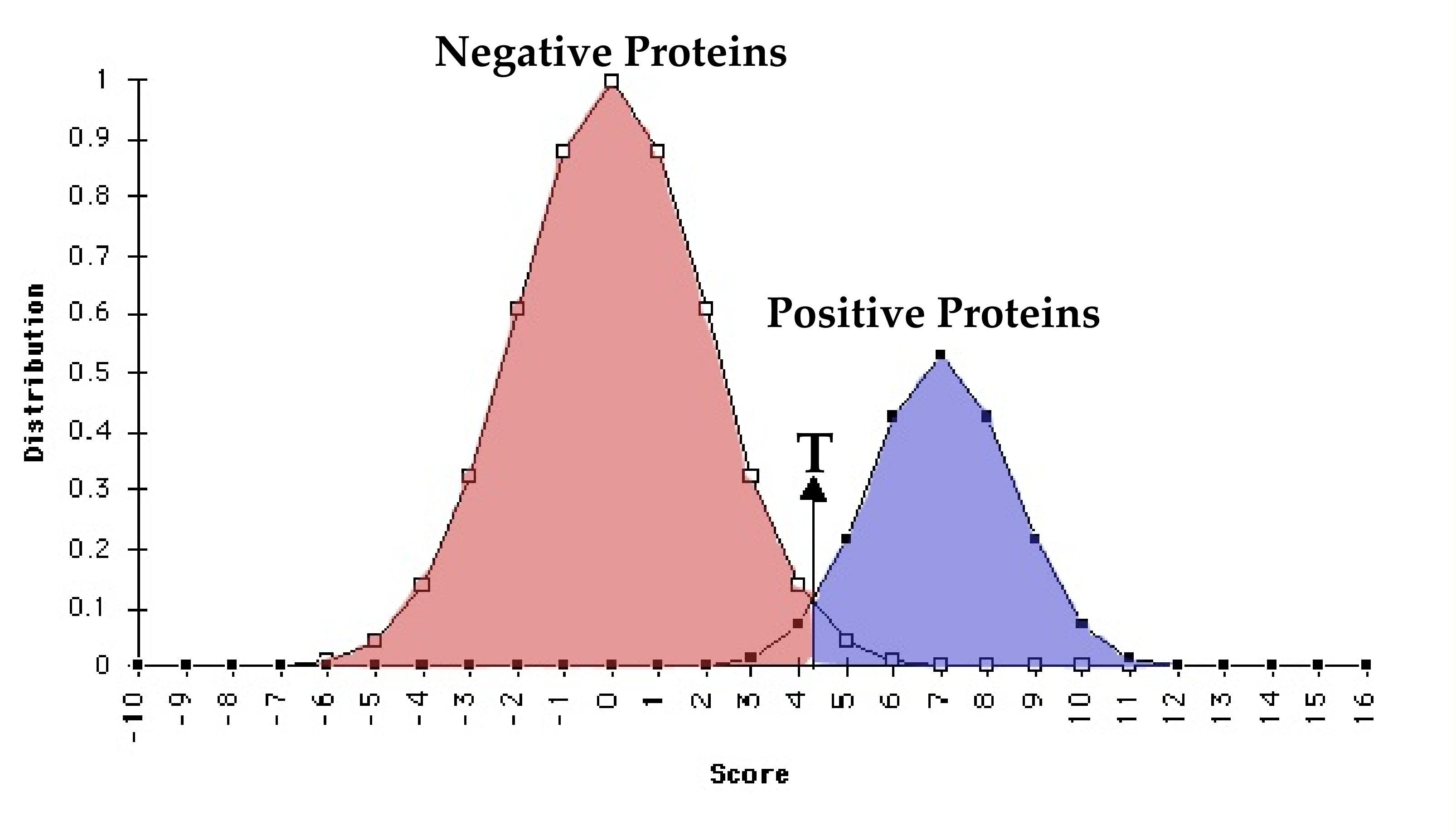
Profiles, PSI-BLAST Hidden Markov Models



Sequence Similarity

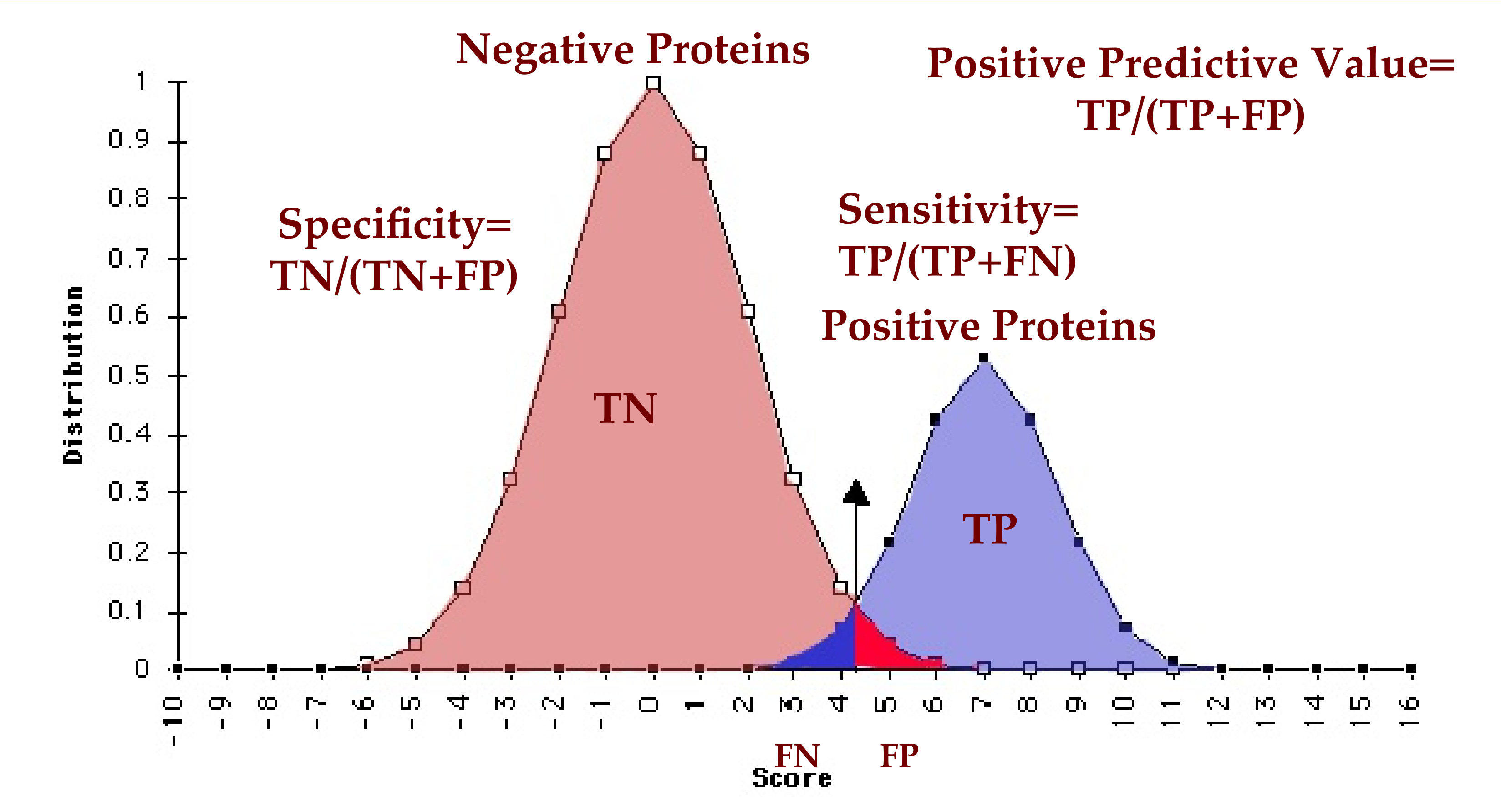
```
                     10        20        30        40              50
Query         HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
               |:| :|: | |:||||     | |:||| |: : :|:|  :|  |      |:  | 
Database      VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
                      10          20        30        40        50
```

Evaluation of PSSMs, Profiles and HMMs



© Doug Brutlag 2015

Evaluation of Profiles



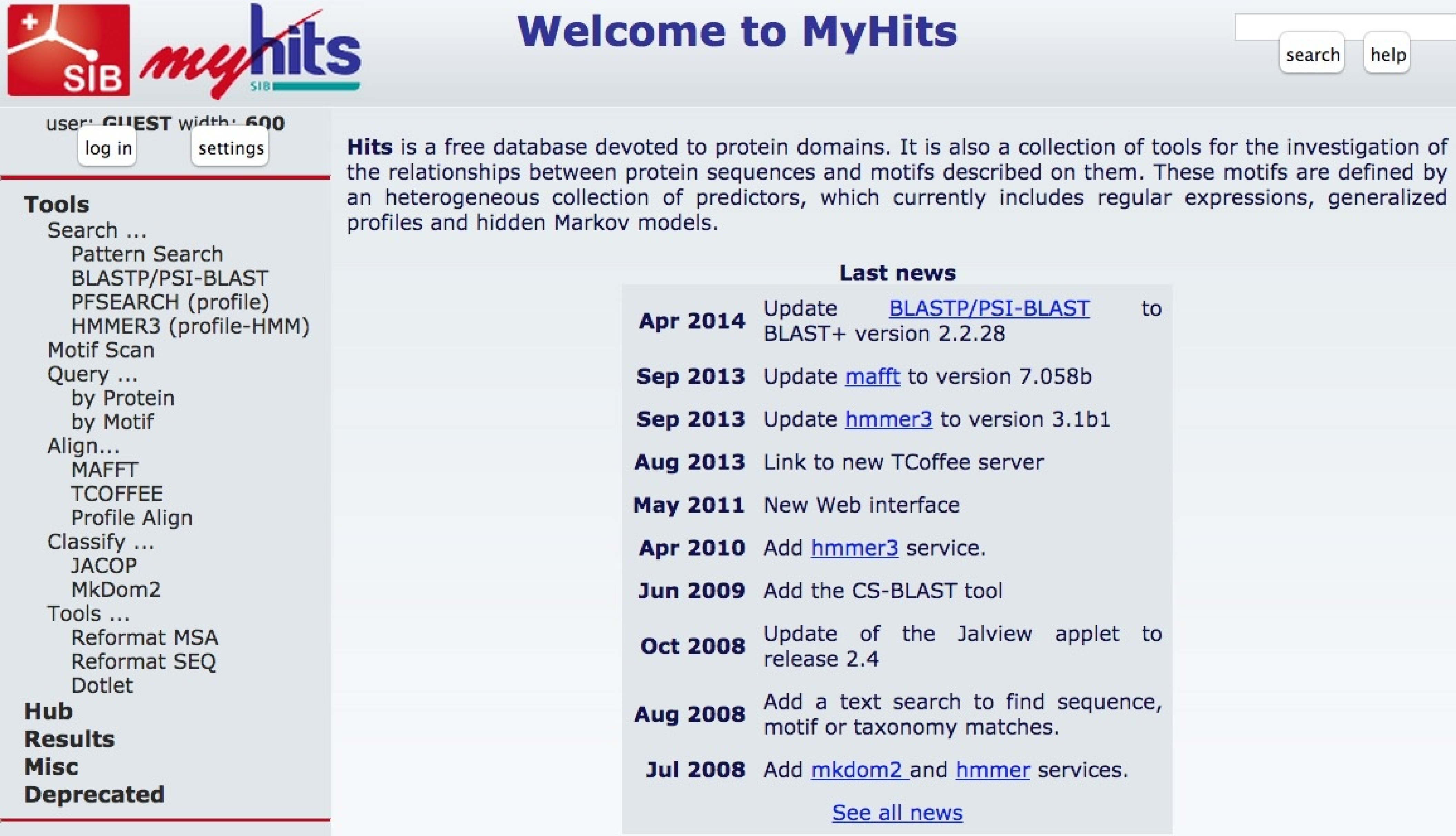
Specificity= TN/(TN+FP)

Sensitivity= TP/(TP+FN)

Positive Predictive Value= TP/(TP+FP)

MyHits Local Motifs Search

http://myhits.isb-sib.ch/



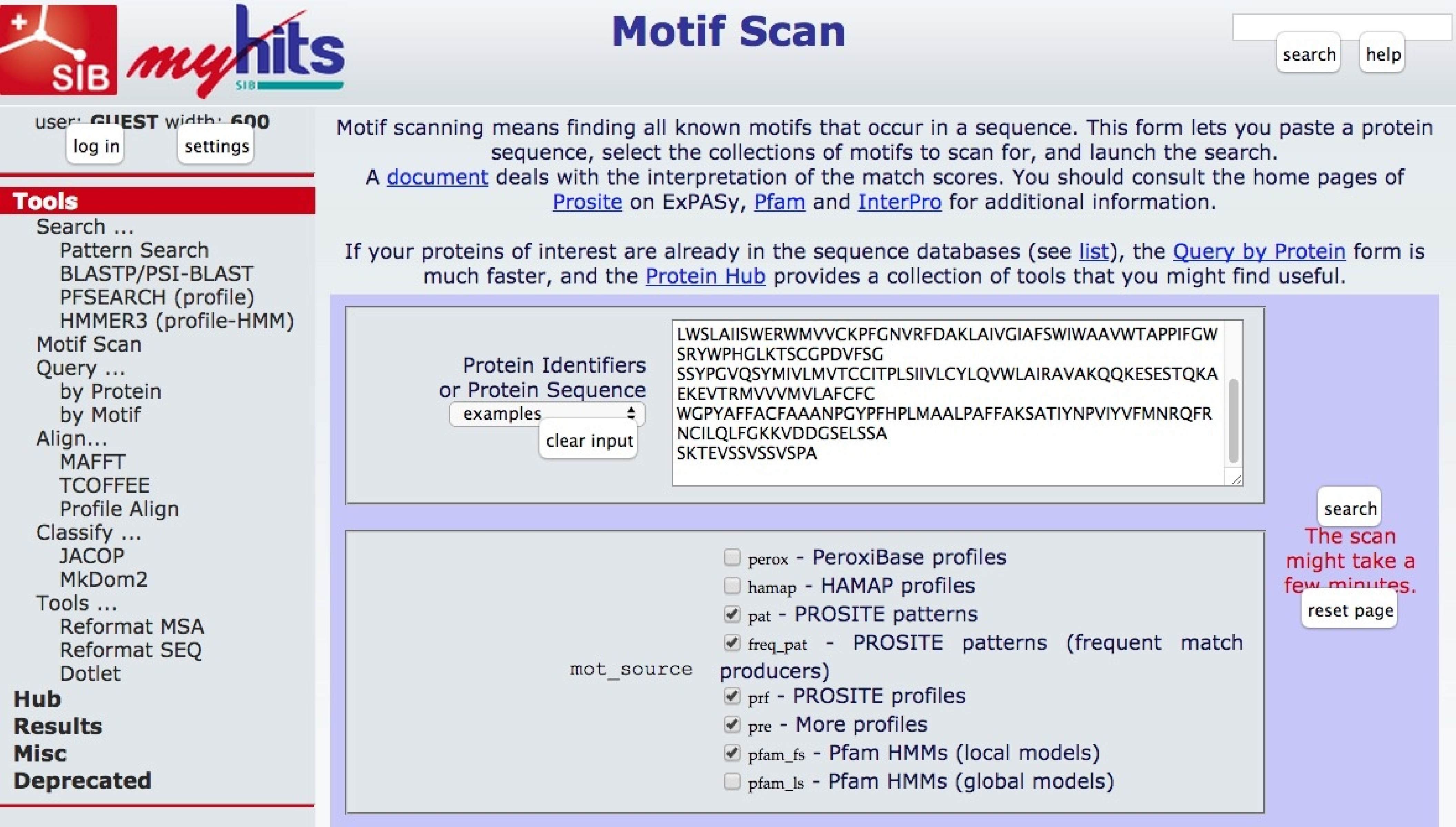
Hits is a free database devoted to protein domains. It is also a collection of tools for the investigation of the relationships between protein sequences and motifs described on them. These motifs are defined by an heterogeneous collection of predictors, which currently includes regular expressions, generalized profiles and hidden Markov models.

Last news

Apr 2014	Update BLASTP/PSI-BLAST to BLAST+ version 2.2.28
Sep 2013	Update mafft to version 7.058b
Sep 2013	Update hmmer3 to version 3.1b1
Aug 2013	Link to new TCoffee server
May 2011	New Web interface
Apr 2010	Add hmmer3 service.
Jun 2009	Add the CS-BLAST tool
Oct 2008	Update of the Jalview applet to release 2.4
Aug 2008	Add a text search to find sequence, motif or taxonomy matches.
Jul 2008	Add mkdom2 and hmmer services.

See all news

MyHits Local Motifs Query
http://myhits.isb-sib.ch/



Motif Scan

Motif scanning means finding all known motifs that occur in a sequence. This form lets you paste a protein sequence, select the collections of motifs to scan for, and launch the search.
A document deals with the interpretation of the match scores. You should consult the home pages of Prosite on ExPASy, Pfam and InterPro for additional information.

If your proteins of interest are already in the sequence databases (see list), the Query by Protein form is much faster, and the Protein Hub provides a collection of tools that you might find useful.

Protein Identifiers or Protein Sequence

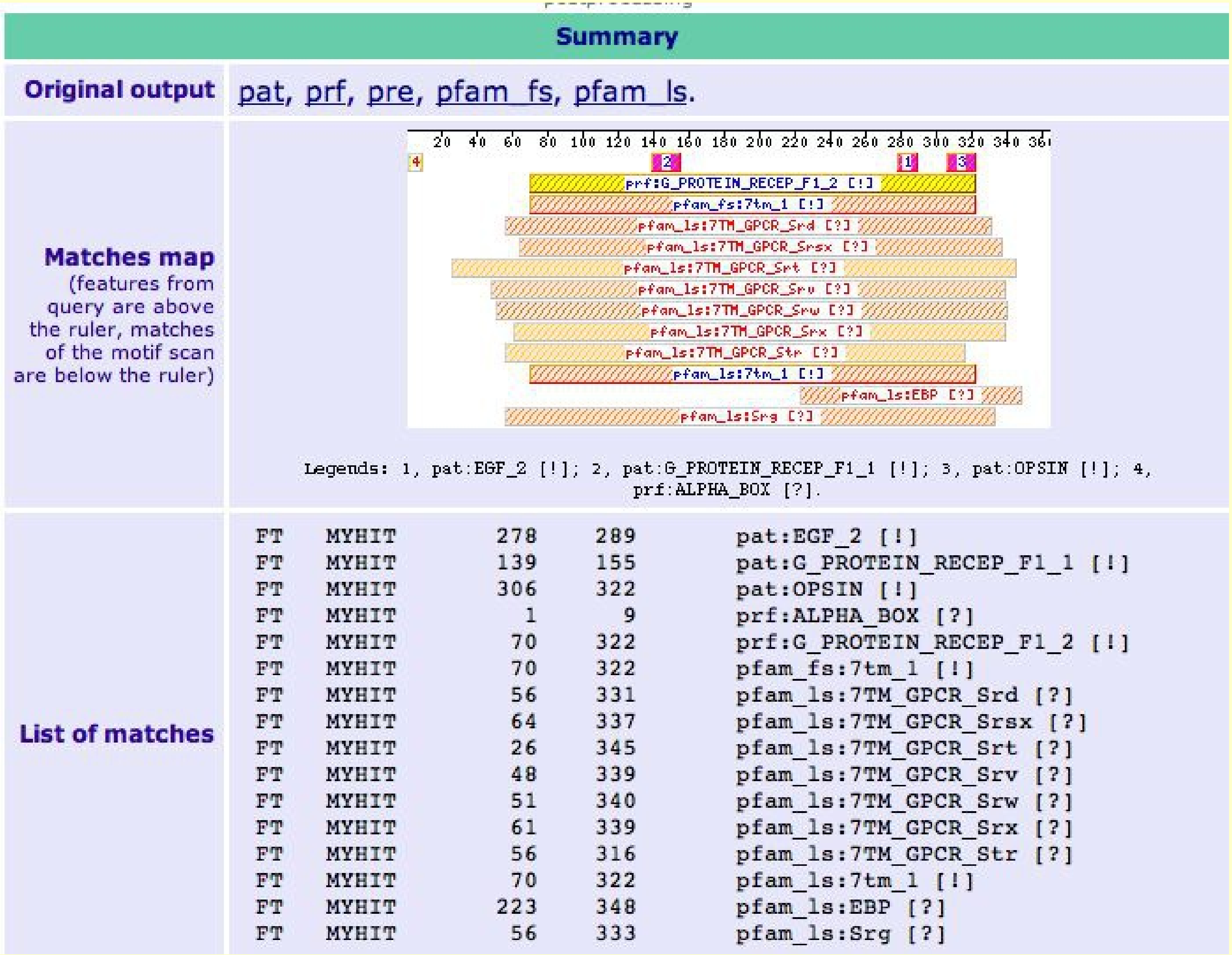
```
LWSLAIISWERWMVVCKPFGNVRFDAKLAIVGIAFSWIWAAVWTAPPIFGW
SRYWPHGLKTSCGPDVFSG
SSYPGVQSYMIVLMVTCCITPLSIIVLCYLQVWLAIRAVAKQQKESESTQKA
EKEVTRMVVVMVLAFCFC
WGPYAFFACFAAANPGYPFHPLMAALPAFFAKSATIYNPVIYVFMNRQFR
NCILQLFGKKVDDGSELSSA
SKTEVSSVSSVSPA
```

The scan might take a few minutes.

mot_source

- ☐ perox - PeroxiBase profiles
- ☐ hamap - HAMAP profiles
- ☑ pat - PROSITE patterns
- ☑ freq_pat - PROSITE patterns (frequent match producers)
- ☑ prf - PROSITE profiles
- ☑ pre - More profiles
- ☑ pfam_fs - Pfam HMMs (local models)
- ☐ pfam_ls - Pfam HMMs (global models)

MyHits Local Motifs Summary
http://myhits.isb-sib.ch/



Summary

Original output pat, prf, pre, pfam_fs, pfam_ls.

Matches map (features from query are above the ruler, matches of the motif scan are below the ruler)

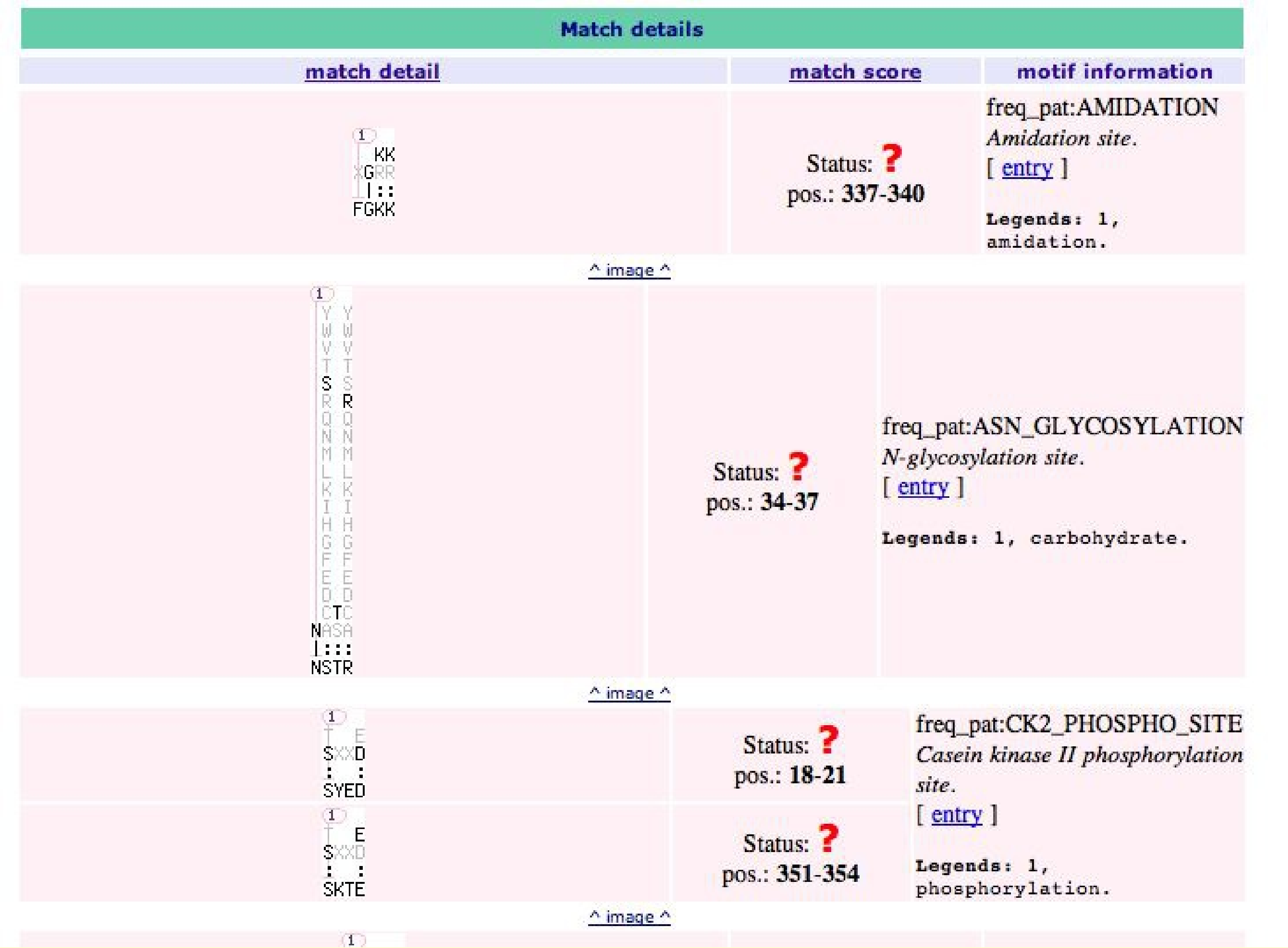
Legends: 1, pat:EGF_2 [!]; 2, pat:G_PROTEIN_RECEP_F1_1 [!]; 3, pat:OPSIN [!]; 4, prf:ALPHA_BOX [?].

List of matches

```
FT   MYHIT      278   289      pat:EGF_2 [!]
FT   MYHIT      139   155      pat:G_PROTEIN_RECEP_F1_1 [!]
FT   MYHIT      306   322      pat:OPSIN [!]
FT   MYHIT        1     9      prf:ALPHA_BOX [?]
FT   MYHIT       70   322      prf:G_PROTEIN_RECEP_F1_2 [!]
FT   MYHIT       70   322      pfam_fs:7tm_1 [!]
FT   MYHIT       56   331      pfam_ls:7TM_GPCR_Srd [?]
FT   MYHIT       64   337      pfam_ls:7TM_GPCR_Srsx [?]
FT   MYHIT       26   345      pfam_ls:7TM_GPCR_Srt [?]
FT   MYHIT       48   339      pfam_ls:7TM_GPCR_Srv [?]
FT   MYHIT       51   340      pfam_ls:7TM_GPCR_Srw [?]
FT   MYHIT       61   339      pfam_ls:7TM_GPCR_Srx [?]
FT   MYHIT       56   316      pfam_ls:7TM_GPCR_Str [?]
FT   MYHIT       70   322      pfam_ls:7tm_1 [!]
FT   MYHIT      223   348      pfam_ls:EBP [?]
FT   MYHIT       56   333      pfam_ls:Srg [?]
```

MyHits Local Motif Hits

http://myhits.isb-sib.ch/

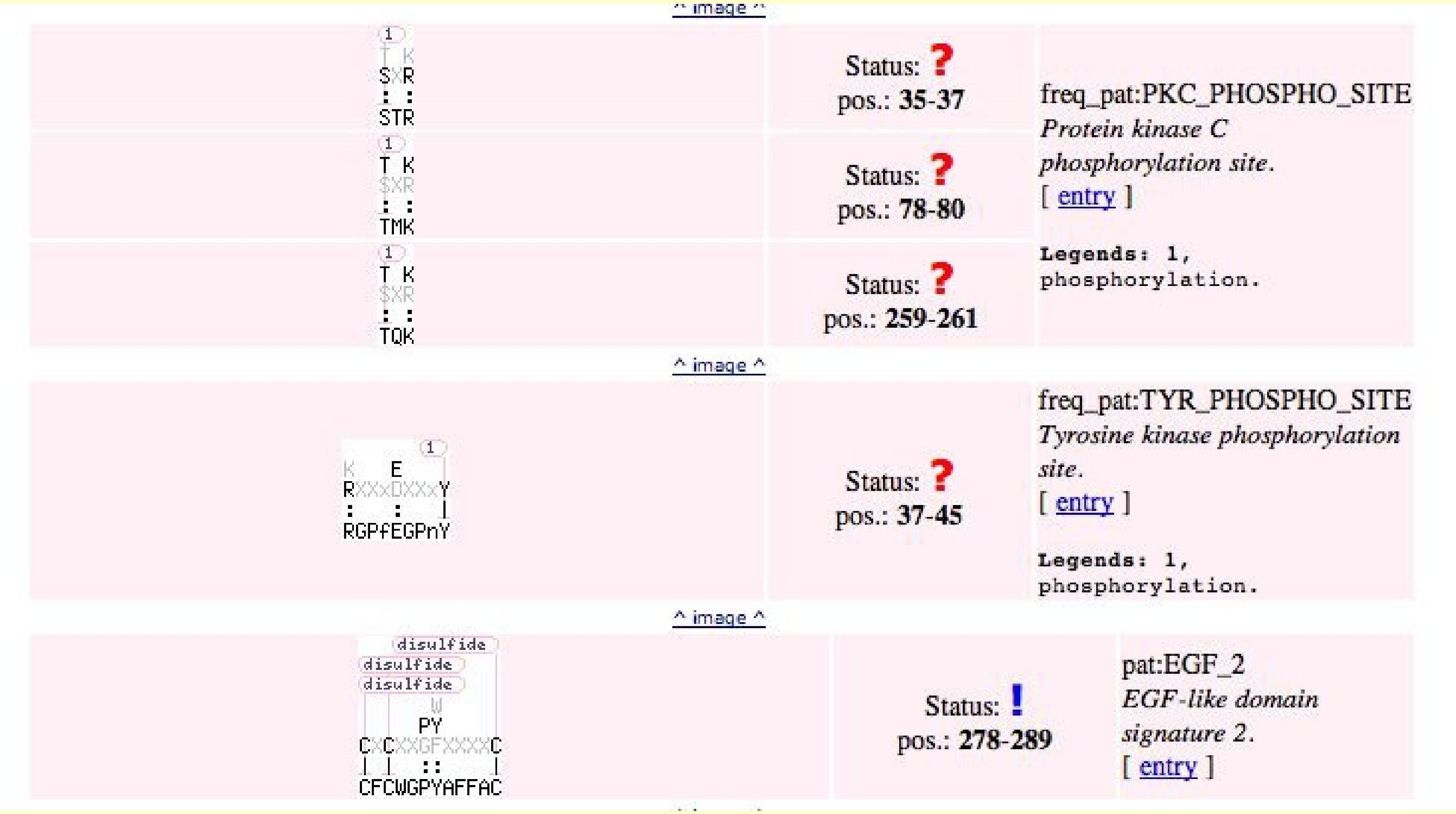


Match details

match detail	match score	motif information
	Status: ? pos.: **337-340**	freq_pat:AMIDATION *Amidation site.* [entry] **Legends:** 1, amidation.
^ image ^		
	Status: ? pos.: **34-37**	freq_pat:ASN_GLYCOSYLATION *N-glycosylation site.* [entry] **Legends:** 1, carbohydrate.
^ image ^		
	Status: ? pos.: **18-21**	freq_pat:CK2_PHOSPHO_SITE *Casein kinase II phosphorylation site.* [entry] **Legends:** 1, phosphorylation.
	Status: ? pos.: **351-354**	
^ image ^		

MyHits Local Motifs Hits (Cont.)

http://myhits.isb-sib.ch/



^ image ^

1 T K SXR : : STR	Status: ? pos.: **35-37**	freq_pat:PKC_PHOSPHO_SITE *Protein kinase C phosphorylation site.* [entry] **Legends:** 1, phosphorylation.
1 T K $XR : : TMK	Status: ? pos.: **78-80**	
1 T K $XR : : TQK	Status: ? pos.: **259-261**	

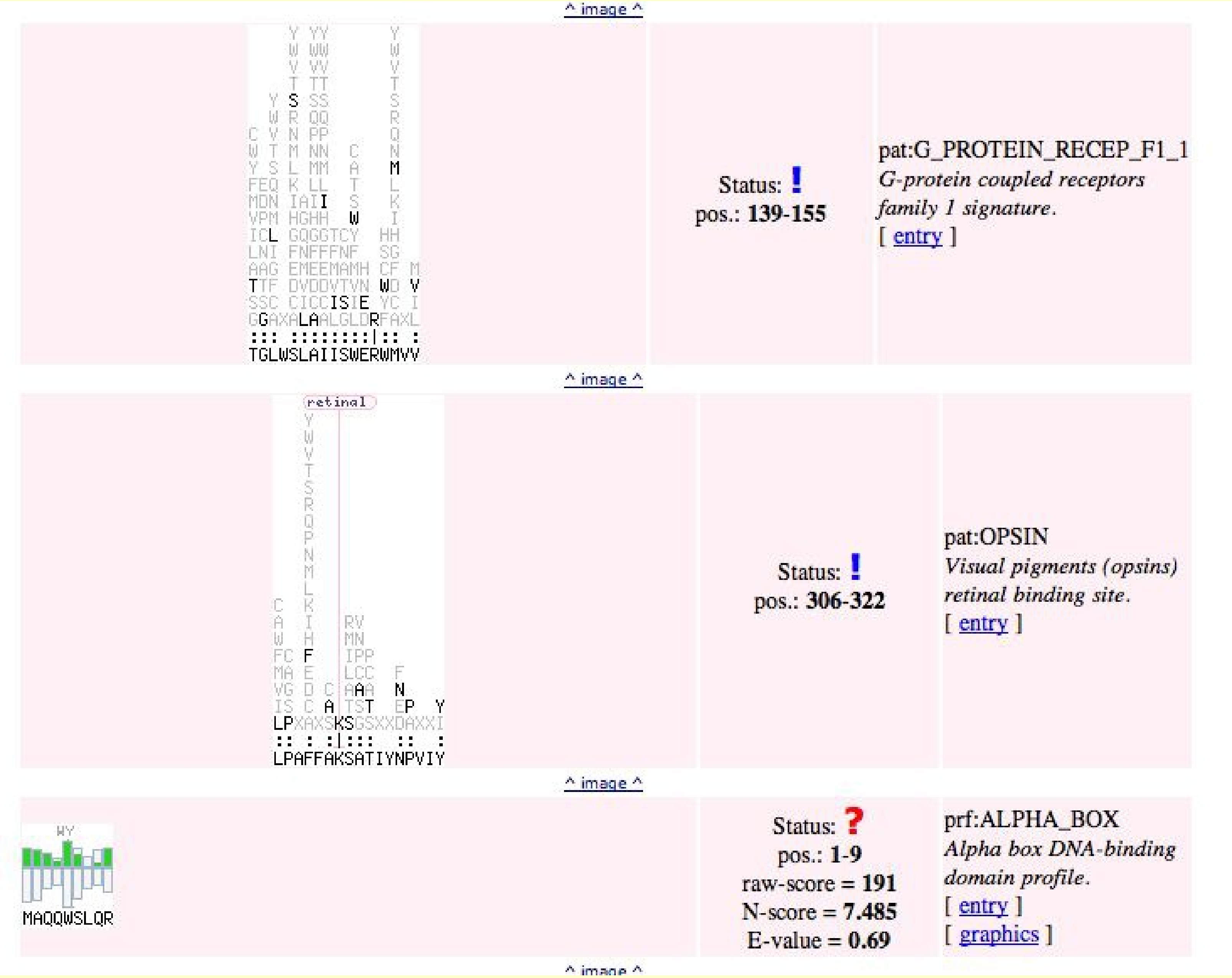
^ image ^

1 K E RXXxDXXxY : : | RGPfEGPnY	Status: ? pos.: **37-45**	freq_pat:TYR_PHOSPHO_SITE *Tyrosine kinase phosphorylation site.* [entry] **Legends:** 1, phosphorylation.

^ image ^

disulfide disulfide disulfide W PY CXCXXGFXXXXC | | :: | CFCWGPYAFFAC	Status: ! pos.: **278-289**	pat:EGF_2 *EGF-like domain signature 2.* [entry]

MyHits Local Motifs Hits (Cont.)



Status: !
pos.: 139-155

pat:G_PROTEIN_RECEP_F1_1
G-protein coupled receptors family 1 signature.
[entry]

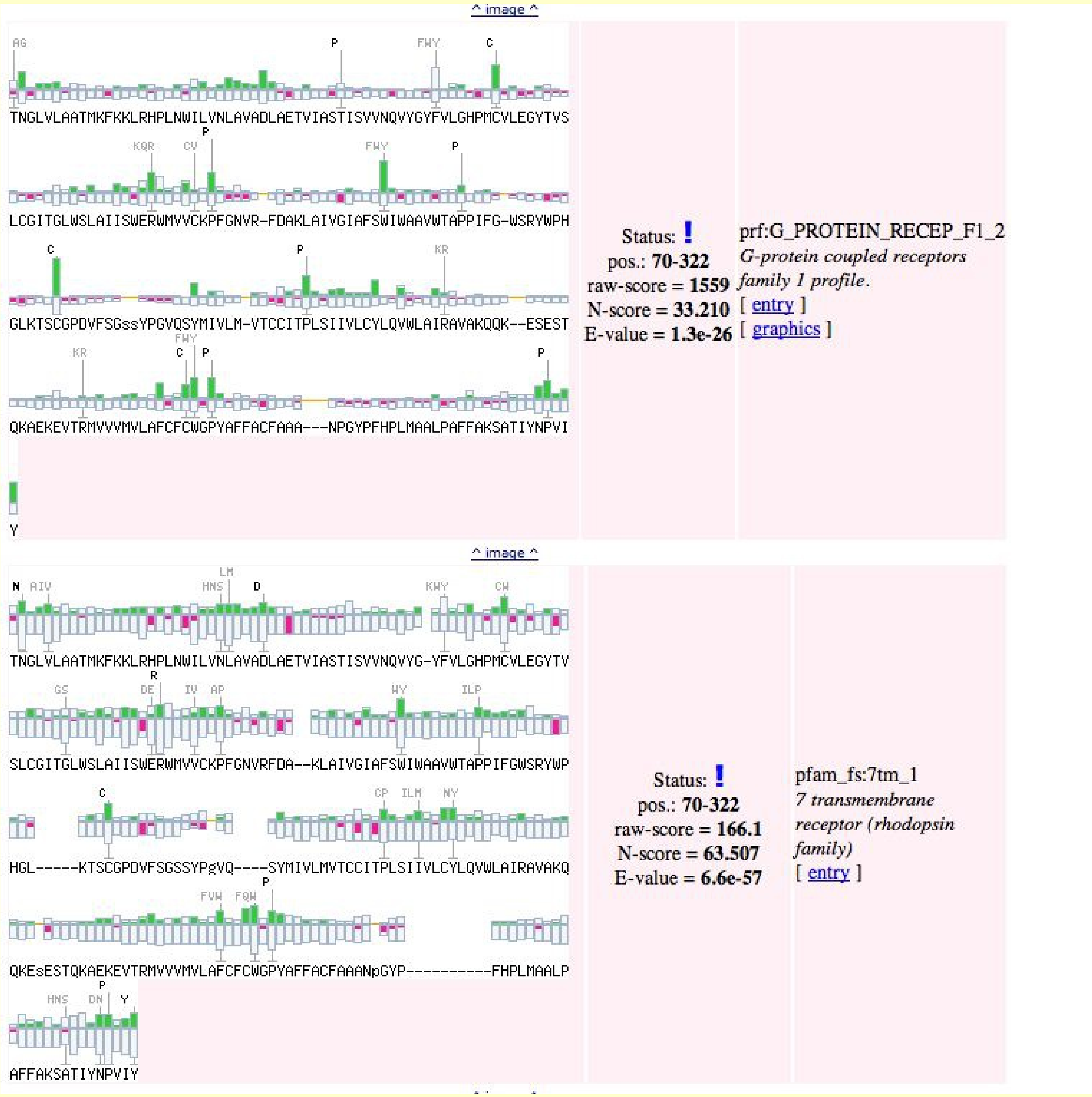
Status: !
pos.: 306-322

pat:OPSIN
Visual pigments (opsins) retinal binding site.
[entry]

Status: ?
pos.: 1-9
raw-score = 191
N-score = 7.485
E-value = 0.69

prf:ALPHA_BOX
Alpha box DNA-binding domain profile.
[entry]
[graphics]

MyHits Local Motifs Hits (Cont.)



```
TNGLVLAATMKFKKLRHPLNWILVNLAVADLAETVIASTISVVNQVYGYFVLGHPMCVLEGYTVS
LCGITGLWSLAIISWERWMVVCKPFGNVR-FDAKLAIVGIAFSWIWAAVWTAPPIFG-WSRYWPH
GLKTSCGPDVFSGssYPGVQSYMIVLM-VTCCITPLSIIVLCYLQVWLAIRAVAKQQK--ESEST
QKAEKEVTRMVVVMVLAFCFCWGPYAFFACFAAA---NPGYPFHPLMAALPAFFAKSATIYNPVI
Y
```

Status: **!**
pos.: **70-322**
raw-score = **1559**
N-score = **33.210**
E-value = **1.3e-26**

prf:G_PROTEIN_RECEP_F1_2
G-protein coupled receptors family 1 profile.
[entry]
[graphics]

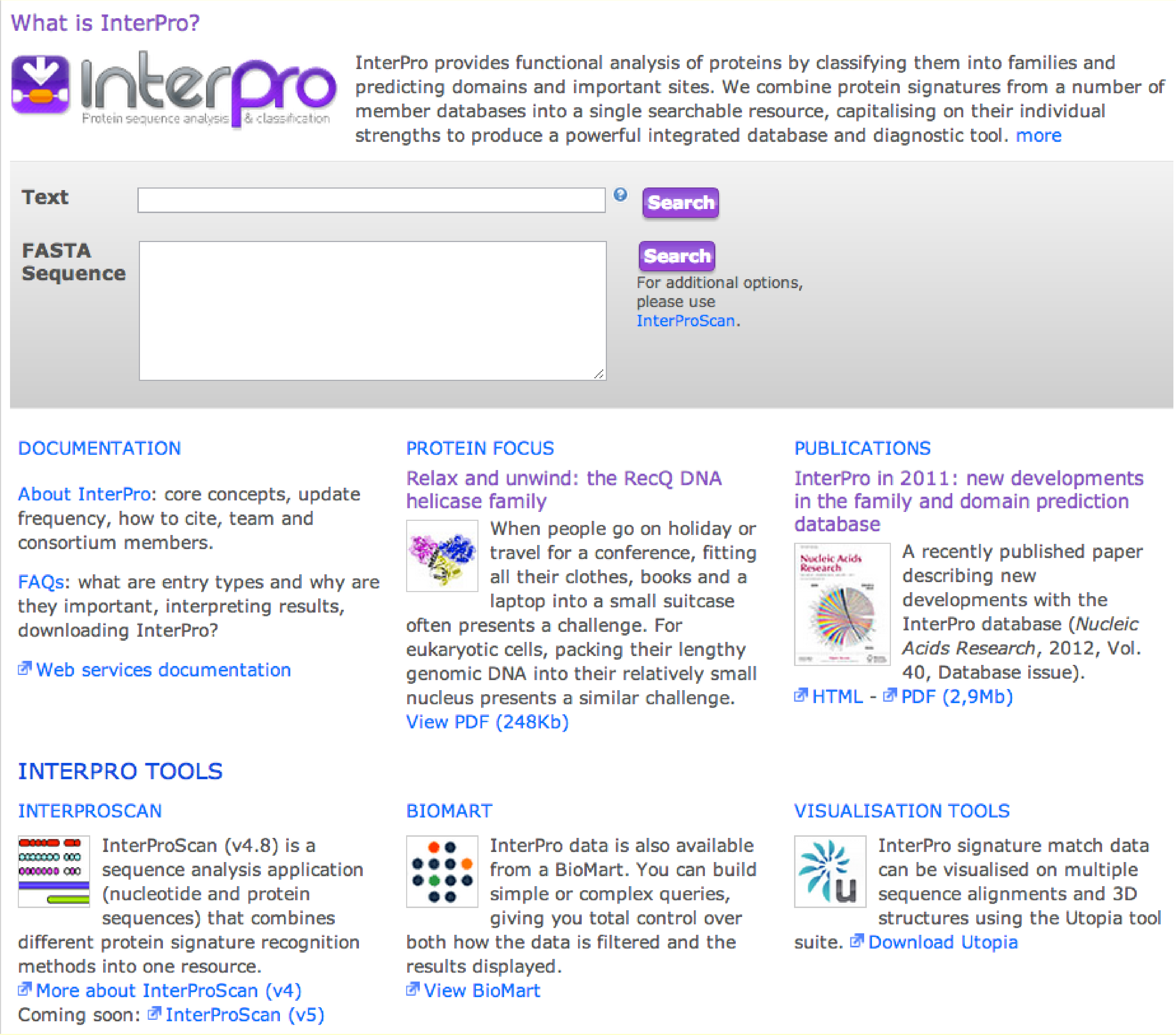
```
TNGLVLAATMKFKKLRHPLNWILVNLAVADLAETVIASTISVVNQVYG-YFVLGHPMCVLEGYTV
SLCGITGLWSLAIISWERWMVVCKPFGNVRFDA--KLAIVGIAFSWIWAAVWTAPPIFGWSRYWP
HGL-----KTSCGPDVFSGSSYPgVQ----SYMIVLMVTCCITPLSIIVLCYLQVWLAIRAVAKQ
QKEsESTQKAEKEVTRMVVVMVLAFCFCWGPYAFFACFAAANpGYP----------FHPLMAALP
AFFAKSATIYNPVIY
```

Status: **!**
pos.: **70-322**
raw-score = **166.1**
N-score = **63.507**
E-value = **6.6e-57**

pfam_fs:7tm_1
7 transmembrane receptor (rhodopsin family)
[entry]

InterPro

http://www.ebi.ac.uk/interpro/



What is InterPro?

InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. We combine protein signatures from a number of member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool. more

For additional options, please use InterProScan.

DOCUMENTATION

About InterPro: core concepts, update frequency, how to cite, team and consortium members.

FAQs: what are entry types and why are they important, interpreting results, downloading InterPro?

Web services documentation

PROTEIN FOCUS

Relax and unwind: the RecQ DNA helicase family

When people go on holiday or travel for a conference, fitting all their clothes, books and a laptop into a small suitcase often presents a challenge. For eukaryotic cells, packing their lengthy genomic DNA into their relatively small nucleus presents a similar challenge. View PDF (248Kb)

PUBLICATIONS

InterPro in 2011: new developments in the family and domain prediction database

A recently published paper describing new developments with the InterPro database (*Nucleic Acids Research*, 2012, Vol. 40, Database issue).

HTML - PDF (2,9Mb)

INTERPRO TOOLS

INTERPROSCAN

InterProScan (v4.8) is a sequence analysis application (nucleotide and protein sequences) that combines different protein signature recognition methods into one resource.

More about InterProScan (v4)

Coming soon: InterProScan (v5)

BIOMART

InterPro data is also available from a BioMart. You can build simple or complex queries, giving you total control over both how the data is filtered and the results displayed.

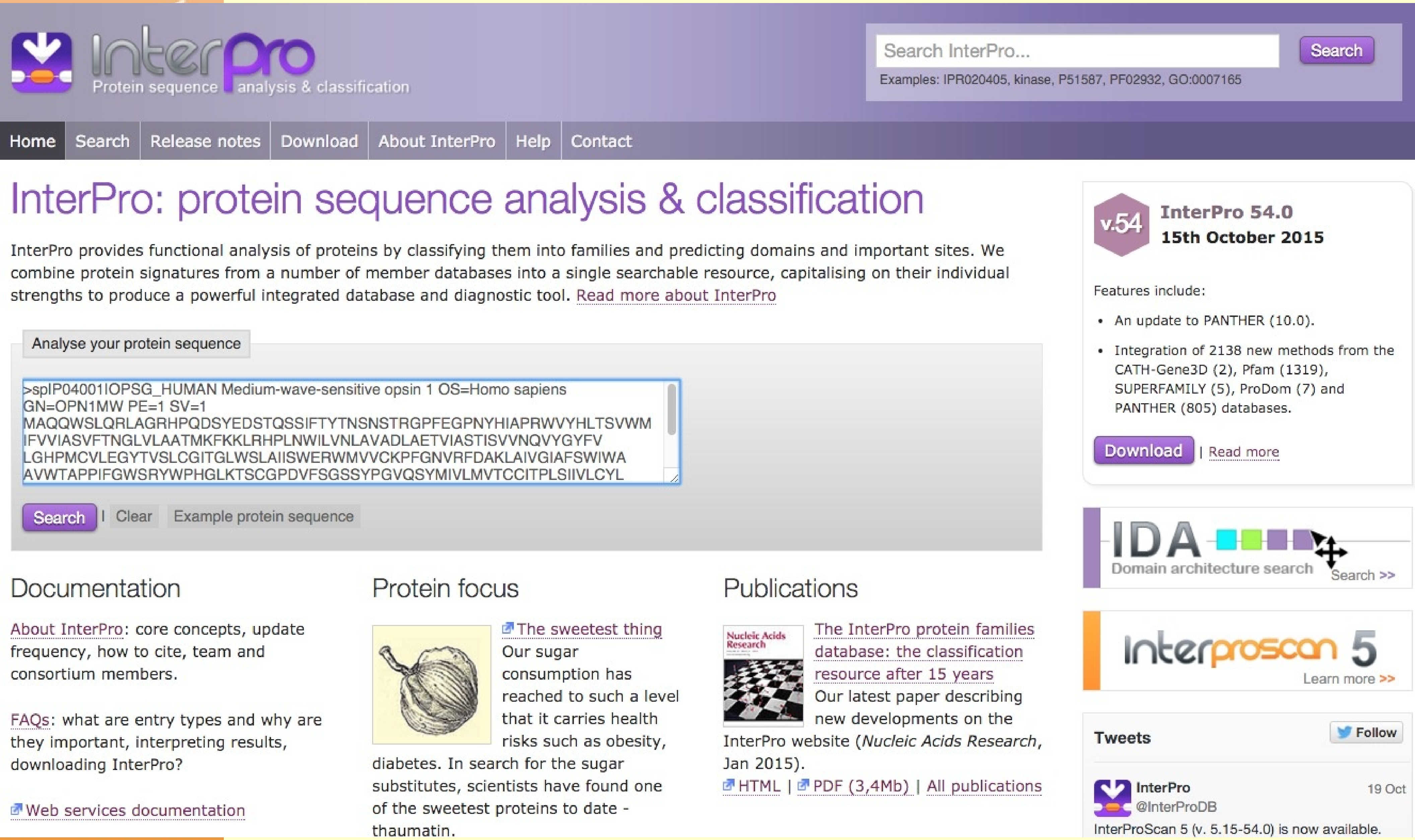
View BioMart

VISUALISATION TOOLS

InterPro signature match data can be visualised on multiple sequence alignments and 3D structures using the Utopia tool suite. Download Utopia

InterPro

http://www.ebi.ac.uk/interpro/index.html



Examples: IPR020405, kinase, P51587, PF02932, GO:0007165

InterPro: protein sequence analysis & classification

InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. We combine protein signatures from a number of member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool. Read more about InterPro

Analyse your protein sequence

```
>spIP04001IOPSG_HUMAN Medium-wave-sensitive opsin 1 OS=Homo sapiens
GN=OPN1MW PE=1 SV=1
MAQQWSLQRLAGRHPQDSYEDSTQSSIFTYTNSNSTRGPFEGPNYHIAPRWVYHLTSVWM
IFVVIASVFTNGLVLAATMKFKKLRHPLNWILVNLAVADLAETVIASTISVVNQVYGYFV
LGHPMCVLEGYTVSLCGITGLWSLAIISWERWMVVCKPFGNVRFDAKLAIVGIAFSWIWA
AVWTAPPIFGWSRYWPHGLKTSCGPDVFSGSSYPGVQSYMIVLMVTCCITPLSIIVLCYL
```

Documentation

About InterPro: core concepts, update frequency, how to cite, team and consortium members.

FAQs: what are entry types and why are they important, interpreting results, downloading InterPro?

Web services documentation

Protein focus

The sweetest thing

Our sugar consumption has reached to such a level that it carries health risks such as obesity, diabetes. In search for the sugar substitutes, scientists have found one of the sweetest proteins to date - thaumatin.

Publications

The InterPro protein families database: the classification resource after 15 years

Our latest paper describing new developments on the InterPro website (*Nucleic Acids Research*, Jan 2015).

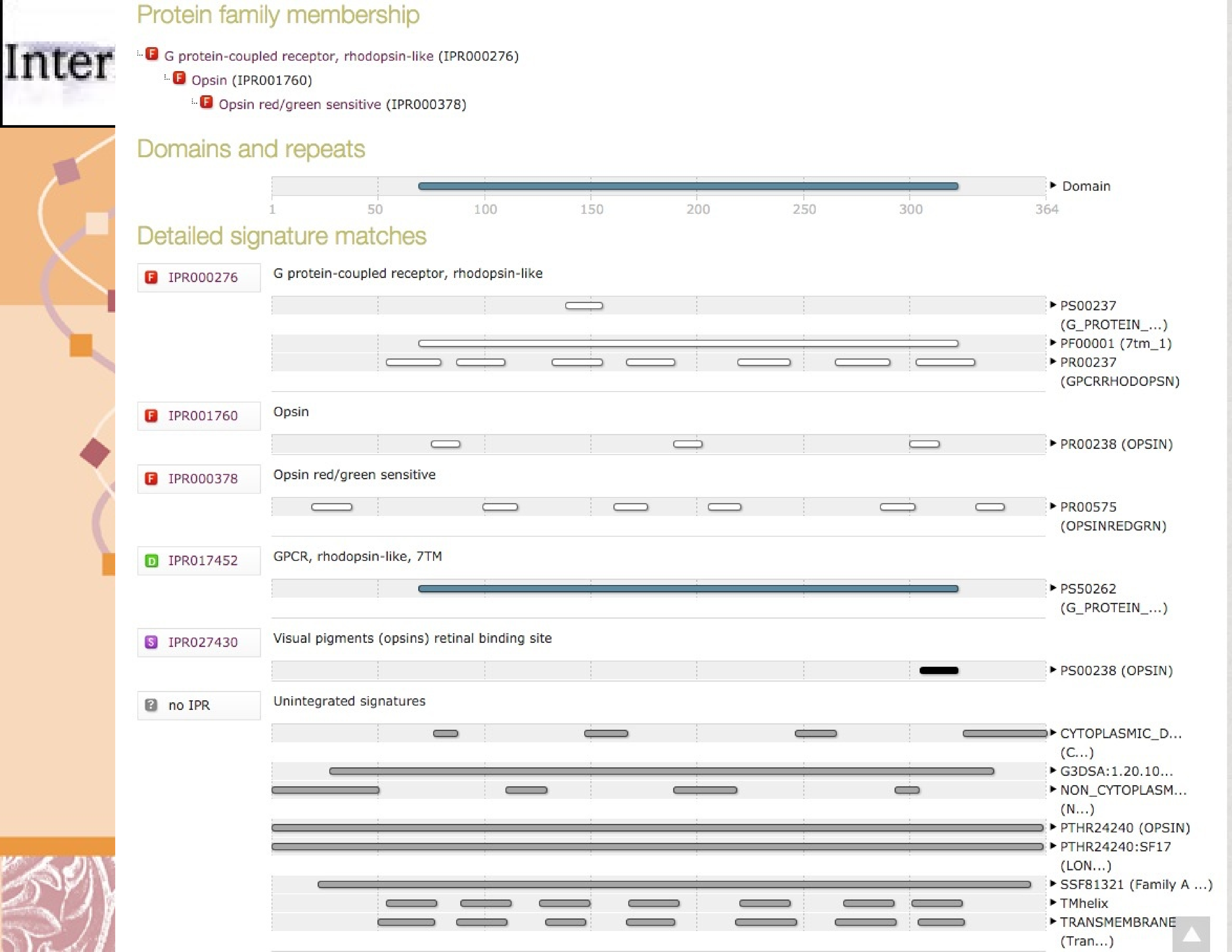
HTML | PDF (3,4Mb) | All publications

InterPro 54.0
15th October 2015

Features include:

- An update to PANTHER (10.0).
- Integration of 2138 new methods from the CATH-Gene3D (2), Pfam (1319), SUPERFAMILY (5), ProDom (7) and PANTHER (805) databases.

InterPro @InterProDB 19 Oct
InterProScan 5 (v. 5.15-54.0) is now available.



Protein family membership

- F G protein-coupled receptor, rhodopsin-like (IPR000276)
  - F Opsin (IPR001760)
    - F Opsin red/green sensitive (IPR000378)

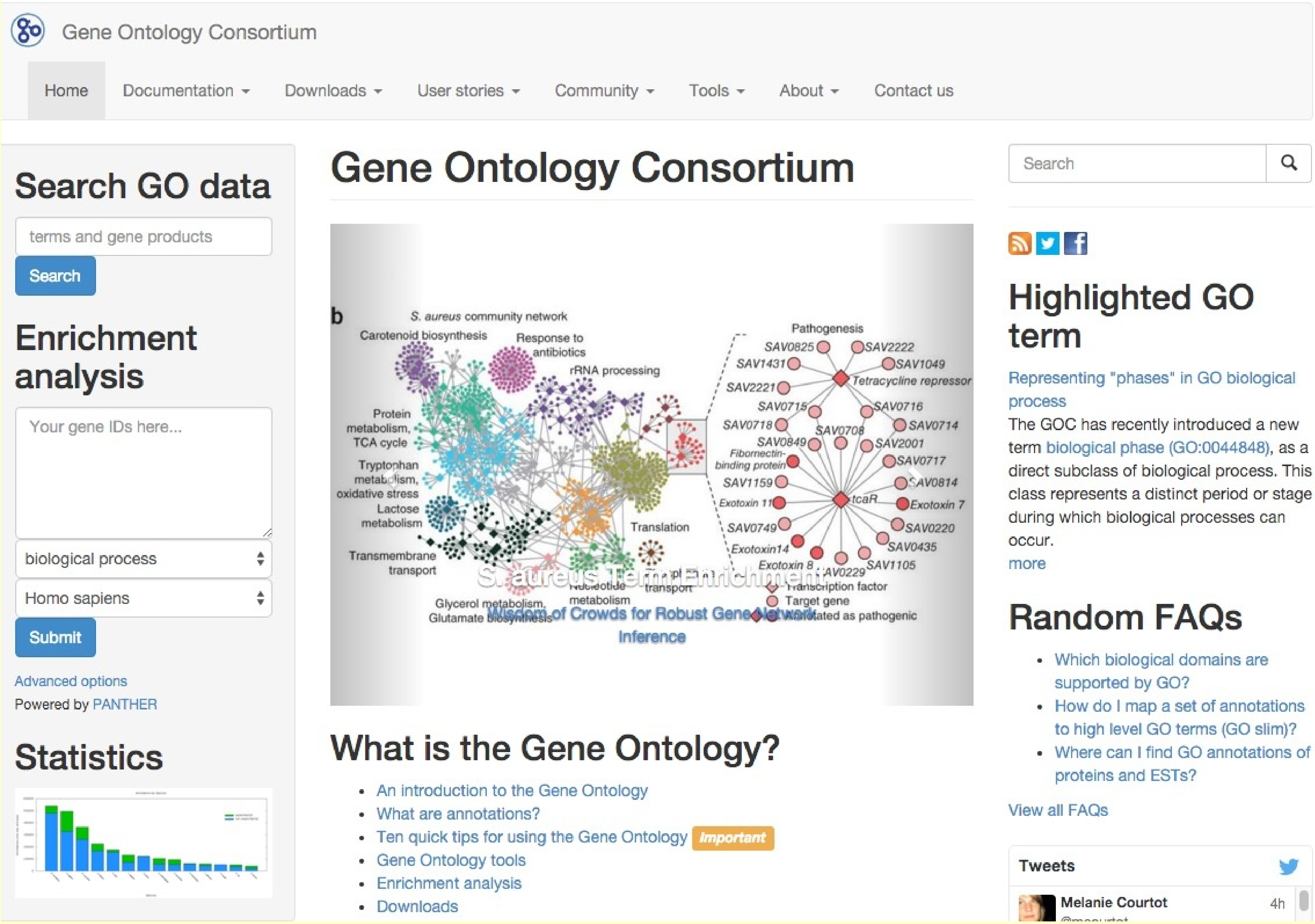
Domains and repeats

Detailed signature matches

F IPR000276	G protein-coupled receptor, rhodopsin-like	PS00237 (G_PROTEIN_...) PF00001 (7tm_1) PR00237 (GPCRRHODOPSN)
F IPR001760	Opsin	PR00238 (OPSIN)
F IPR000378	Opsin red/green sensitive	PR00575 (OPSINREDGRN)
D IPR017452	GPCR, rhodopsin-like, 7TM	PS50262 (G_PROTEIN_...)
S IPR027430	Visual pigments (opsins) retinal binding site	PS00238 (OPSIN)
? no IPR	Unintegrated signatures	CYTOPLASMIC_D... (C...) G3DSA:1.20.10... NON_CYTOPLASM... (N...) PTHR24240 (OPSIN) PTHR24240:SF17 (LON...) SSF81321 (Family A ...) TMhelix TRANSMEMBRANE (Tran...)

GO: Gene Ontology Database

http://www.geneontology.org/



Search GO data

terms and gene products

Search

Enrichment analysis

Your gene IDs here...

biological process

Homo sapiens

Submit

Advanced options

Powered by PANTHER

Statistics

Gene Ontology Consortium

S. aureus Term Enrichment

Wisdom of Crowds for Robust Gene Network Inference

What is the Gene Ontology?

- An introduction to the Gene Ontology
- What are annotations?
- Ten quick tips for using the Gene Ontology *Important*
- Gene Ontology tools
- Enrichment analysis
- Downloads

Highlighted GO term

Representing "phases" in GO biological process

The GOC has recently introduced a new term biological phase (GO:0044848), as a direct subclass of biological process. This class represents a distinct period or stage during which biological processes can occur.

more

Random FAQs

- Which biological domains are supported by GO?
- How do I map a set of annotations to high level GO terms (GO slim)?
- Where can I find GO annotations of proteins and ESTs?

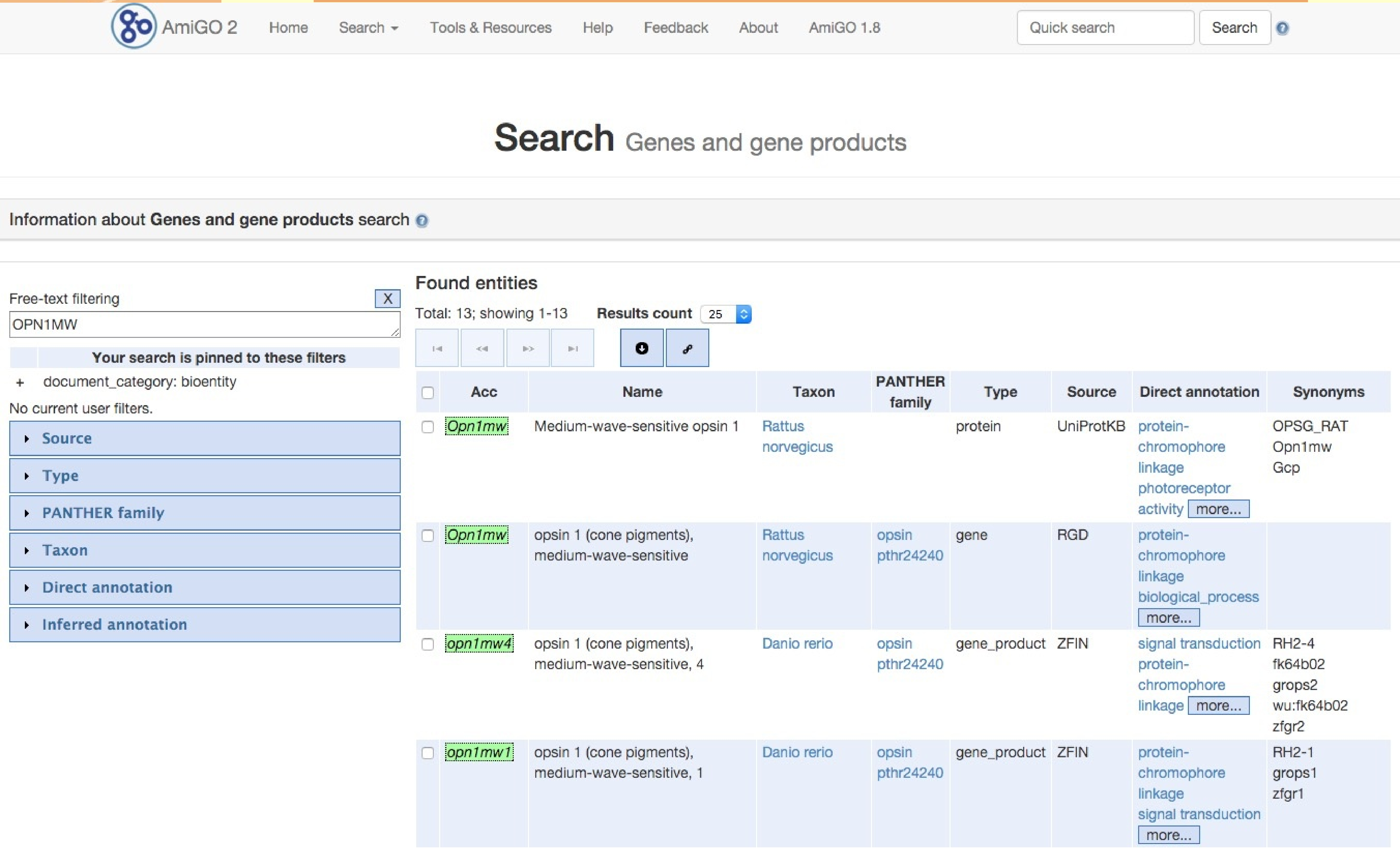
View all FAQs

Tweets

Melanie Courtot 4h

GO: Gene Ontology for Opsin OPN1MW

http://www.geneontology.org/



Search Genes and gene products

Information about **Genes and gene products** search

Free-text filtering X

OPN1MW

Your search is pinned to these filters

\+ document_category: bioentity

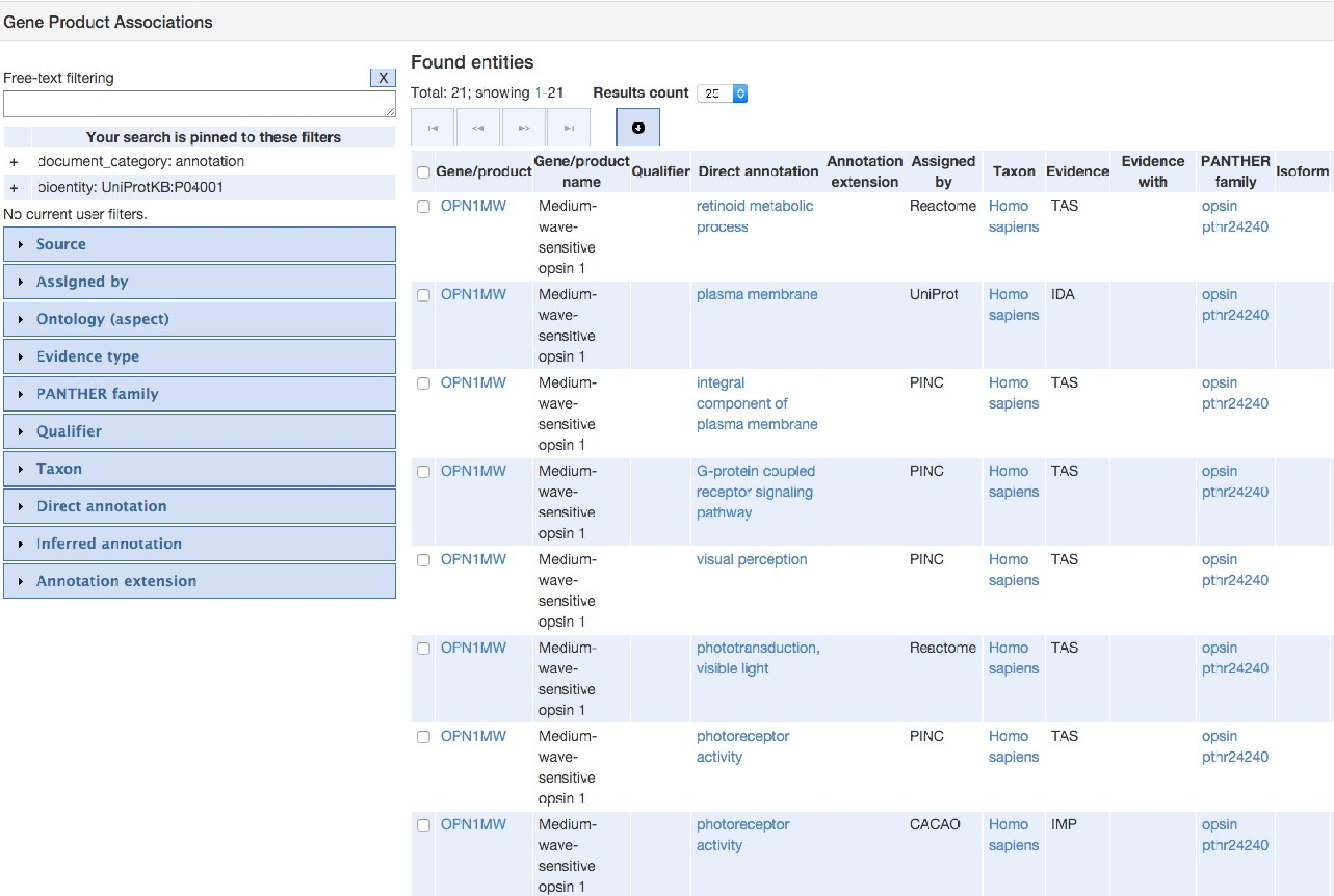
No current user filters.

Found entities

Total: 13; showing 1-13 **Results count** 25

☐	Acc	Name	Taxon	PANTHER family	Type	Source	Direct annotation	Synonyms
☐	Opn1mw	Medium-wave-sensitive opsin 1	Rattus norvegicus		protein	UniProtKB	protein-chromophore linkage photoreceptor activity more...	OPSG_RAT Opn1mw Gcp
☐	Opn1mw	opsin 1 (cone pigments), medium-wave-sensitive	Rattus norvegicus	opsin pthr24240	gene	RGD	protein-chromophore linkage biological_process more...	
☐	opn1mw4	opsin 1 (cone pigments), medium-wave-sensitive, 4	Danio rerio	opsin pthr24240	gene_product	ZFIN	signal transduction protein-chromophore linkage more...	RH2-4 fk64b02 grops2 wu:fk64b02 zfgr2
☐	opn1mw1	opsin 1 (cone pigments), medium-wave-sensitive, 1	Danio rerio	opsin pthr24240	gene_product	ZFIN	protein-chromophore linkage signal transduction more...	RH2-1 grops1 zfgr1

GO: Annotations for OPN1MW



Gene Product Associations

Free-text filtering X

Your search is pinned to these filters

+ document_category: annotation

+ bioentity: UniProtKB:P04001

No current user filters.

Found entities

Total: 21; showing 1-21 Results count 25

☐	Gene/product	Gene/product name	Qualifier	Direct annotation	Annotation extension	Assigned by	Taxon	Evidence	Evidence with	PANTHER family	Isoform
☐	OPN1MW	Medium-wave-sensitive opsin 1		retinoid metabolic process		Reactome	Homo sapiens	TAS		opsin pthr24240	
☐	OPN1MW	Medium-wave-sensitive opsin 1		plasma membrane		UniProt	Homo sapiens	IDA		opsin pthr24240	
☐	OPN1MW	Medium-wave-sensitive opsin 1		integral component of plasma membrane		PINC	Homo sapiens	TAS		opsin pthr24240	
☐	OPN1MW	Medium-wave-sensitive opsin 1		G-protein coupled receptor signaling pathway		PINC	Homo sapiens	TAS		opsin pthr24240	
☐	OPN1MW	Medium-wave-sensitive opsin 1		visual perception		PINC	Homo sapiens	TAS		opsin pthr24240	
☐	OPN1MW	Medium-wave-sensitive opsin 1		phototransduction, visible light		Reactome	Homo sapiens	TAS		opsin pthr24240	
☐	OPN1MW	Medium-wave-sensitive opsin 1		photoreceptor activity		PINC	Homo sapiens	TAS		opsin pthr24240	
☐	OPN1MW	Medium-wave-sensitive opsin 1		photoreceptor activity		CACAO	Homo sapiens	IMP		opsin pthr24240	

Bioinformatics Homework

http://biochem118.stanford.edu/bioinformatics.html

Homework Assignment

1) Select a **protein** from OMIM or from Entrez Gene or from UniProt concerning the disease of interest to you. Copy and save the FASTA format of the protein file.

2) Search your protein for motifs with the MyHits Motif Scan Query. Be sure to Include Prosite Patterns, Prosite Frequent Patterns, Prosite Profiles, Profiles, Pfam HMMSs (local Models) in your search. Please send me the MyHits you think are biologically significant and at least 1 or 2 hits which you think are not statistically or biologically significant. Please note that only the Profiles have expectation values. The Patterns do not have a measure of statistical significance.

3) Search your protein for blocks using the InterPro database. Please send me a few of the InterPro domains hits you think are significant and at least 1 or 2 hits which you think are not statistically or biologically significant. Please note that the default graphic output of InterPro does not list expectation values. You must switch to the Tabular view to obtain the statistical significance.

4) Search your protein for homology using the BLAST method. Please report two or three hits which are both statistically and biologically significant. Also report two or three hits which you think are neither statistically nor biologically significant. If your protein family is very large, you may have to ask BLAST to return more hits to find statistically insignificant hits.

Statistical vs. Biological Significance

Assignment

First, for each search (MyHits, InterPro and BLAST hit), I would like you to report some significance hits and describe why you think they are significant both statistically and biologically; also report some statistically insignificant hits (and why) and are any of your statistically insignificant hits, still significant biologically). To remind you what I said in class: a statistically significant find in the database search is always biologically significant, but a biologically significant result in the search is not necessarily always statistically significant.

Statistical significance and expectation values.

Statistical significance is determined by the expectation value which gives you a measure of how likely this finding is based on pure chance. A finding with an E-value of 1 or greater is not significant because it could occur by pure chance. A finding with an E-value less than 10^{-3} (one chance in a thousand) is generally considered statistically significant (unless of course you are doing a 1,000 searches!). So the lower the expectation value, the more significant the finding. Findings between 10^{-3} and 1 are in the so called twilight zone and require some further analysis or experiments to determine their validity.

Statistical vs. Biological Significance (cont.)

InterPro

Unlike most of the other methods, InterPro sets a very high level of significance for a finding before it will report it. This means that you will usually not find any statistically insignificant hits for this particular search.

Biological Significance

In order to determine biological significance you must read the biological properties (ontology terms are the most useful) of your protein and the biological properties of your findings. The findings may be significant because the finding defines a very closely related protein family (opsins for example) or a very broad family (G-coupled protein receptors or 7-transmembrane proteins) or a common structure (protein fold) or a specific function (retinal binding site) or a very specific catalytic activity. You should describe in words the level of the biological significance.

Statistical vs. Biological Significance (cont.)

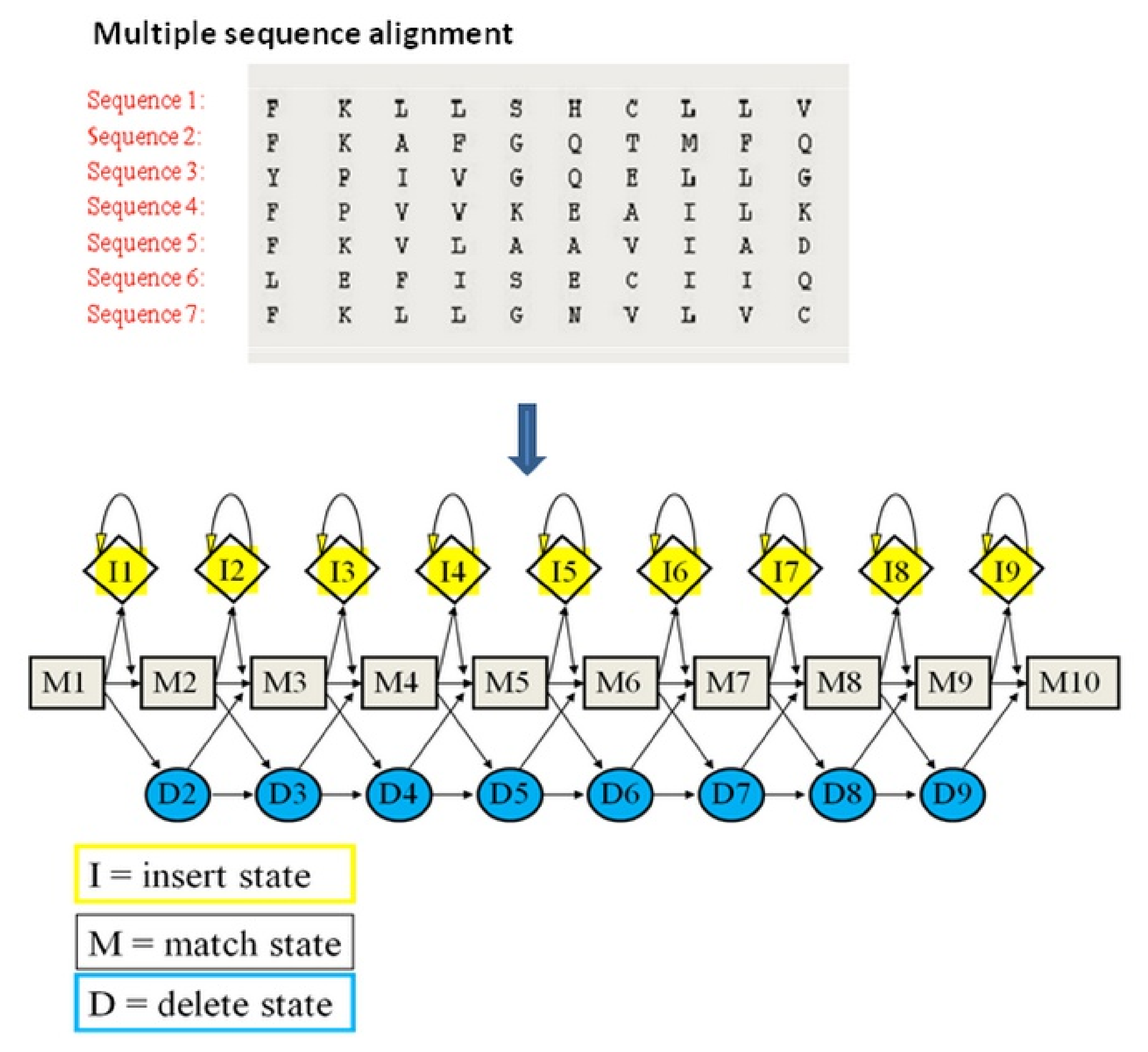
MyHits

If you ask MyHits to return PATTERNs as well as motifs, you will notice that PATTERNs do not have E-values associated with them so there is no easy way to judge statistical significance. With pattern findings you are left only with judging biological significance. Also none of the Frequent patterns from MyHits are statistically significant.

BLAST

If you do not have any insignificant hits from the BLAST search, it means that your protein family is very large and you have to ask BLAST to return more results using the Advanced Options at the bottom of the form. Only when you see hits with E-values > 0.001 do you have insignificant findings.

Hidden Markov Models from Multiple Sequence Alignments

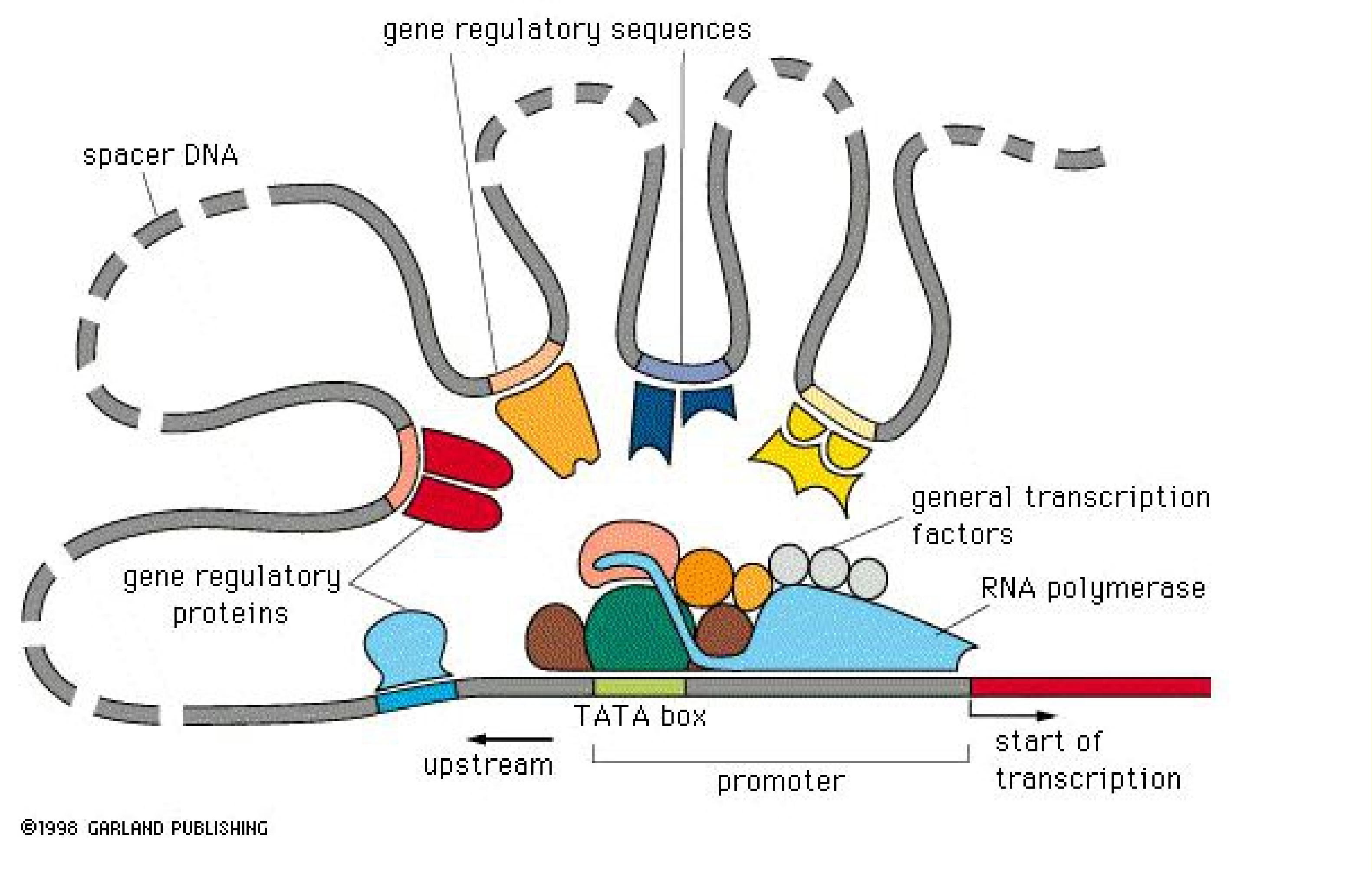


Multiple sequence alignment

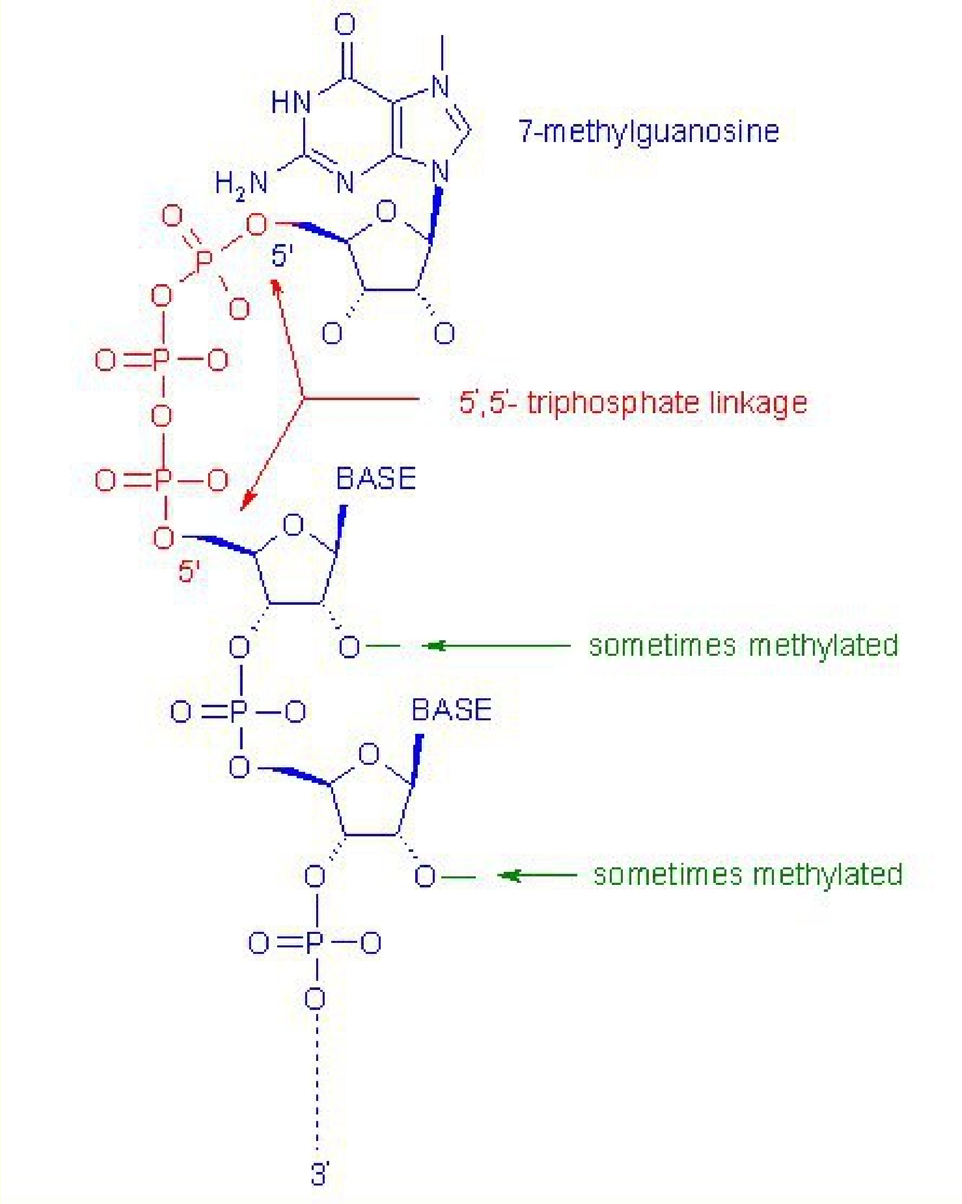
Sequence 1:	F	K	L	L	S	H	C	L	L	V
Sequence 2:	F	K	A	F	G	Q	T	M	F	Q
Sequence 3:	Y	P	I	V	G	Q	E	L	L	G
Sequence 4:	F	P	V	V	K	E	A	I	L	K
Sequence 5:	F	K	V	L	A	A	V	I	A	D
Sequence 6:	L	E	F	I	S	E	C	I	I	Q
Sequence 7:	F	K	L	L	G	N	V	L	V	C

EBI Course on Protein Motifs/Signatures
http://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi

Multiple Enhancer Sequences

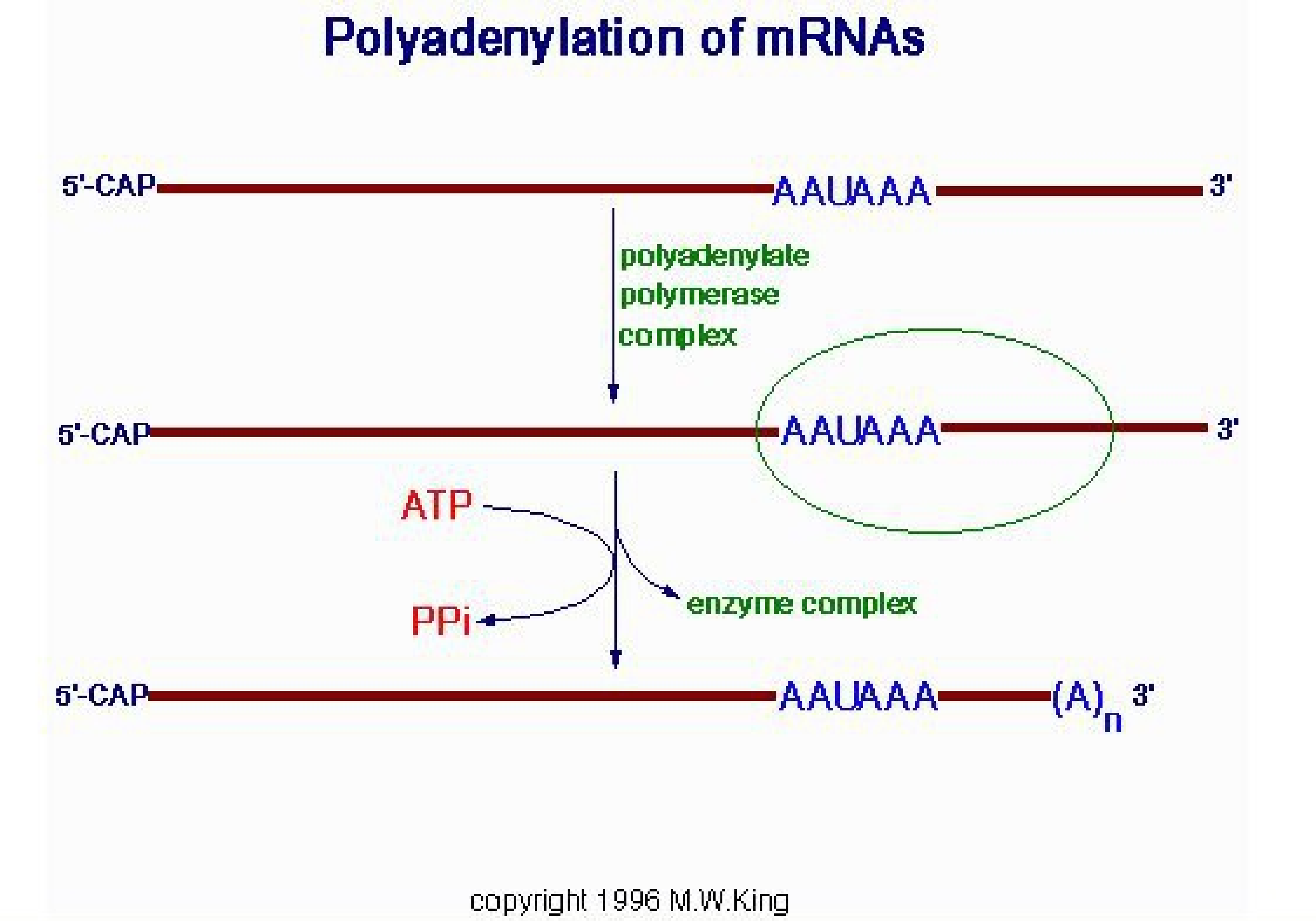


Structure of 5′ CAP



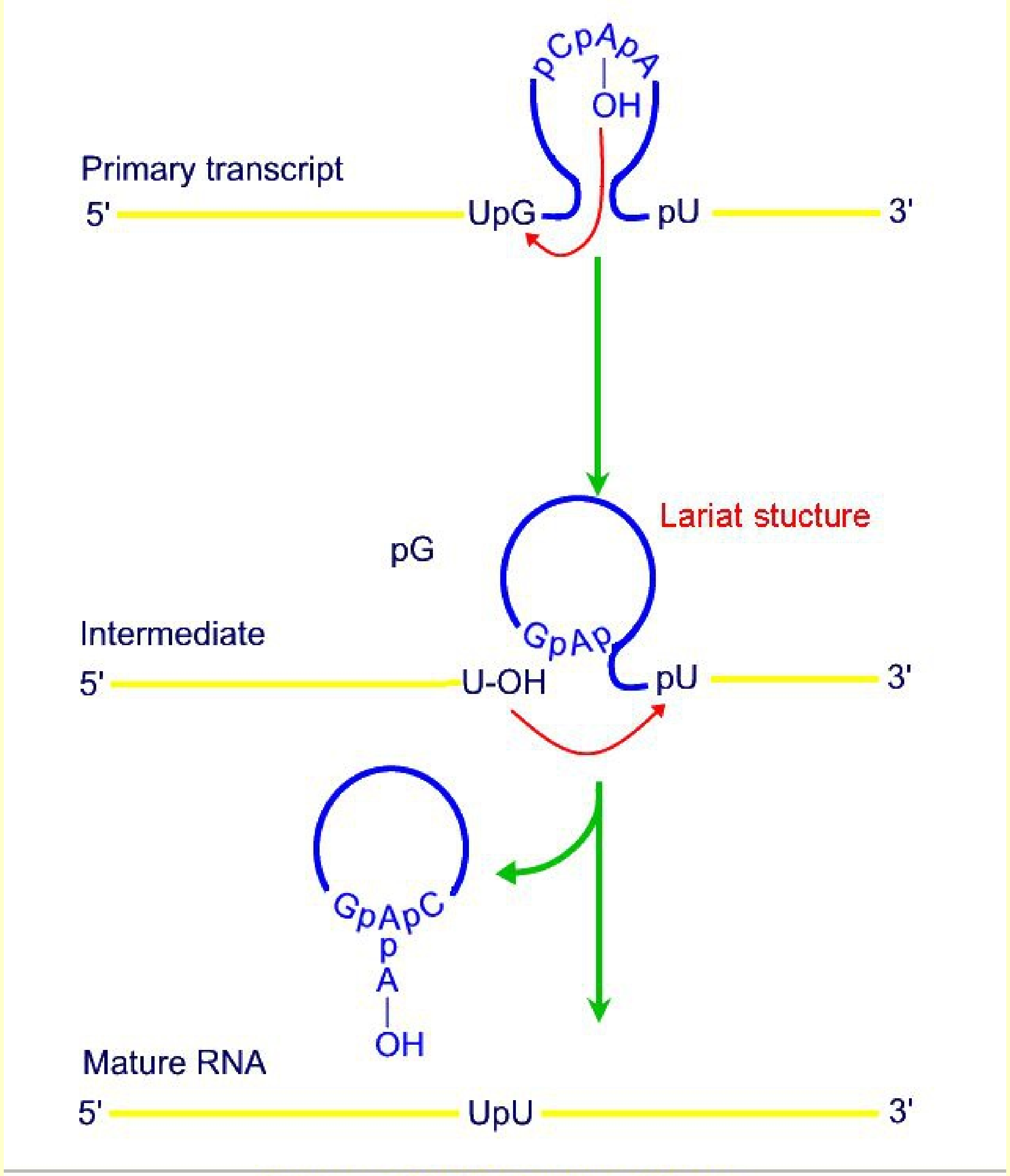
Structure of the 5'-Cap of Eukaryotic mRNAs

PolyAdenylation of mRNAs



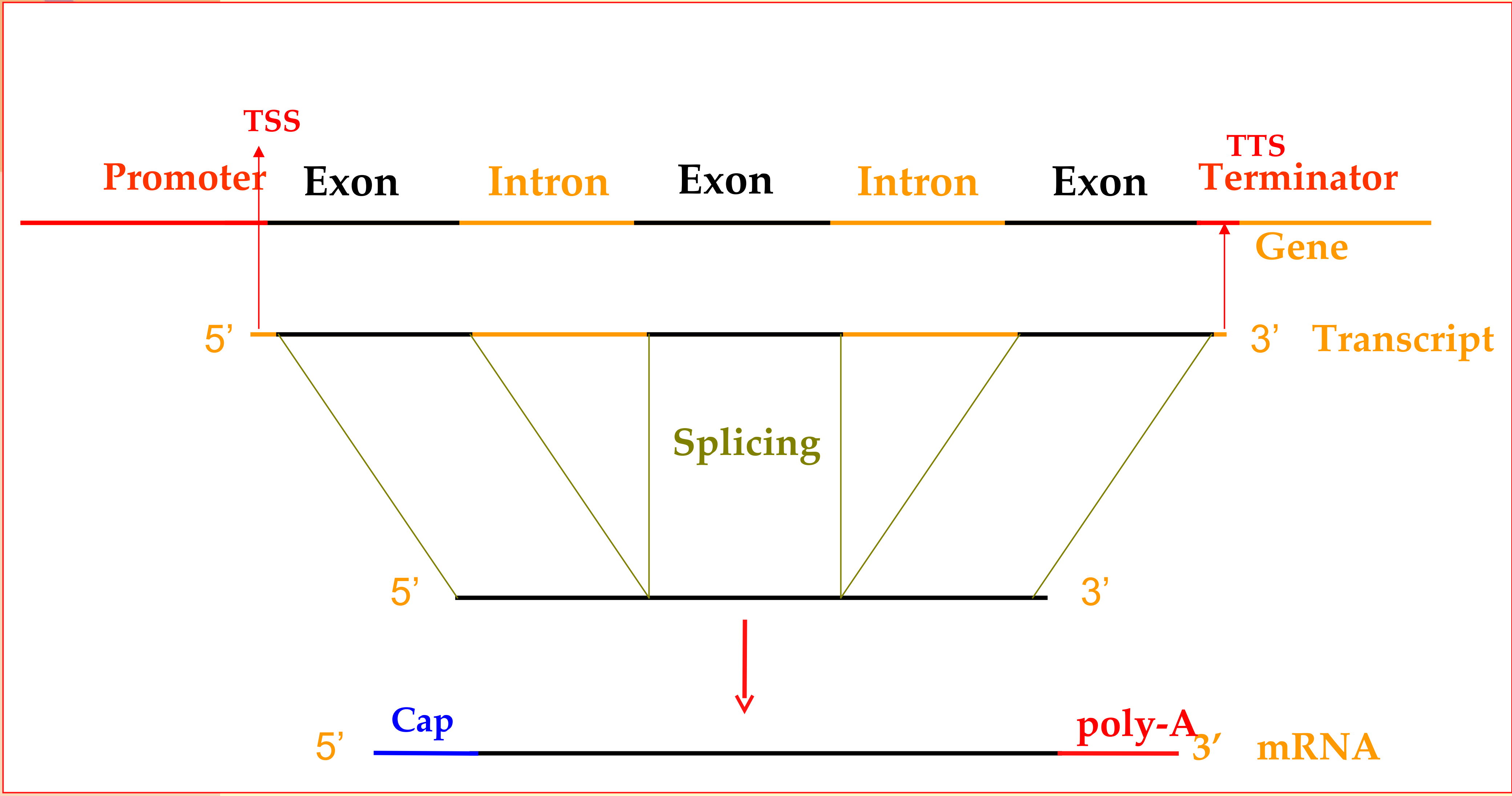
Processes of Polyadenylation

Intron Splicing Mechanism



Processes of Splicing by Group 2 Introns

Splicing, Capping & polyAdenylation Yields Mature mRNA

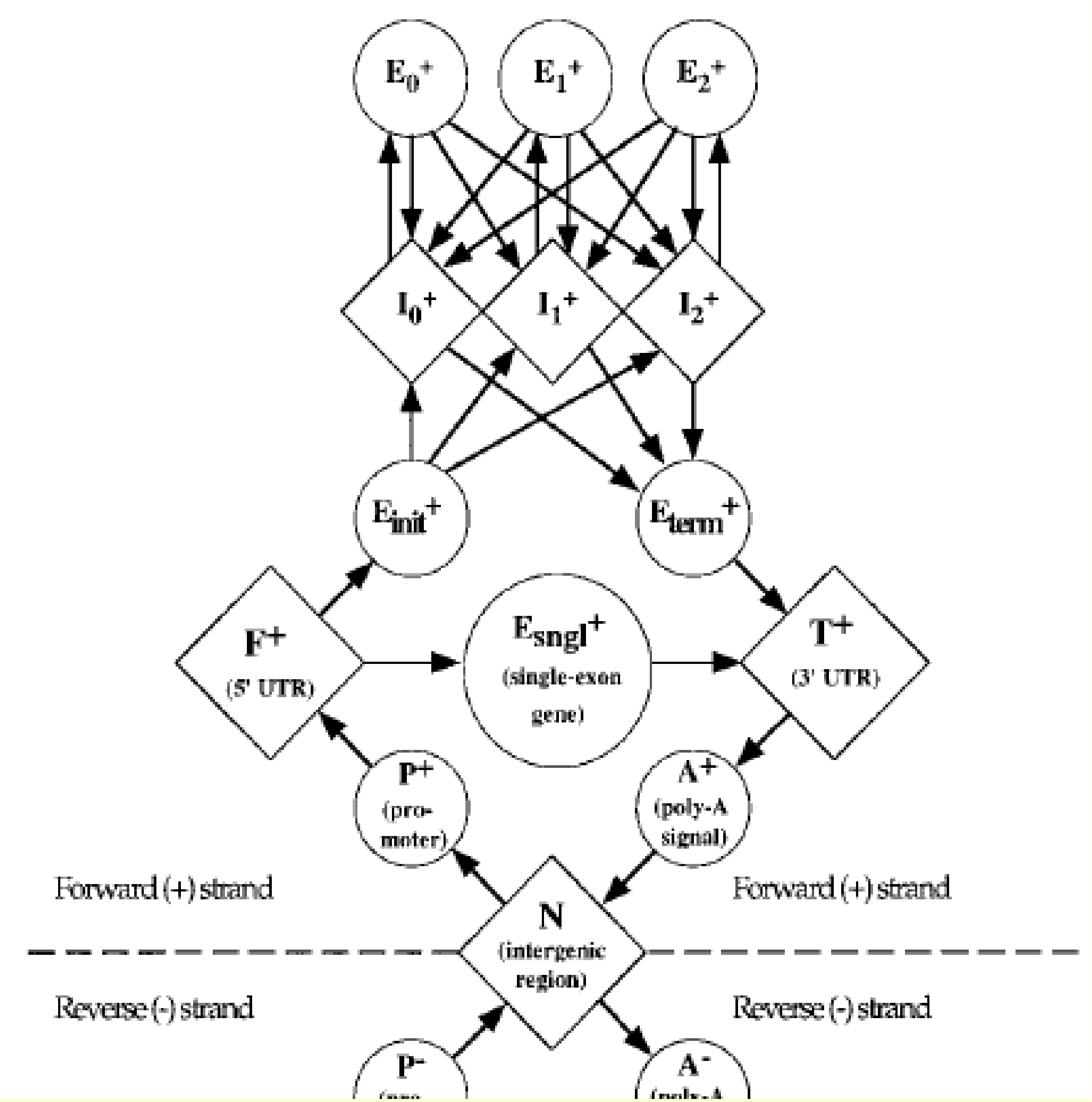


© Doug Brutlag 2015

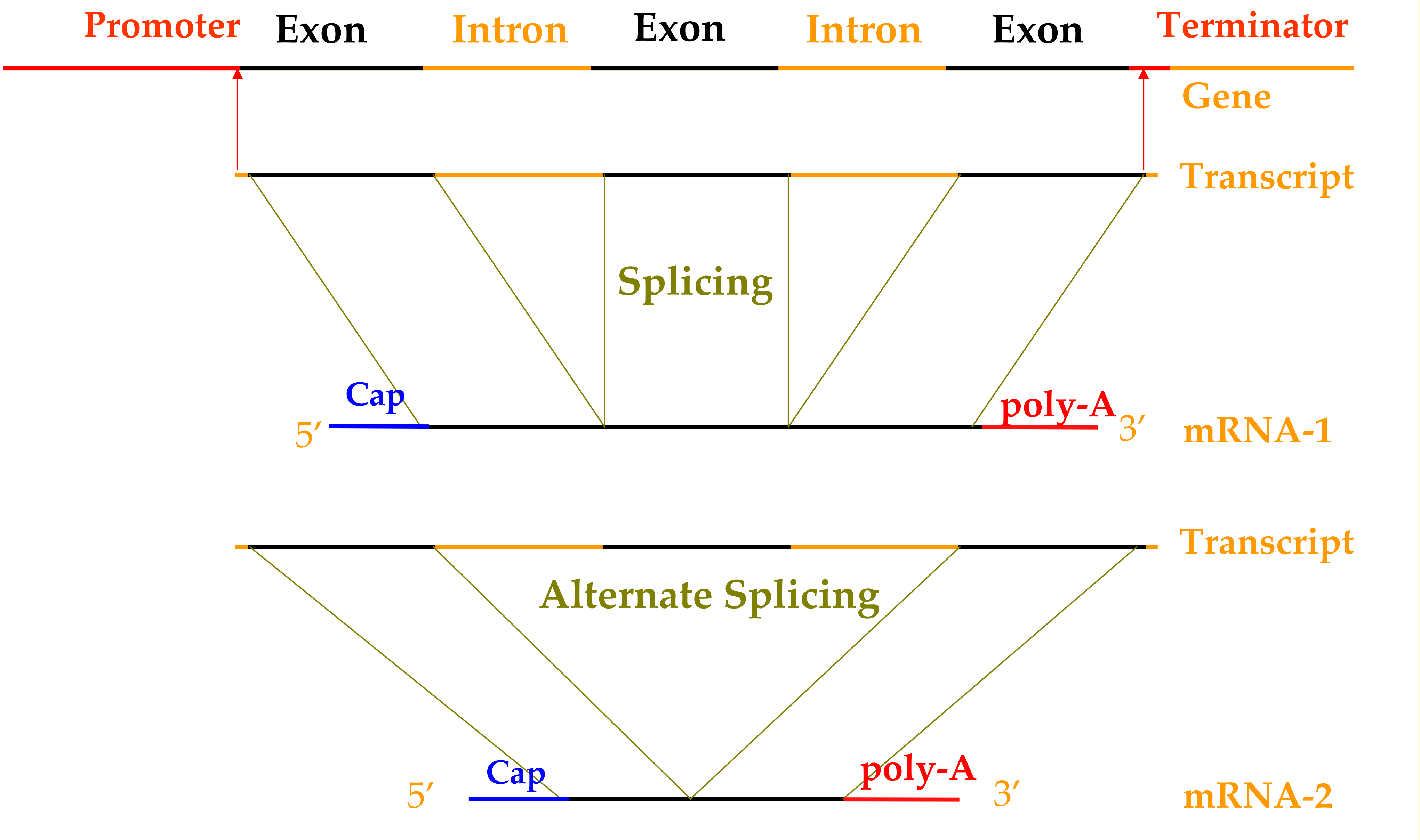
GENSCAN Gene Model

http://genes.mit.edu/GENSCAN.html

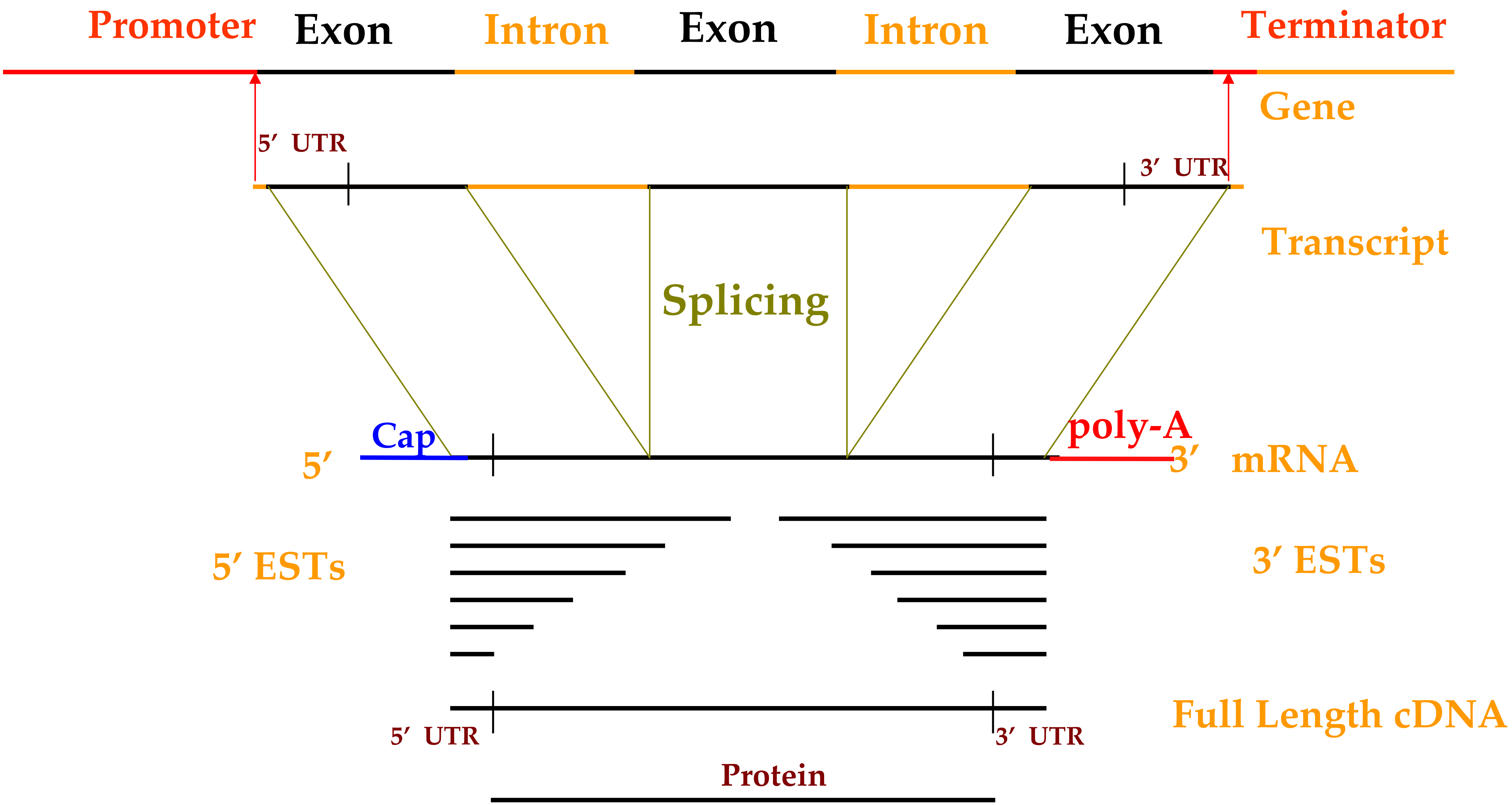
Hidden Markov models of gene structure



Alternative Splicing Generates Distinct Proteins

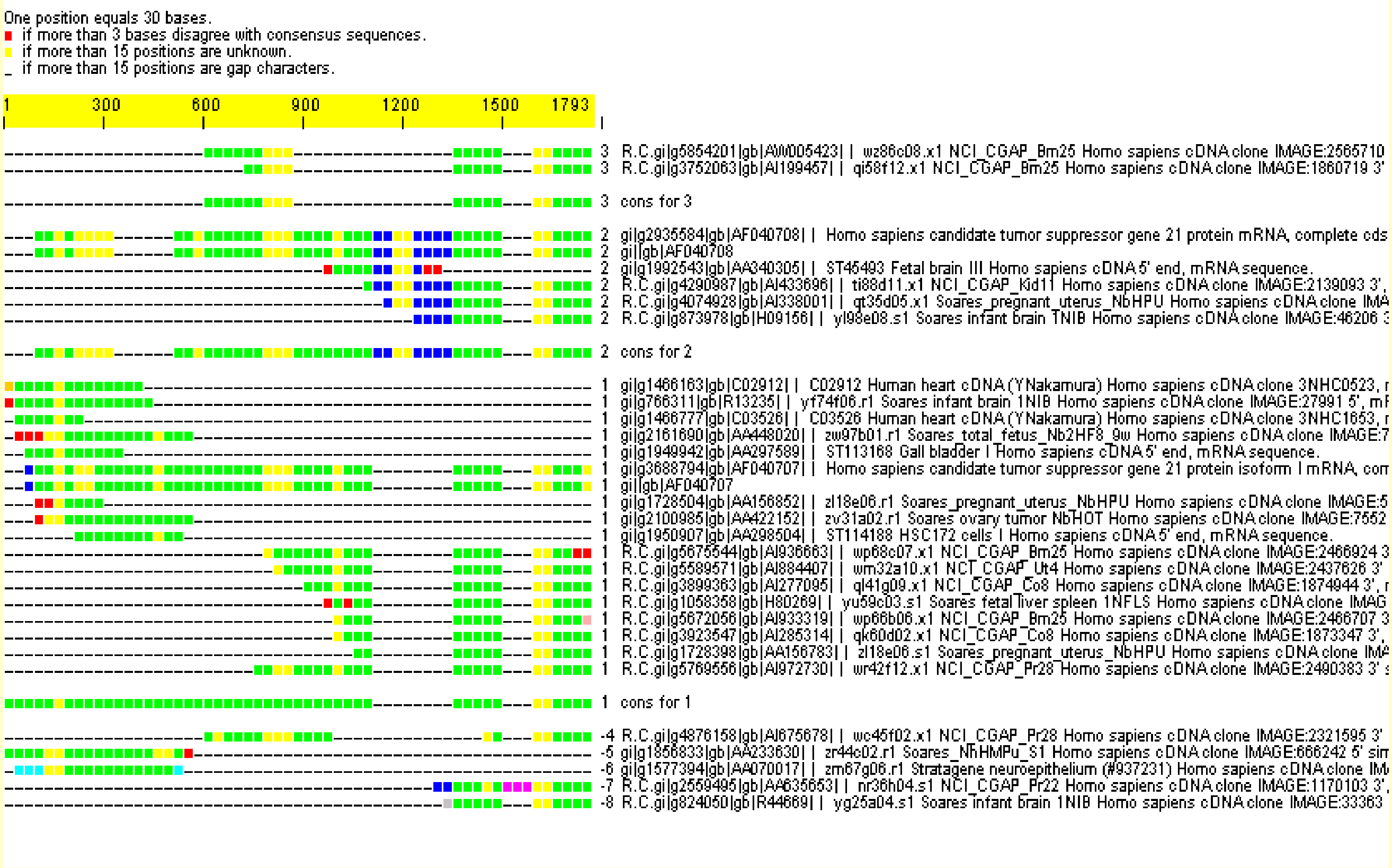


ESTs, Full Length cDNA
UniGene & RefSeq Databases



© Doug Brutlag 2015

Alternative Splicing Detected in EST Libraries

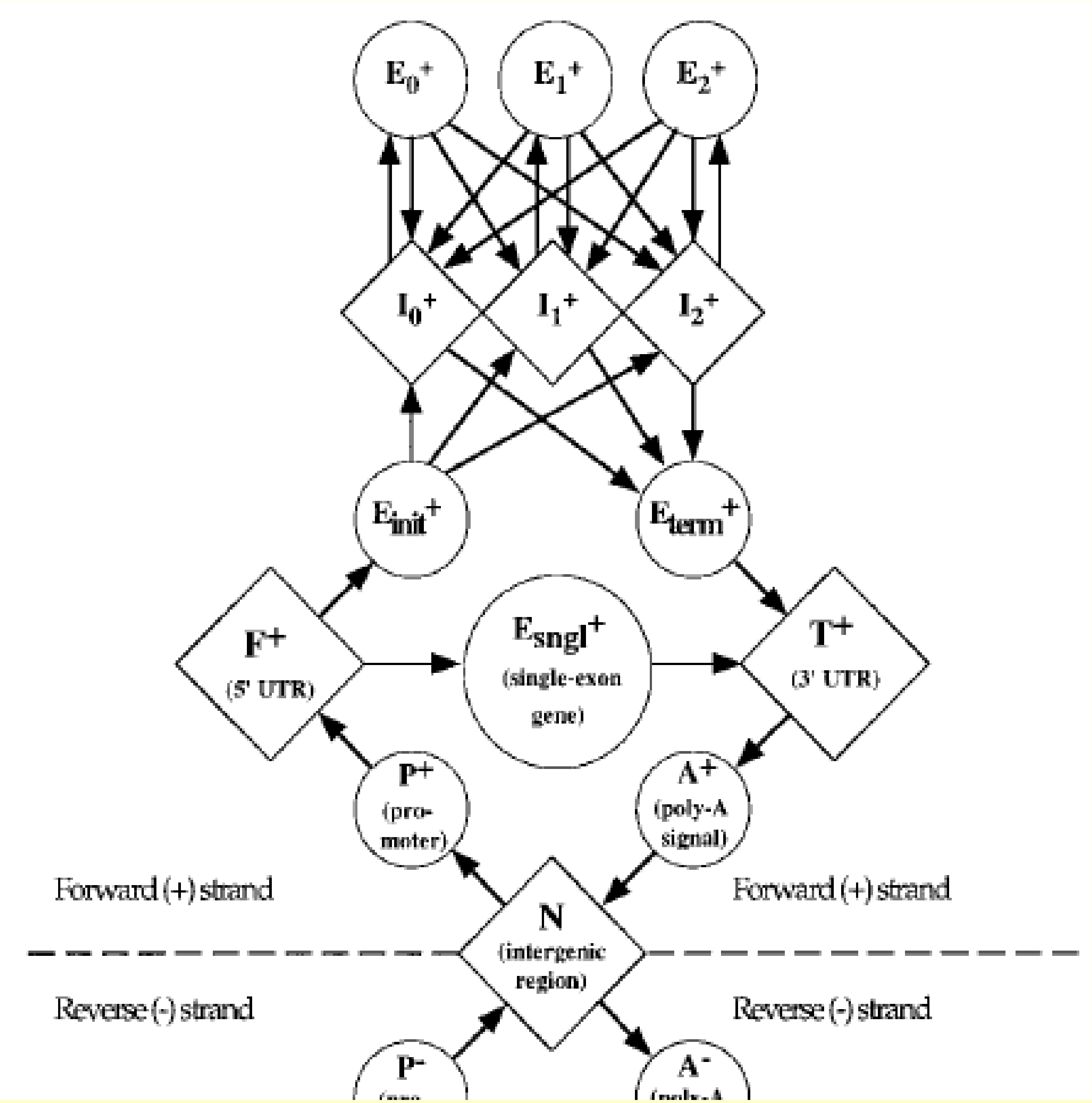


```
One position equals 30 bases.
🟥 if more than 3 bases disagree with consensus sequences.
🟨 if more than 15 positions are unknown.
_ if more than 15 positions are gap characters.

1    300    600    900    1200    1500    1793

3  R.C.gi|g5854201|gb|AW005423| |  wz86c08.x1 NCI_CGAP_Bm25 Homo sapiens cDNA clone IMAGE:2565710
3  R.C.gi|g3752063|gb|AI199457| |  qi58f12.x1 NCI_CGAP_Bm25 Homo sapiens cDNA clone IMAGE:1860719 3'

3  cons for 3

2  gi|g2935584|gb|AF040708| |  Homo sapiens candidate tumor suppressor gene 21 protein mRNA, complete cds
2  gi||gb|AF040708
2  gi|g1992543|gb|AA340305| |  ST45493 Fetal brain III Homo sapiens cDNA 5' end, mRNA sequence.
2  R.C.gi|g4290987|gb|AI433696| |  ti88d11.x1 NCI_CGAP_Kid11 Homo sapiens cDNA clone IMAGE:2139093 3',
2  R.C.gi|g4074928|gb|AI338001| |  qt35d05.x1 Soares_pregnant_uterus_NbHPU Homo sapiens cDNA clone IMA
2  R.C.gi|g873978|gb|H09156| |  yl98e08.s1 Soares infant brain 1NIB Homo sapiens cDNA clone IMAGE:46206

2  cons for 2

1  gi|g1466163|gb|C02912| |  C02912 Human heart cDNA (Y Nakamura) Homo sapiens cDNA clone 3NHC0523,
1  gi|g766311|gb|R13235| |  yf74f06.r1 Soares infant brain 1NIB Homo sapiens cDNA clone IMAGE:27991 5', m
1  gi|g1466777|gb|C03526| |  C03526 Human heart cDNA (Y Nakamura) Homo sapiens cDNA clone 3NHC1653,
1  gi|g2161690|gb|AA448020| |  zw97b01.r1 Soares_total_fetus_Nb2HF8_9w Homo sapiens cDNA clone IMAGE:7
1  gi|g1949942|gb|AA297589| |  ST113168 Gall bladder I Homo sapiens cDNA 5' end, mRNA sequence.
1  gi|g3688794|gb|AF040707| |  Homo sapiens candidate tumor suppressor gene 21 protein isoform I mRNA, com
1  gi||gb|AF040707
1  gi|g1728504|gb|AA156852| |  zl18e06.r1 Soares_pregnant_uterus_NbHPU Homo sapiens cDNA clone IMAGE:5
1  gi|g2100985|gb|AA422152| |  zv31a02.r1 Soares ovary tumor NbHOT Homo sapiens cDNA clone IMAGE:7552
1  gi|g1950907|gb|AA298504| |  ST114188 HSC172 cells I Homo sapiens cDNA 5' end, mRNA sequence.
1  R.C.gi|g5675544|gb|AI936663| |  wp68c07.x1 NCI_CGAP_Bm25 Homo sapiens cDNA clone IMAGE:2466924 3
1  R.C.gi|g5589571|gb|AI884407| |  wm32a10.x1 NCI_CGAP_Ut4 Homo sapiens cDNA clone IMAGE:2437626 3'
1  R.C.gi|g3899363|gb|AI277095| |  ql41g09.x1 NCI_CGAP_Co8 Homo sapiens cDNA clone IMAGE:1874944 3',
1  R.C.gi|g1058358|gb|H80269| |  yu59c03.s1 Soares fetal liver spleen 1NFLS Homo sapiens cDNA clone IMAG
1  R.C.gi|g5672056|gb|AI933319| |  wp66b06.x1 NCI_CGAP_Bm25 Homo sapiens cDNA clone IMAGE:2466707 3
1  R.C.gi|g3923547|gb|AI285314| |  qk60d02.x1 NCI_CGAP_Co8 Homo sapiens cDNA clone IMAGE:1873347 3',
1  R.C.gi|g1728398|gb|AA156783| |  zl18e06.s1 Soares_pregnant_uterus_NbHPU Homo sapiens cDNA clone IMA
1  R.C.gi|g5769556|gb|AI972730| |  wr42f12.x1 NCI_CGAP_Pr28 Homo sapiens cDNA clone IMAGE:2490383 3'

1  cons for 1

-4 R.C.gi|g4876158|gb|AI675678| |  wc45f02.x1 NCI_CGAP_Pr28 Homo sapiens cDNA clone IMAGE:2321595 3'
-5 gi|g1856833|gb|AA233630| |  zr44c02.r1 Soares_NhHMPu_S1 Homo sapiens cDNA clone IMAGE:666242 5' sim
-6 gi|g1577394|gb|AA070017| |  zm67g06.r1 Stratagene neuroepithelium (#937231) Homo sapiens cDNA clone IM
-7 R.C.gi|g2559495|gb|AA635653| |  nr36h04.s1 NCI_CGAP_Pr22 Homo sapiens cDNA clone IMAGE:1170103 3',
-8 R.C.gi|g824050|gb|R44669| |  yg25a04.s1 Soares infant brain 1NIB Homo sapiens cDNA clone IMAGE:33363
```

GENSCAN Gene Model

http://genes.mit.edu/GENSCAN.html

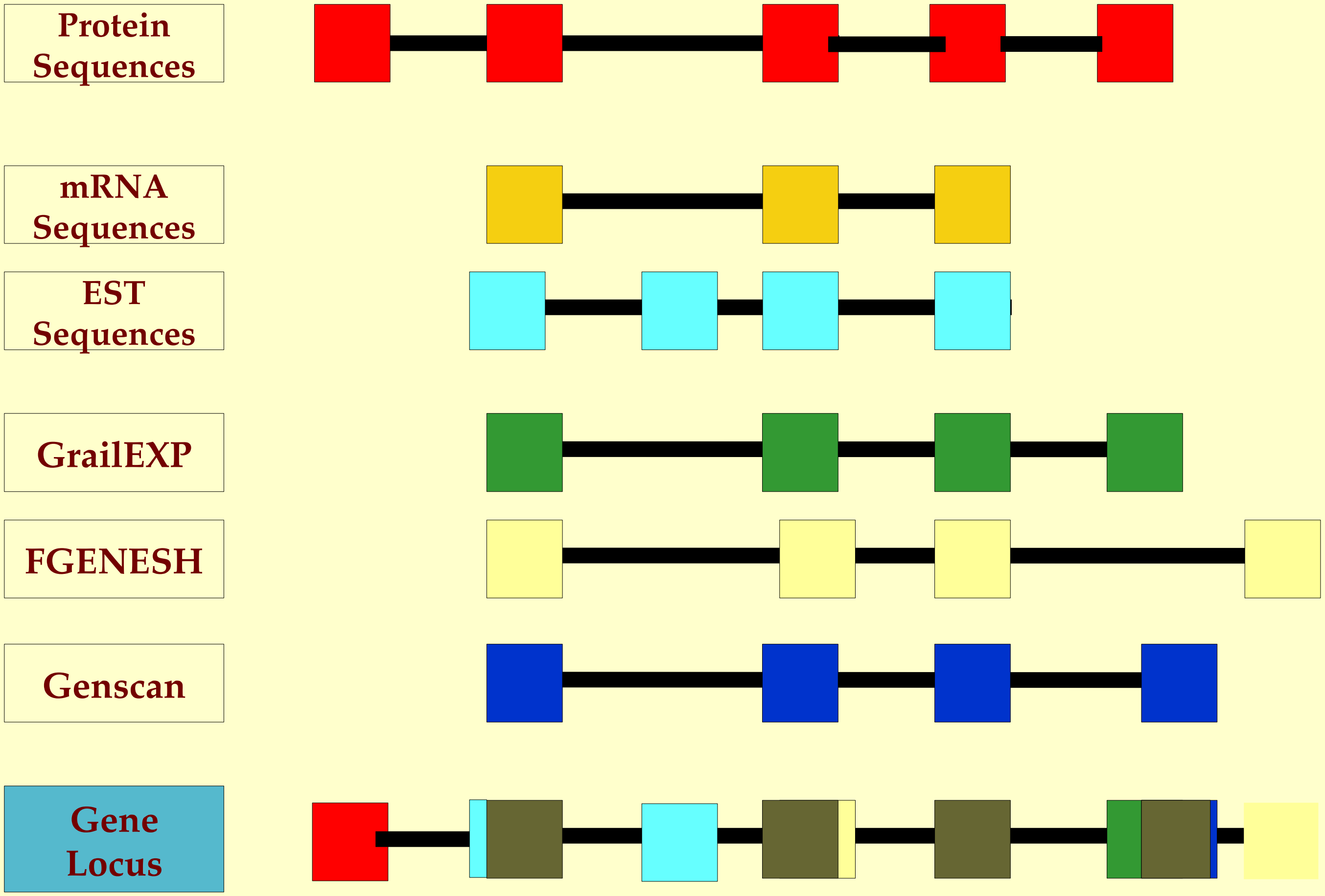
Hidden Markov models of gene structure



© Doug Brutlag 2015

Gene Loci

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



Protein Sequences

mRNA Sequences

EST Sequences

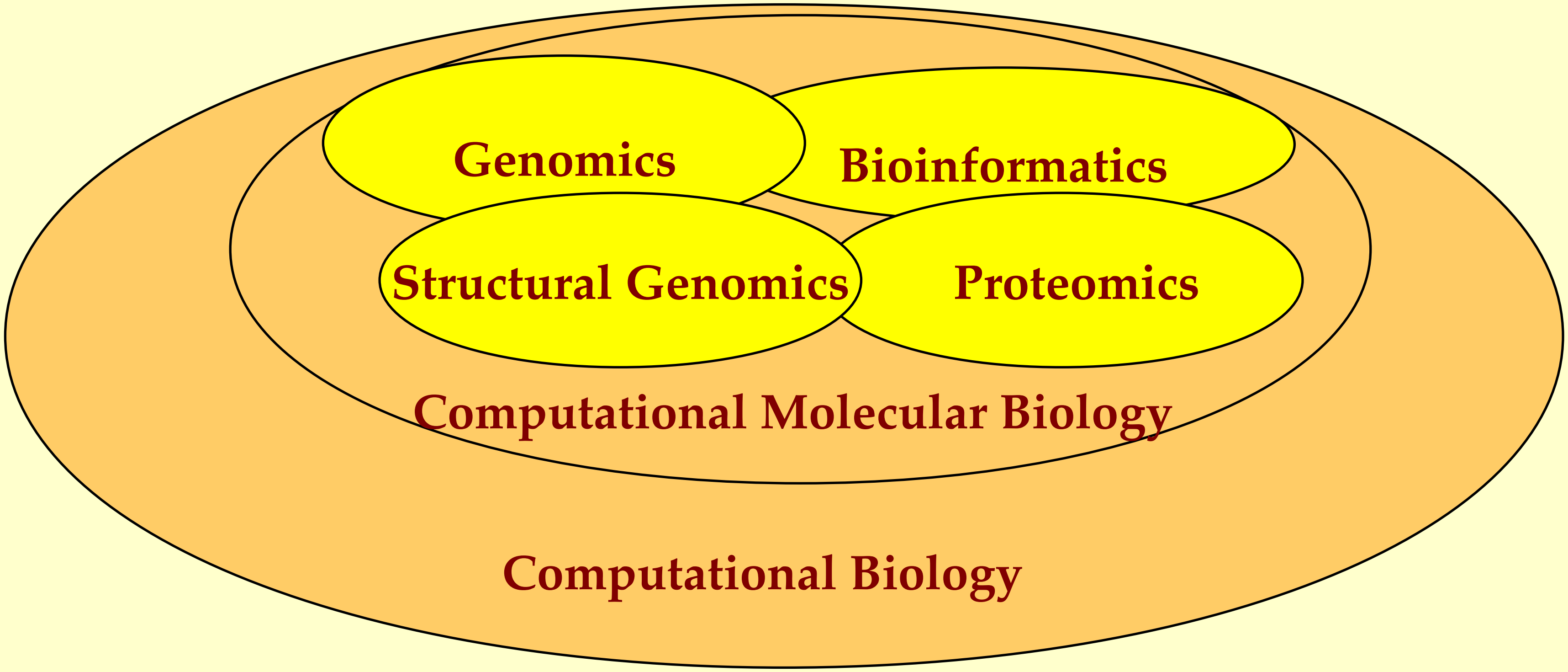
GrailEXP

FGENESH

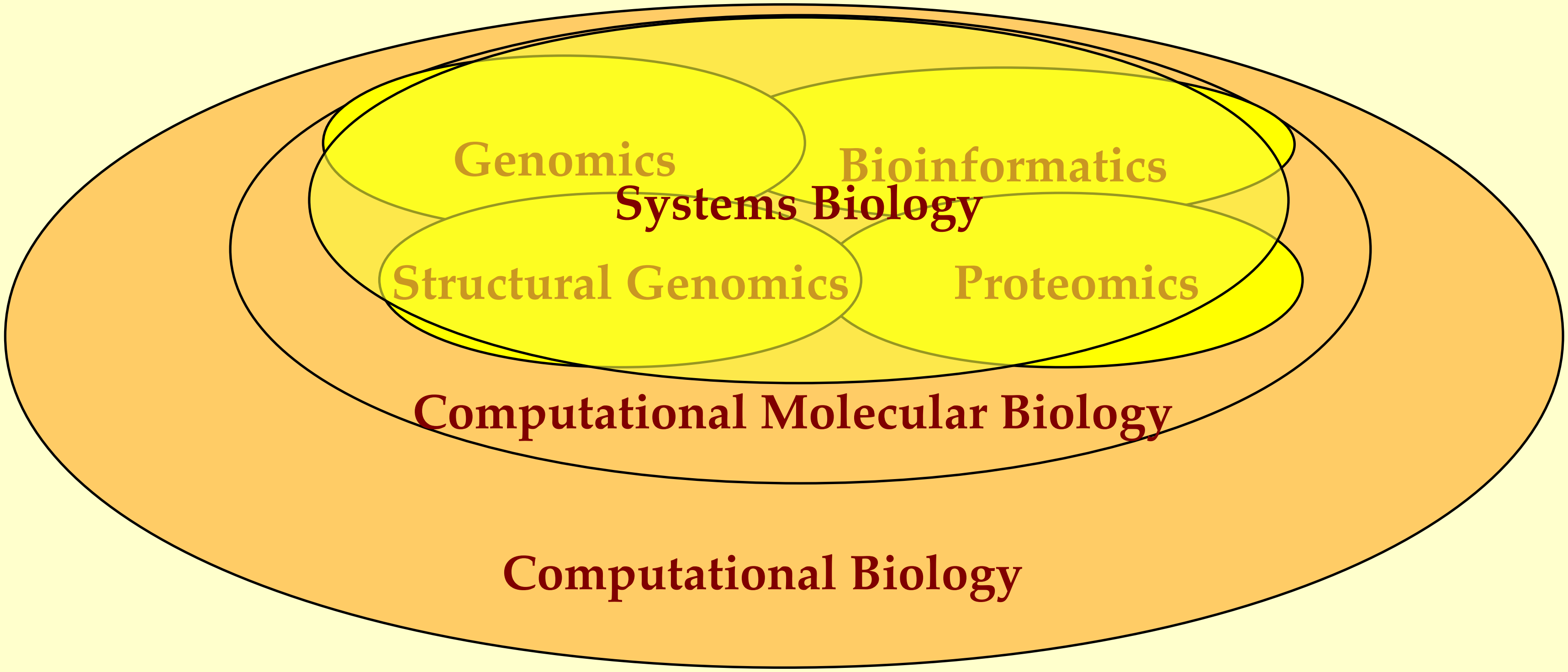
Genscan

Gene Locus

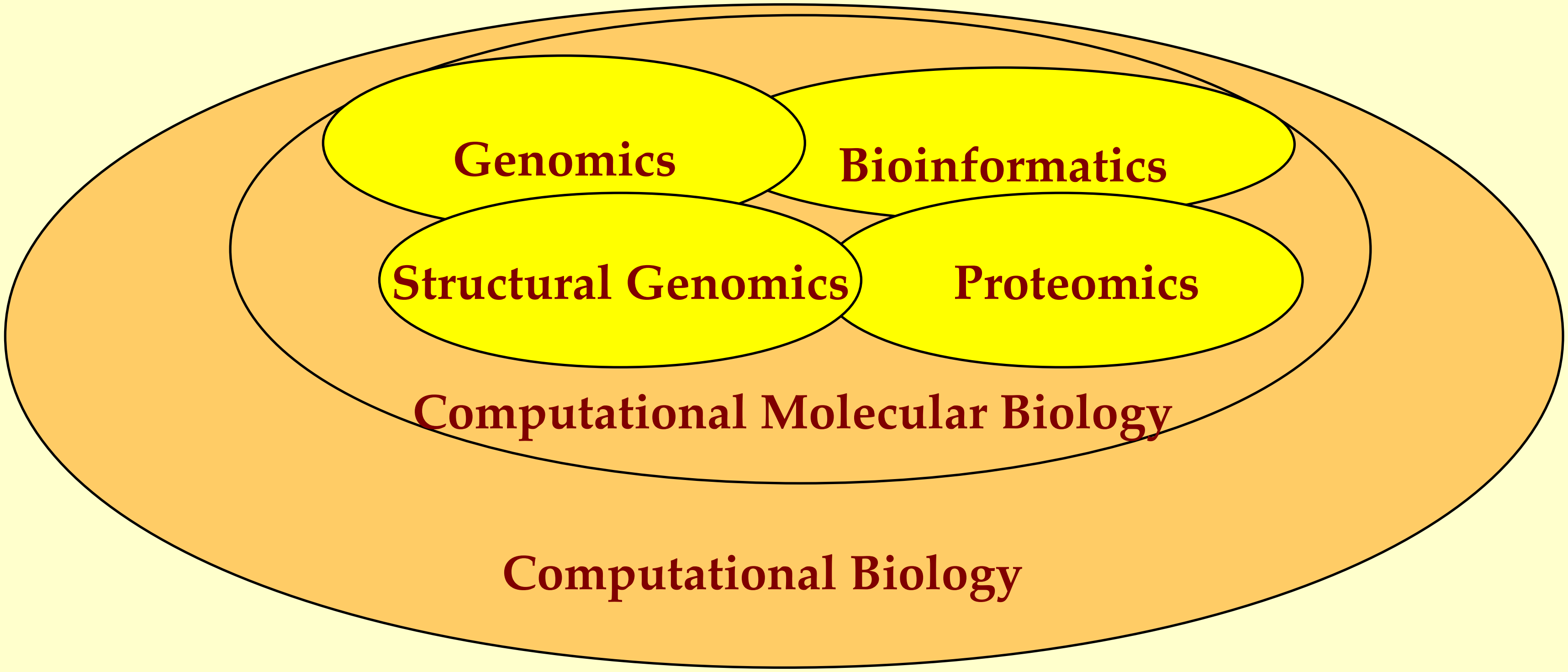
Genomics, Bioinformatics & Computational Biology



Genomics, Bioinformatics & Computational Biology



Genomics, Bioinformatics & Computational Biology



Genomics

Bioinformatics

Structural Genomics

Proteomics

Computational Molecular Biology

Computational Biology

Machine Learning

Robotics

Databases

Artificial Intelligence

Statistics & Probability

Information Theory

Algorithms

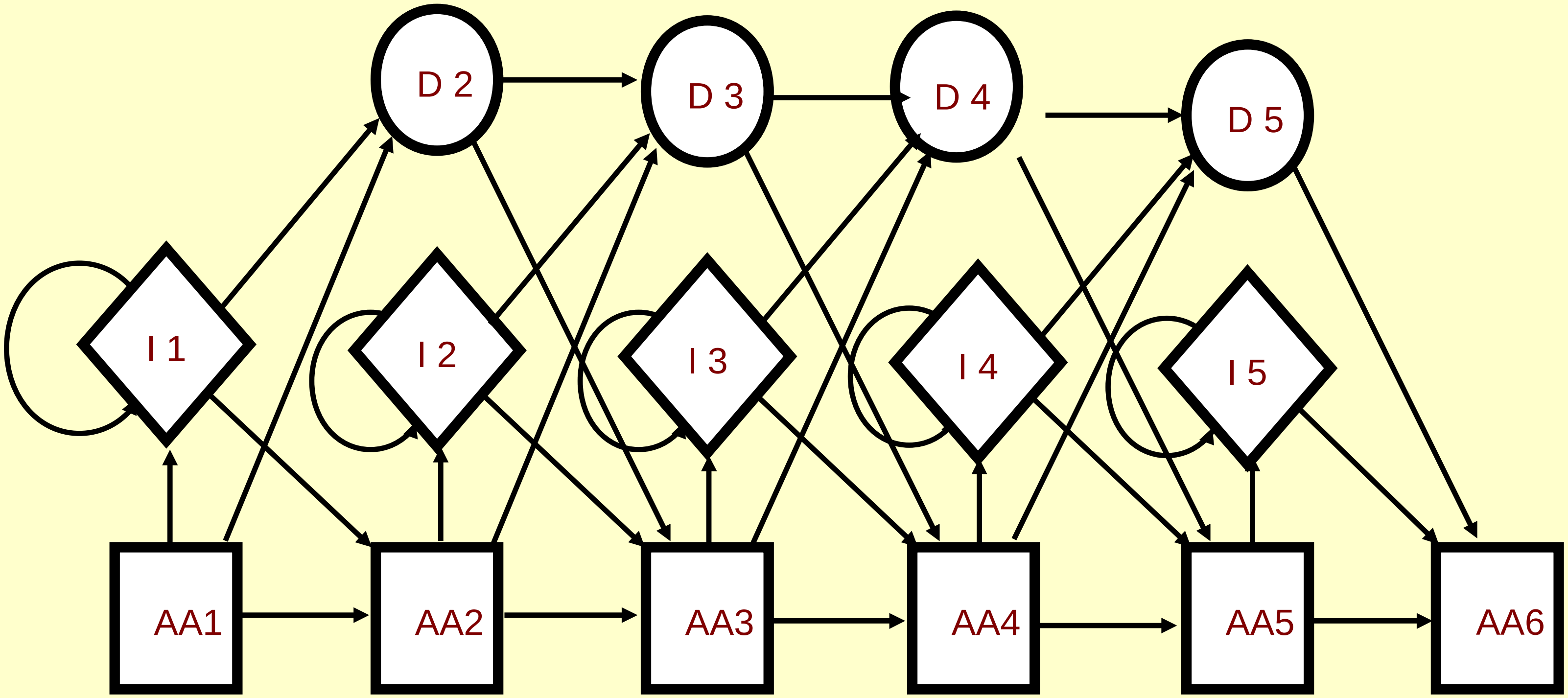
Graph Theory

Redundancy in Genomic & Protein Sequences

- DNA is double-stranded
- Genetic code
- Acceptable amino-acid replacements
- Intron-exon variation
- Alternative splicing
- Strain variations (SNPs)
- Sequencing errors

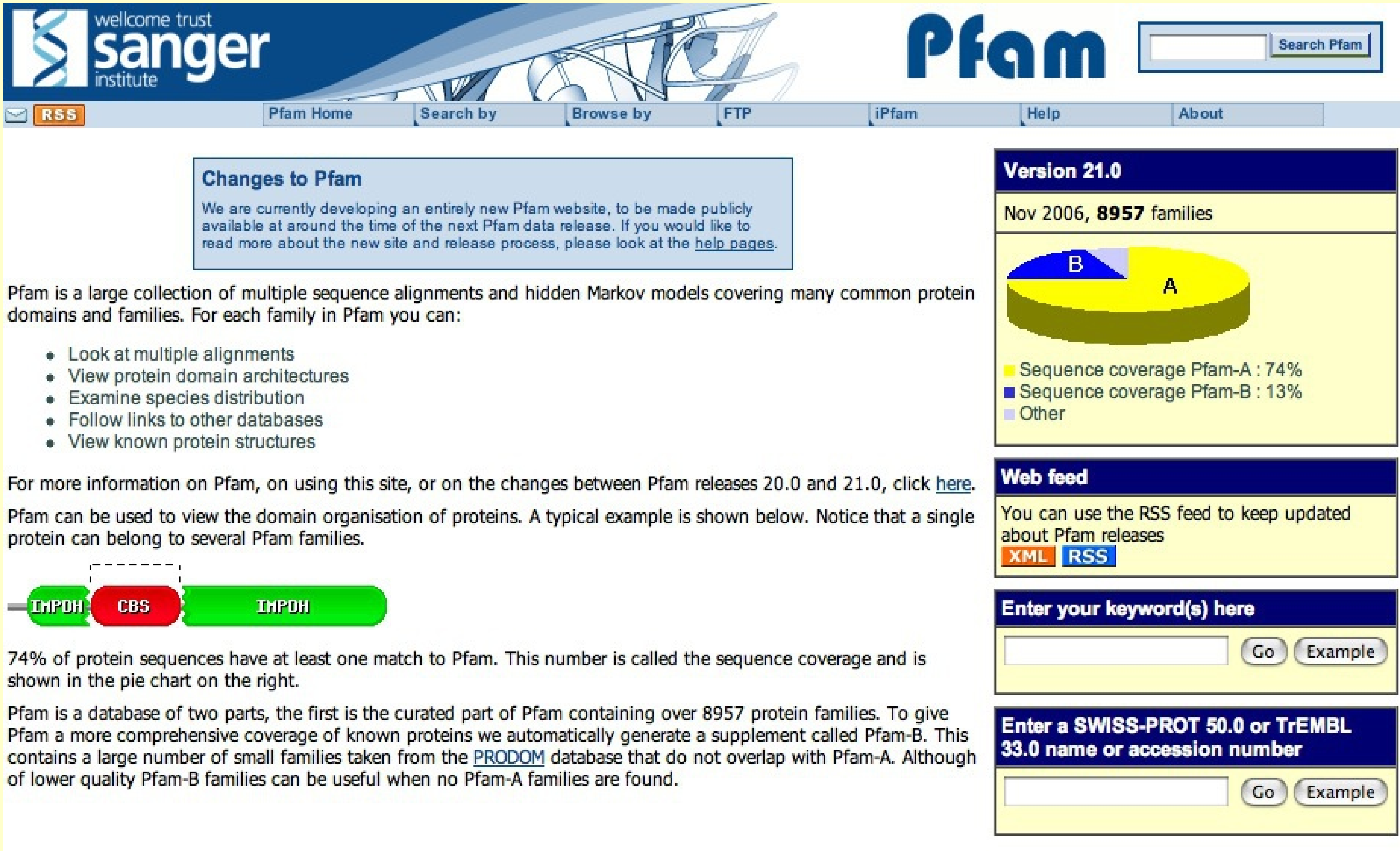
Hidden Markov Models (after Haussler)

http://www.cse.ucsc.edu/compbio/sam.html



pFAM at Sanger Center (UK)

http://www.sanger.ac.uk/Software/Pfam/



Changes to Pfam

We are currently developing an entirely new Pfam website, to be made publicly available at around the time of the next Pfam data release. If you would like to read more about the new site and release process, please look at the help pages.

Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families. For each family in Pfam you can:

- Look at multiple alignments
- View protein domain architectures
- Examine species distribution
- Follow links to other databases
- View known protein structures

For more information on Pfam, on using this site, or on the changes between Pfam releases 20.0 and 21.0, click here.

Pfam can be used to view the domain organisation of proteins. A typical example is shown below. Notice that a single protein can belong to several Pfam families.

74% of protein sequences have at least one match to Pfam. This number is called the sequence coverage and is shown in the pie chart on the right.

Pfam is a database of two parts, the first is the curated part of Pfam containing over 8957 protein families. To give Pfam a more comprehensive coverage of known proteins we automatically generate a supplement called Pfam-B. This contains a large number of small families taken from the PRODOM database that do not overlap with Pfam-A. Although of lower quality Pfam-B families can be useful when no Pfam-A families are found.

Version 21.0

Nov 2006, **8957** families

- Sequence coverage Pfam-A : 74%
- Sequence coverage Pfam-B : 13%
- Other

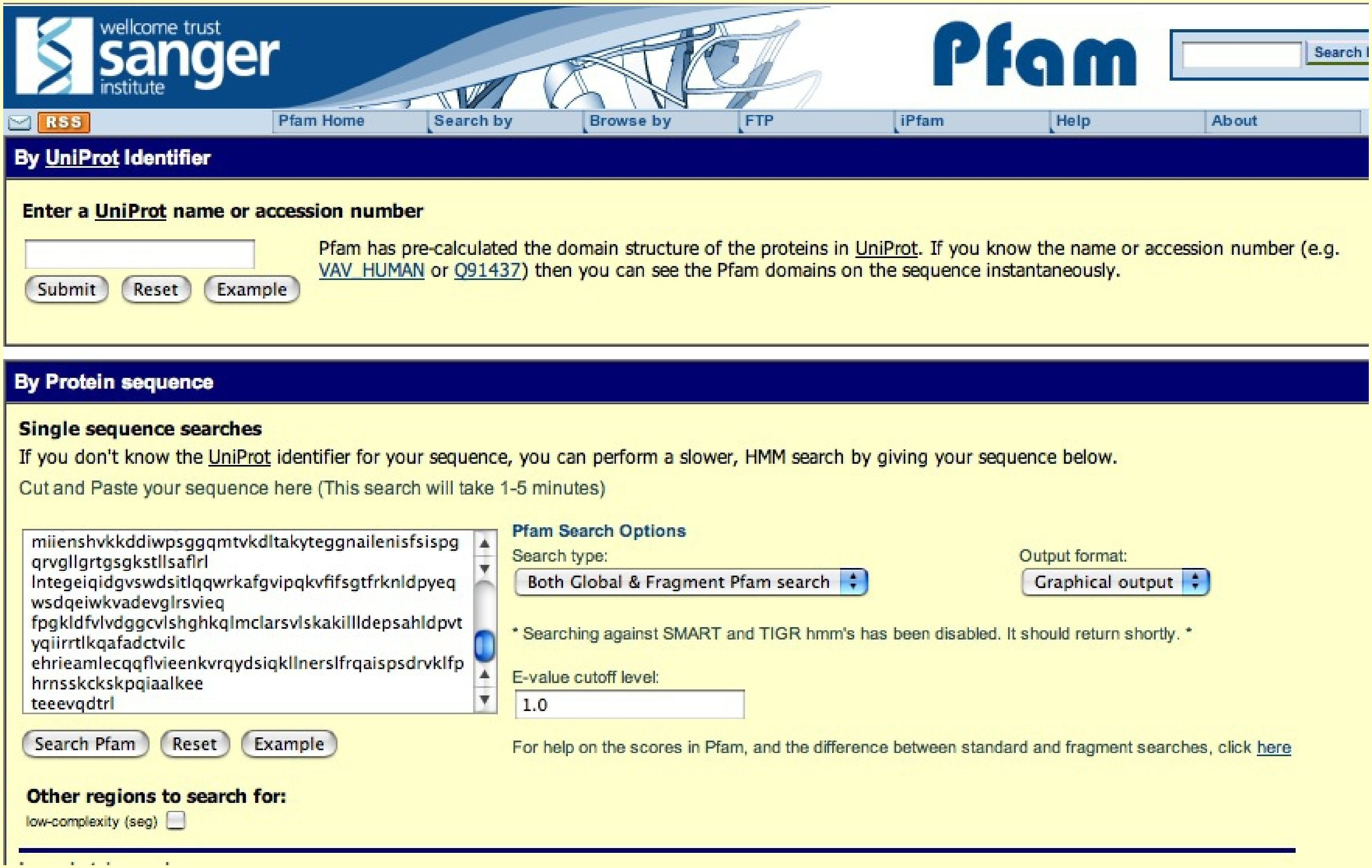
Web feed

You can use the RSS feed to keep updated about Pfam releases

Enter your keyword(s) here

Enter a SWISS-PROT 50.0 or TrEMBL 33.0 name or accession number

pFAM at Sanger Center (UK)
http://pfam.sanger.ac.uk/



By UniProt Identifier

Enter a UniProt name or accession number

Pfam has pre-calculated the domain structure of the proteins in UniProt. If you know the name or accession number (e.g. VAV_HUMAN or Q91437) then you can see the Pfam domains on the sequence instantaneously.

By Protein sequence

Single sequence searches

If you don't know the UniProt identifier for your sequence, you can perform a slower, HMM search by giving your sequence below.

Cut and Paste your sequence here (This search will take 1-5 minutes)

```
miienshvkkddiwpsggqmtvkdltakyteggnailenisfsispg
qrvgllgrtgsgkstllsaflrl
lntegeiqidgvswdsitlqqwrkafgvipqkvfifsgtfrknldpyeq
wsdqeiwkvadevglrsvieq
fpgkldfvlvdggcvlshghkqlmclarsvlskakillldepsahldpvt
yqiirrtlkqafadctvilc
ehrieamlecqqflvieenkvrqydsiqkllnerslfrqaispsdrvklfp
hrnsskckskpqiaalkee
teeevqdtrl
```

Pfam Search Options

Search type: Both Global & Fragment Pfam search

Output format: Graphical output

* Searching against SMART and TIGR hmm's has been disabled. It should return shortly. *

E-value cutoff level: 1.0

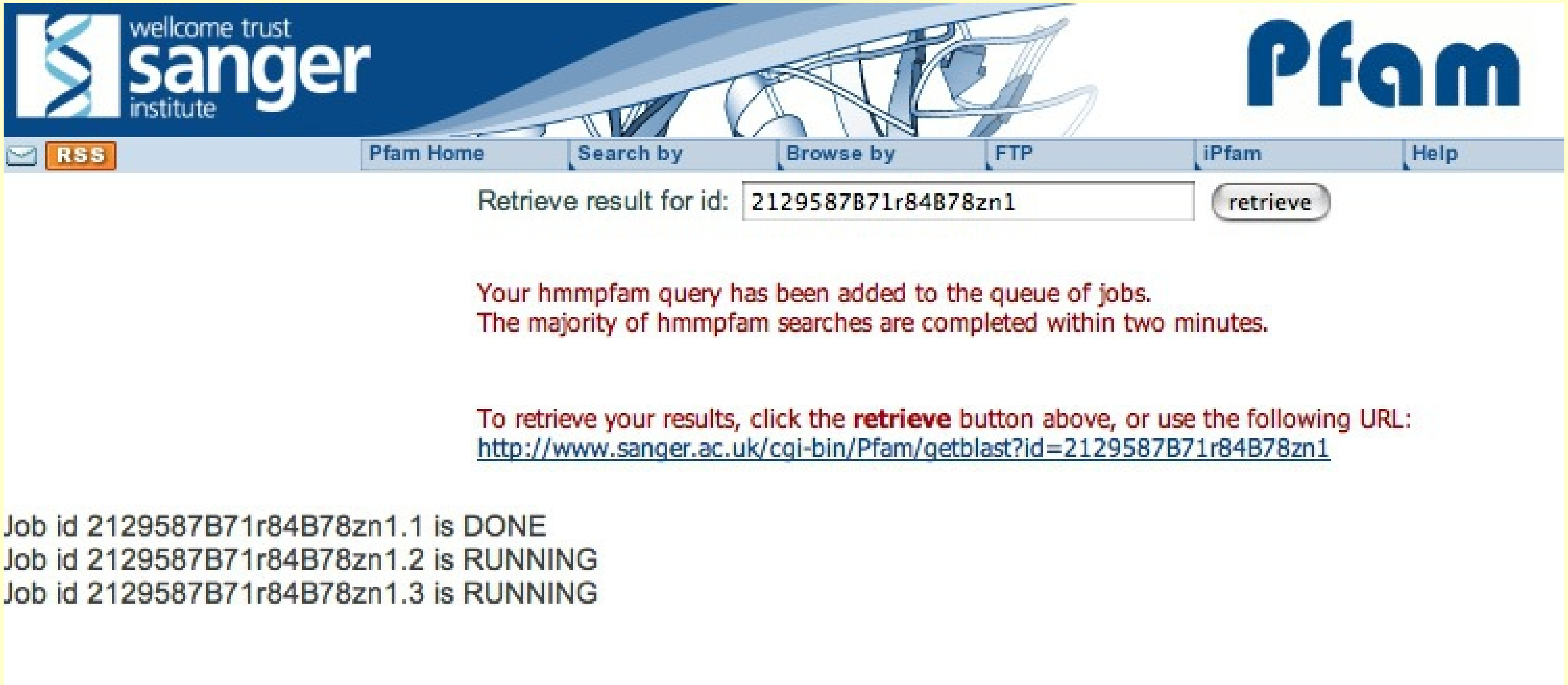
For help on the scores in Pfam, and the difference between standard and fragment searches, click here

Other regions to search for:

low-complexity (seg) ☐

pFAM at Sanger Center (UK)

http://pfam.sanger.ac.uk/



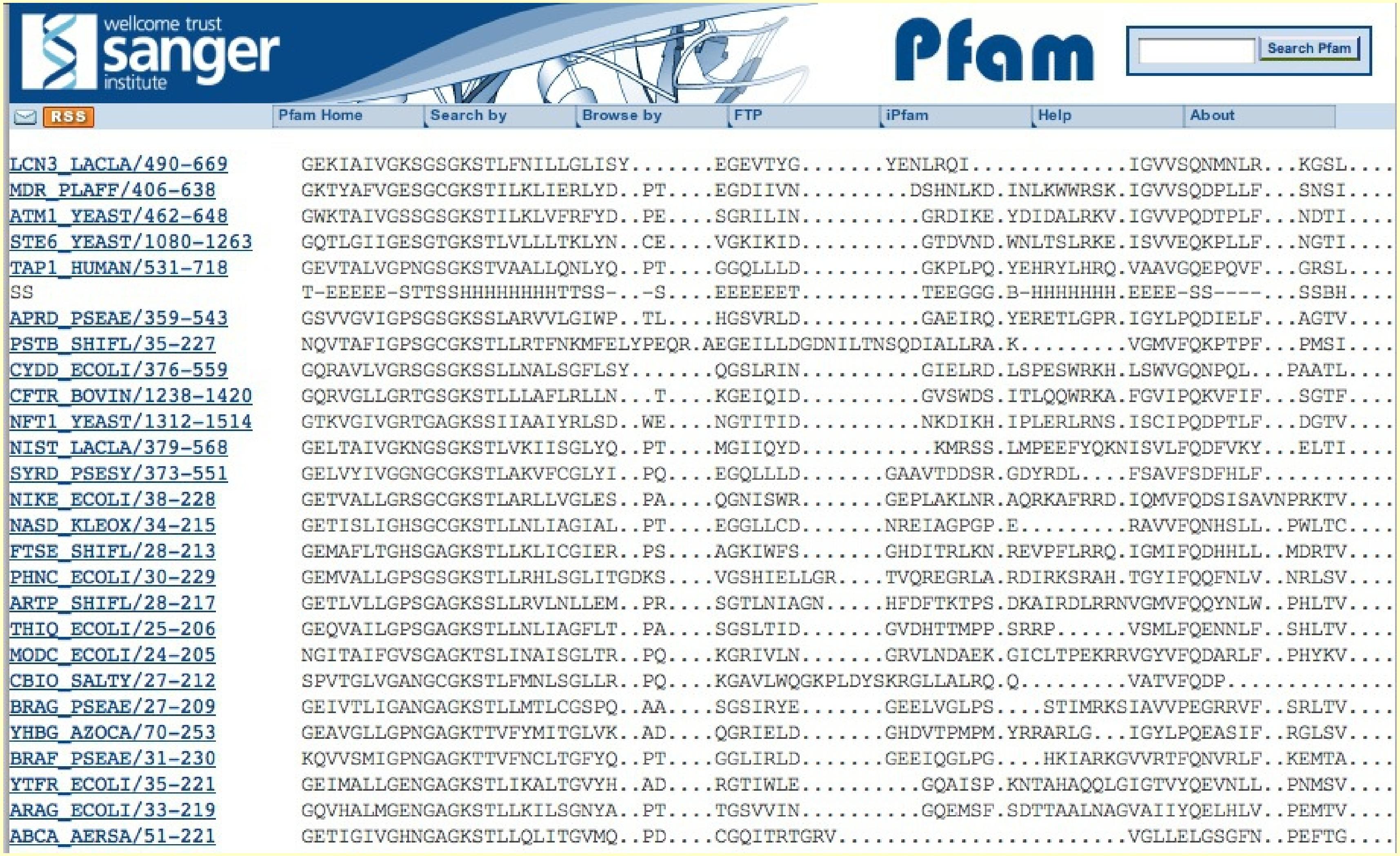
Retrieve result for id: 2129587B71r84B78zn1

Your hmmpfam query has been added to the queue of jobs.
The majority of hmmpfam searches are completed within two minutes.

To retrieve your results, click the **retrieve** button above, or use the following URL:
http://www.sanger.ac.uk/cgi-bin/Pfam/getblast?id=2129587B71r84B78zn1

```
Job id 2129587B71r84B78zn1.1 is DONE
Job id 2129587B71r84B78zn1.2 is RUNNING
Job id 2129587B71r84B78zn1.3 is RUNNING
```

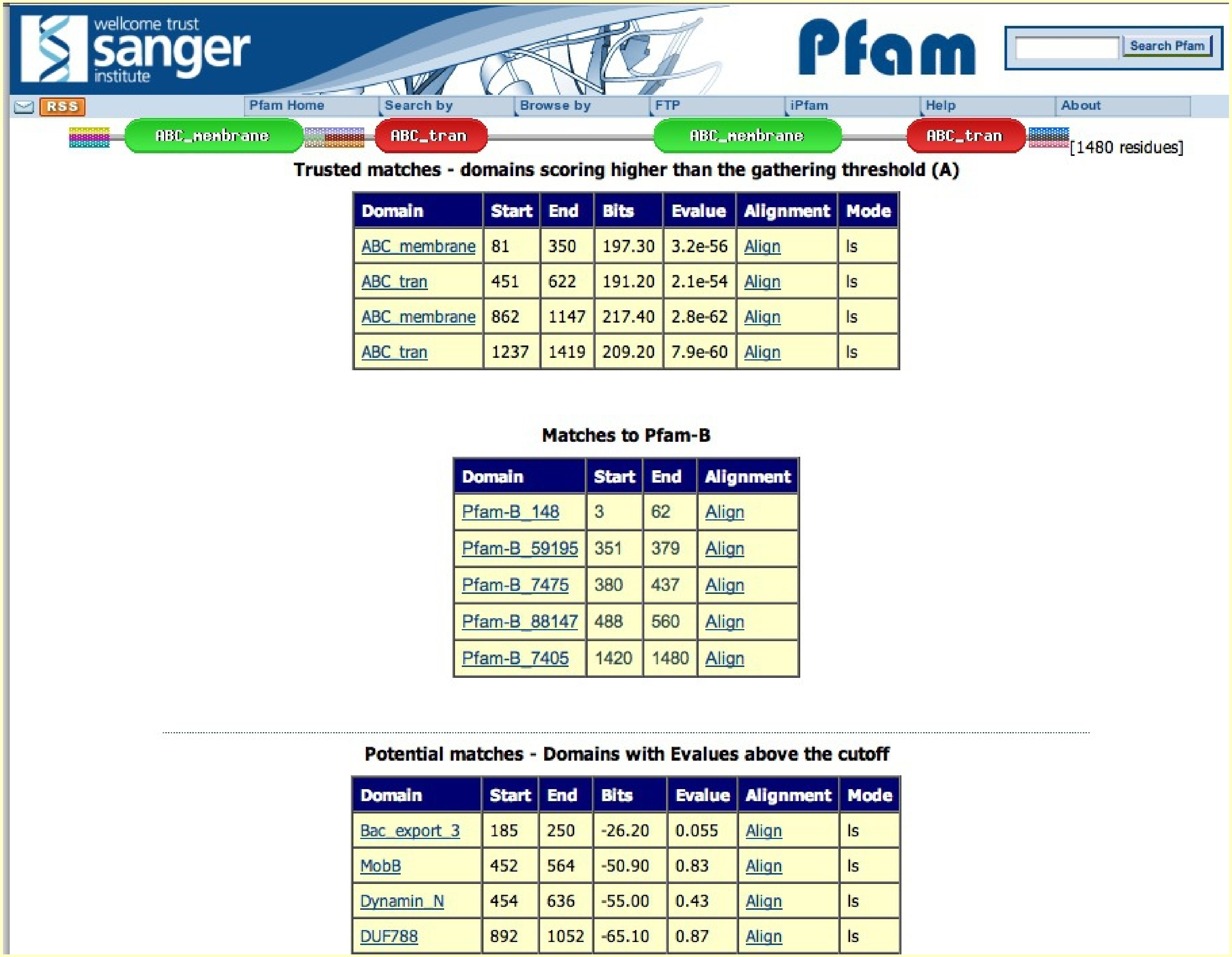
pFAM at Sanger Center (UK)
http://pfam.sanger.ac.uk/



```
LCN3_LACLA/490-669      GEKIAIVGKSGSGKSTLFNILLGLISY.......EGEVTYG.......YENLRQI.............IGVVSQNMNLR...KGSL....
MDR_PLAFF/406-638       GKTYAFVGESGCGKSTILKLIERLYD..PT....EGDIIVN.........DSHNLKD.INLKWWRSK.IGVVSQDPLLF...SNSI....
ATM1_YEAST/462-648      GWKTAIVGSSGSGKSTILKLVFRFYD..PE....SGRILIN..........GRDIKE.YDIDALRKV.IGVVPQDTPLF...NDTI....
STE6_YEAST/1080-1263    GQTLGIIGESGTGKSTLVLLLTKLYN..CE....VGKIKID..........GTDVND.WNLTSLRKE.ISVVEQKPLLF...NGTI....
TAP1_HUMAN/531-718      GEVTALVGPNGSGKSTVAALLQNLYQ..PT....GGQLLLD..........GKPLPQ.YEHRYLHRQ.VAAVGQEPQVF...GRSL....
SS                      T-EEEEE-STTSSHHHHHHHHTTSS-..-S....EEEEEET..........TEEGGG.B-HHHHHHH.EEEE-SS----...SSBH....
APRD_PSEAE/359-543      GSVVGVIGPSGSGKSSLARVVLGIWP..TL....HGSVRLD..........GAEIRQ.YERETLGPR.IGYLPQDIELF...AGTV....
PSTB_SHIFL/35-227       NQVTAFIGPSGCGKSTLLRTFNKMFELYPEQR.AEGEILLDGDNILTNSQDIALLRA.K.........VGMVFQKPTPF...PMSI....
CYDD_ECOLI/376-559      GQRAVLVGRSGSGKSSLLNALSGFLSY.......QGSLRIN..........GIELRD.LSPESWRKH.LSWVGQNPQL...PAATL....
CFTR_BOVIN/1238-1420    GQRVGLLGRTGSGKSTLLLAFLRLLN...T....KGEIQID..........GVSWDS.ITLQQWRKA.FGVIPQKVFIF...SGTF....
NFT1_YEAST/1312-1514    GTKVGIVGRTGAGKSSIIAAIYRLSD..WE....NGTITID..........NKDIKH.IPLERLRNS.ISCIPQDPTLF...DGTV....
NIST_LACLA/379-568      GELTAIVGKNGSGKSTLVKIISGLYQ..PT....MGIIQYD...........KMRSS.LMPEEFYQKNISVLFQDFVKY...ELTI....
SYRD_PSESY/373-551      GELVYIVGGNGCGKSTLAKVFCGLYI..PQ....EGQLLLD.......GAAVTDDSR.GDYRDL....FSAVFSDFHLF...........
NIKE_ECOLI/38-228       GETVALLGRSGCGKSTLARLLVGLES..PA....QGNISWR.......GEPLAKLNR.AQRKAFRRD.IQMVFQDSISAVNPRKTV....
NASD_KLEOX/34-215       GETISLIGHSGCGKSTLLNLIAGIAL..PT....EGGLLCD.......NREIAGPGP.E.........RAVVFQNHSLL..PWLTC....
FTSE_SHIFL/28-213       GEMAFLTGHSGAGKSTLLKLICGIER..PS....AGKIWFS.......GHDITRLKN.REVPFLRRQ.IGMIFQDHHLL..MDRTV....
PHNC_ECOLI/30-229       GEMVALLGPSGSGKSTLLRHLSGLITGDKS....VGSHIELLGR....TVQREGRLA.RDIRKSRAH.TGYIFQQFNLV..NRLSV....
ARTP_SHIFL/28-217       GETLVLLGPSGAGKSSLLRVLNLLEM..PR....SGTLNIAGN.....HFDFTKTPS.DKAIRDLRRNVGMVFQQYNLW..PHLTV....
THIQ_ECOLI/25-206       GEQVAILGPSGAGKSTLLNLIAGFLT..PA....SGSLTID.......GVDHTTMPP.SRRP......VSMLFQENNLF..SHLTV....
MODC_ECOLI/24-205       NGITAIFGVSGAGKTSLINAISGLTR..PQ....KGRIVLN.......GRVLNDAEK.GICLTPEKRRVGYVFQDARLF..PHYKV....
CBIO_SALTY/27-212       SPVTGLVGANGCGKSTLFMNLSGLLR..PQ....KGAVLWQGKPLDYSKRGLLALRQ.Q.........VATVFQDP..............
BRAG_PSEAE/27-209       GEIVTLIGANGAGKSTLLMTLCGSPQ..AA....SGSIRYE.......GEELVGLPS....STIMRKSIAVVPEGRRVF..SRLTV....
YHBG_AZOCA/70-253       GEAVGLLGPNGAGKTTVFYMITGLVK..AD....QGRIELD.......GHDVTPMPM.YRRARLG...IGYLPQEASIF..RGLSV....
BRAF_PSEAE/31-230       KQVVSMIGPNGAGKTTVFNCLTGFYQ..PT....GGLIRLD.......GEEIQGLPG....HKIARKGVVRTFQNVRLF..KEMTA....
YTFR_ECOLI/35-221       GEIMALLGENGAGKSTLIKALTGVYH..AD....RGTIWLE..........GQAISP.KNTAHAQQLGIGTVYQEVNLL..PNMSV....
ARAG_ECOLI/33-219       GQVHALMGENGAGKSTLLKILSGNYA..PT....TGSVVIN..........GQEMSF.SDTTAALNAGVAIIYQELHLV..PEMTV....
ABCA_AERSA/51-221       GETIGIVGHNGAGKSTLLQLITGVMQ..PD....CGQITRTGRV........................VGLLELGSGFN..PEFTG....
```

pFAM at Sanger Center (UK)

http://pfam.sanger.ac.uk/



Trusted matches - domains scoring higher than the gathering threshold (A)

Domain	Start	End	Bits	Evalue	Alignment	Mode
ABC_membrane	81	350	197.30	3.2e-56	Align	ls
ABC_tran	451	622	191.20	2.1e-54	Align	ls
ABC_membrane	862	1147	217.40	2.8e-62	Align	ls
ABC_tran	1237	1419	209.20	7.9e-60	Align	ls

Matches to Pfam-B

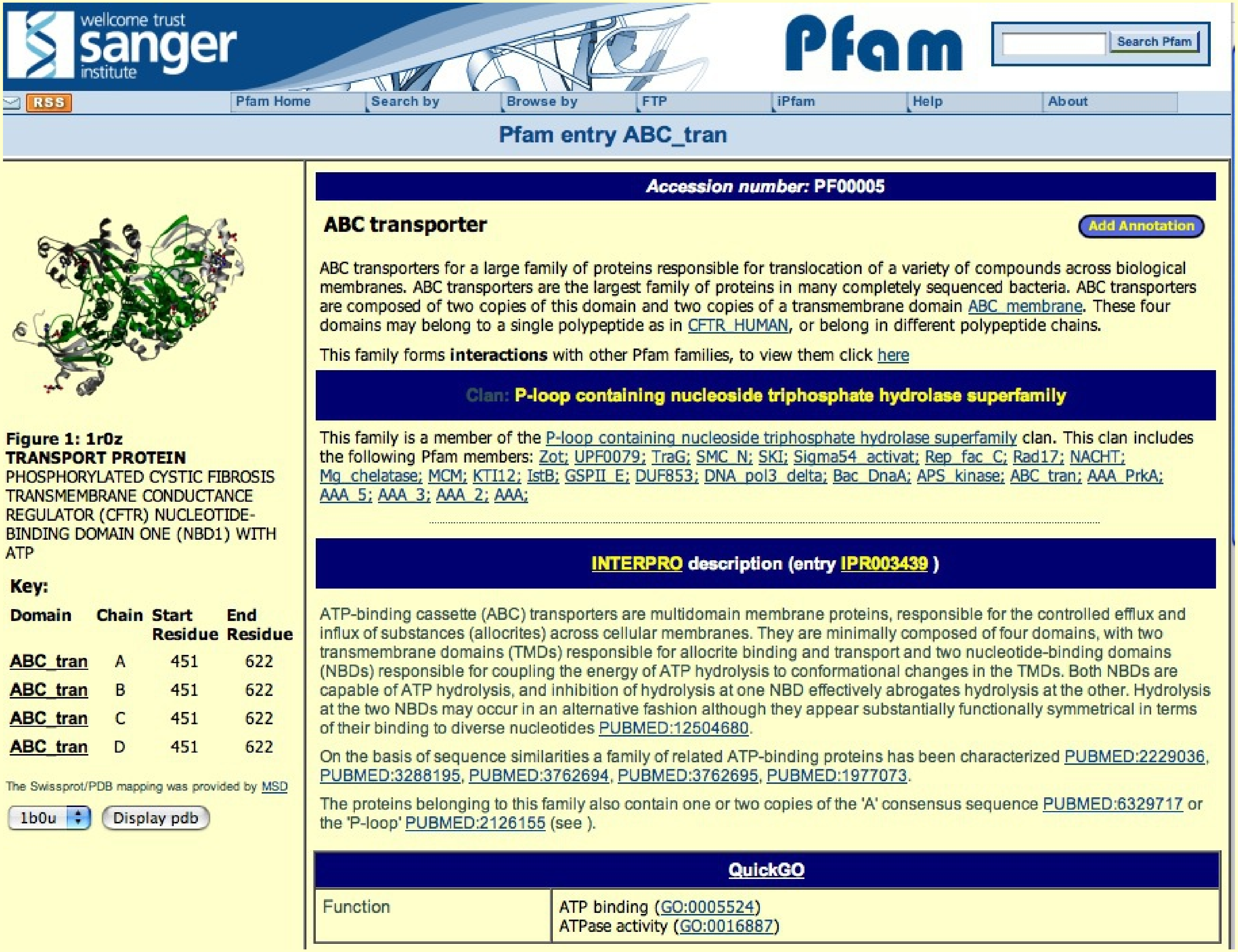
Domain	Start	End	Alignment
Pfam-B_148	3	62	Align
Pfam-B_59195	351	379	Align
Pfam-B_7475	380	437	Align
Pfam-B_88147	488	560	Align
Pfam-B_7405	1420	1480	Align

Potential matches - Domains with Evalues above the cutoff

Domain	Start	End	Bits	Evalue	Alignment	Mode
Bac_export_3	185	250	-26.20	0.055	Align	ls
MobB	452	564	-50.90	0.83	Align	ls
Dynamin_N	454	636	-55.00	0.43	Align	ls
DUF788	892	1052	-65.10	0.87	Align	ls

pFAM at Sanger Center (UK)

http://pfam.sanger.ac.uk/



Pfam entry ABC_tran

Figure 1: 1r0z
TRANSPORT PROTEIN
PHOSPHORYLATED CYSTIC FIBROSIS TRANSMEMBRANE CONDUCTANCE REGULATOR (CFTR) NUCLEOTIDE-BINDING DOMAIN ONE (NBD1) WITH ATP

Key:

Domain	Chain	Start Residue	End Residue
ABC_tran	A	451	622
ABC_tran	B	451	622
ABC_tran	C	451	622
ABC_tran	D	451	622

The Swissprot/PDB mapping was provided by MSD

Accession number: PF00005

ABC transporter

ABC transporters for a large family of proteins responsible for translocation of a variety of compounds across biological membranes. ABC transporters are the largest family of proteins in many completely sequenced bacteria. ABC transporters are composed of two copies of this domain and two copies of a transmembrane domain ABC_membrane. These four domains may belong to a single polypeptide as in CFTR_HUMAN, or belong in different polypeptide chains.

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

Clan: P-loop containing nucleoside triphosphate hydrolase superfamily

This family is a member of the P-loop containing nucleoside triphosphate hydrolase superfamily clan. This clan includes the following Pfam members: Zot; UPF0079; TraG; SMC_N; SKI; Sigma54_activat; Rep_fac_C; Rad17; NACHT; Mg_chelatase; MCM; KTI12; IstB; GSPII_E; DUF853; DNA_pol3_delta; Bac_DnaA; APS_kinase; ABC_tran; AAA_PrkA; AAA_5; AAA_3; AAA_2; AAA;

INTERPRO description (entry IPR003439)

ATP-binding cassette (ABC) transporters are multidomain membrane proteins, responsible for the controlled efflux and influx of substances (allocrites) across cellular membranes. They are minimally composed of four domains, with two transmembrane domains (TMDs) responsible for allocrite binding and transport and two nucleotide-binding domains (NBDs) responsible for coupling the energy of ATP hydrolysis to conformational changes in the TMDs. Both NBDs are capable of ATP hydrolysis, and inhibition of hydrolysis at one NBD effectively abrogates hydrolysis at the other. Hydrolysis at the two NBDs may occur in an alternative fashion although they appear substantially functionally symmetrical in terms of their binding to diverse nucleotides PUBMED:12504680.

On the basis of sequence similarities a family of related ATP-binding proteins has been characterized PUBMED:2229036, PUBMED:3288195, PUBMED:3762694, PUBMED:3762695, PUBMED:1977073.

The proteins belonging to this family also contain one or two copies of the 'A' consensus sequence PUBMED:6329717 or the 'P-loop' PUBMED:2126155 (see).

QuickGO

Function	ATP binding (GO:0005524) ATPase activity (GO:0016887)