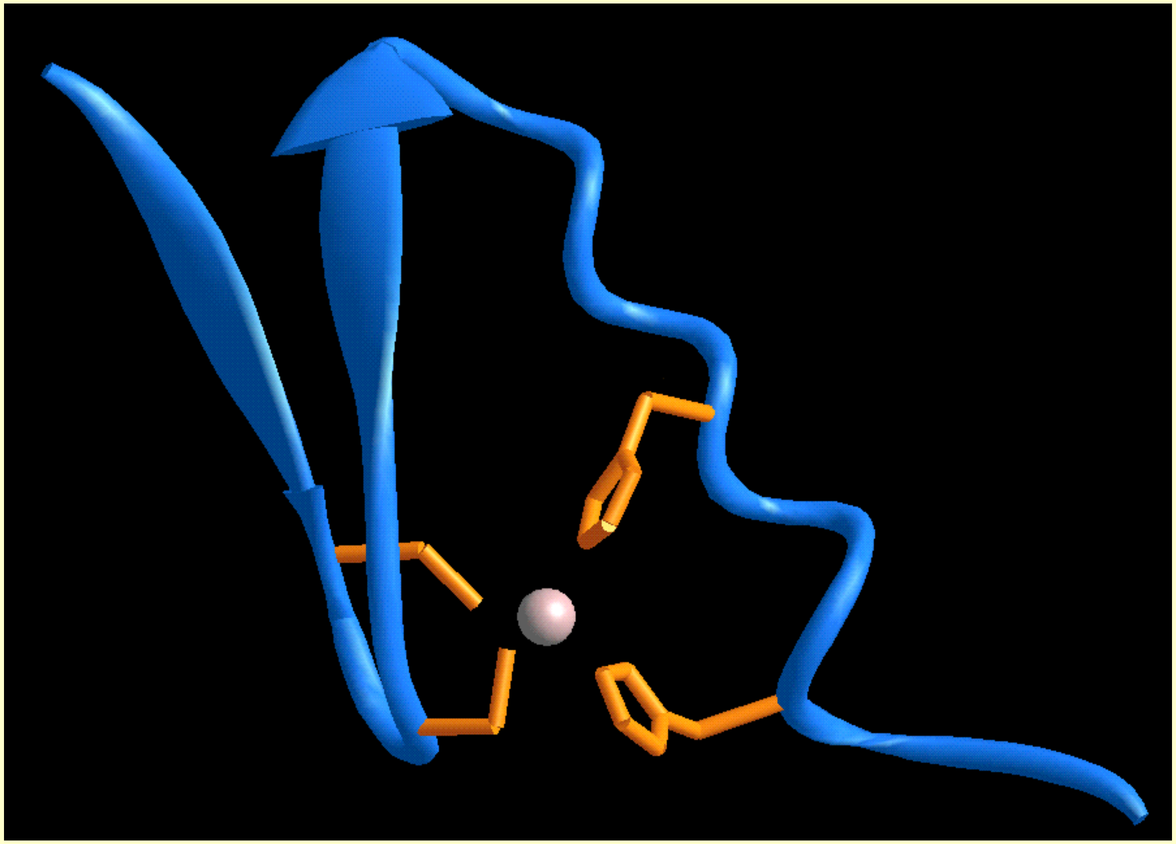
Bioinformatics

http://biochem158.stanford.edu/bioinformatics.html

Genomics, Bioinformatics & Medicine

http://biochem158.stanford.edu/



Doug Brutlag
Professor Emeritus of Biochemistry & Medicine
Stanford University School of Medicine

Beckman Center
FOR MOLECULAR AND GENETIC MEDICINE

25th Anniversary Beckman Symposium

http://beckman.stanford.edu/symposium-schedule-review.html

Innovation in the Biosphere

AM	8:45	Lucy Shapiro	*Opening Remarks*
	9:00	Adam Arkin	*Biomolecular Response to Biosphere Challenges*
	9:50	Jay Keasling	*Reengineering Life*
	10:50	Uri Alon	*Laws of Genetic Microcircuits*
	11:40	Jonathan Weissman	*Deep Reading of Genomes*
PM	12:30	Lunch	
	1:30	Stephen Quake	*Precision Measurement in Biology*
	2:20	Karl Deisseroth	*Optogenetics Manipulating Brain Circuitry*
	3:20	Carla Shatz	*New Synapses in Old Brains*
	4:10	J. Craig Venter	*Life from Scratch*
	5:00	Paul Berg	*Closing Remarks*

25th Anniversary Beckman Symposium

Feb 23, 2015, 8:45AM - 5:15PM

Berg Hall
Li Ka Shing Conference Center

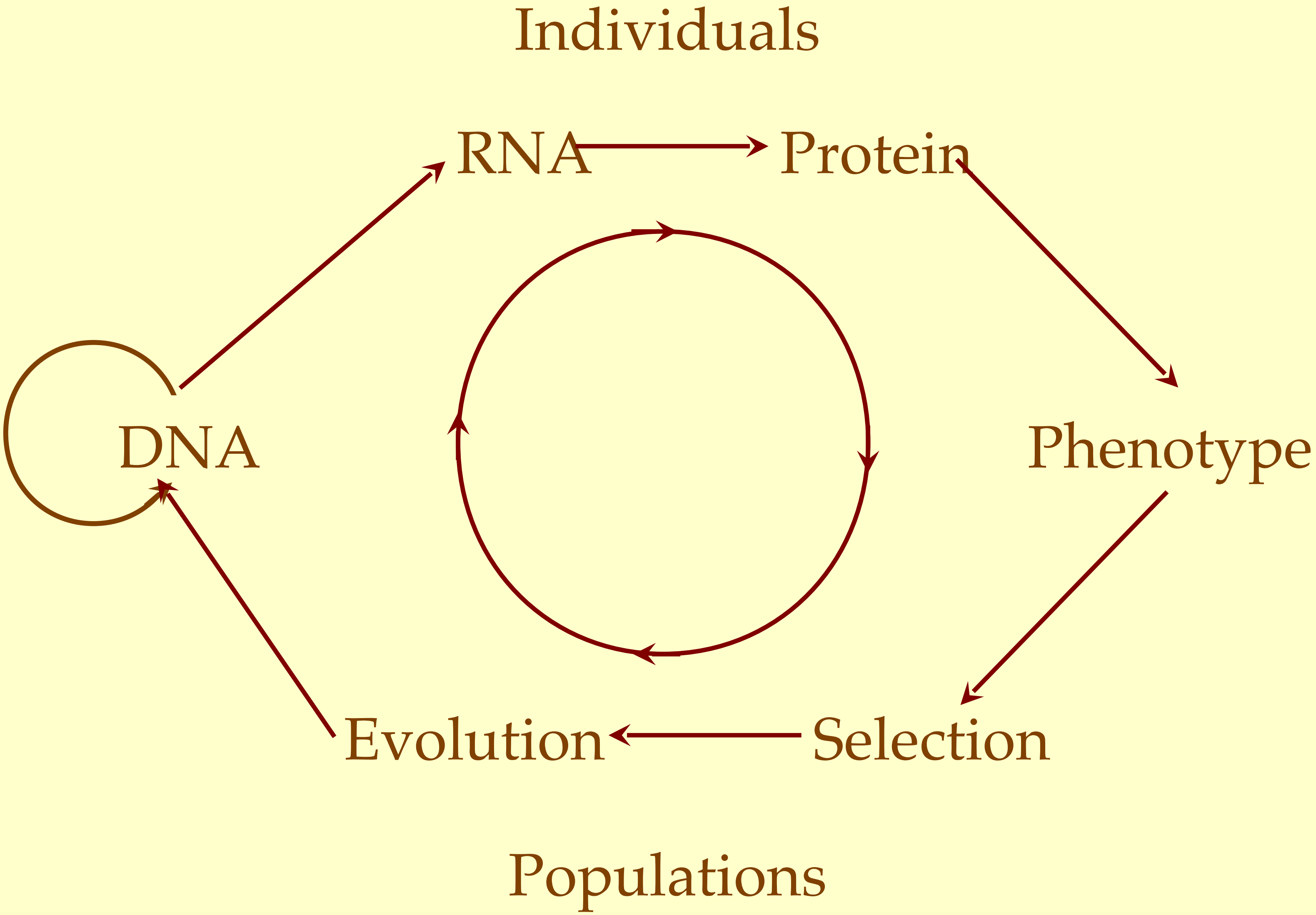
Co-hosts:
Lucy Shapiro & Stephen Quake

More speaker and program info:
http://beckman.stanford.edu

Design: ghostdog.com

Thanks To Matthew Lum

What is Bioinformatics?



Individuals

RNA

Protein

DNA

Phenotype

Evolution

Selection

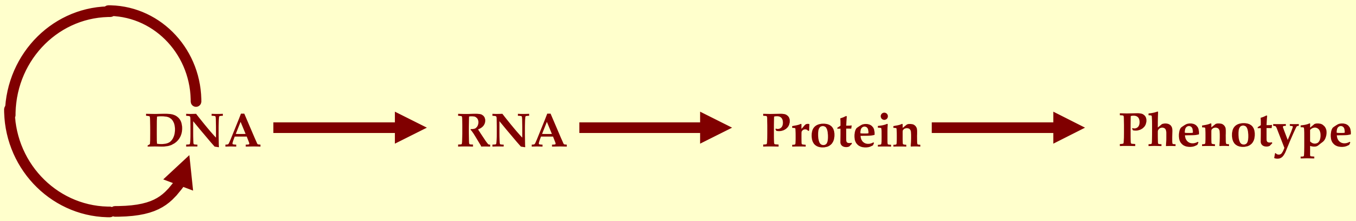
Populations

Biological Information

Computational Goals of Bioinformatics

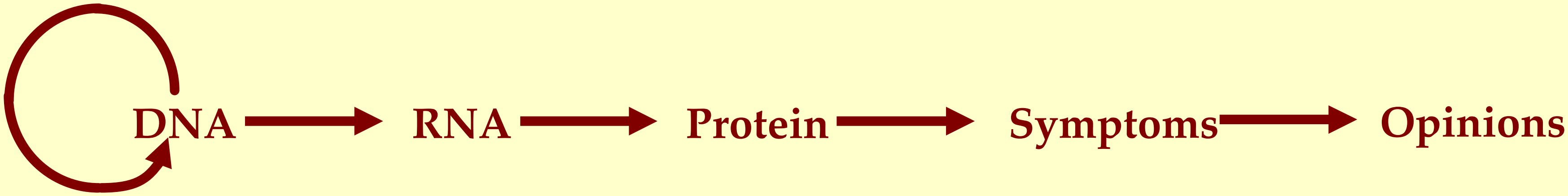
- Learn & Generalize: Discover conserved patterns (models) of sequences, structures, metabolism & chemistries from well-studied examples.
- Prediction: Infer function or structure of newly sequenced genes, genomes, proteomes or proteins from these generalizations.
- Organize & Integrate: Develop a systematic 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

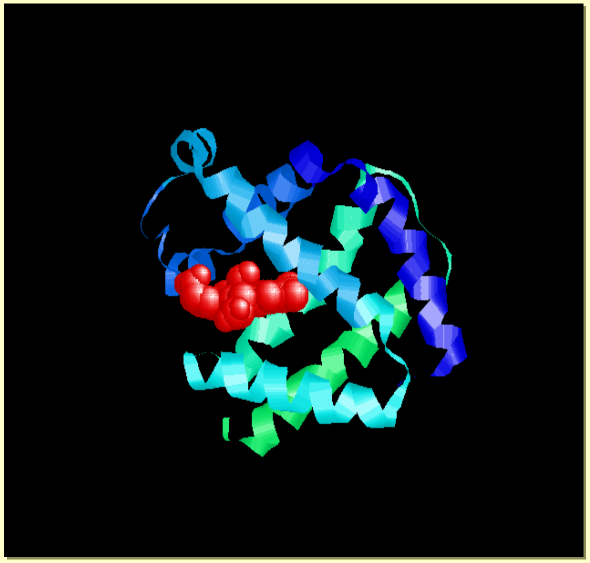


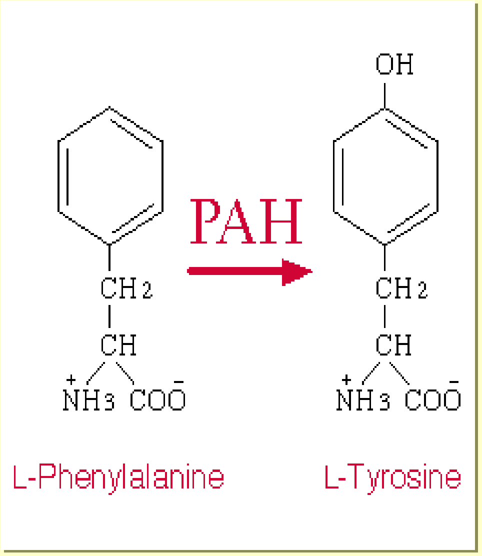
DNA → RNA → Protein → Symptoms → Opinions

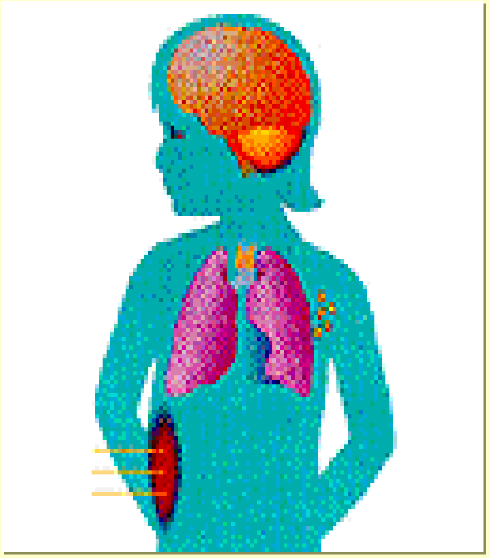
Central Paradigm of Bioinformatics

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

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



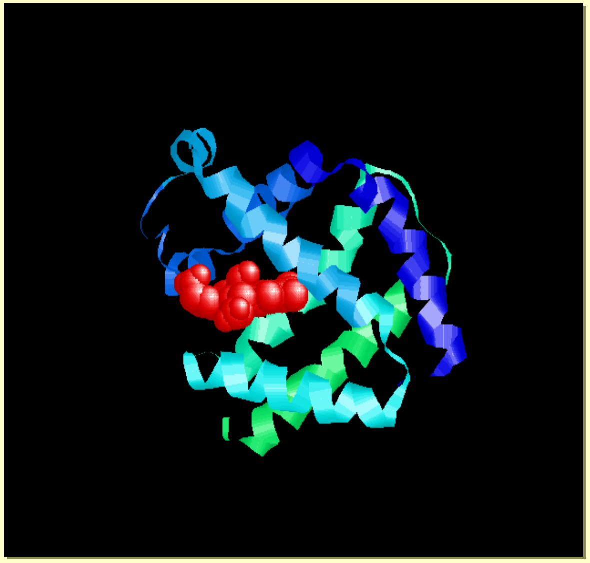


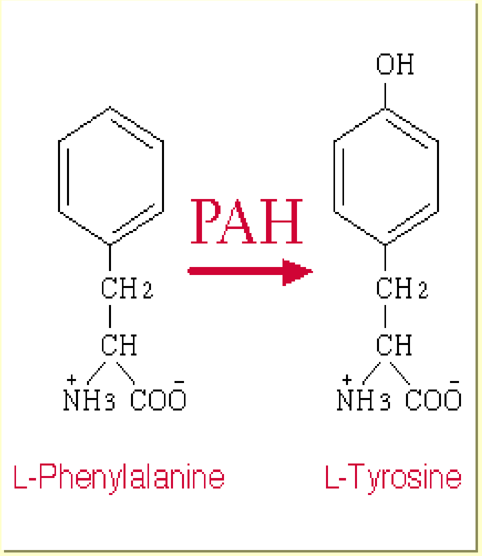


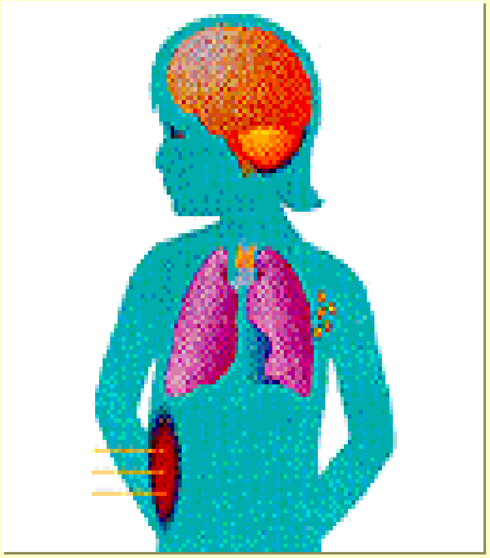
Central Paradigm of Bioinformatics

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

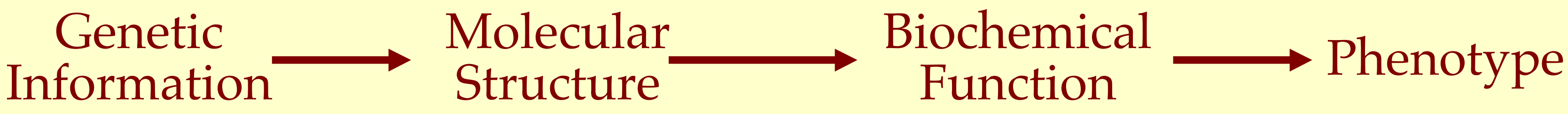
MVHLTPEEKT
AVNALWGKVN
VDAVGGEALG
RLLVVYPWTQ
RFFESFGDLS
SPDAVMGNPK
VKAHGKKVLG
AFSDGLAHLD
NLKGTFSQLS
ELHCDKLHVD
PENFRLLGNV
LVCVLARNFG
KEFTPQMQAA
YQKVVAGVAN
ALAHKYH







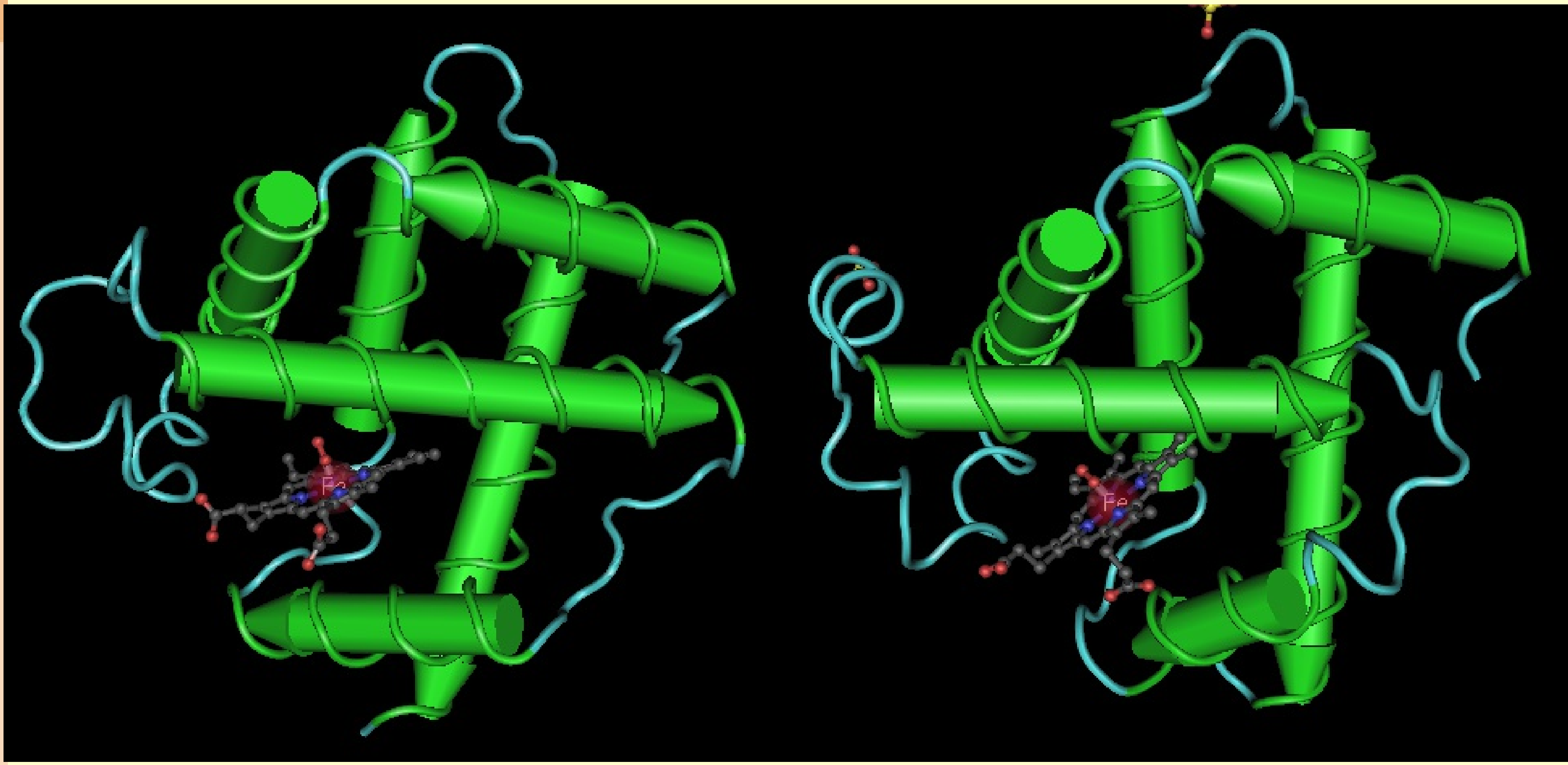
Challenges Understanding Genetic Information



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 meta-stable

Challenges Understanding Genetic Information

Genetic Information → Molecular Structure → Biochemical Function → Phenotype

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

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 250 - Log Odds
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

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

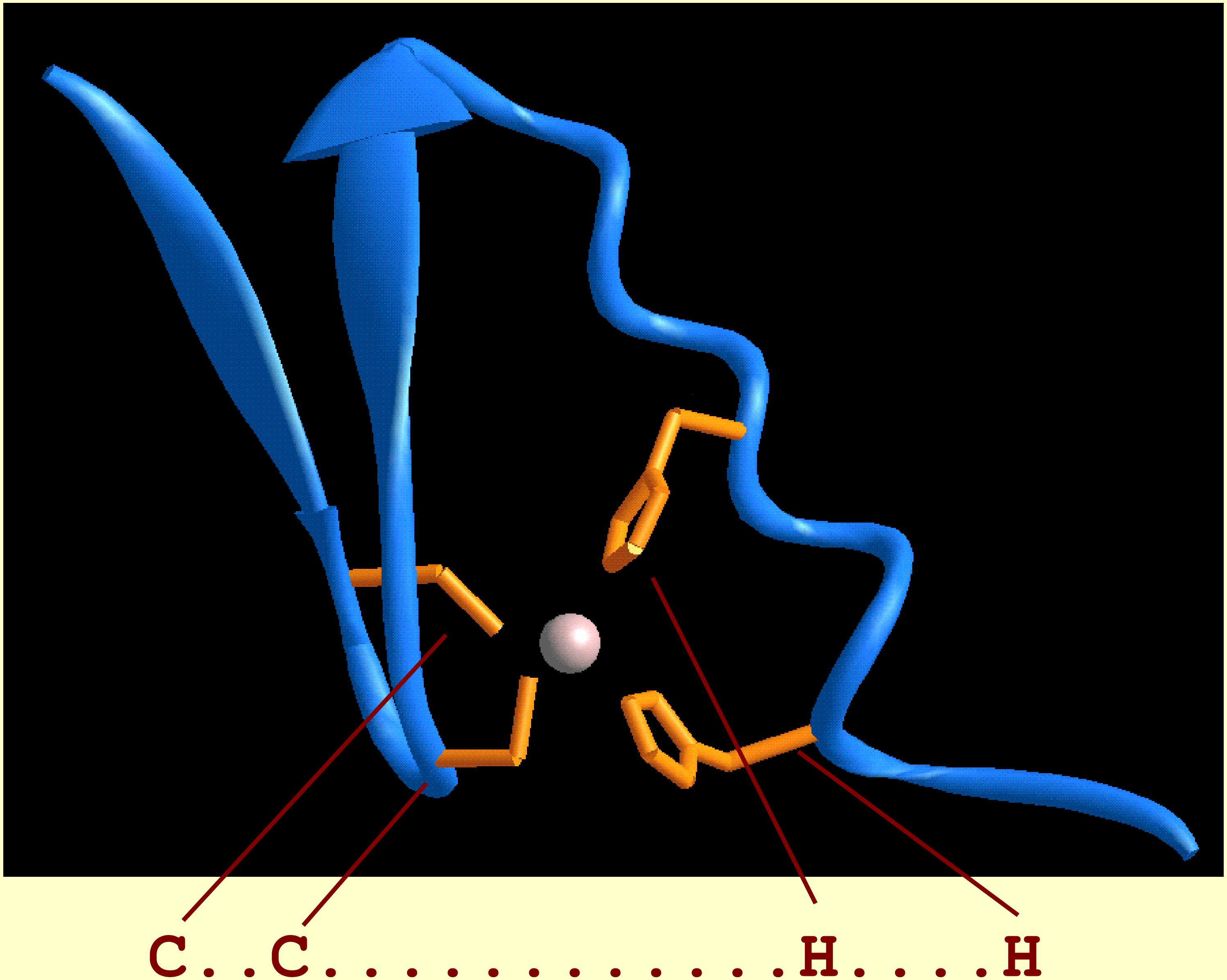
Zinc Finger (C2H2 type)
C X{2,4} C X{12} H X{3,5} H

Sequences of
Common
Structure or Function

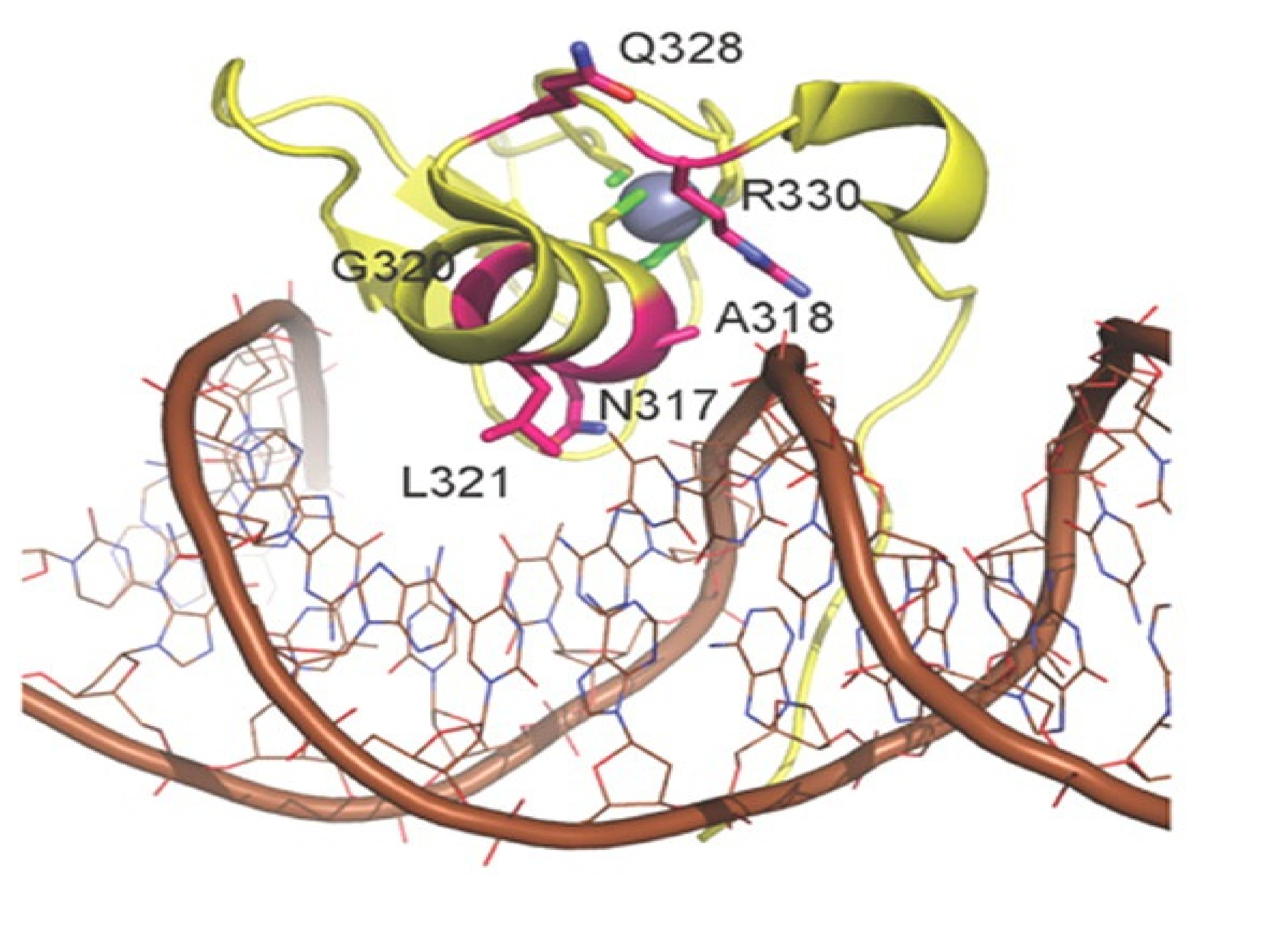
Sequence Similarity

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

A Typical Motif:
Zinc Finger DNA Binding Motif

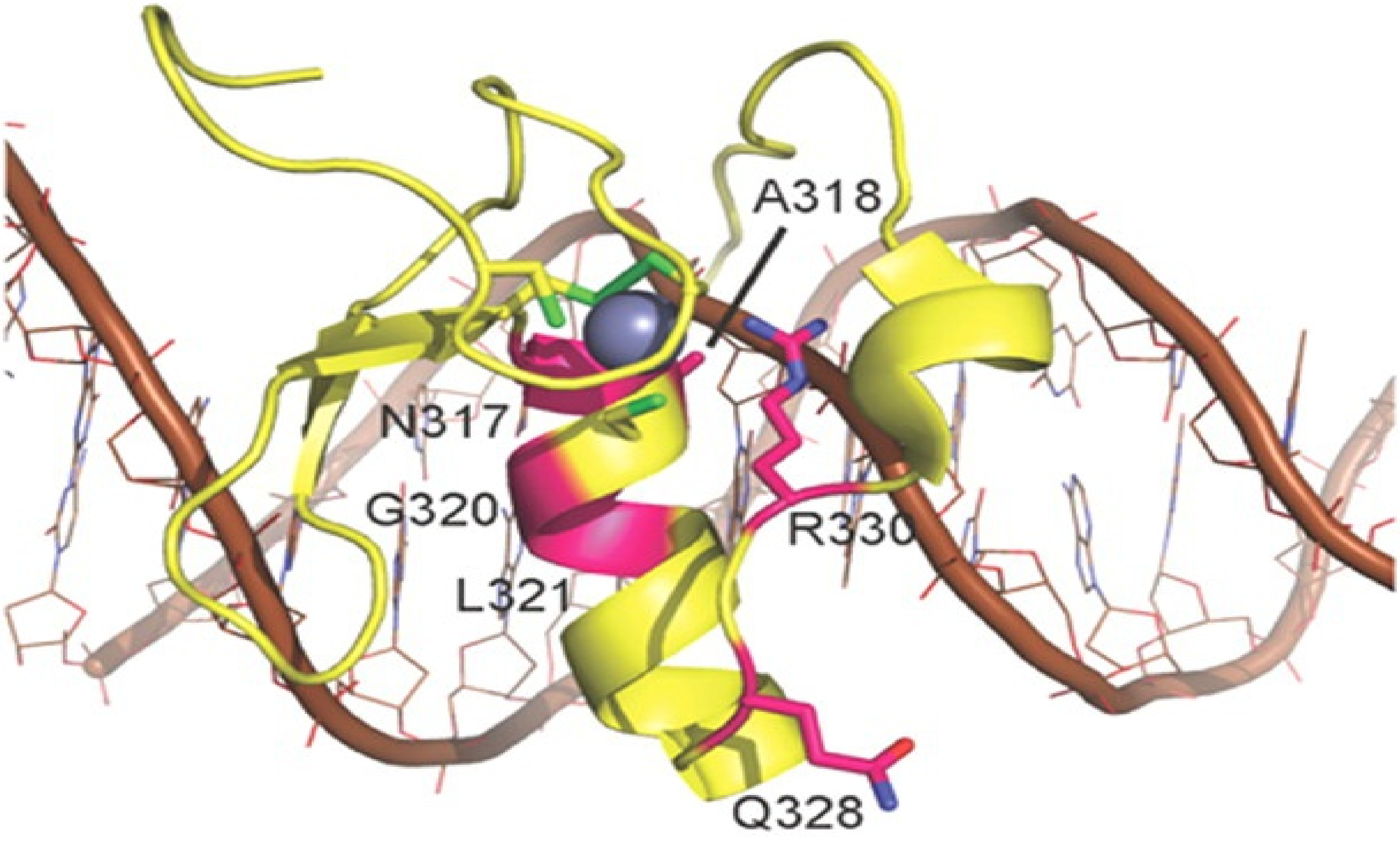


Zinc Finger DNA Binding Motif



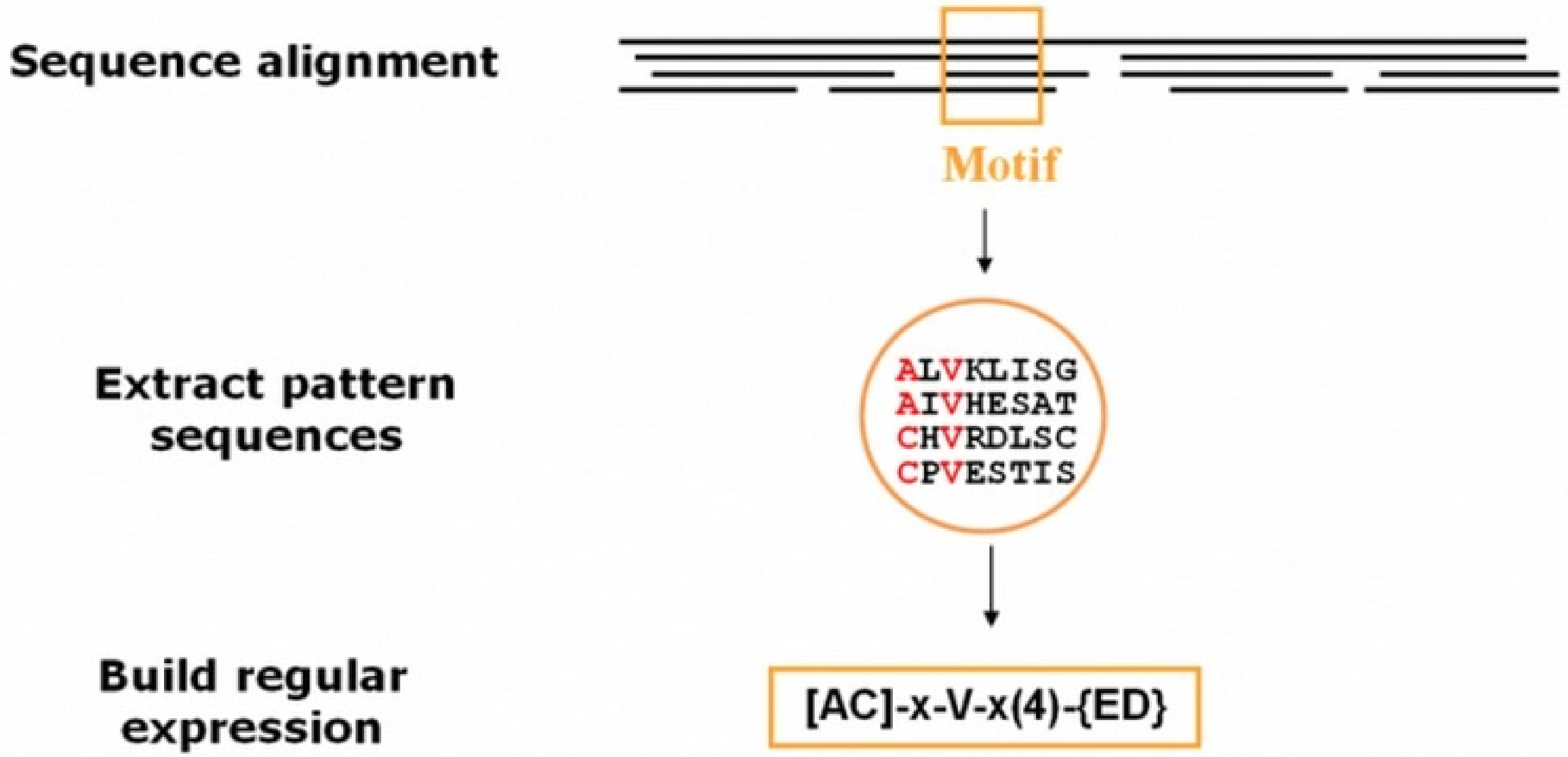
Greif et al, Blood (July 12, 2012) 120

Zinc Finger DNA Binding Motif



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

Protein Motifs from Multiple Sequence Alignments



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_2H_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

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

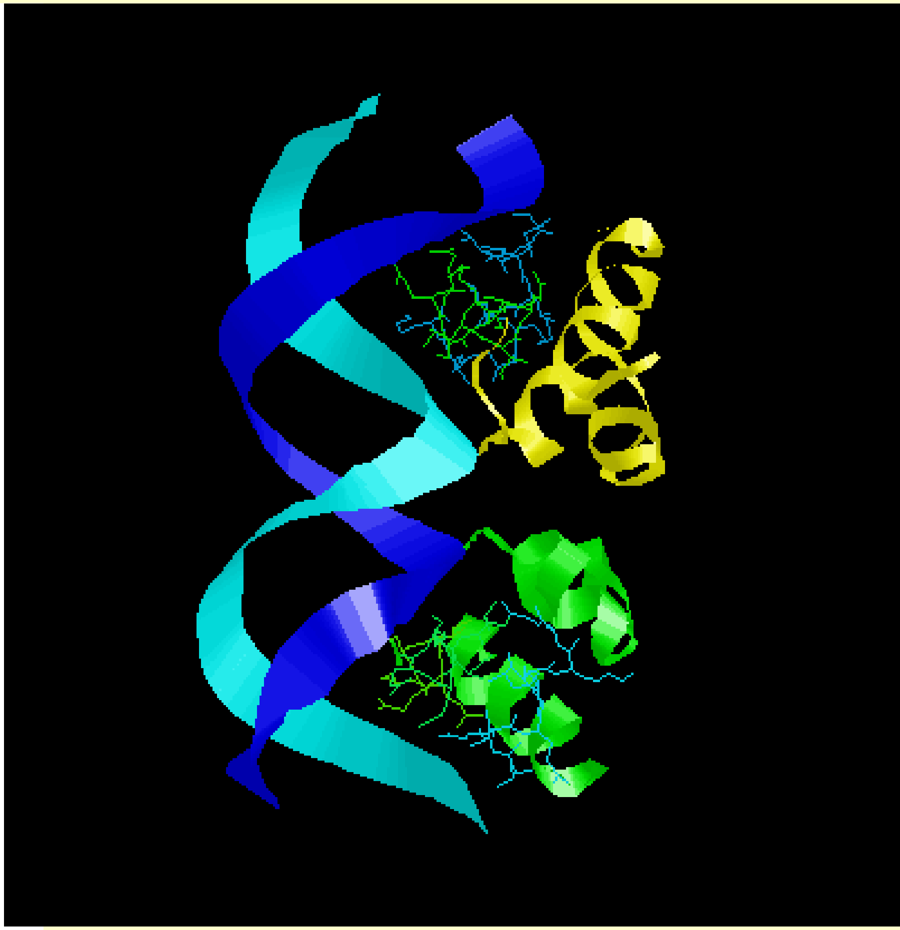
Sequences of Common Structure or Function

Sequence Similarity

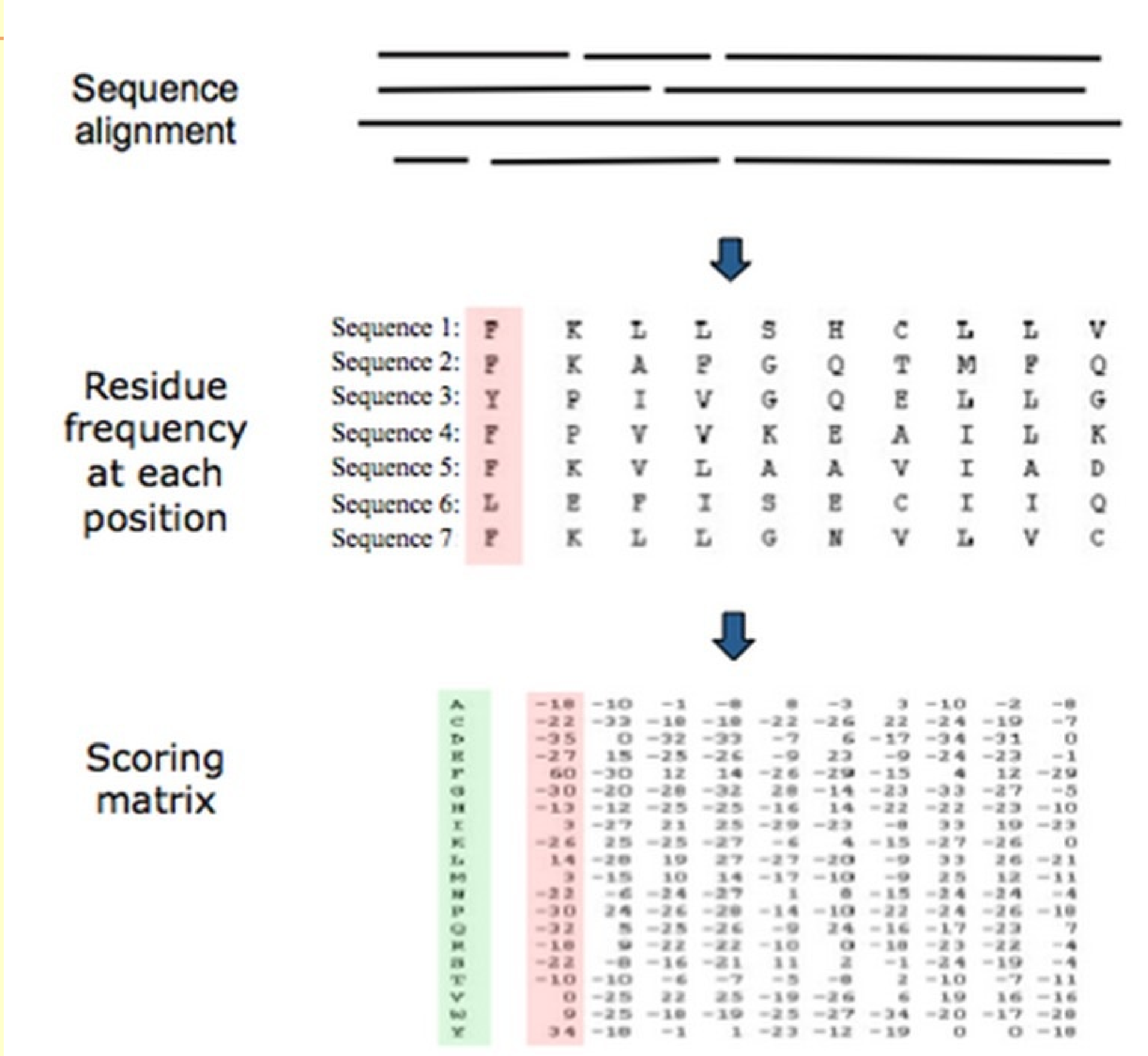
```
                     10        20        30        40              50
Query         HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
               |:| :|: | |:||||     | |:||| |: : :|:|  :|  |      |:  |
Database      VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
        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

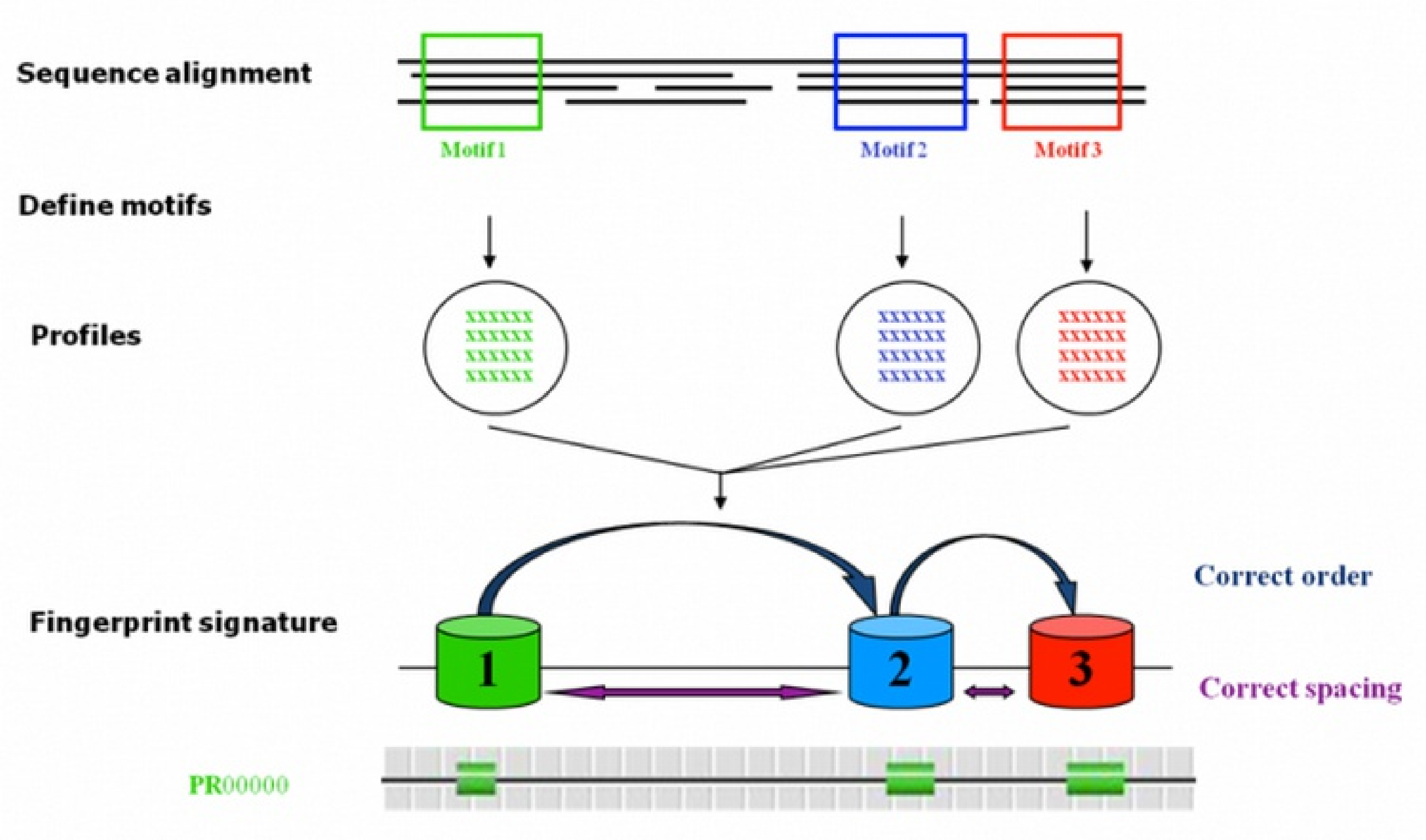
Sequence 1:	F	K	L	L	S	H	C	L	L	V
Sequence 2:	F	K	A	F	G	Q	T	M	F	Q
Sequence 3:	Y	P	I	V	G	Q	E	L	L	G
Sequence 4:	F	P	V	V	K	E	A	I	L	K
Sequence 5:	F	K	V	L	A	A	V	I	A	D
Sequence 6:	L	E	F	I	S	E	C	I	I	Q
Sequence 7	F	K	L	L	G	N	V	L	V	C

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

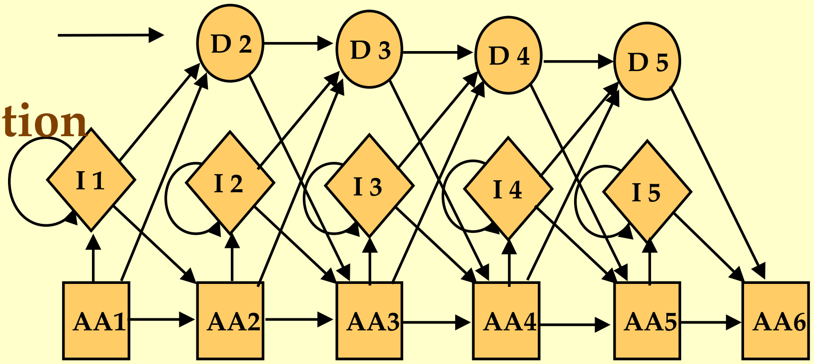
	1	2	3	4	5	6	7	8	9	10	11	12
A	2	1	3	13	10	12	67	4	13	9	1	2
R	7	5	8	9	4	0	1	16	7	0	1	0
N	0	8	0	1	0	0	0	2	1	1	10	0
D	0	1	0	1	13	0	0	12	1	0	4	0
C	0	0	1	0	0	0	0	0	0	2	2	1
Q	1	1	21	8	10	0	0	7	6	0	0	2
E	2	0	0	9	21	0	0	15	7	3	3	0
G	9	7	1	4	0	0	8	0	0	0	46	0
H	4	3	1	1	2	0	0	2	2	0	5	0
I	10	0	11	1	2	10	0	4	9	3	0	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

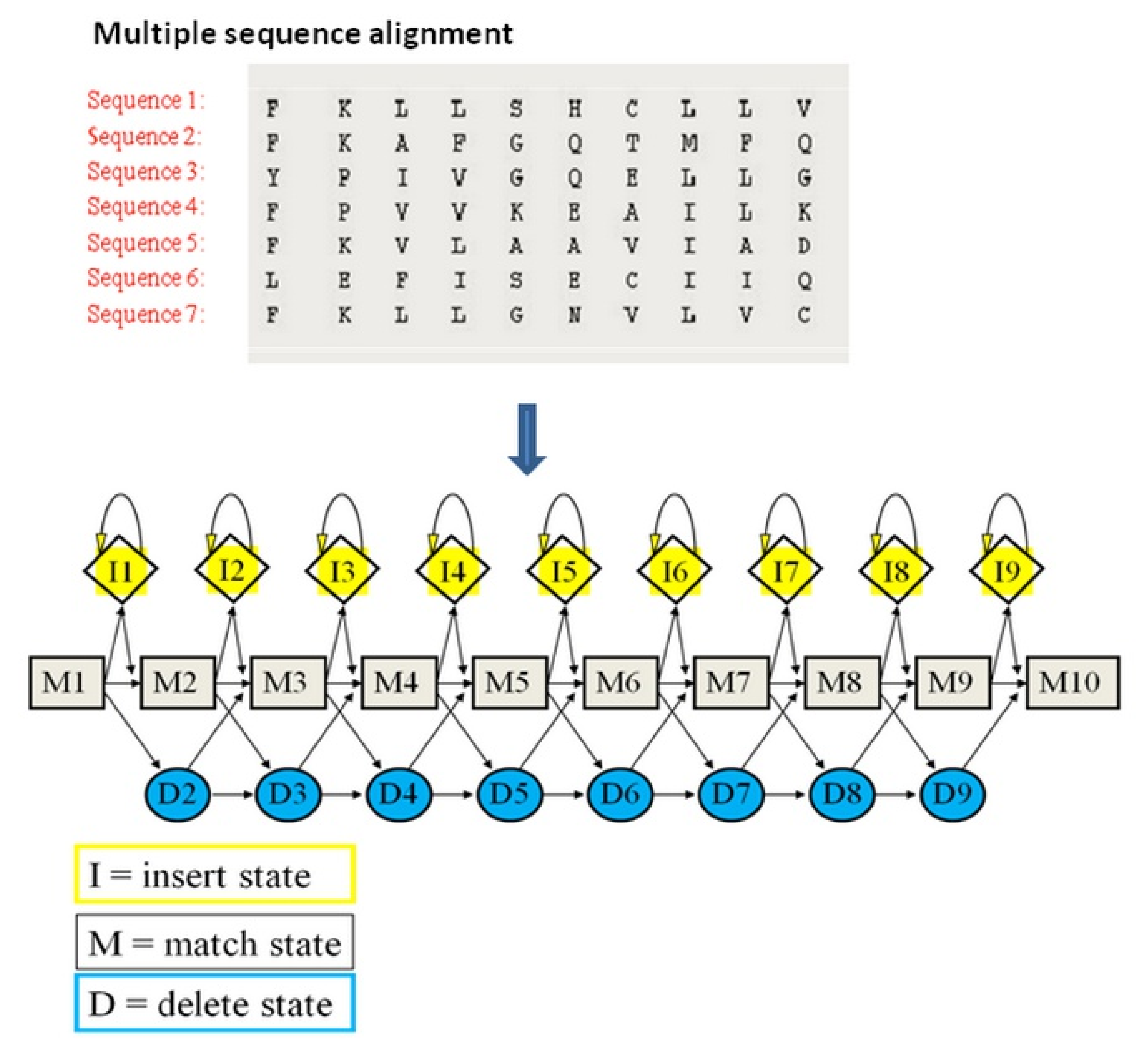
Profiles, PSI-BLAST Hidden Markov Models



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



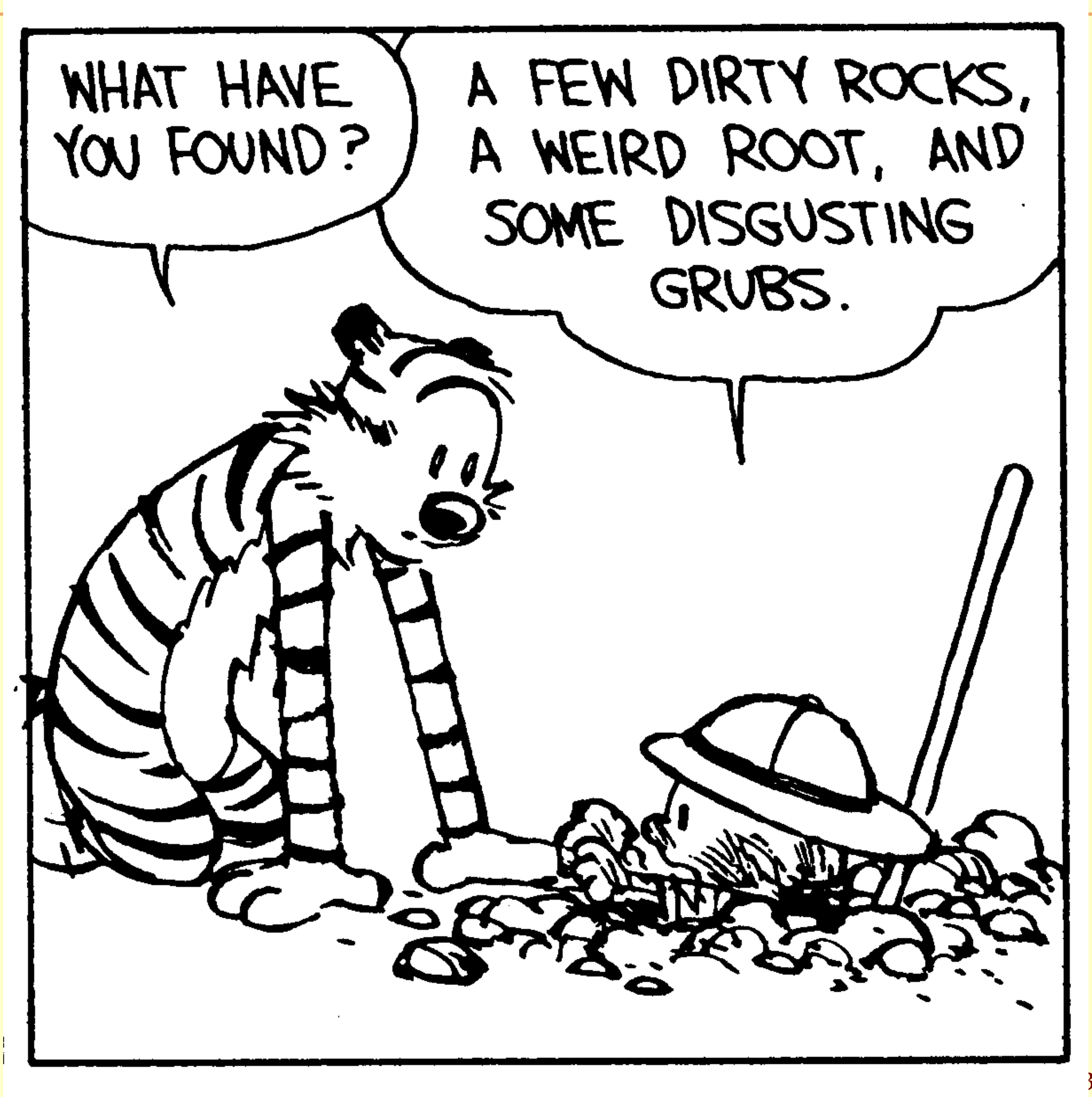
Multiple sequence alignment

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

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

Discovering Function from Protein Sequence

BLOCKs, PRINTs, PSSMS or Weight Matrices

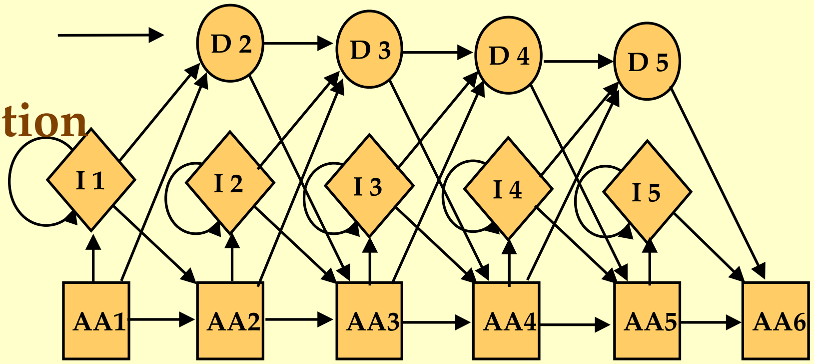
	Position											
	1	2	3	4	5	6	7	8	9	10	11	12
A	2	1	3	13	10	12	67	4	13	9	1	2
R	7	5	8	9	4	0	1	16	7	0	1	0
N	0	8	0	1	0	0	0	2	1	1	10	0
D	0	1	0	1	13	0	0	12	1	0	4	0
C	0	0	1	0	0	0	0	0	0	2	2	1
Q	1	1	21	8	10	0	0	7	6	0	0	2
E	2	0	0	9	21	0	0	15	7	3	3	0
G	9	7	1	4	0	0	8	0	0	0	46	0
H	4	3	1	1	2	0	0	2	2	0	5	0
I	10	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

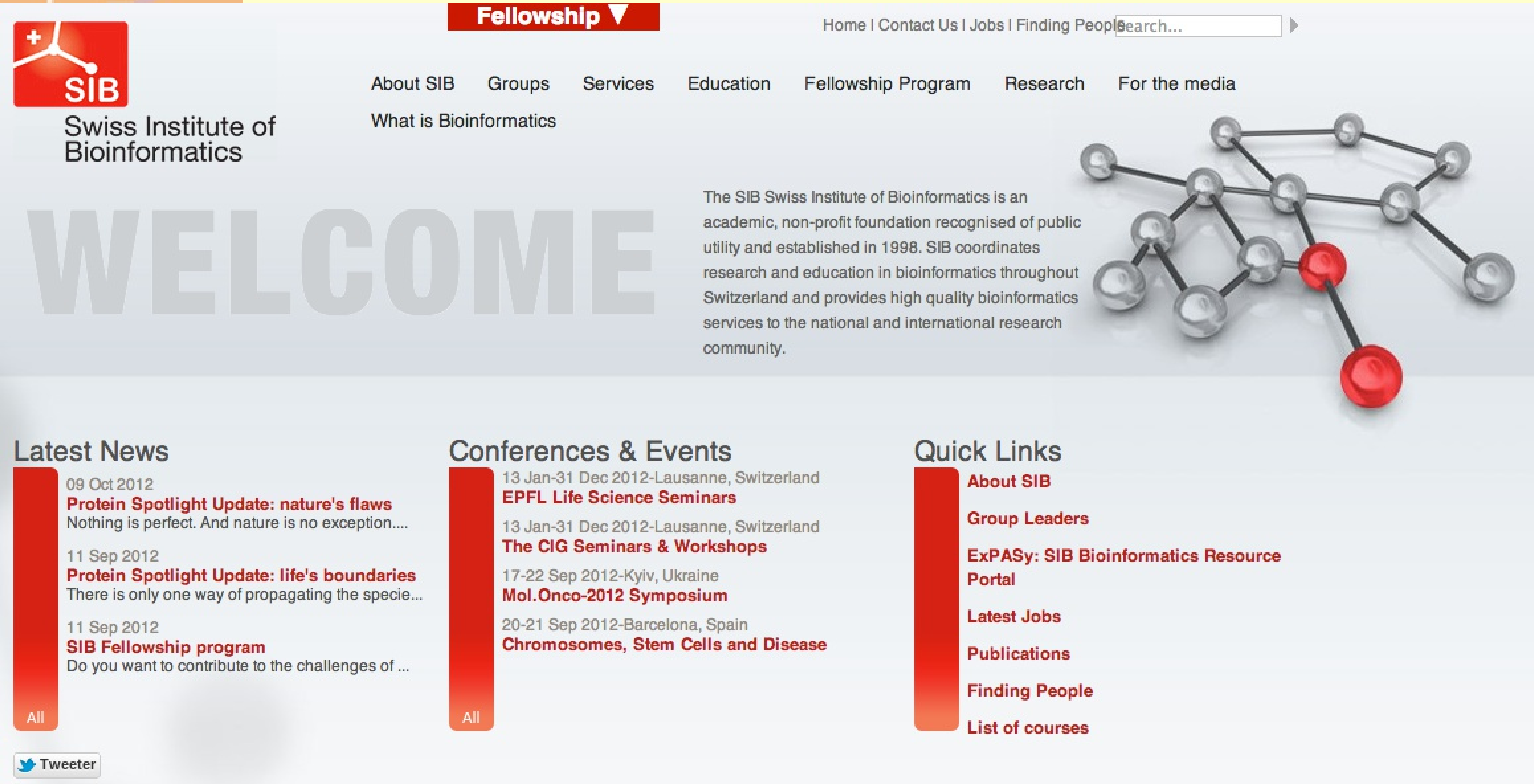
Profiles, PSI-BLAST Hidden Markov Models



Sequence Similarity

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

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

09 Oct 2012
Protein Spotlight Update: nature's flaws
Nothing is perfect. And nature is no exception....

11 Sep 2012
Protein Spotlight Update: life's boundaries
There is only one way of propagating the specie...

11 Sep 2012
SIB Fellowship program
Do you want to contribute to the challenges of ...

Conferences & Events

13 Jan-31 Dec 2012-Lausanne, Switzerland
EPFL Life Science Seminars

13 Jan-31 Dec 2012-Lausanne, Switzerland
The CIG Seminars & Workshops

17-22 Sep 2012-Kyiv, Ukraine
Mol.Onco-2012 Symposium

20-21 Sep 2012-Barcelona, Spain
Chromosomes, Stem Cells and Disease

Quick Links

- **About SIB**
- **Group Leaders**
- **ExPASy: SIB Bioinformatics Resource Portal**
- **Latest Jobs**
- **Publications**
- **Finding People**
- **List of courses**

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

HCD/CID spectra merger

A tool to combine HCD and CID MS/MS spectra
[details]

How to use this portal?

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

Popular resources

- UniProtKB
- SWISS-MODEL
- STRING
- PROSITE

Latest News

Updates to HAMAP and its website - 2012-10-12

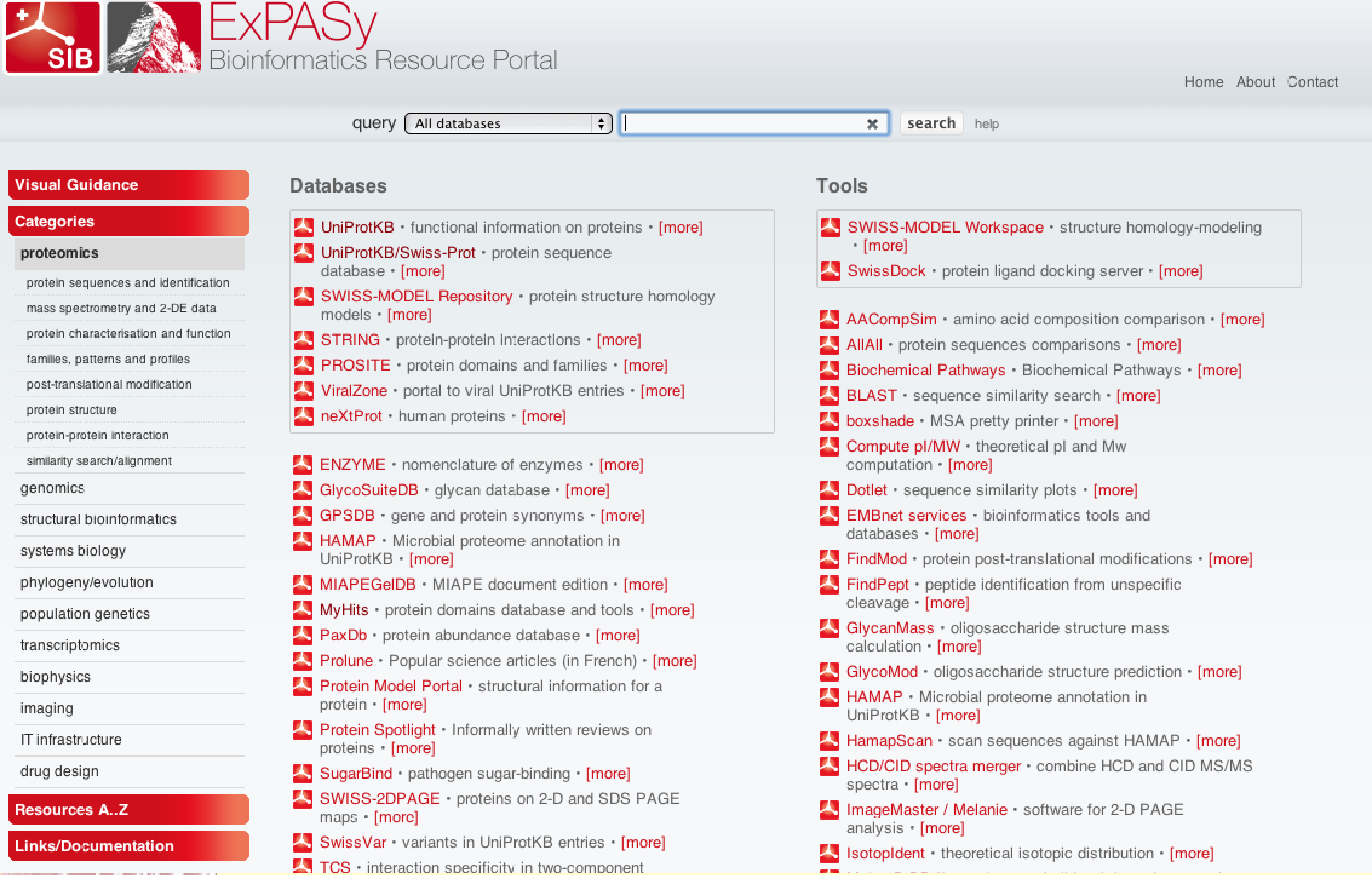
The scope of HAMAP has been extended to classify and annotate also proteins from eukaryotic species. Additionally, our website has been updated and we provide now distinct pages for HAMAP family profiles on the HAMAP website.
More information on: http://hamap.expasy.org.

UniProt Knowledgebase release 2012_09 - 2012-10-03

UniProtKB/SwissProt Release of 03-Oct-2012 contains 538,010 sequence entries...More.
UniProtKB/TrEMBL Release of 03-Oct-2012 contains 26,079,526 sequence entries...More

Expasy Bioinformatics Resource Portal

http://expasy.org/



Databases

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

- ENZYME • nomenclature of enzymes • [more]
- GlycoSuiteDB • glycan database • [more]
- GPSDB • gene and protein synonyms • [more]
- HAMAP • Microbial proteome annotation in UniProtKB • [more]
- MIAPEGelDB • MIAPE document edition • [more]
- MyHits • protein domains database and tools • [more]
- PaxDb • protein abundance database • [more]
- Prolune • Popular science articles (in French) • [more]
- Protein Model Portal • structural information for a protein • [more]
- Protein Spotlight • Informally written reviews on proteins • [more]
- SugarBind • pathogen sugar-binding • [more]
- SWISS-2DPAGE • proteins on 2-D and SDS PAGE maps • [more]
- SwissVar • variants in UniProtKB entries • [more]
- TCS • interaction specificity in two-component

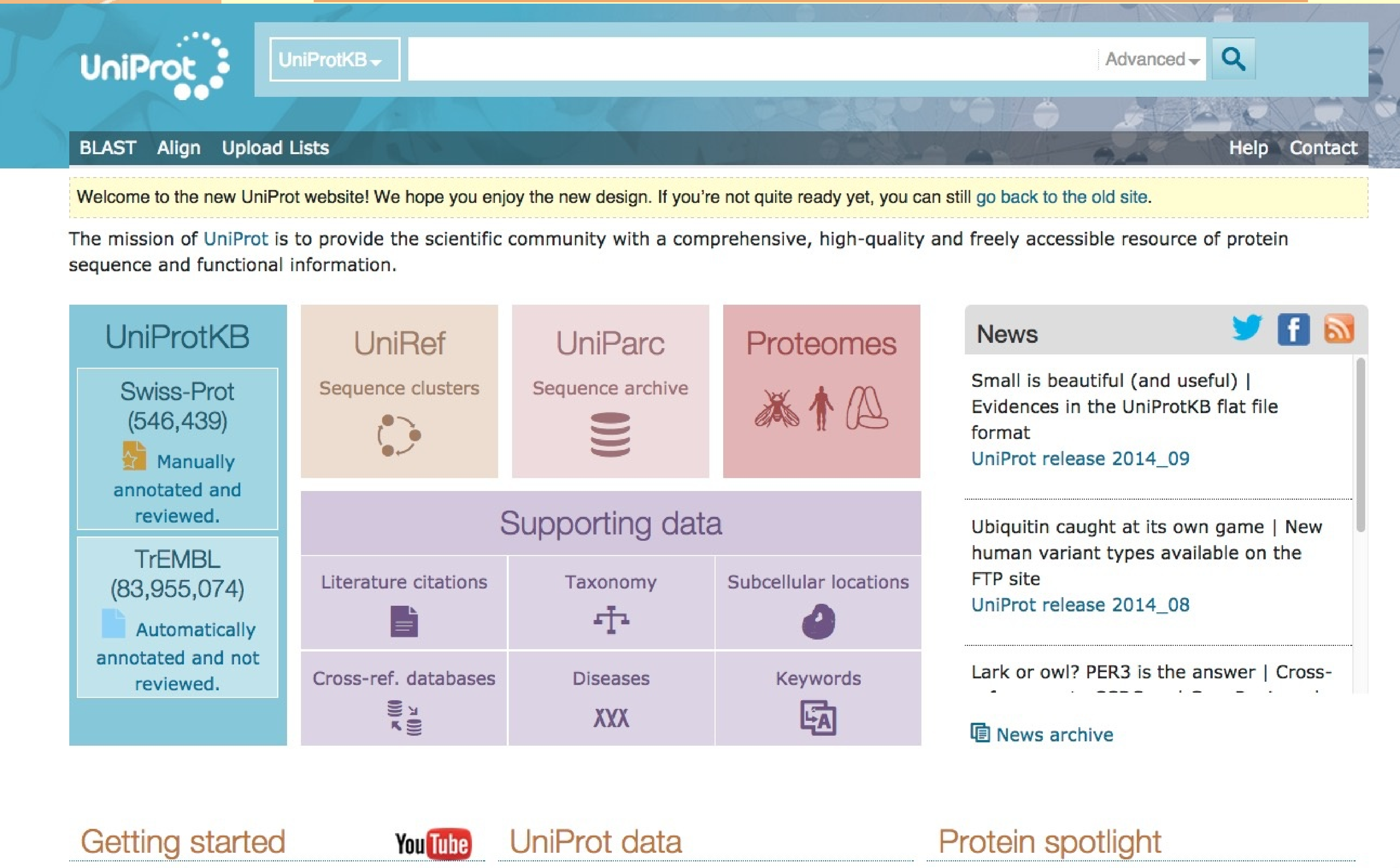
Tools

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

- AACompSim • amino acid composition comparison • [more]
- AllAll • protein sequences comparisons • [more]
- Biochemical Pathways • Biochemical Pathways • [more]
- BLAST • sequence similarity search • [more]
- boxshade • MSA pretty printer • [more]
- Compute pI/MW • theoretical pI and Mw computation • [more]
- Dotlet • sequence similarity plots • [more]
- EMBnet services • bioinformatics tools and databases • [more]
- FindMod • protein post-translational modifications • [more]
- FindPept • peptide identification from unspecific cleavage • [more]
- GlycanMass • oligosaccharide structure mass calculation • [more]
- GlycoMod • oligosaccharide structure prediction • [more]
- HAMAP • Microbial proteome annotation in UniProtKB • [more]
- HamapScan • scan sequences against HAMAP • [more]
- HCD/CID spectra merger • combine HCD and CID MS/MS spectra • [more]
- ImageMaster / Melanie • software for 2-D PAGE analysis • [more]
- IsotopIdent • theoretical isotopic distribution • [more]

UniProt Knowledge Base

http://www.uniprot.org/



Welcome to the new UniProt website! We hope you enjoy the new design. If you're not quite ready yet, you can still go back to the old site.

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.

UniProtKB

Swiss-Prot (546,439)
Manually annotated and reviewed.

TrEMBL (83,955,074)
Automatically annotated and not reviewed.

UniRef
Sequence clusters

UniParc
Sequence archive

Proteomes

Supporting data

Literature citations	Taxonomy	Subcellular locations
Cross-ref. databases	Diseases	Keywords

News

Small is beautiful (and useful) | Evidences in the UniProtKB flat file format
UniProt release 2014_09

Ubiquitin caught at its own game | New human variant types available on the FTP site
UniProt release 2014_08

Lark or owl? PER3 is the answer | Cross-

News archive

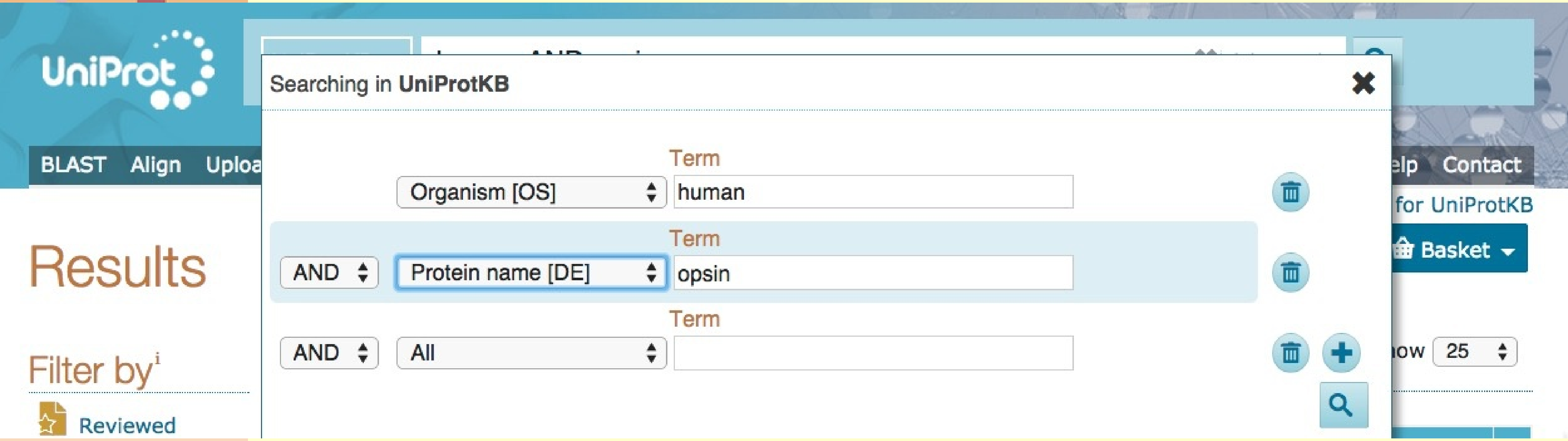
Getting started

UniProt data

Protein spotlight

UniProt Knowledge Base Advanced Search

http://www.uniprot.org/



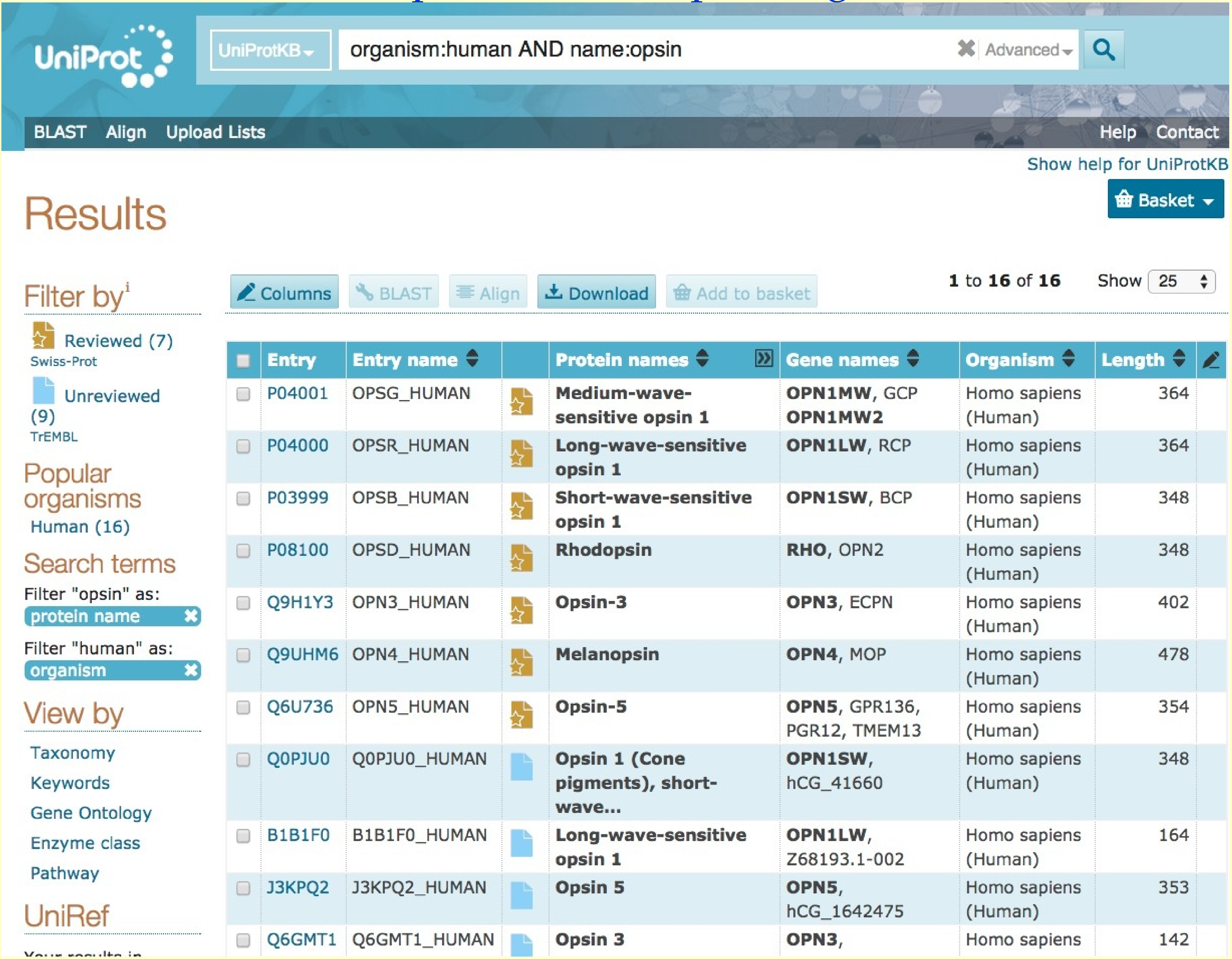
Searching in **UniProtKB**

		Term
	Organism [OS]	human
AND	Protein name [DE]	opsin
AND	All	

© Doug Brutlag 2015

UniProt Human Opsin Entries

http://www.uniprot.org/



organism:human AND name:opsin

Results

Filter by

- Reviewed (7) Swiss-Prot
- Unreviewed (9) TrEMBL

Popular organisms

- Human (16)

Search terms

Filter "opsin" as: protein name ✖

Filter "human" as: organism ✖

View by

- Taxonomy
- Keywords
- Gene Ontology
- Enzyme class
- Pathway

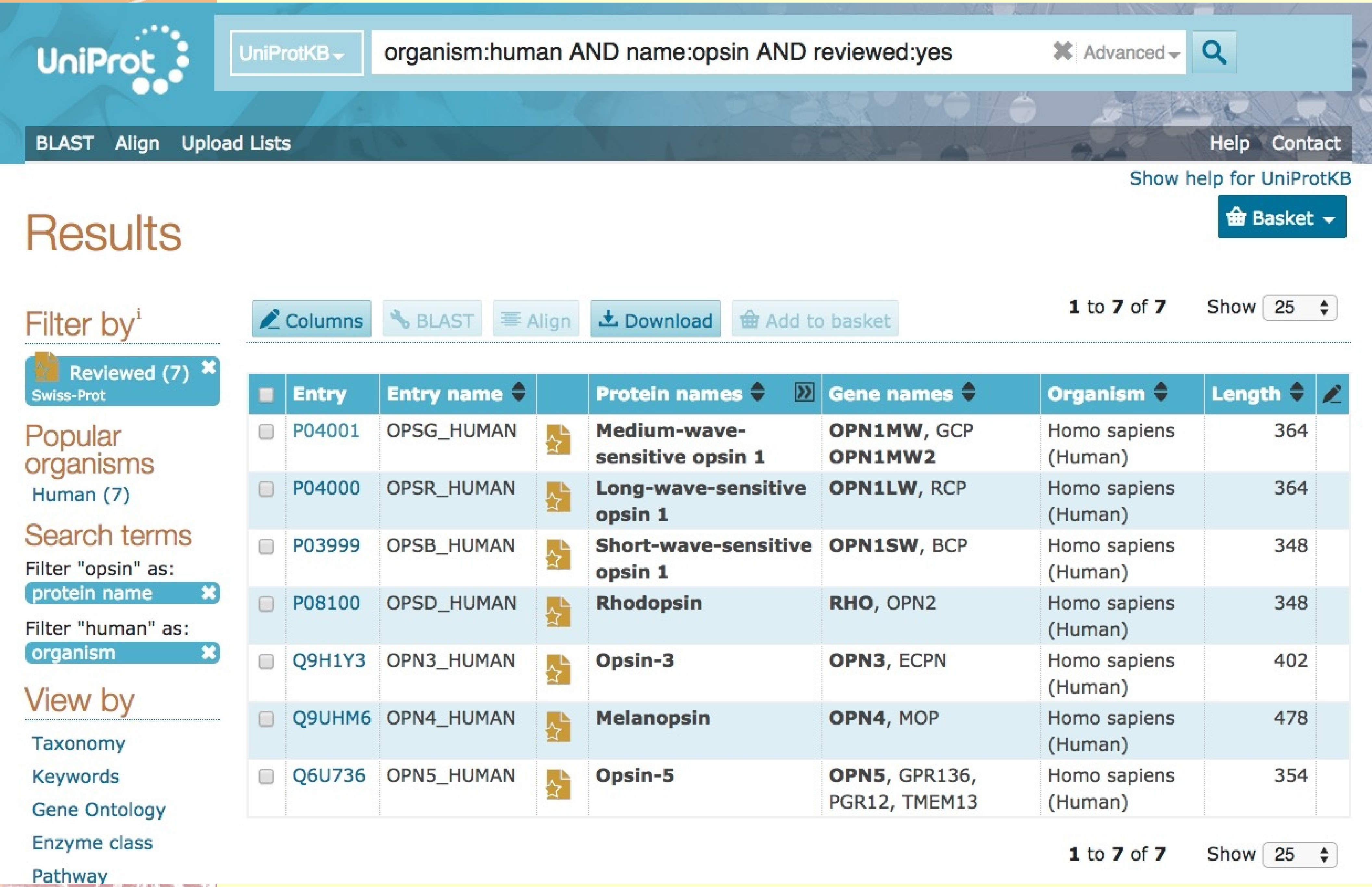
UniRef

1 to 16 of 16 Show 25

☐	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
☐	Q0PJU0	Q0PJU0_HUMAN	▯	**Opsin 1 (Cone pigments), short-wave...**	**OPN1SW**, hCG_41660	Homo sapiens (Human)	348
☐	B1B1F0	B1B1F0_HUMAN	▯	**Long-wave-sensitive opsin 1**	**OPN1LW**, Z68193.1-002	Homo sapiens (Human)	164
☐	J3KPQ2	J3KPQ2_HUMAN	▯	**Opsin 5**	**OPN5**, hCG_1642475	Homo sapiens (Human)	353
☐	Q6GMT1	Q6GMT1_HUMAN	▯	**Opsin 3**	**OPN3**,	Homo sapiens	142

UniProt Human Opsin Entries Reviewed

http://www.uniprot.org/



organism:human AND name:opsin AND reviewed:yes

Results

Filter by

Reviewed (7) Swiss-Prot

Popular organisms

Human (7)

Search terms

Filter "opsin" as: protein name

Filter "human" as: organism

View by

1 to 7 of 7 Show 25

☐	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 Show 25

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

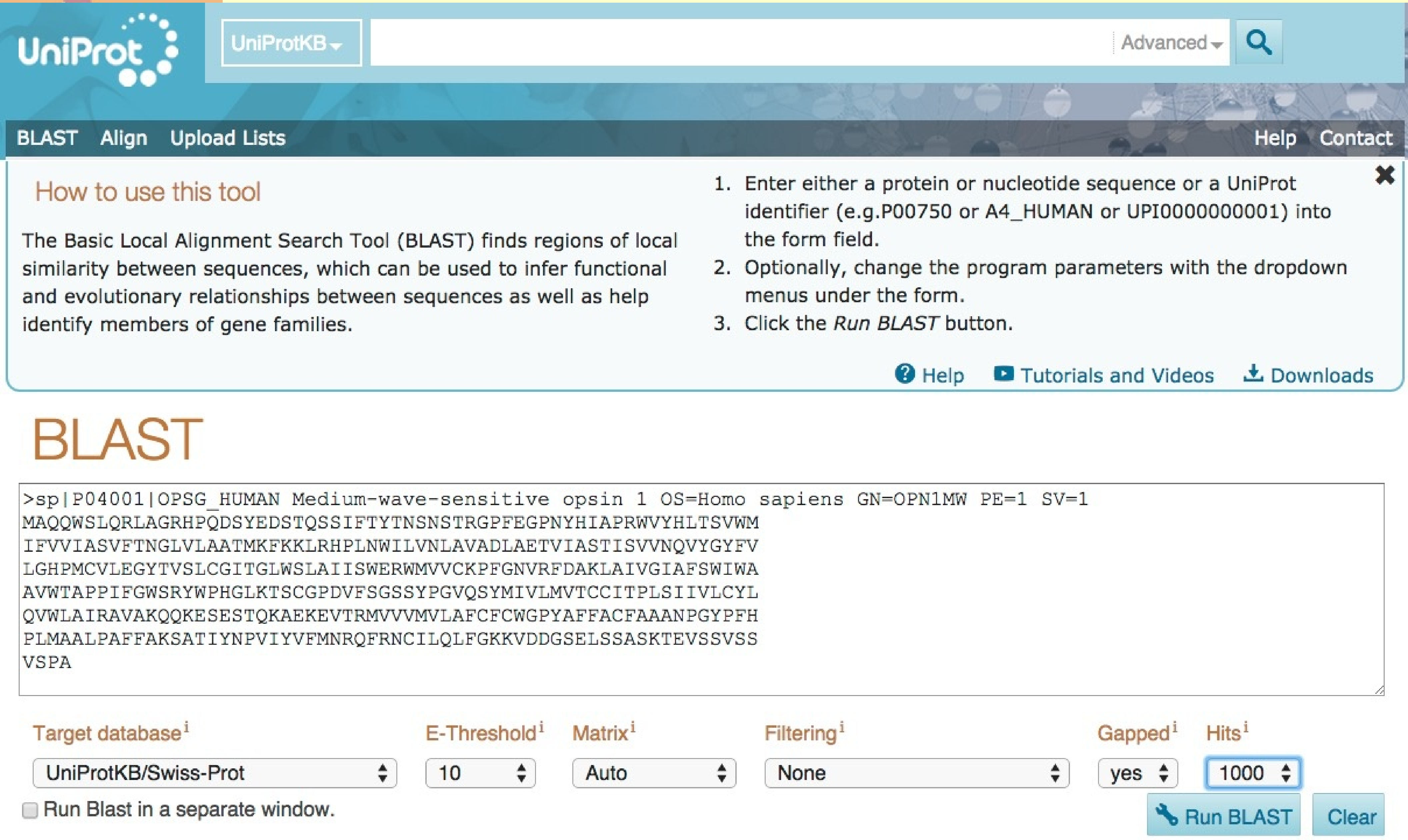
Sensory transduction, Vision

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
QVWLAIRAVAKQQKESESTQKAEKEVTRMVVVMLAFCFCWGPYAFFACFAAANPGYPFH
PLMAALPAFFAKSATIYNPVIYVFMNRQFRNCILQLFGKKVDDGSELSSASKTEVSSVSS
VSPA
```

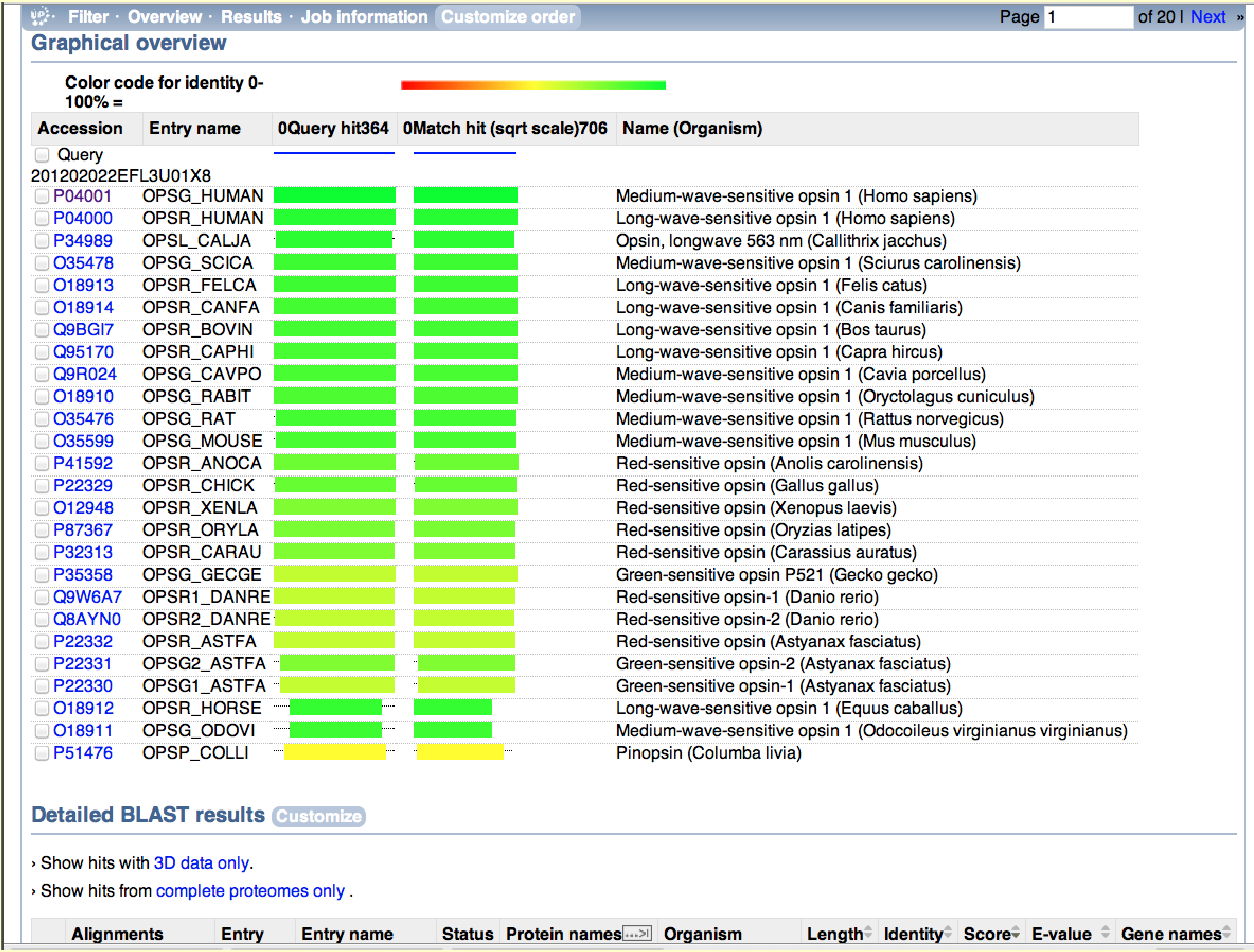
Target database	E-Threshold	Matrix	Filtering	Gapped	Hits
UniProtKB/Swiss-Prot	10	Auto	None	yes	1000

☐ Run Blast in a separate window.

© Doug Brutlag 2015

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)

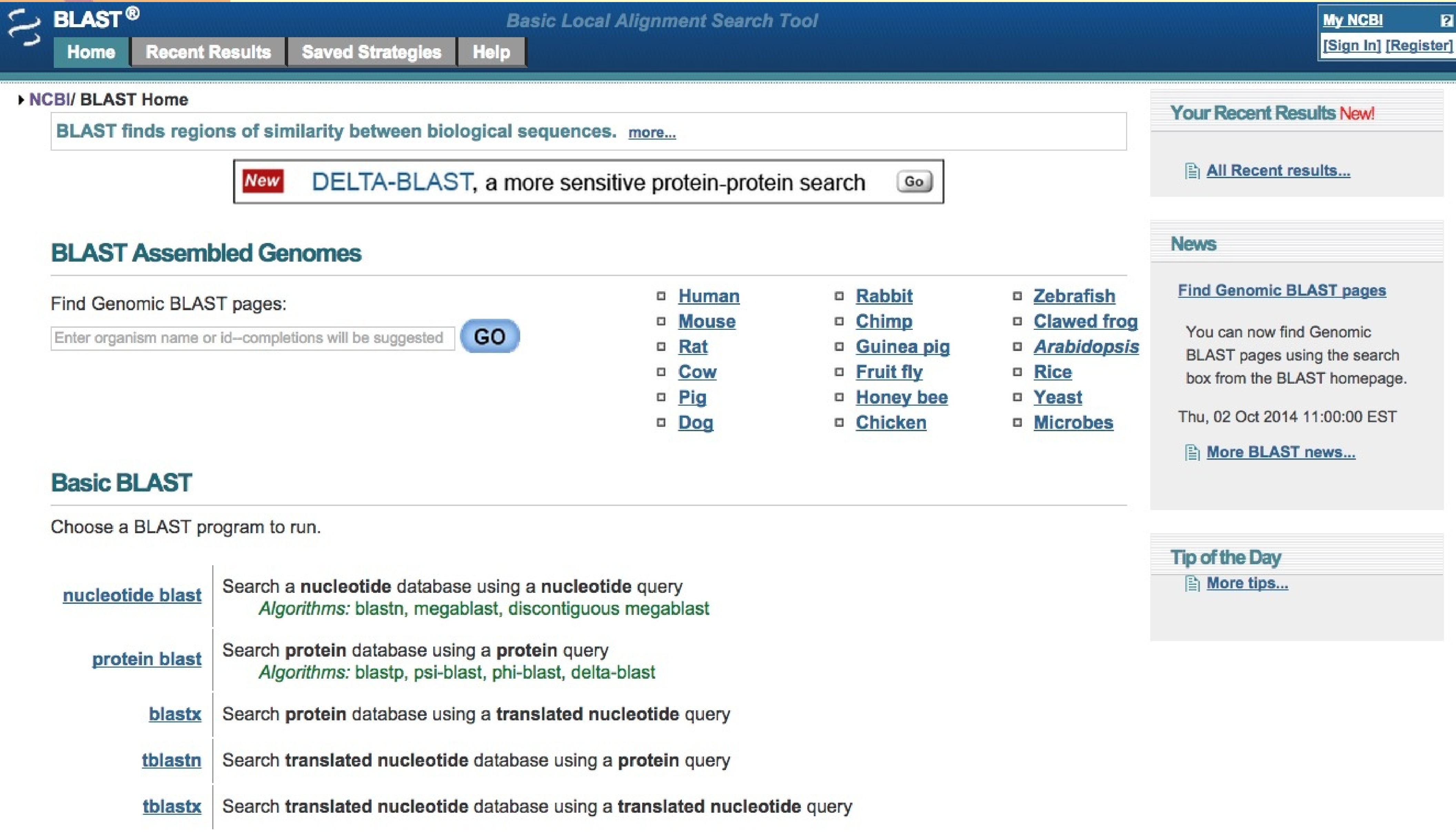
Detailed BLAST results

- › 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/



NCBI/ BLAST Home

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

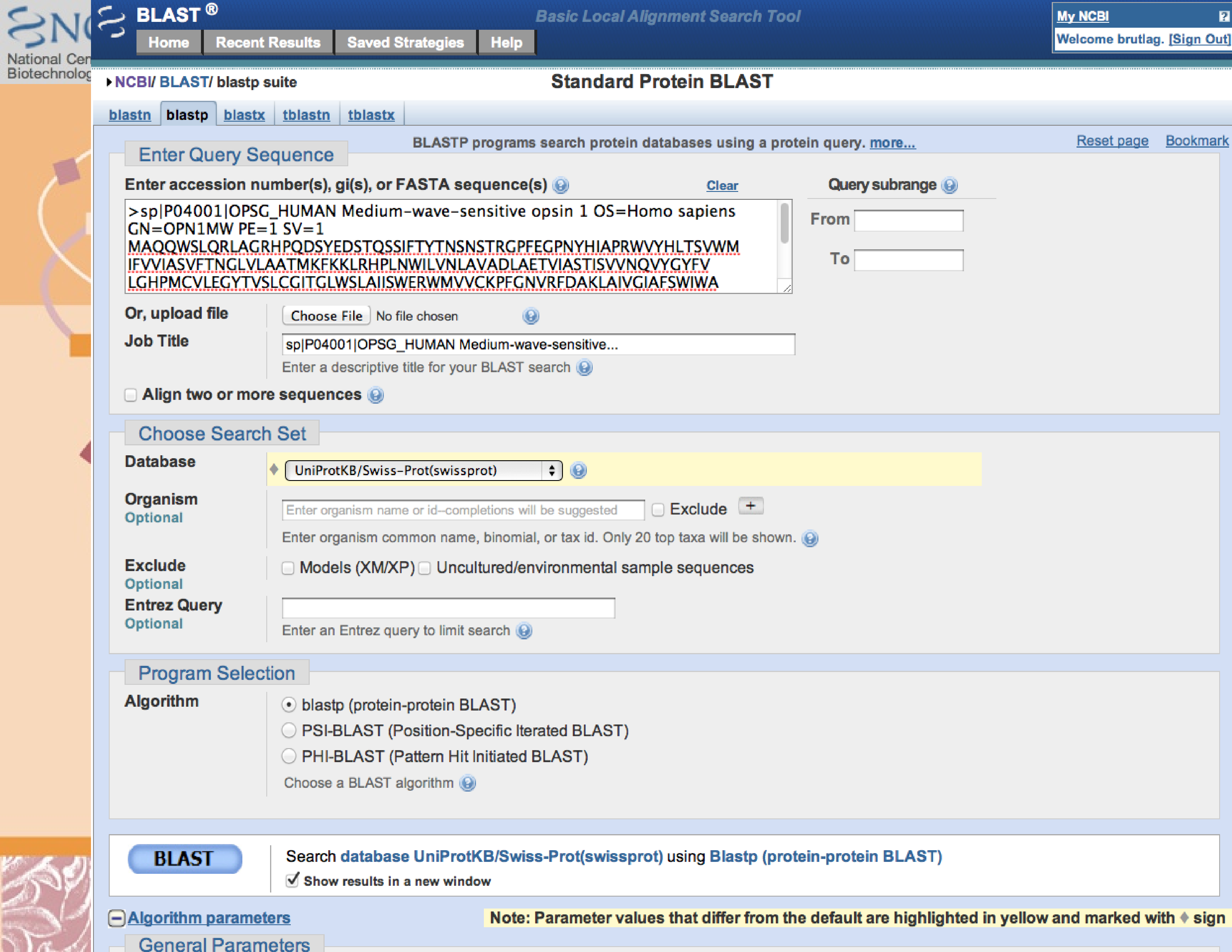
You can now find Genomic BLAST pages using the search box from the BLAST homepage.

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) Clear

```
>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: Choose 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

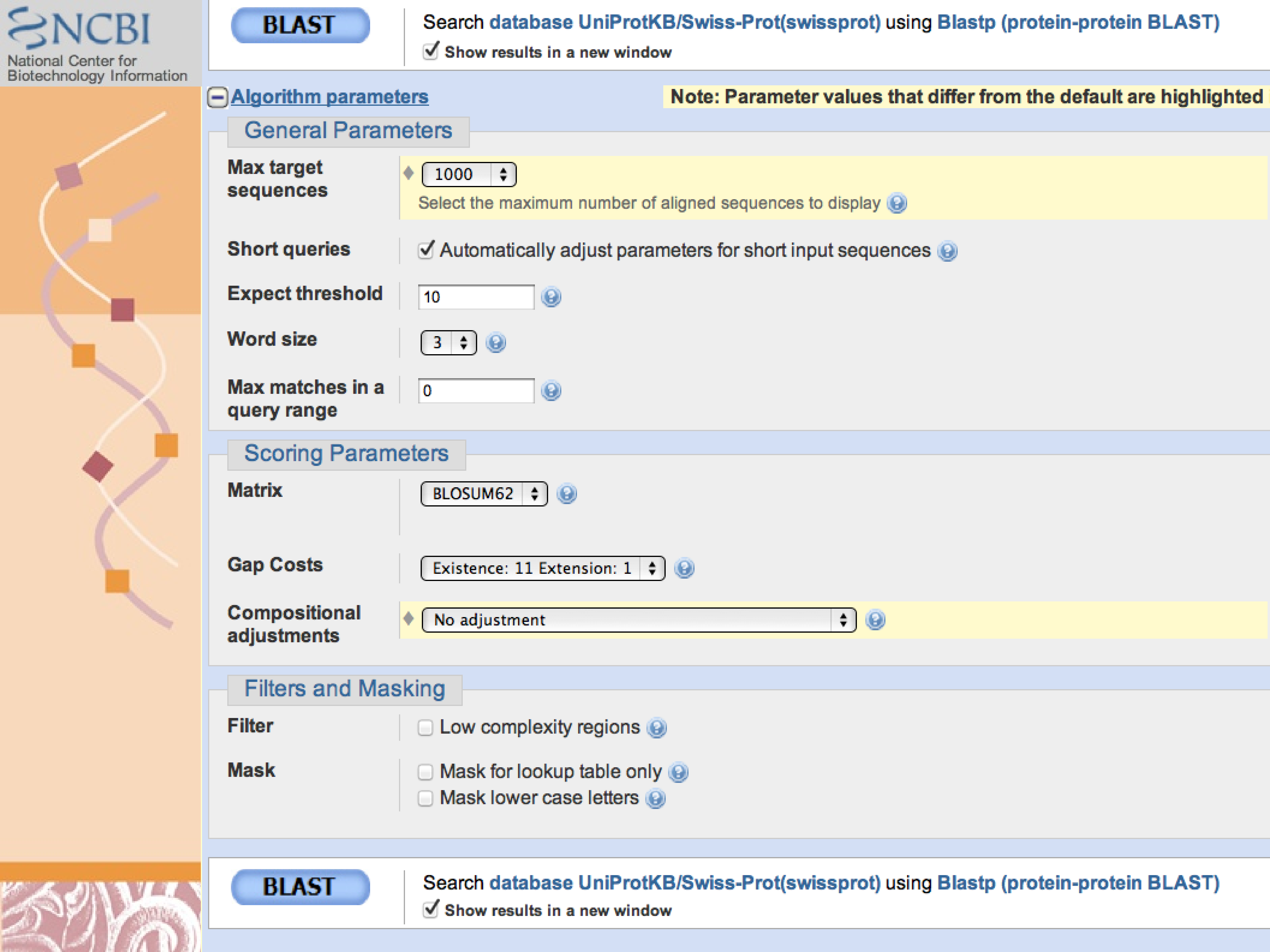
Algorithm
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
Choose a BLAST algorithm

BLAST 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 in yellow and marked with ♦ sign

General Parameters



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

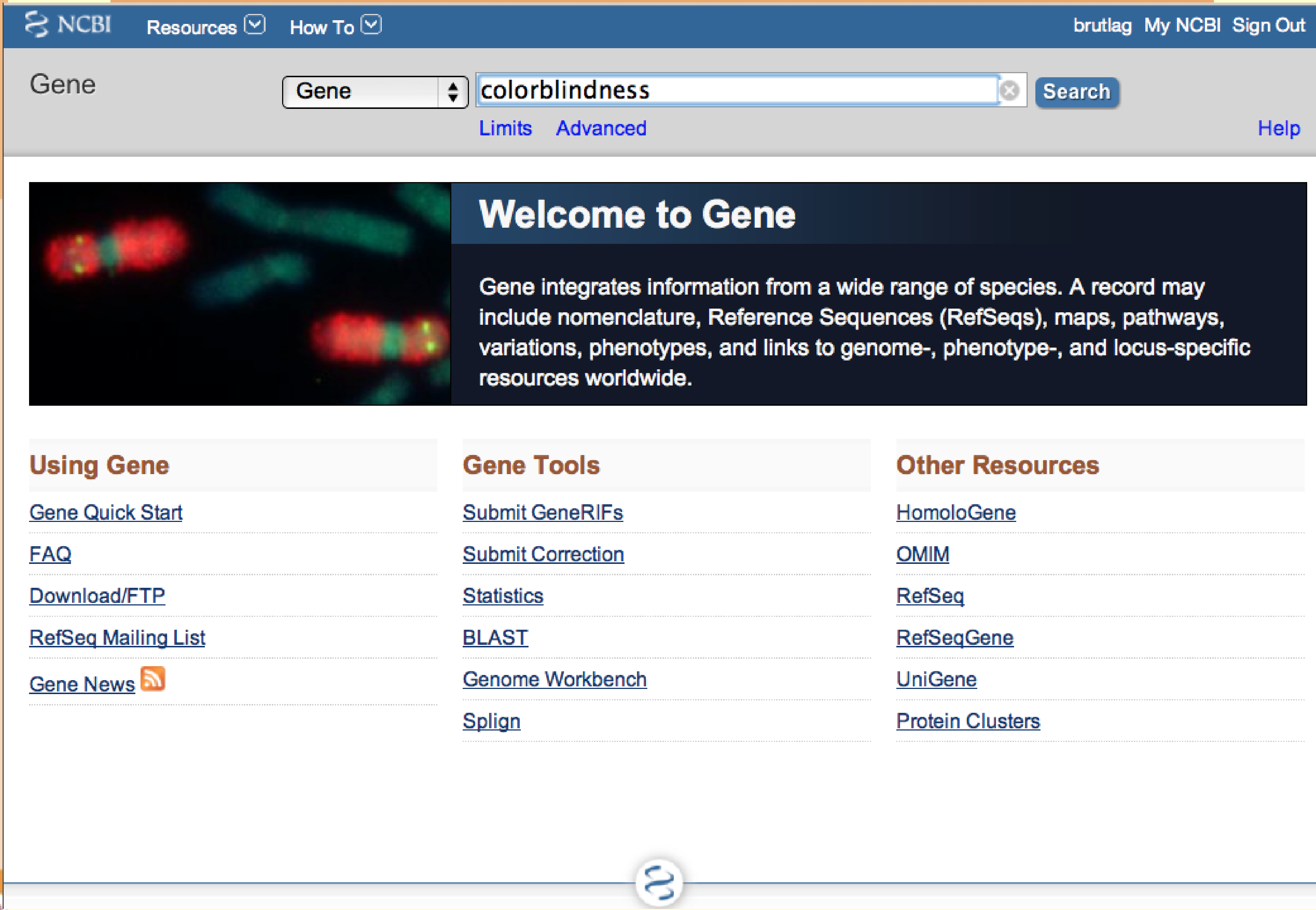
Filter: ☐ Low complexity regions

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

Entrez Gene search for Colorblindness



Welcome to Gene

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.

Using Gene

- Gene Quick Start
- FAQ
- Download/FTP
- RefSeq Mailing List
- Gene News

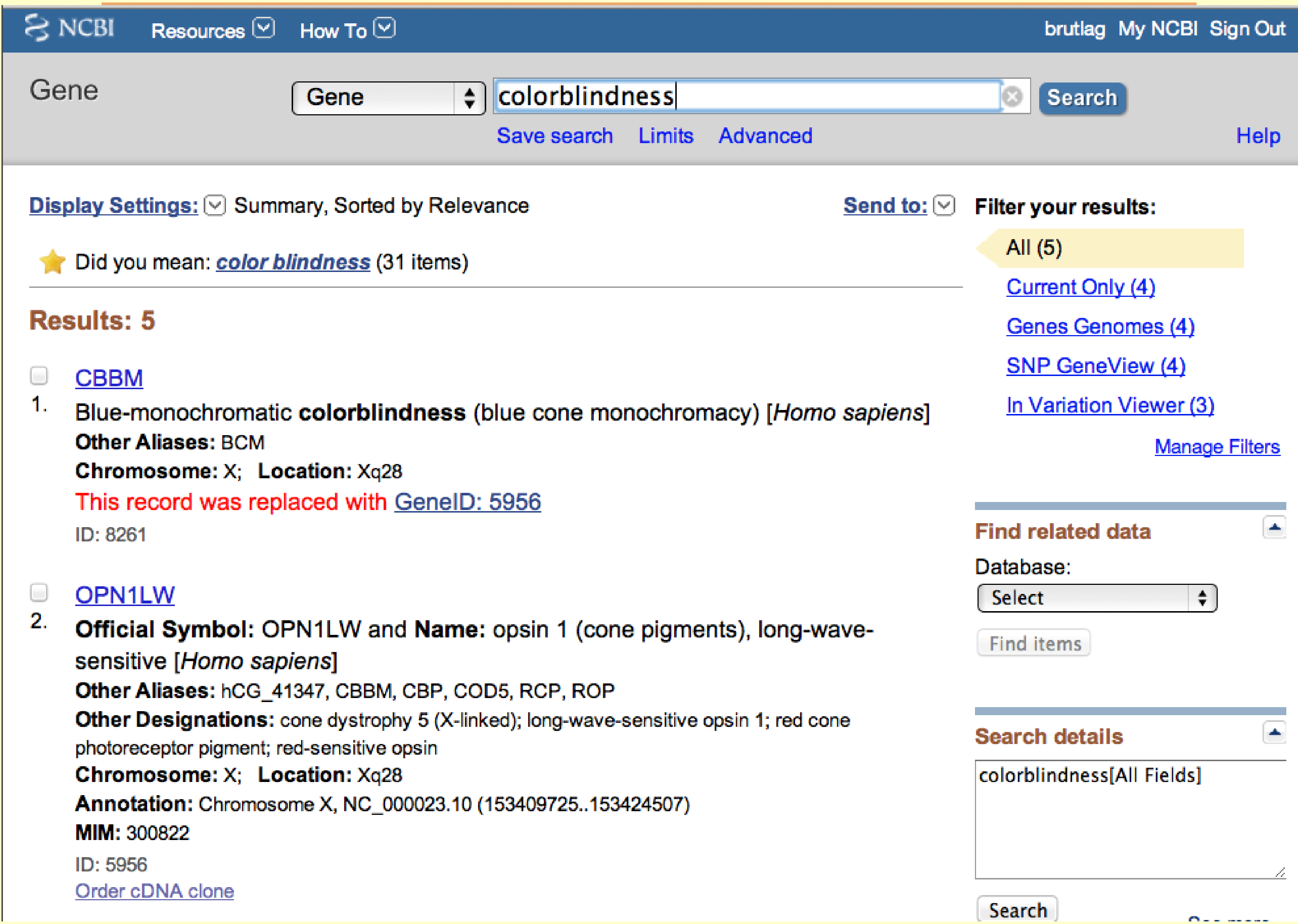
Gene Tools

- Submit GeneRIFs
- Submit Correction
- Statistics
- BLAST
- Genome Workbench
- Splign

Other Resources

- HomoloGene
- OMIM
- RefSeq
- RefSeqGene
- UniGene
- Protein Clusters

Entrez Gene search for Colorblindness



Display Settings: Summary, Sorted by Relevance

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

Results: 5

☐ CBBM

1. Blue-monochromatic **colorblindness** (blue cone monochromacy) [*Homo sapiens*]
Other Aliases: BCM
Chromosome: X; **Location:** Xq28
This record was replaced with GeneID: 5956
ID: 8261

☐ OPN1LW

2. **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

Search details

colorblindness[All Fields]

Entrez Gene search for Colorblindness

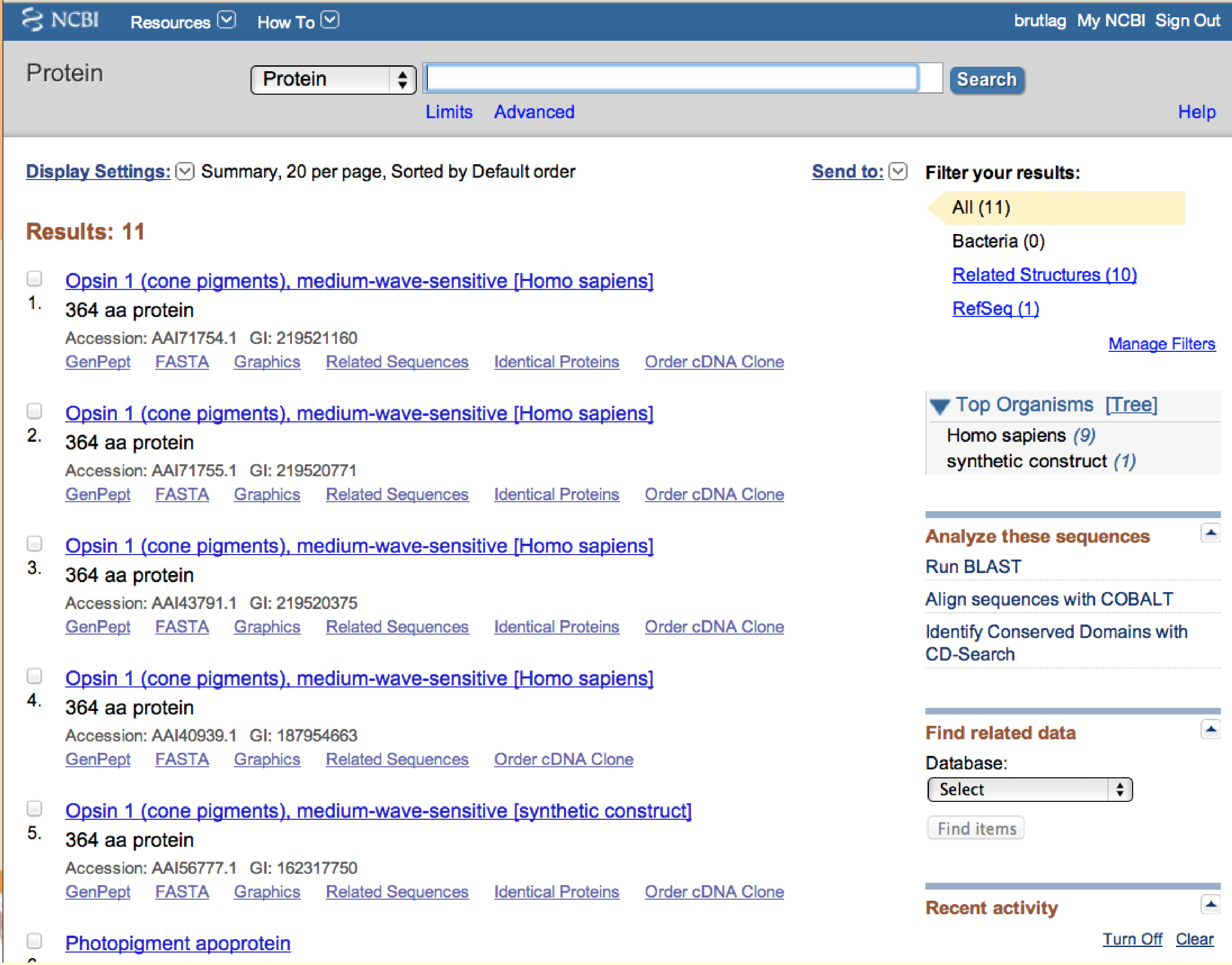
OPN1MW opsin 1 (cone pigments), medium-wave-sensitive [*Homo sapiens*]

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

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)

Homo sapiens *(9)*
synthetic construct *(1)*

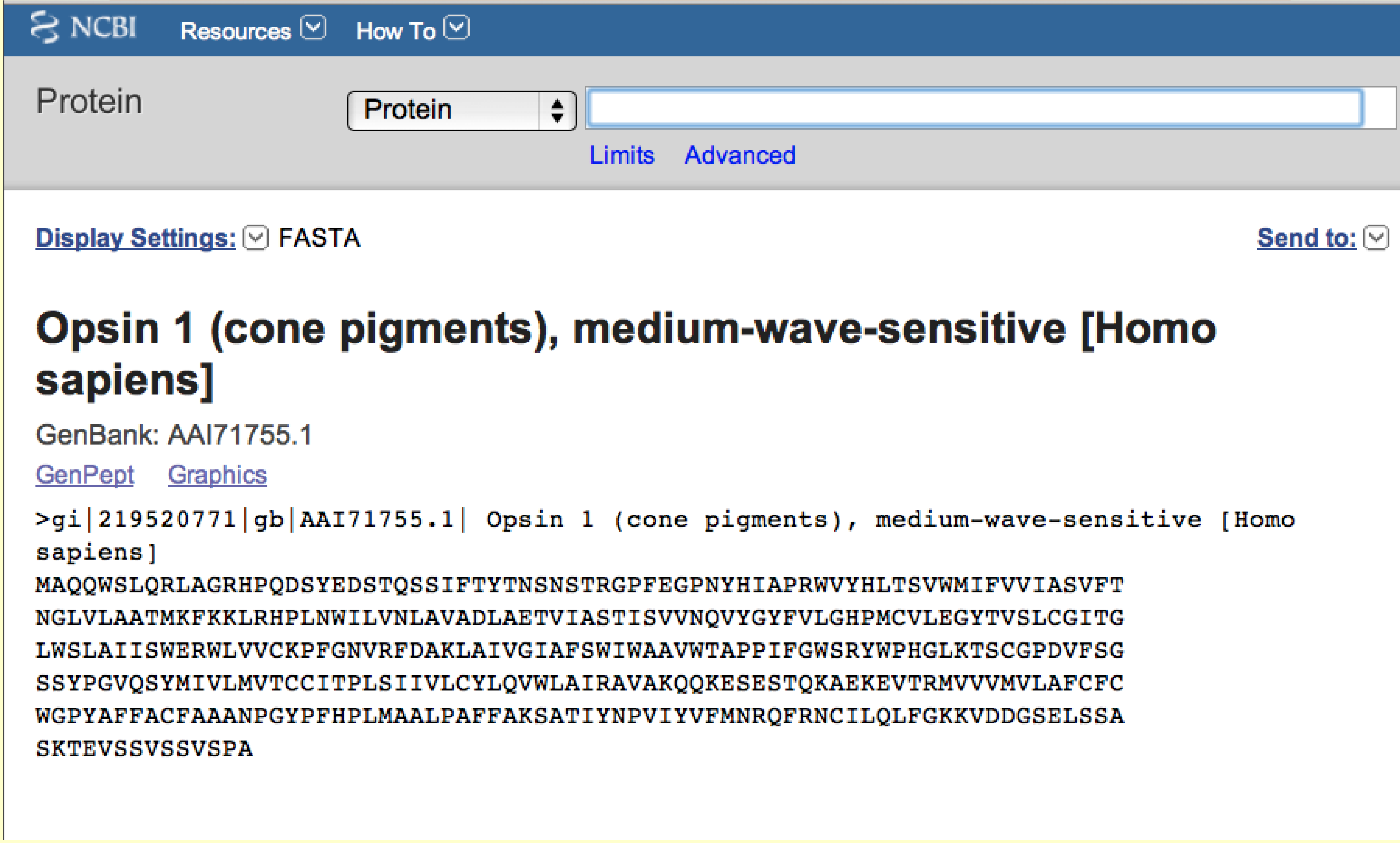
Analyze these sequences

Find related data

Database:

Recent activity

Entrez Gene search for Opsins



Display Settings: FASTA

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

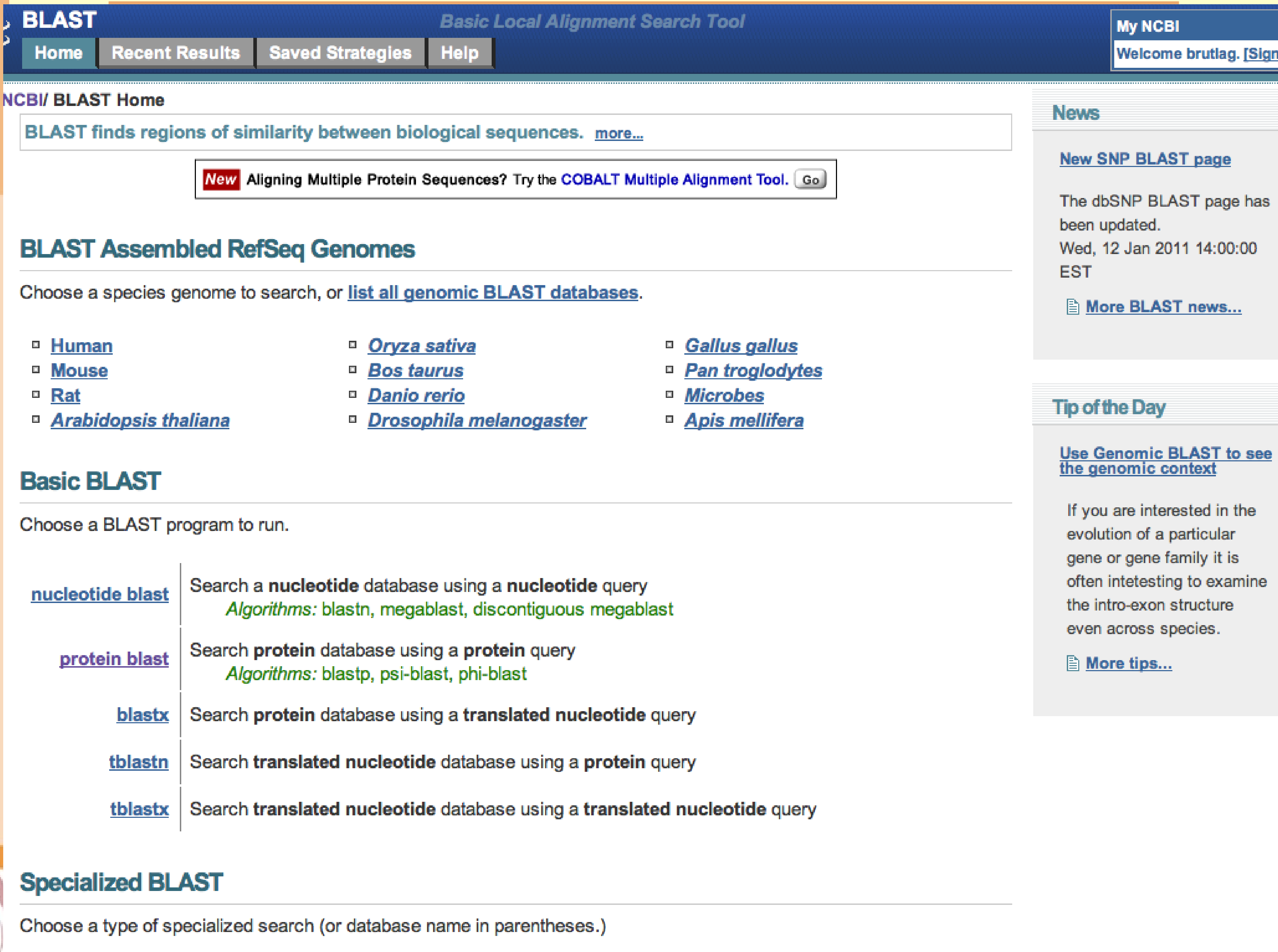
GenBank: AAI71755.1

GenPept Graphics

```
>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.

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...

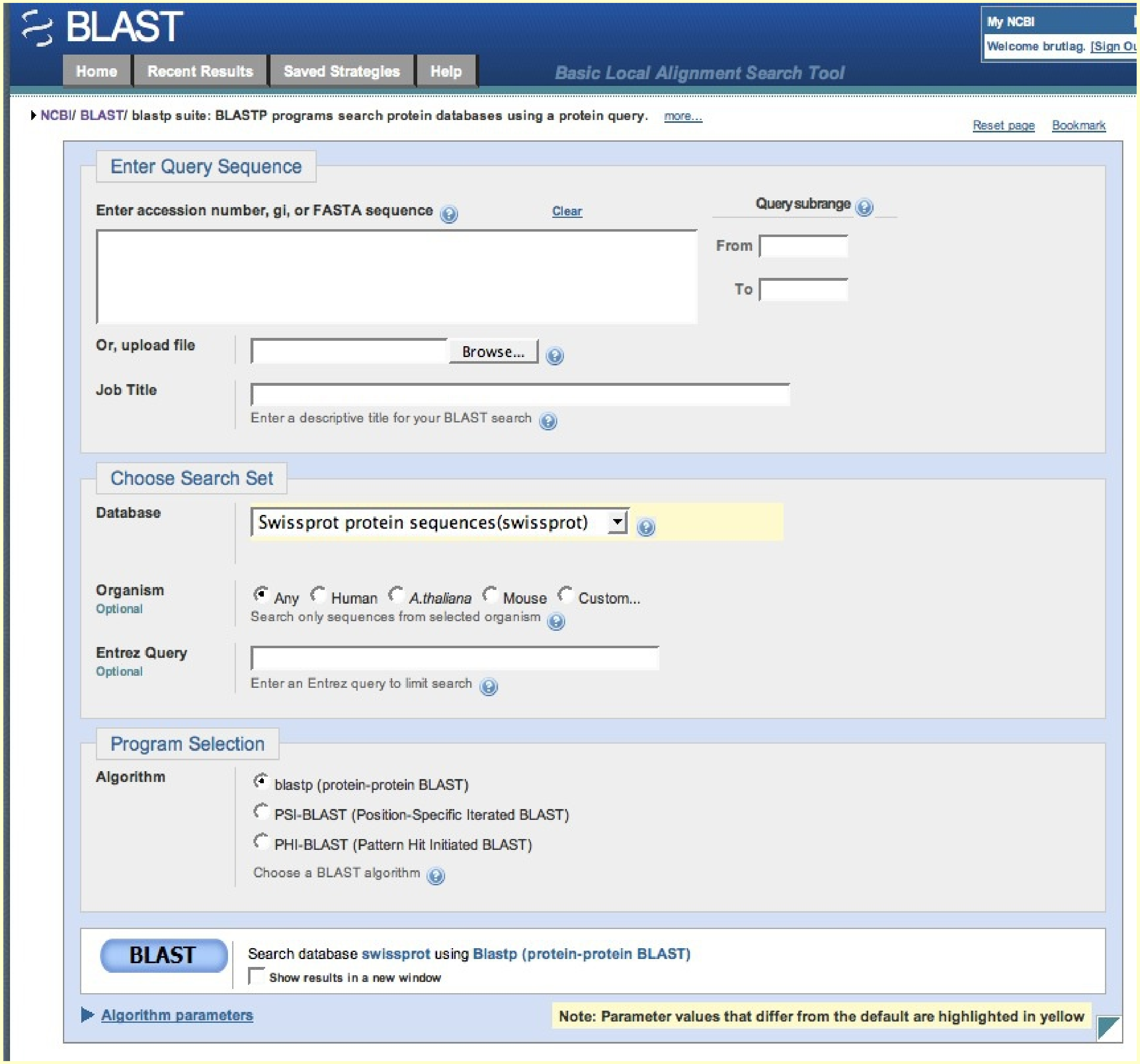
Tip of the Day

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)

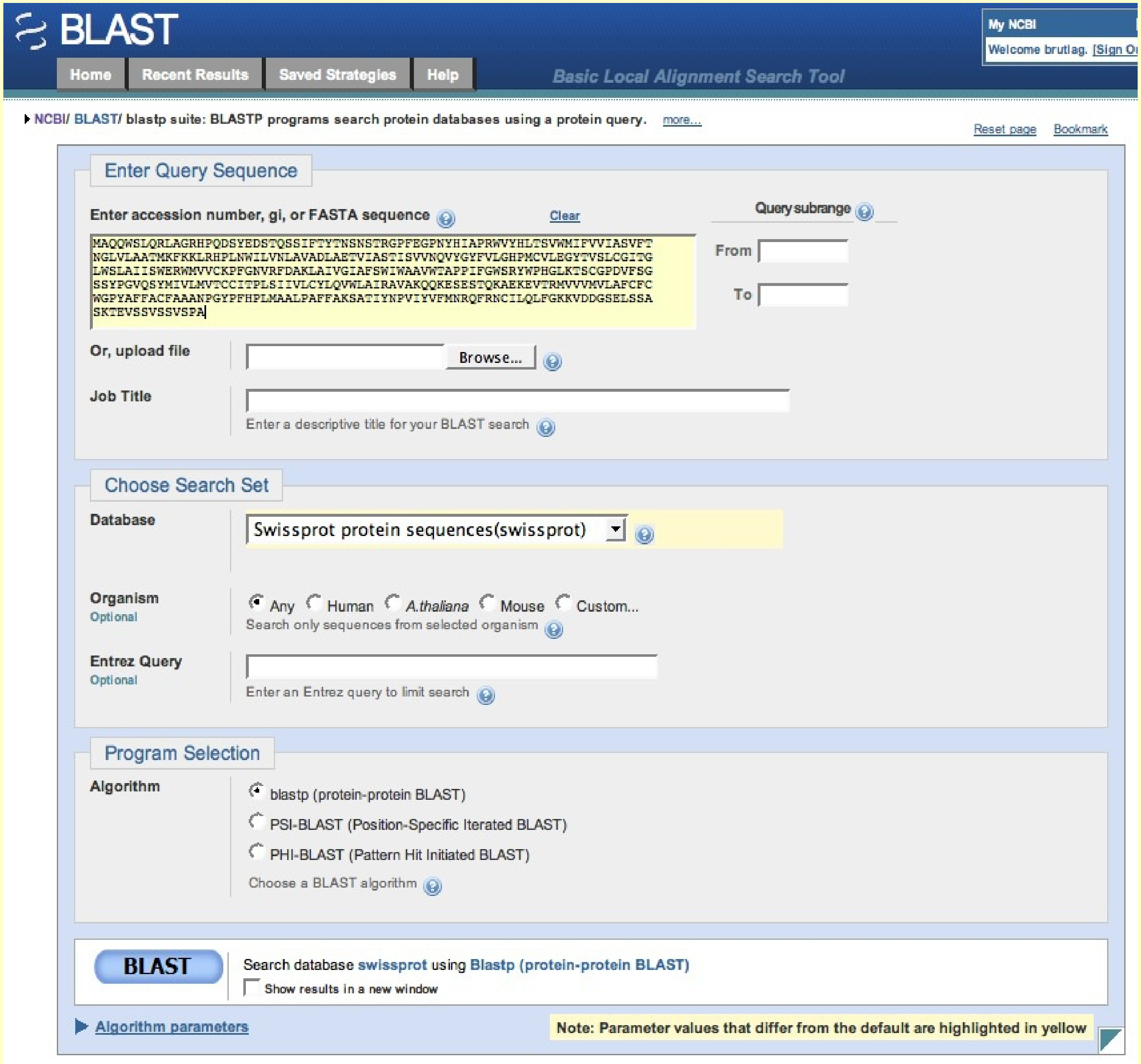
Choose a BLAST algorithm

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

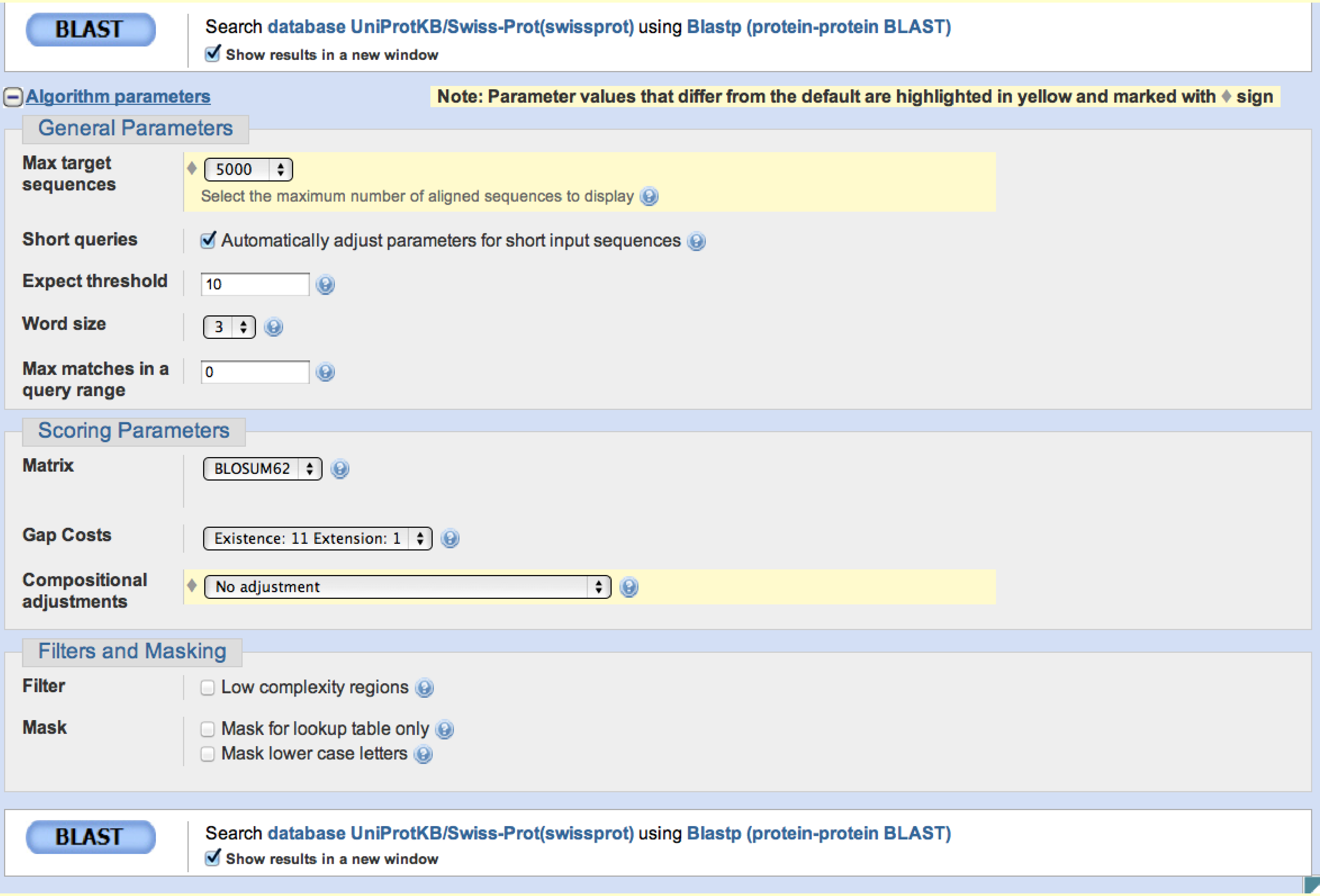
Program Selection

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

⊟ Algorithm parameters

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

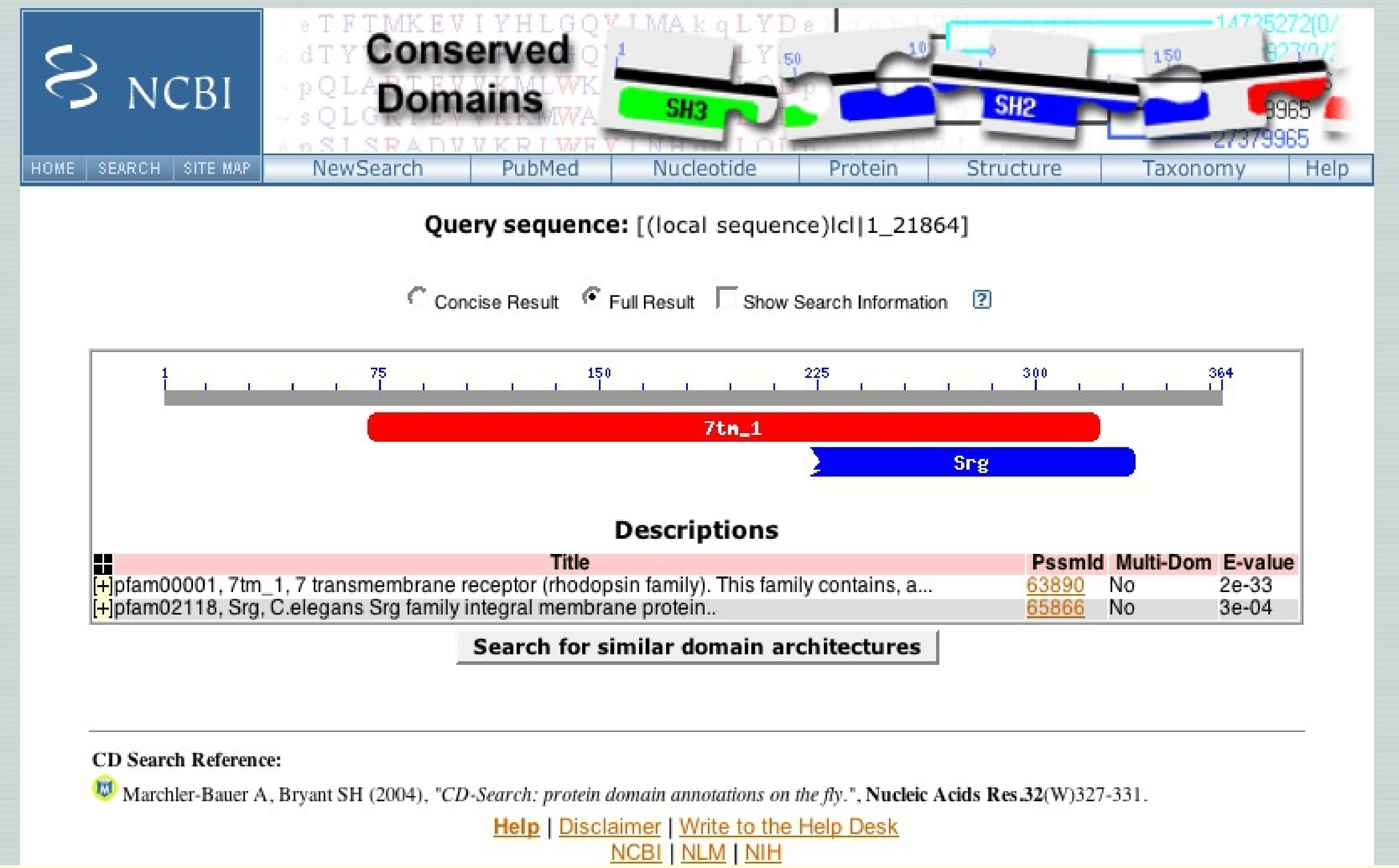
Mask ☐ Mask for lookup table only

☐ 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

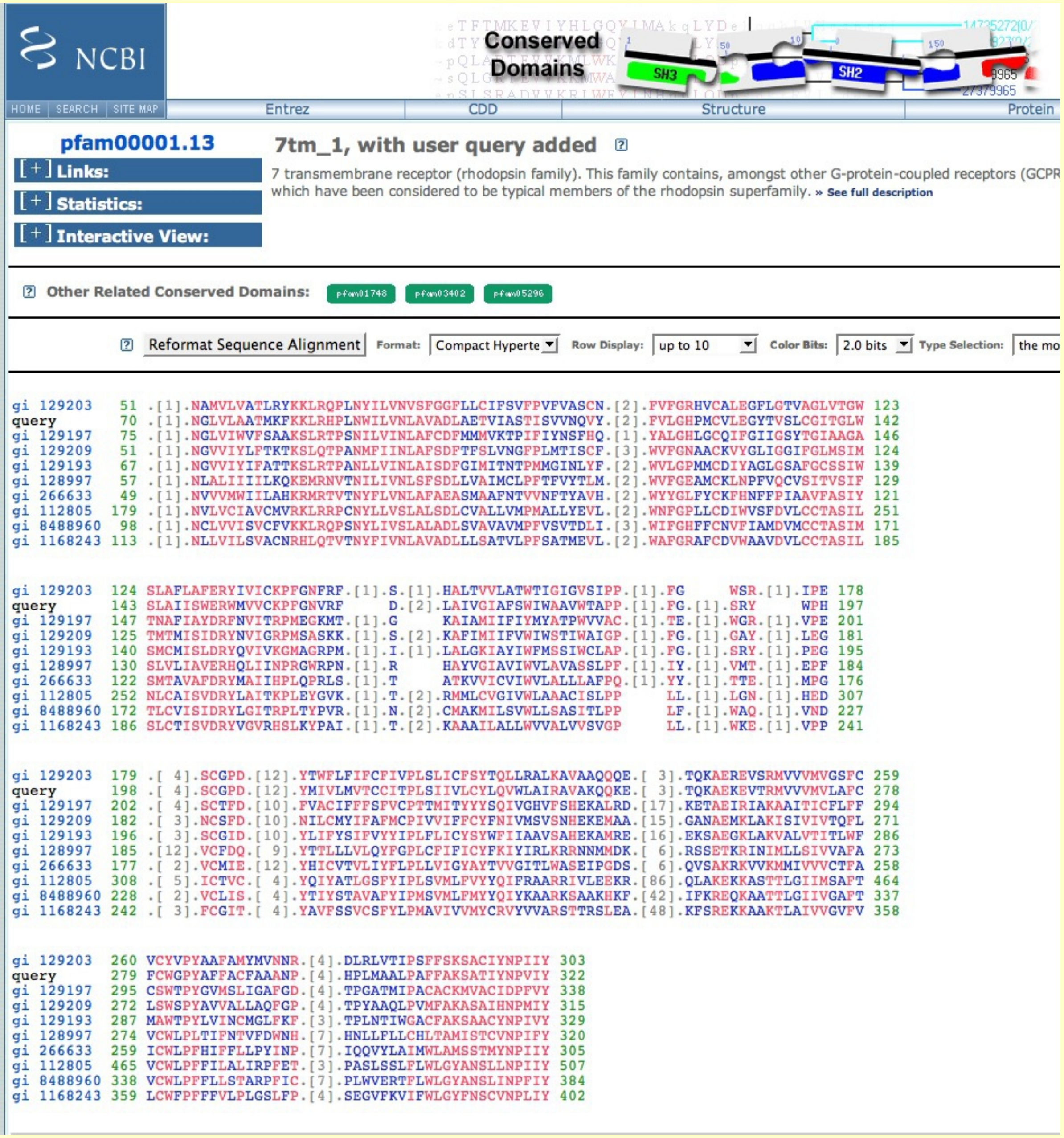
Title	PssmId	Multi-Dom	E-value
[+]pfam00001, 7tm_1, 7 transmembrane receptor (rhodopsin family). This family contains, a...	63890	No	2e-33
[+]pfam02118, Srg, C.elegans Srg family integral membrane protein..	65866	No	3e-04

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

[+] Links:

[+] Statistics:

[+] Interactive View:

7tm_1, with user query added

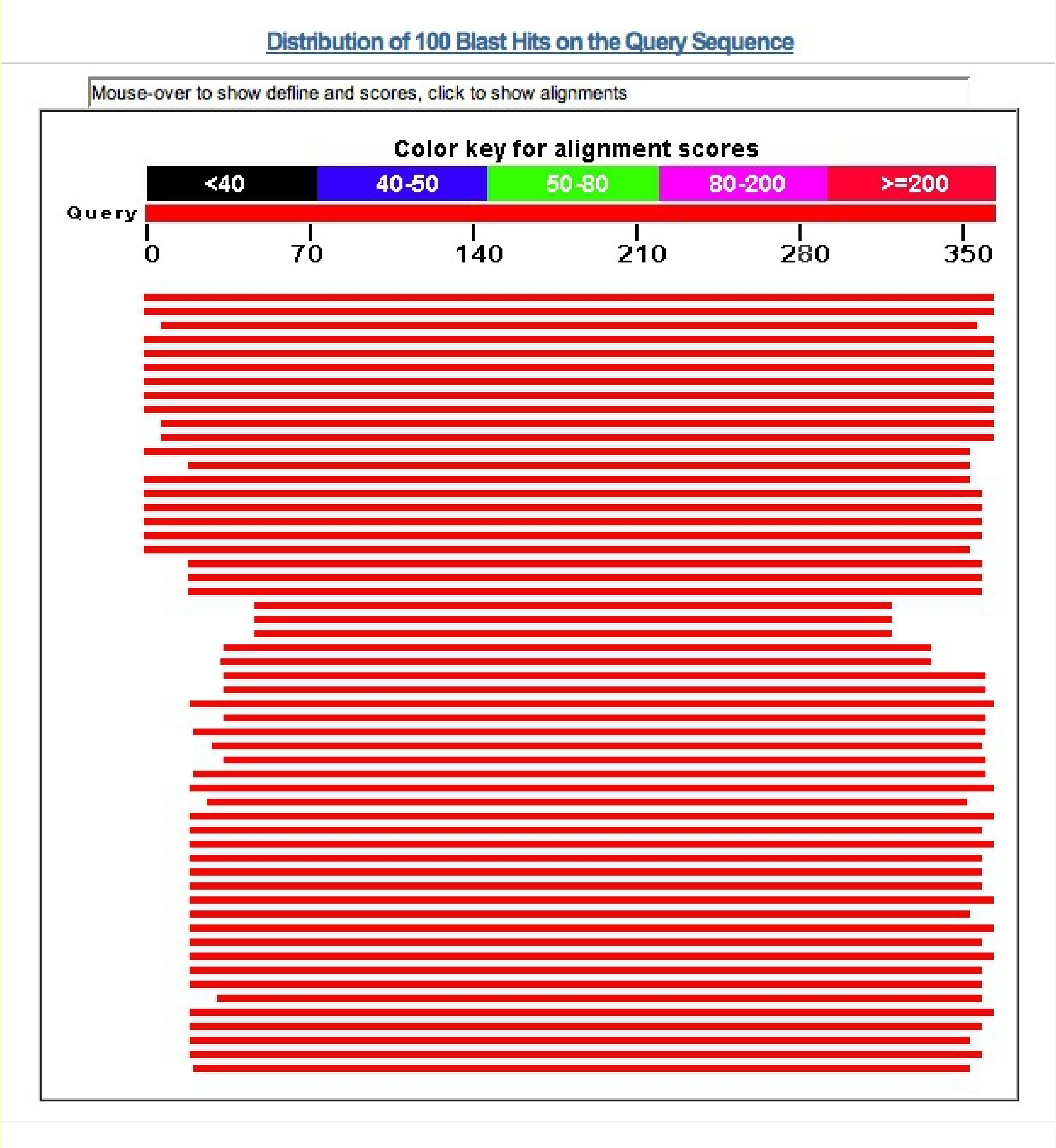
7 transmembrane receptor (rhodopsin family). This family contains, amongst other G-protein-coupled receptors (GCPR which have been considered to be typical members of the rhodopsin superfamily. » **See full description**

Other Related Conserved Domains: pfam01748 pfam03402 pfam05296

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

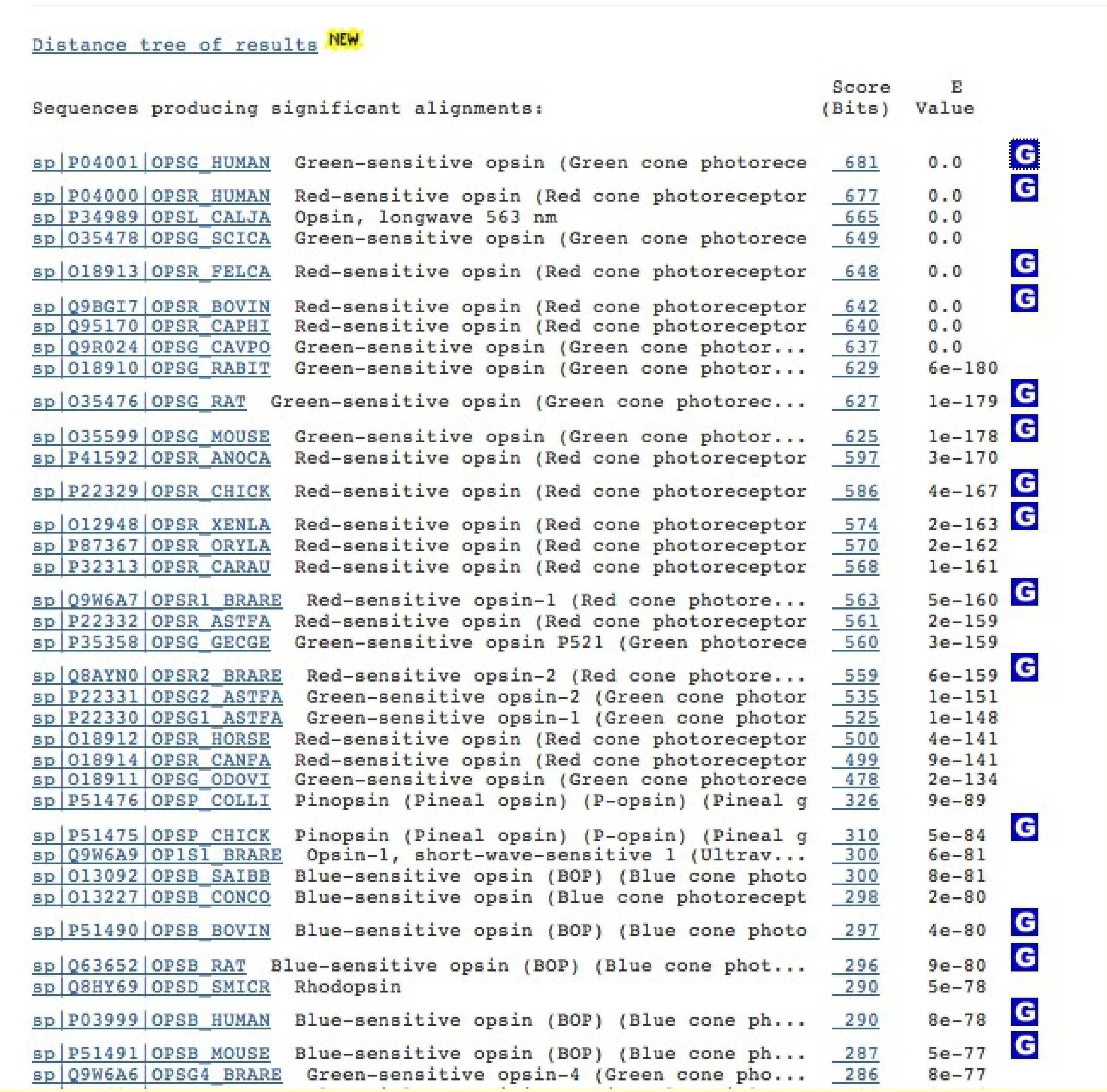
```
gi 129203   51 .[1].NAMVLVATLRYKKLRQPLNYILVNVSFGGFLLCIFSVFPVFVASCN.[2].FVFGRHVCALEGFLGTVAGLVTGW 123
query       70 .[1].NGLVLAATMKFKKLRHPLNWILVNLAVADLAETVIASTISVVNQVY.[2].FVLGHPMCVLEGYTVSLCGITGLW 142
gi 129197   75 .[1].NGLVIWVFSAAKSLRTPSNILVINLAFCDFMMMVKTPIFIYNSFHQ.[1].YALGHLGCQIFGIIGSYTGIAAGA 146
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

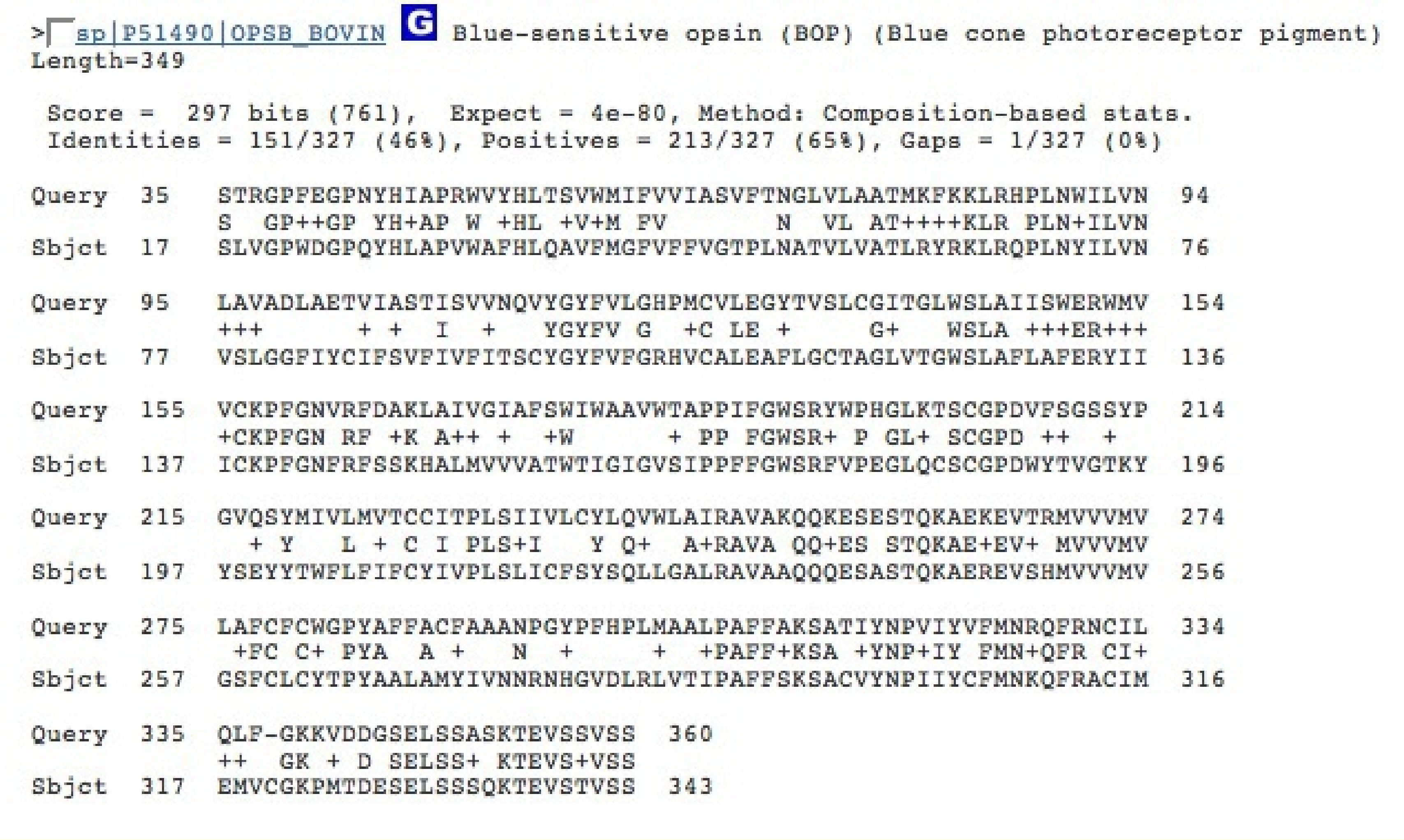
Mouse-over to show defline and scores, click to show alignments

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

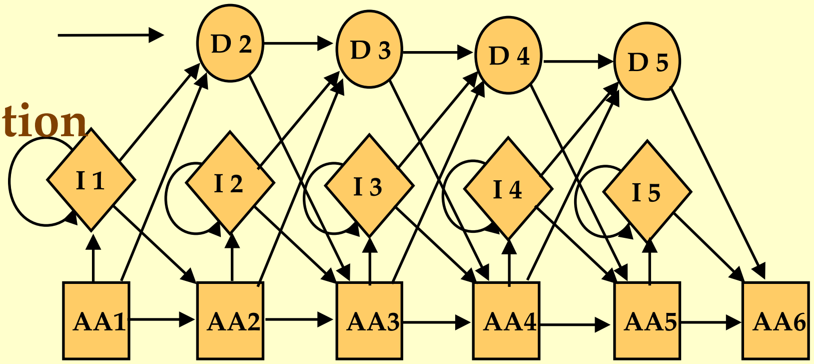
Consensus Sequences or Sequence Motifs

Zinc Finger (C2H2 type)

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

Sequences of Common Structure or Function

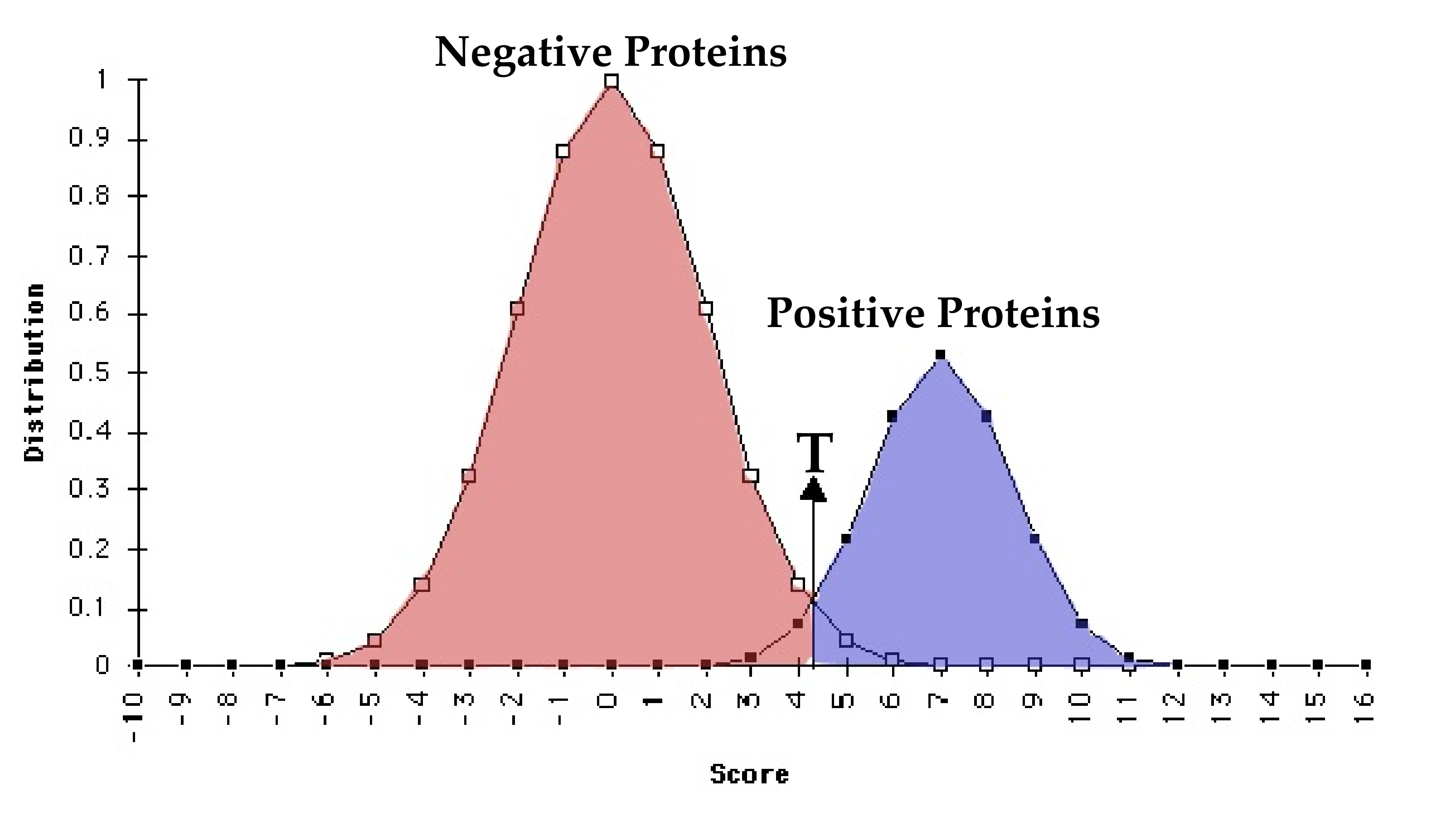
Profiles, PSI-BLAST Hidden Markov Models



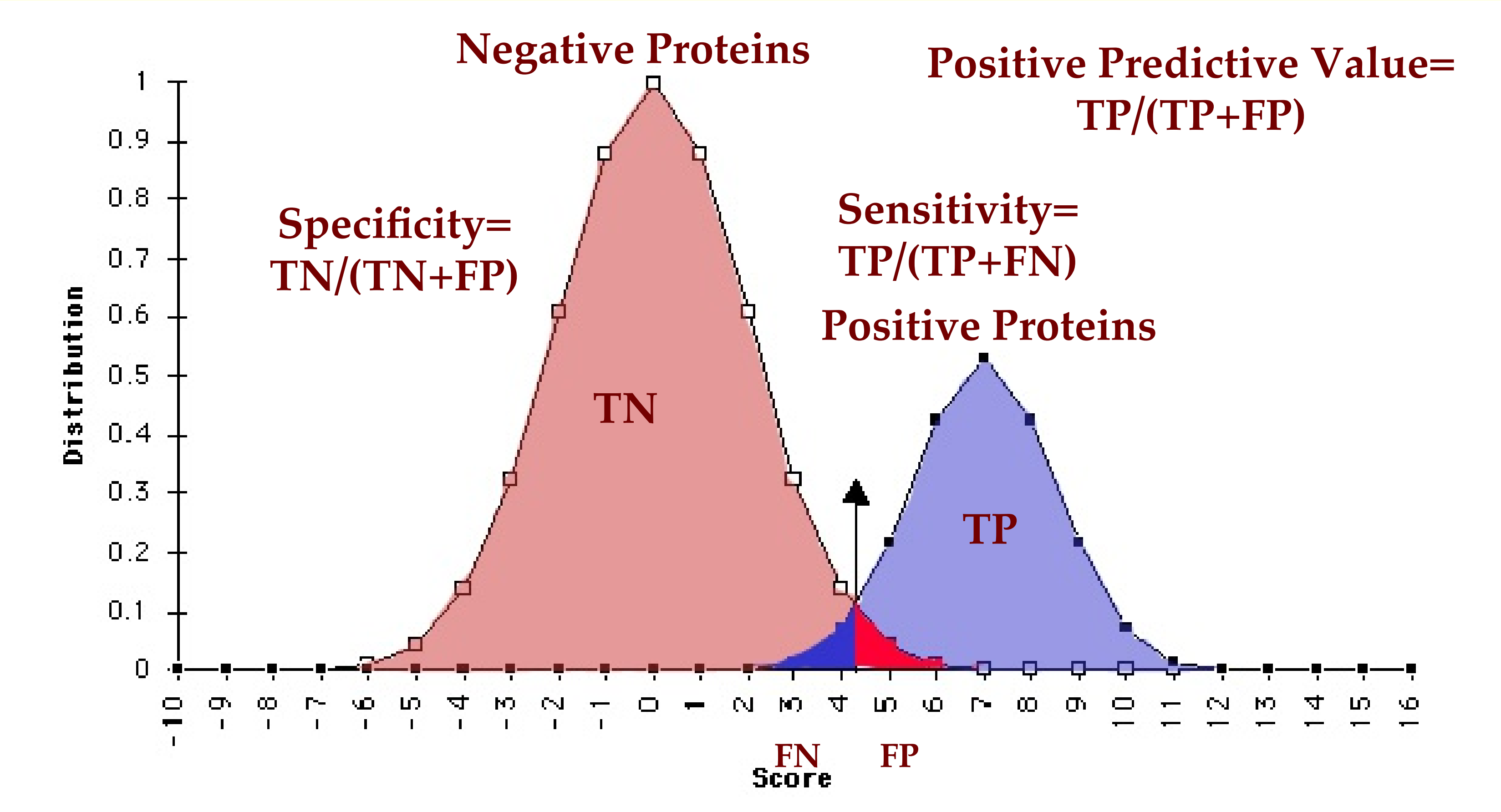
Sequence Similarity

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

Evaluation of PSSMs, Profiles and HMMs



Evaluation of Profiles



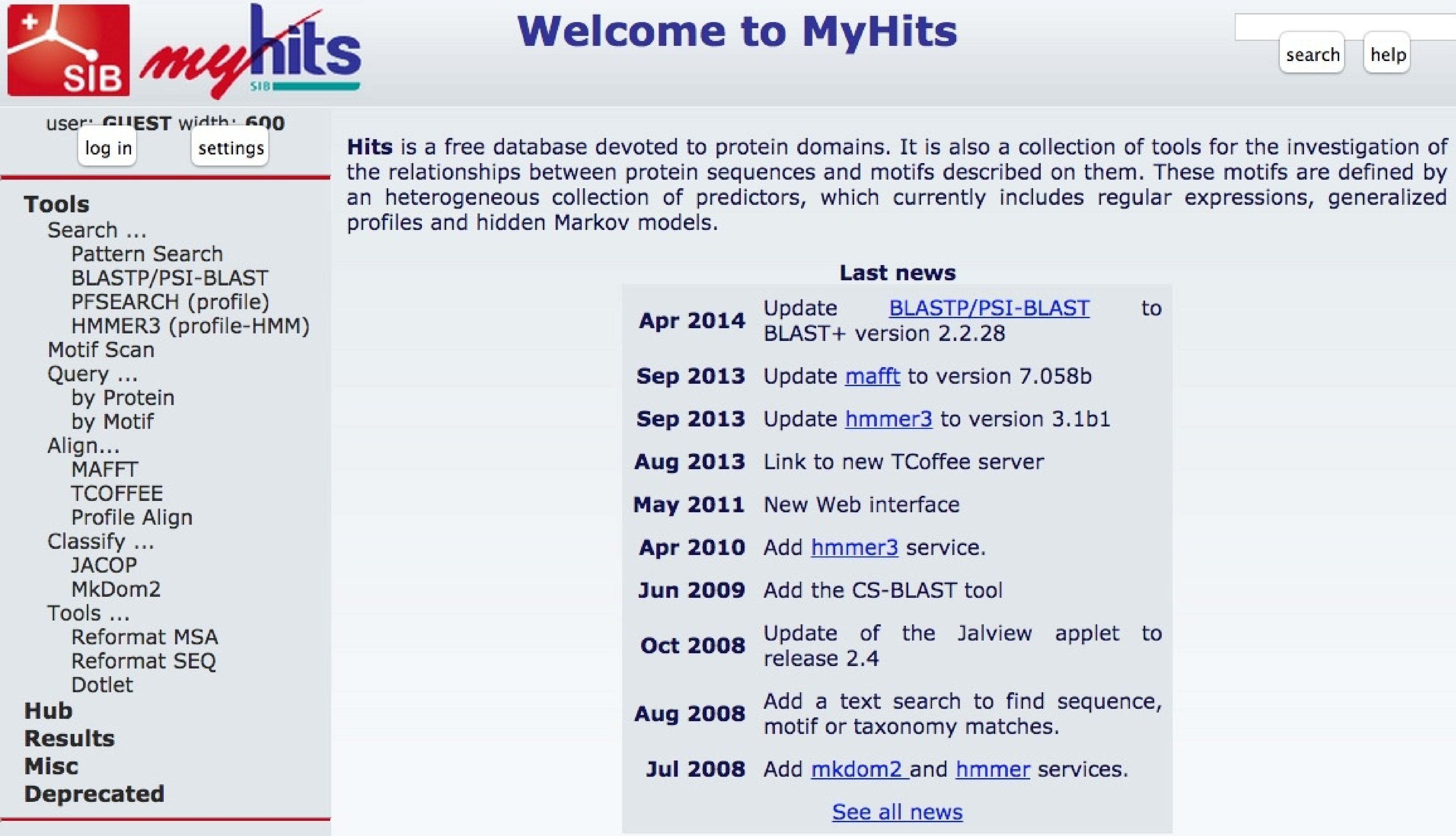
Specificity=
TN/(TN+FP)

Sensitivity=
TP/(TP+FN)

Positive Predictive Value=
TP/(TP+FP)

MyHits Local Motifs Search

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



Welcome to MyHits

Tools
- Search ...
  - Pattern Search
  - BLASTP/PSI-BLAST
  - PFSEARCH (profile)
  - HMMER3 (profile-HMM)
- Motif Scan
- Query ...
  - by Protein
  - by Motif
- Align...
  - MAFFT
  - TCOFFEE
  - Profile Align
- Classify ...
  - JACOP
  - MkDom2
- Tools ...
  - Reformat MSA
  - Reformat SEQ
  - Dotlet

Hub

Results

Misc

Deprecated

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.

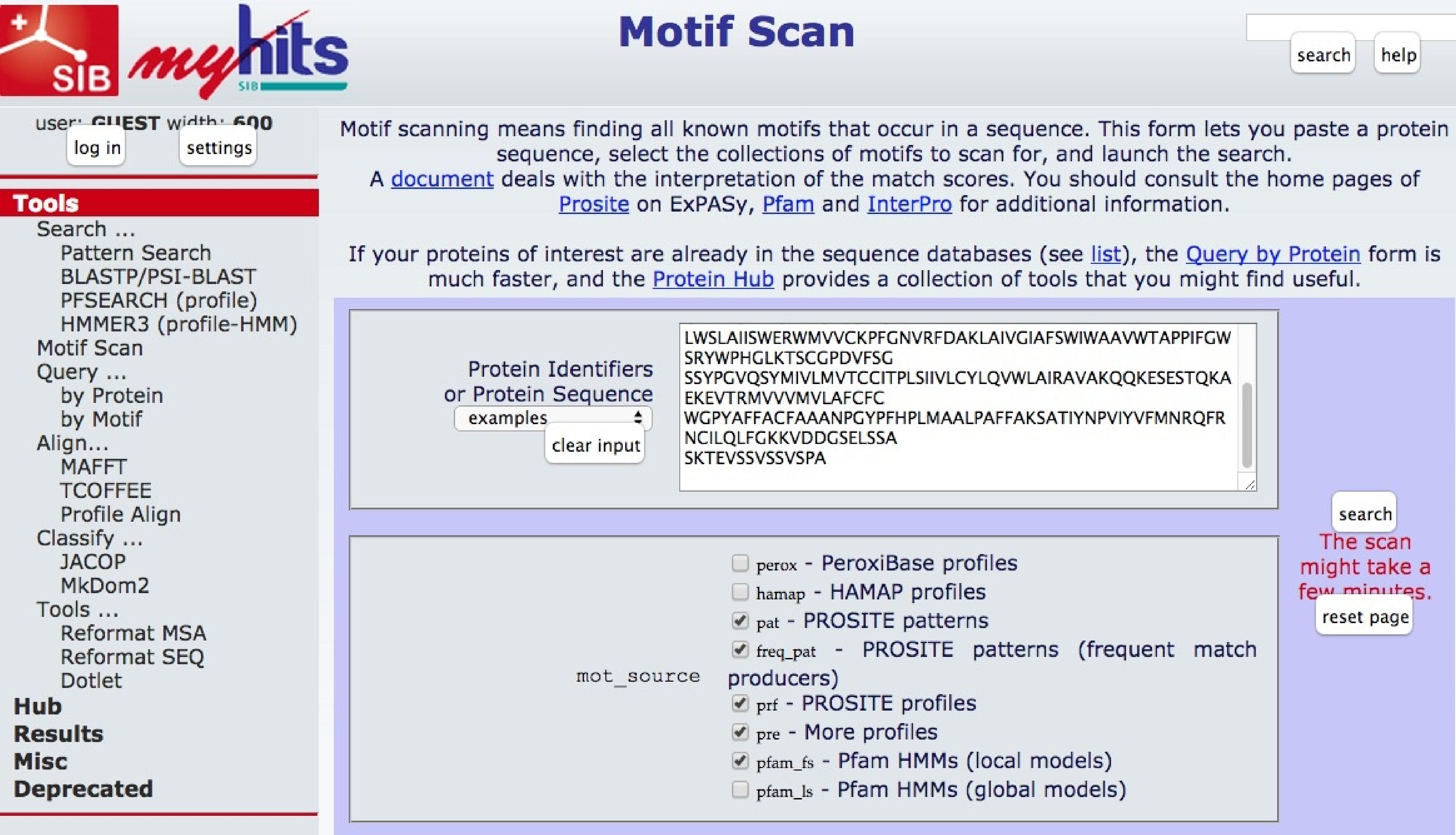
Last news

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

Tools

- Search ...
  - Pattern Search
  - BLASTP/PSI-BLAST
  - PFSEARCH (profile)
  - HMMER3 (profile-HMM)
- Motif Scan
- Query ...
  - by Protein
  - by Motif
- Align...
  - MAFFT
  - TCOFFEE
  - Profile Align
- Classify ...
  - JACOP
  - MkDom2
- Tools ...
  - Reformat MSA
  - Reformat SEQ
  - Dotlet

Hub

Results

Misc

Deprecated

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
```

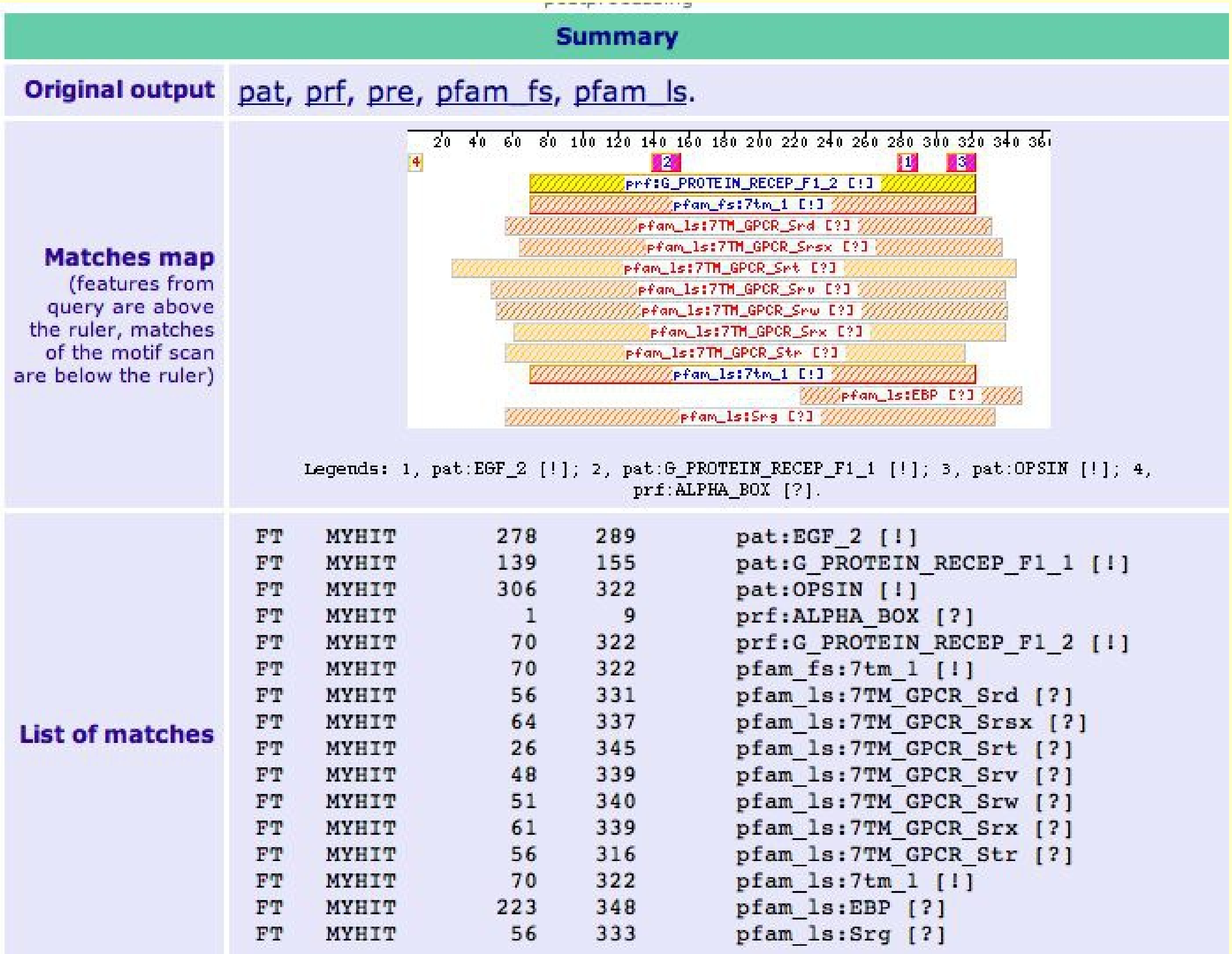
The scan might take a few minutes.

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

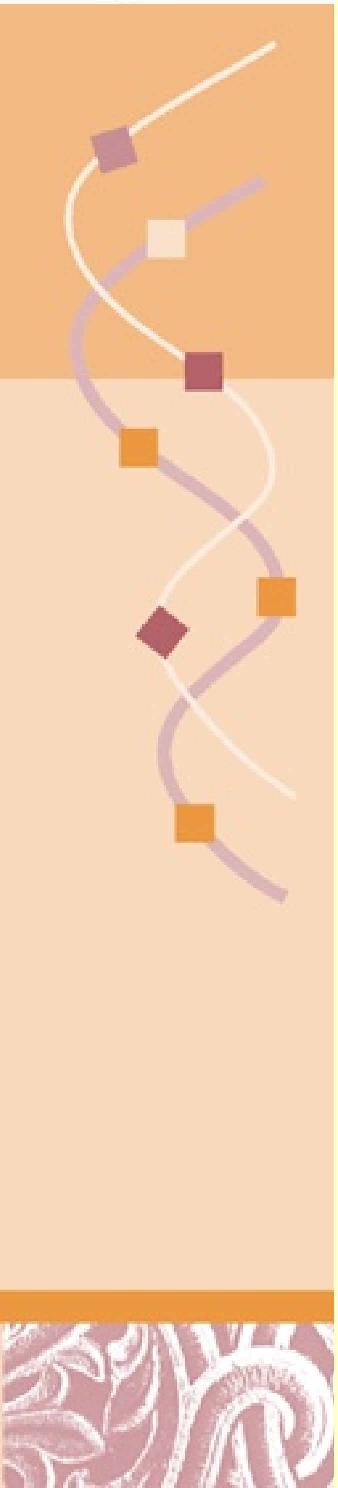
Original output pat, prf, pre, pfam_fs, pfam_ls.

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 [?].

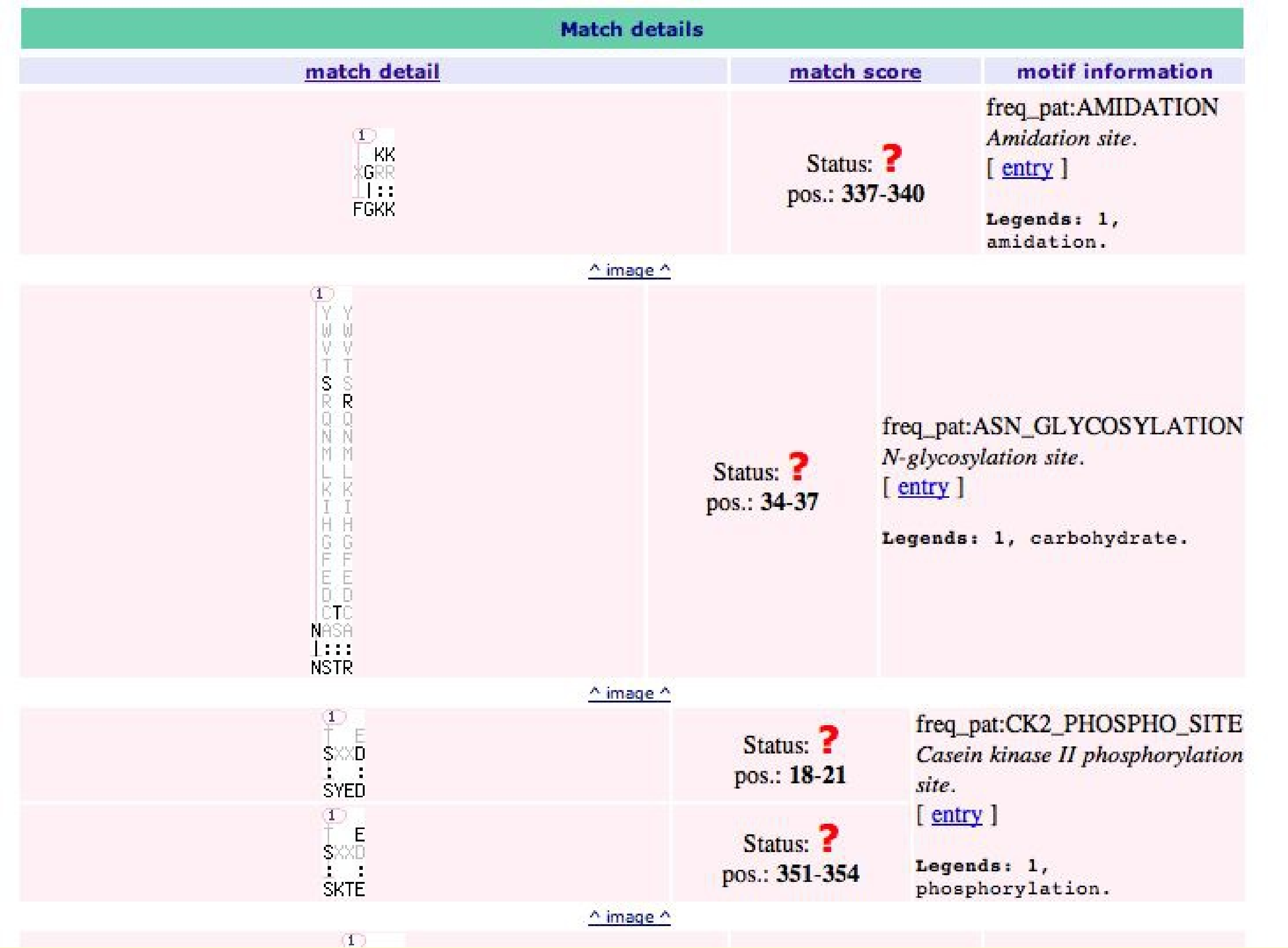
List of matches

```
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

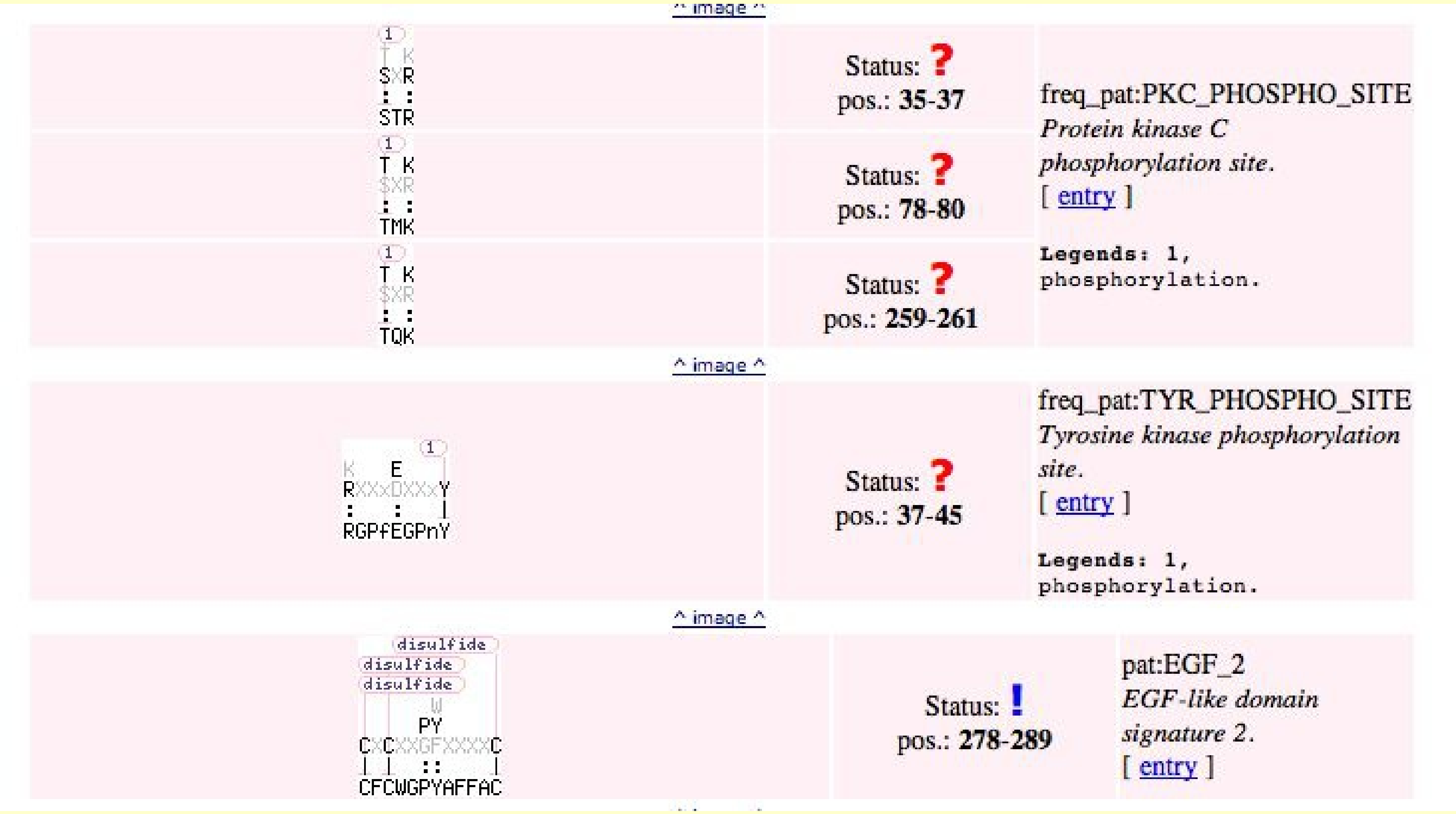
http://myhits.isb-sib.ch/



Match details		
match detail	match score	motif information
1 KK XGRR |:: FGKK	Status: ? pos.: 337-340	freq_pat:AMIDATION *Amidation site.* [entry] **Legends:** 1, amidation.
^ image ^		
1 Y Y W W V V T T S S R R Q Q N N M M L L K K I I H H G G F F E E D D CTC NASA |::: NSTR	Status: ? pos.: 34-37	freq_pat:ASN_GLYCOSYLATION *N-glycosylation site.* [entry] **Legends:** 1, carbohydrate.
^ image ^		
1 T E SXXD : : SYED	Status: ? pos.: 18-21	freq_pat:CK2_PHOSPHO_SITE *Casein kinase II phosphorylation site.* [entry] **Legends:** 1, phosphorylation.
1 T E SXXD : : SKTE	Status: ? pos.: 351-354	
^ image ^		
1		

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**	

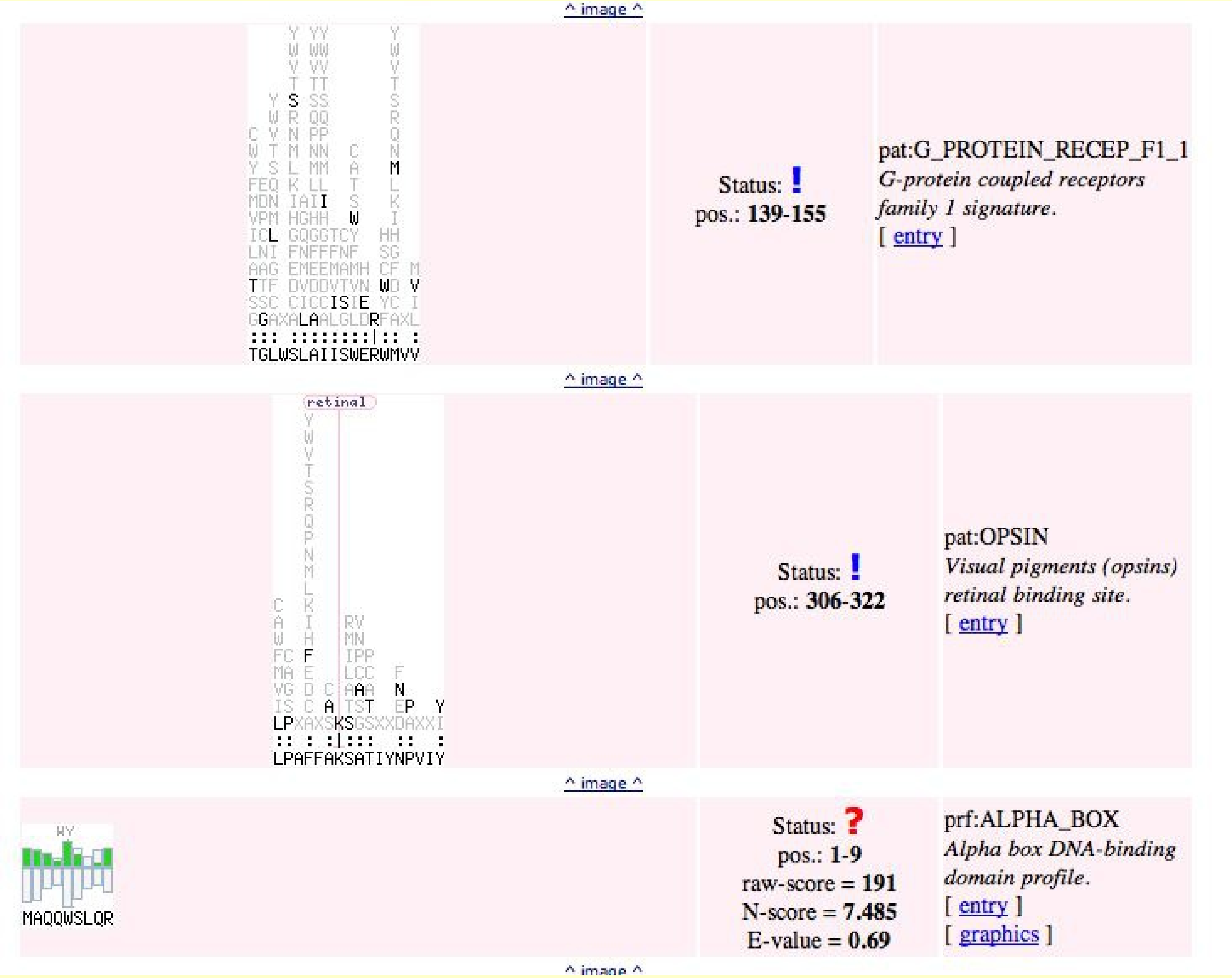
^ image ^

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]

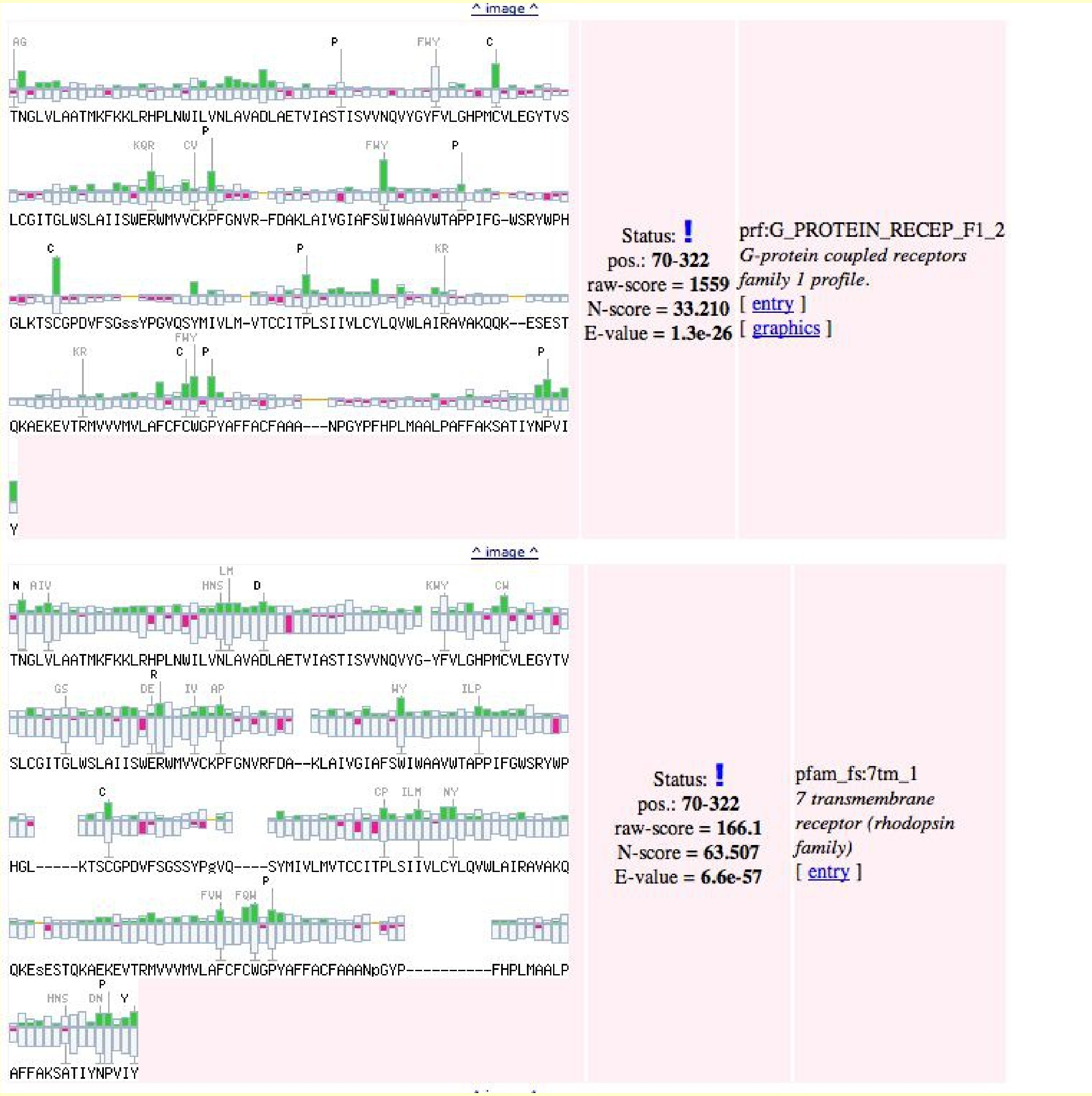
Status: **!**
pos.: **306-322**

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]

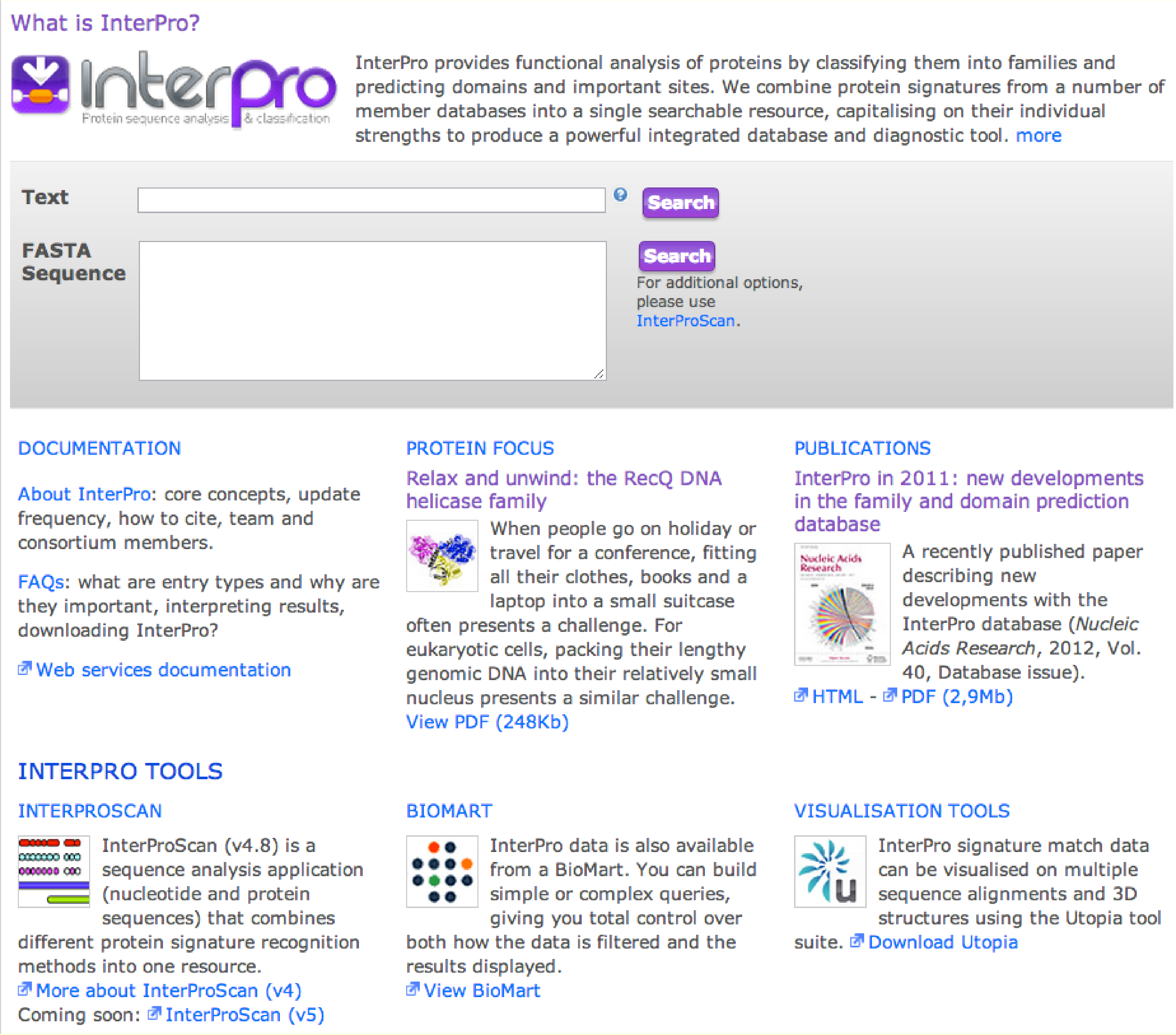
```
TNGLVLAATMKFKKLRHPLNWILVNLAVADLAETVIASTISVVNQVYG-YFVLGHPMCVLEGYTV
SLCGITGLWSLAIISWERWMVVCKPFGNVRFDA--KLAIVGIAFSWIWAAVWTAPPIFGWSRYWP
HGL-----KTSCGPDVFSGSSYPgVQ----SYMIVLMVTCCITPLSIIVLCYLQVWLAIRAVAKQ
QKEsESTQKAEKEVTRMVVVMVLAFCFCWGPYAFFACFAAANpGYP----------FHPLMAALP
AFFAKSATIYNPVIY
```

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.

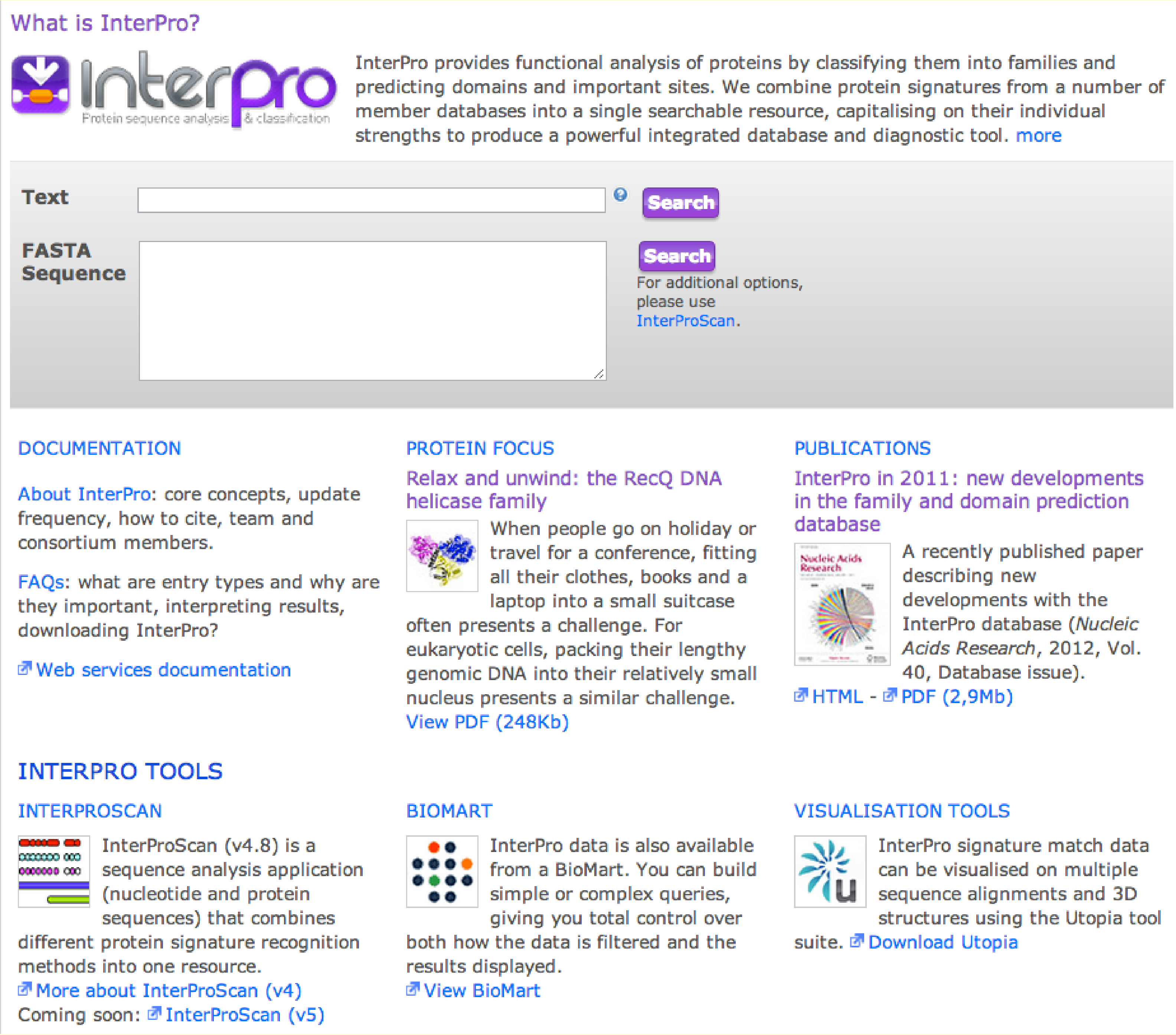
View BioMart

VISUALISATION TOOLS

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

InterPro Scan

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.

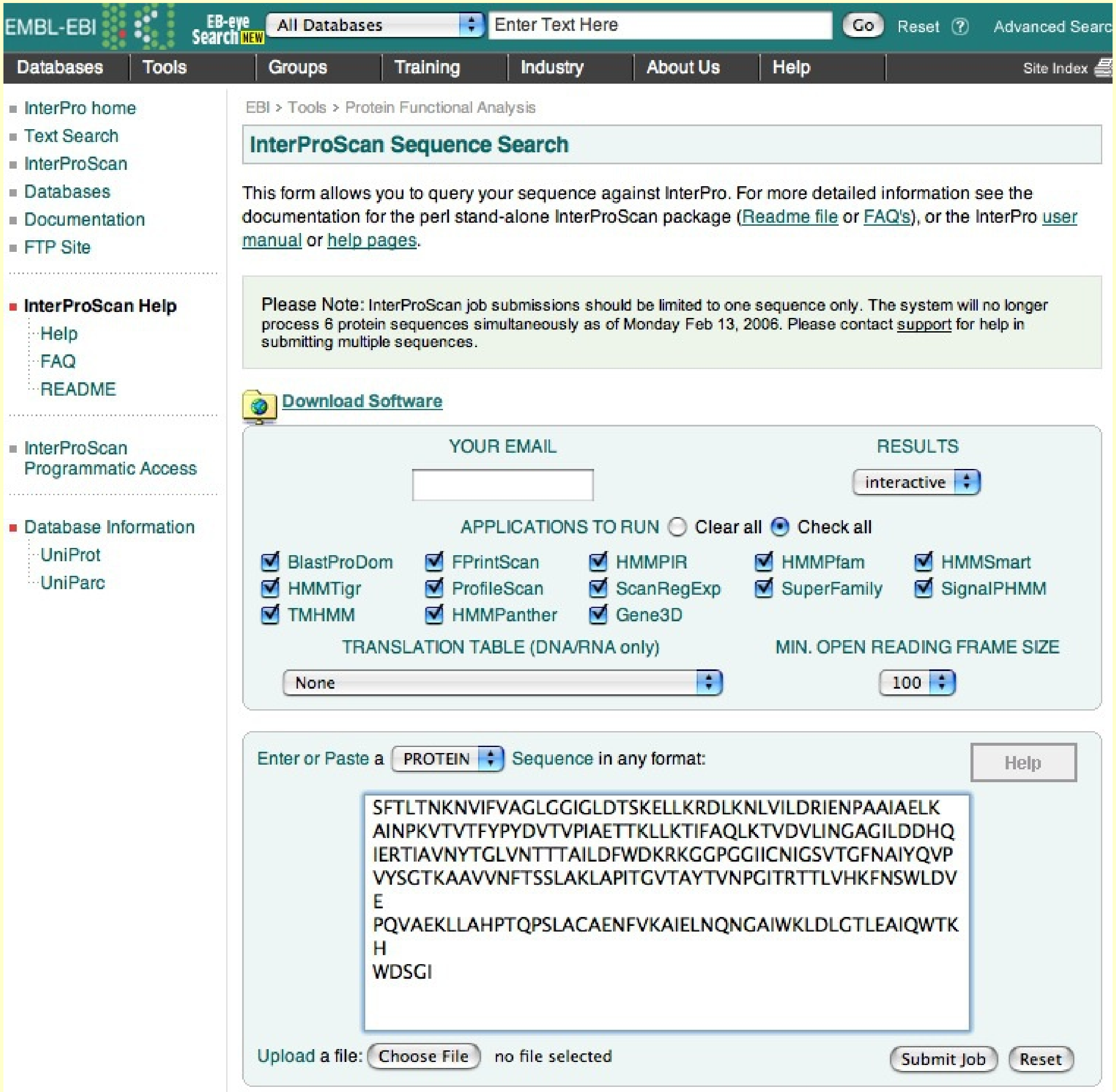
View BioMart

VISUALISATION TOOLS

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

InterPro Scan

http://www.ebi.ac.uk/Tools/pfa/iprscan/



InterProScan Sequence Search

This form allows you to query your sequence against InterPro. For more detailed information see the documentation for the perl stand-alone InterProScan package (Readme file or FAQ's), or the InterPro user manual or help pages.

Please Note: InterProScan job submissions should be limited to one sequence only. The system will no longer process 6 protein sequences simultaneously as of Monday Feb 13, 2006. Please contact support for help in submitting multiple sequences.

Download Software

YOUR EMAIL

RESULTS: interactive

APPLICATIONS TO RUN ○ Clear all ● Check all

☑ BlastProDom ☑ FPrintScan ☑ HMMPIR ☑ HMMPfam ☑ HMMSmart
☑ HMMTigr ☑ ProfileScan ☑ ScanRegExp ☑ SuperFamily ☑ SignalPHMM
☑ TMHMM ☑ HMMPanther ☑ Gene3D

TRANSLATION TABLE (DNA/RNA only): None

MIN. OPEN READING FRAME SIZE: 100

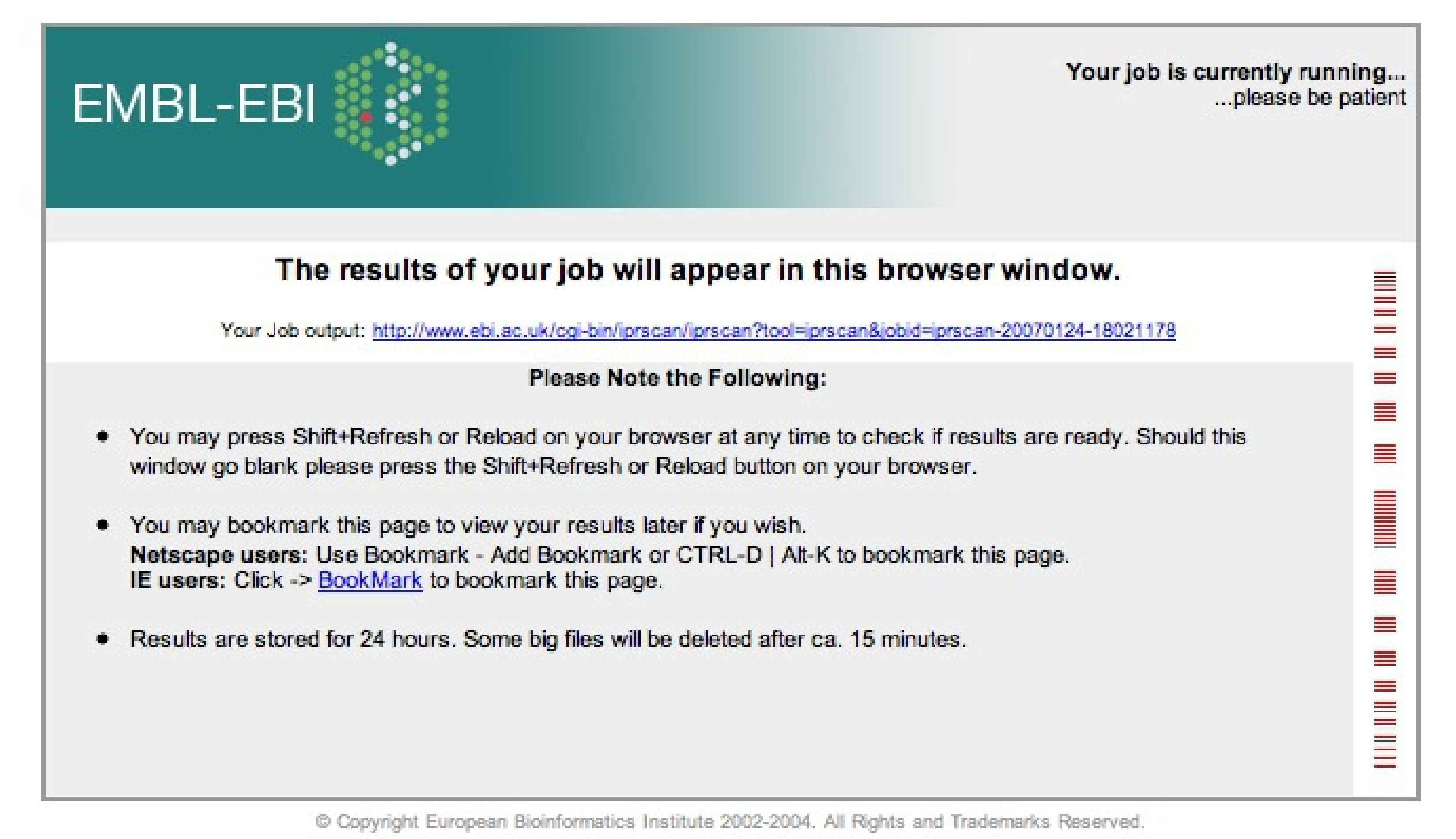
Enter or Paste a PROTEIN Sequence in any format:

```
SFTLTNKNVIFVAGLGGIGLDTSKELLKRDLKNLVILDRIENPAAIAELK
AINPKVTVTFYPYDVTVPIAETTKLLKTIFAQLKTVDVLINGAGILDDHQ
IERTIAVNYTGLVNTTTAILDFWDKRKGGPGGIICNIGSVTGFNAIYQVP
VYSGTKAAVVNFTSSLAKLAPITGVTAYTVNPGITRTTLVHKFNSWLDV
E
PQVAEKLLAHPTQPSLACAENFVKAIELNQNGAIWKLDLGTLEAIQWTK
H
WDSGI
```

Upload a file: no file selected

InterPro Scan Hour Glass

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



Your job is currently running...
...please be patient

The results of your job will appear in this browser window.

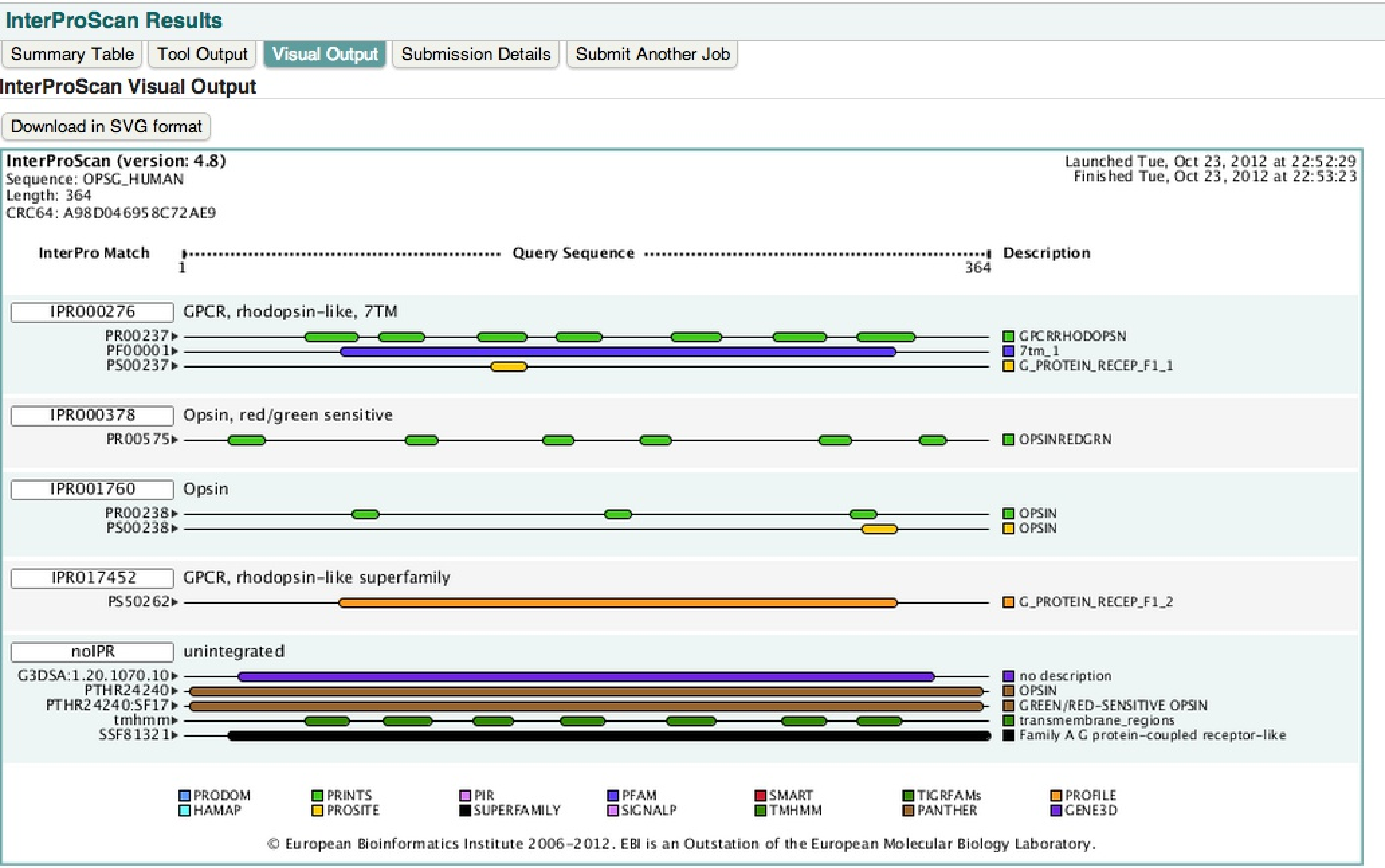
Your Job output: http://www.ebi.ac.uk/cgi-bin/iprscan/iprscan?tool=iprscan&jobid=iprscan-20070124-18021178

Please Note the Following:

- You may press Shift+Refresh or Reload on your browser at any time to check if results are ready. Should this window go blank please press the Shift+Refresh or Reload button on your browser.
- You may bookmark this page to view your results later if you wish.
  Netscape users: Use Bookmark - Add Bookmark or CTRL-D | Alt-K to bookmark this page.
  IE users: Click -> BookMark to bookmark this page.
- Results are stored for 24 hours. Some big files will be deleted after ca. 15 minutes.

InterPro Scan Results

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

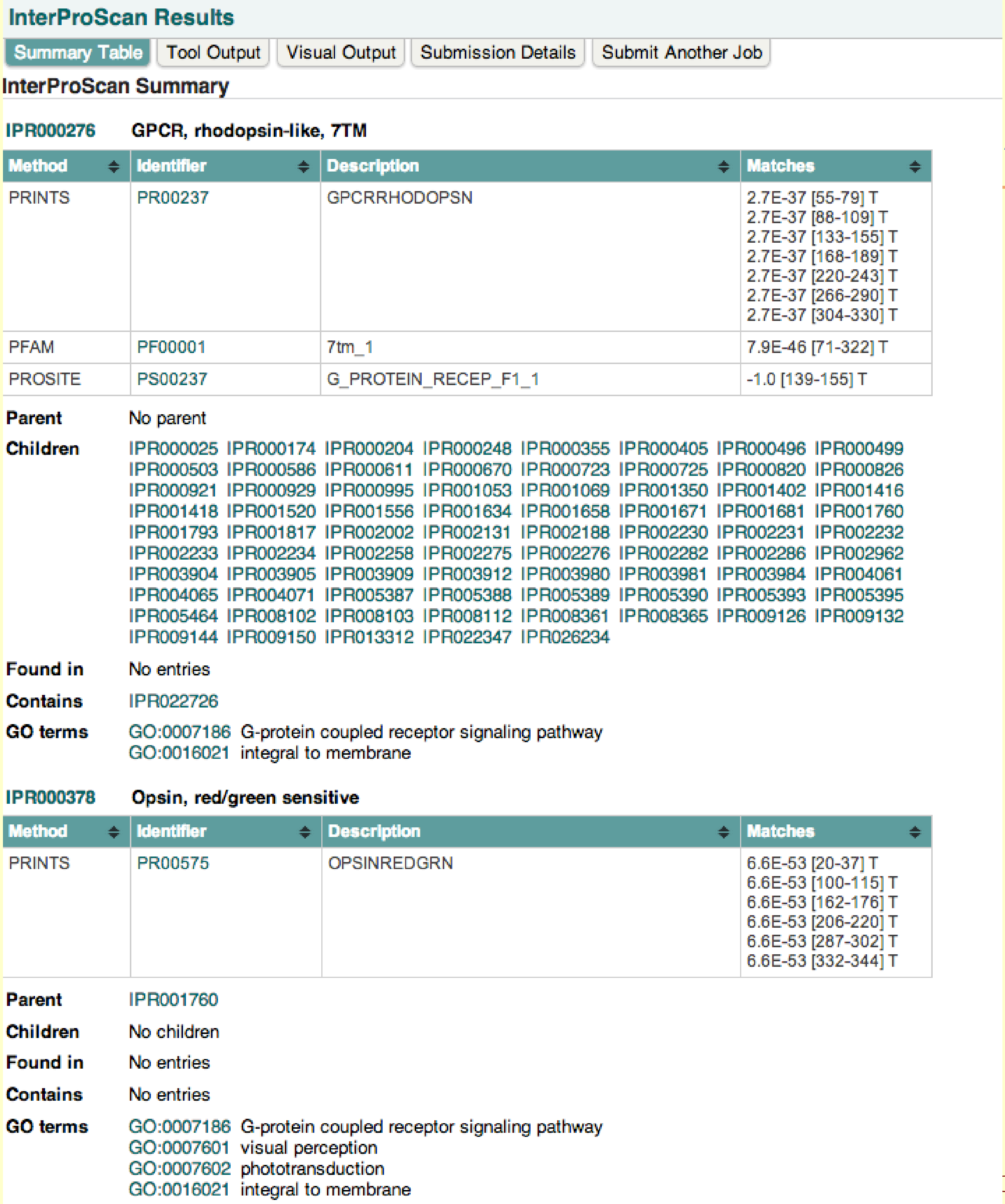


InterProScan Visual Output

InterProScan (version: 4.8)
Sequence: OPSG_HUMAN
Length: 364
CRC64: A98D046958C72AE9

Launched Tue, Oct 23, 2012 at 22:52:29
Finished Tue, Oct 23, 2012 at 22:53:23

InterPro Match	Query Sequence (1 … 364)	Description
IPR000276	GPCR, rhodopsin-like, 7TM	
PR00237		GPCRRHODOPSN
PF00001		7tm_1
PS00237		G_PROTEIN_RECEP_F1_1
IPR000378	Opsin, red/green sensitive	
PR00575		OPSINREDGRN
IPR001760	Opsin	
PR00238		OPSIN
PS00238		OPSIN
IPR017452	GPCR, rhodopsin-like superfamily	
PS50262		G_PROTEIN_RECEP_F1_2
noIPR	unintegrated	
G3DSA:1.20.1070.10		no description
PTHR24240		OPSIN
PTHR24240:SF17		GREEN/RED-SENSITIVE OPSIN
tmhmm		transmembrane_regions
SSF81321		Family A G protein-coupled receptor-like



InterProScan Summary

IPR000276 **GPCR, rhodopsin-like, 7TM**

Method	Identifier	Description	Matches
PRINTS	PR00237	GPCRRHODOPSN	2.7E-37 [55-79] T 2.7E-37 [88-109] T 2.7E-37 [133-155] T 2.7E-37 [168-189] T 2.7E-37 [220-243] T 2.7E-37 [266-290] T 2.7E-37 [304-330] T
PFAM	PF00001	7tm_1	7.9E-46 [71-322] T
PROSITE	PS00237	G_PROTEIN_RECEP_F1_1	-1.0 [139-155] T

Parent No parent

Children IPR000025 IPR000174 IPR000204 IPR000248 IPR000355 IPR000405 IPR000496 IPR000499
IPR000503 IPR000586 IPR000611 IPR000670 IPR000723 IPR000725 IPR000820 IPR000826
IPR000921 IPR000929 IPR000995 IPR001053 IPR001069 IPR001350 IPR001402 IPR001416
IPR001418 IPR001520 IPR001556 IPR001634 IPR001658 IPR001671 IPR001681 IPR001760
IPR001793 IPR001817 IPR002002 IPR002131 IPR002188 IPR002230 IPR002231 IPR002232
IPR002233 IPR002234 IPR002258 IPR002275 IPR002276 IPR002282 IPR002286 IPR002962
IPR003904 IPR003905 IPR003909 IPR003912 IPR003980 IPR003981 IPR003984 IPR004061
IPR004065 IPR004071 IPR005387 IPR005388 IPR005389 IPR005390 IPR005393 IPR005395
IPR005464 IPR008102 IPR008103 IPR008112 IPR008361 IPR008365 IPR009126 IPR009132
IPR009144 IPR009150 IPR013312 IPR022347 IPR026234

Found in No entries

Contains IPR022726

GO terms GO:0007186 G-protein coupled receptor signaling pathway
GO:0016021 integral to membrane

IPR000378 **Opsin, red/green sensitive**

Method	Identifier	Description	Matches
PRINTS	PR00575	OPSINREDGRN	6.6E-53 [20-37] T 6.6E-53 [100-115] T 6.6E-53 [162-176] T 6.6E-53 [206-220] T 6.6E-53 [287-302] T 6.6E-53 [332-344] T

Parent IPR001760

Children No children

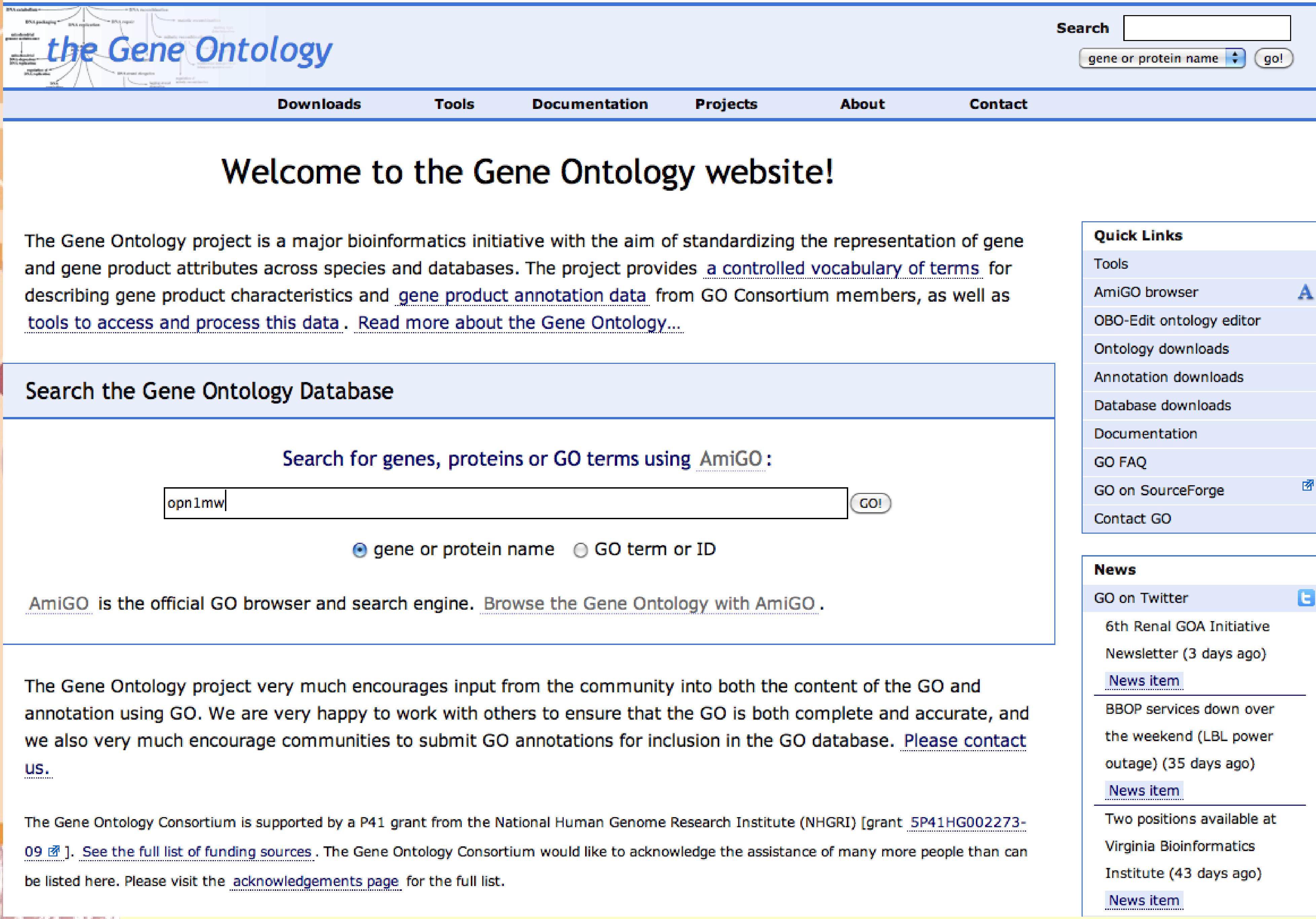
Found in No entries

Contains No entries

GO terms GO:0007186 G-protein coupled receptor signaling pathway
GO:0007601 visual perception
GO:0007602 phototransduction
GO:0016021 integral to membrane

GO: Gene Ontology Database

http://www.geneontology.org/



Welcome to the Gene Ontology website!

The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data from GO Consortium members, as well as tools to access and process this data. Read more about the Gene Ontology...

Search the Gene Ontology Database

Search for genes, proteins or GO terms using AmiGO:

opn1mw

● gene or protein name ○ GO term or ID

AmiGO is the official GO browser and search engine. Browse the Gene Ontology with AmiGO.

The Gene Ontology project very much encourages input from the community into both the content of the GO and annotation using GO. We are very happy to work with others to ensure that the GO is both complete and accurate, and we also very much encourage communities to submit GO annotations for inclusion in the GO database. Please contact us.

The Gene Ontology Consortium is supported by a P41 grant from the National Human Genome Research Institute (NHGRI) [grant 5P41HG002273-09]. See the full list of funding sources. The Gene Ontology Consortium would like to acknowledge the assistance of many more people than can be listed here. Please visit the acknowledgements page for the full list.

6th Renal GOA Initiative Newsletter (3 days ago)
News item

BBOP services down over the weekend (LBL power outage) (35 days ago)
News item

Two positions available at Virginia Bioinformatics Institute (43 days ago)
News item

GO: Gene Ontology for Opsin OPN1MW

http://www.geneontology.org/

Search GO opn1mw ○ GO terms ● genes or proteins ☐ exact match

Gene Product Search Results

7 results for **opn1mw** in genes or proteins fields **symbol, full name(s) and synonyms**

Filter search results

Filter Gene Products

Gene Product Type: All, gene, gene product, mirna

Data source: All, ASAP, AspGD, CGD

Species: All, Arabidopsis thaliana, Bacillus anthraci..., Bacillus subtilis

Filter Gene Products by Associations

Ontology: All, biological process, cellular component, molecular function

Evidence Code: All, IMR, IRD, IC

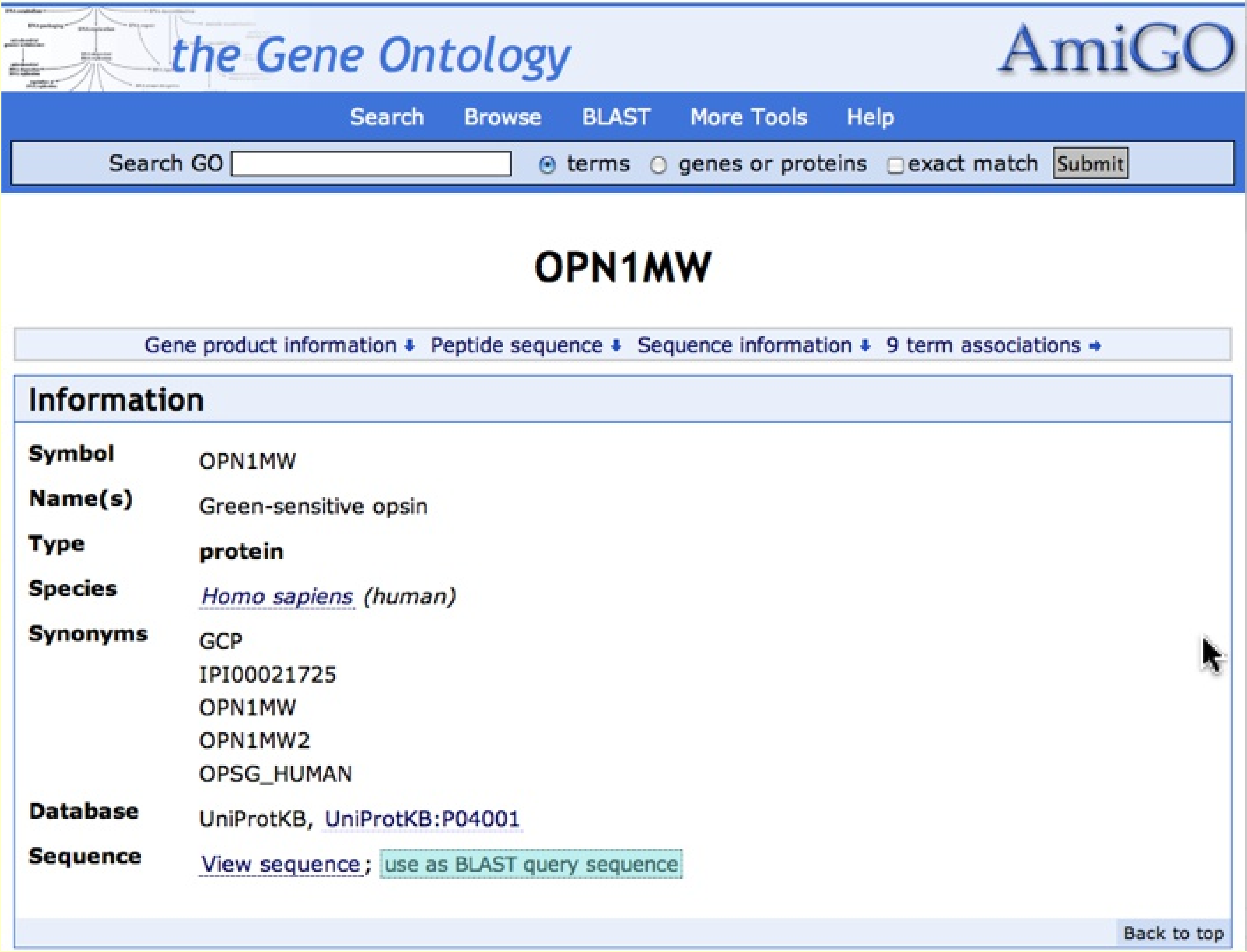
Results are sorted by **relevance**. To change the sort order, click on the column headers.

Perform an action with this page's selected gene products...

rel ↓	Symbol , full name		Species
☐	**OPN1MW** Green-sensitive opsin	9 associations BLAST	**protein** from *Homo sapiens*
☐	**Opn1mw** opsin 1 (cone pigments), medium-wave-sensitive	12 associations BLAST	**gene** from *Rattus norvegicus*
☐	**Opn1mw** opsin 1 (cone pigments), medium-wave-sensitive (color blindness, deutan)	13 associations	**protein** from *Mus musculus*
☐	**opn1mw**1 opsin 1 (cone pigments), medium-wave-sensitive, 1	13 associations BLAST	**gene_product** from *Danio rerio*

GO: Gene Ontology for Opsin OPN1MW

http://www.geneontology.org/



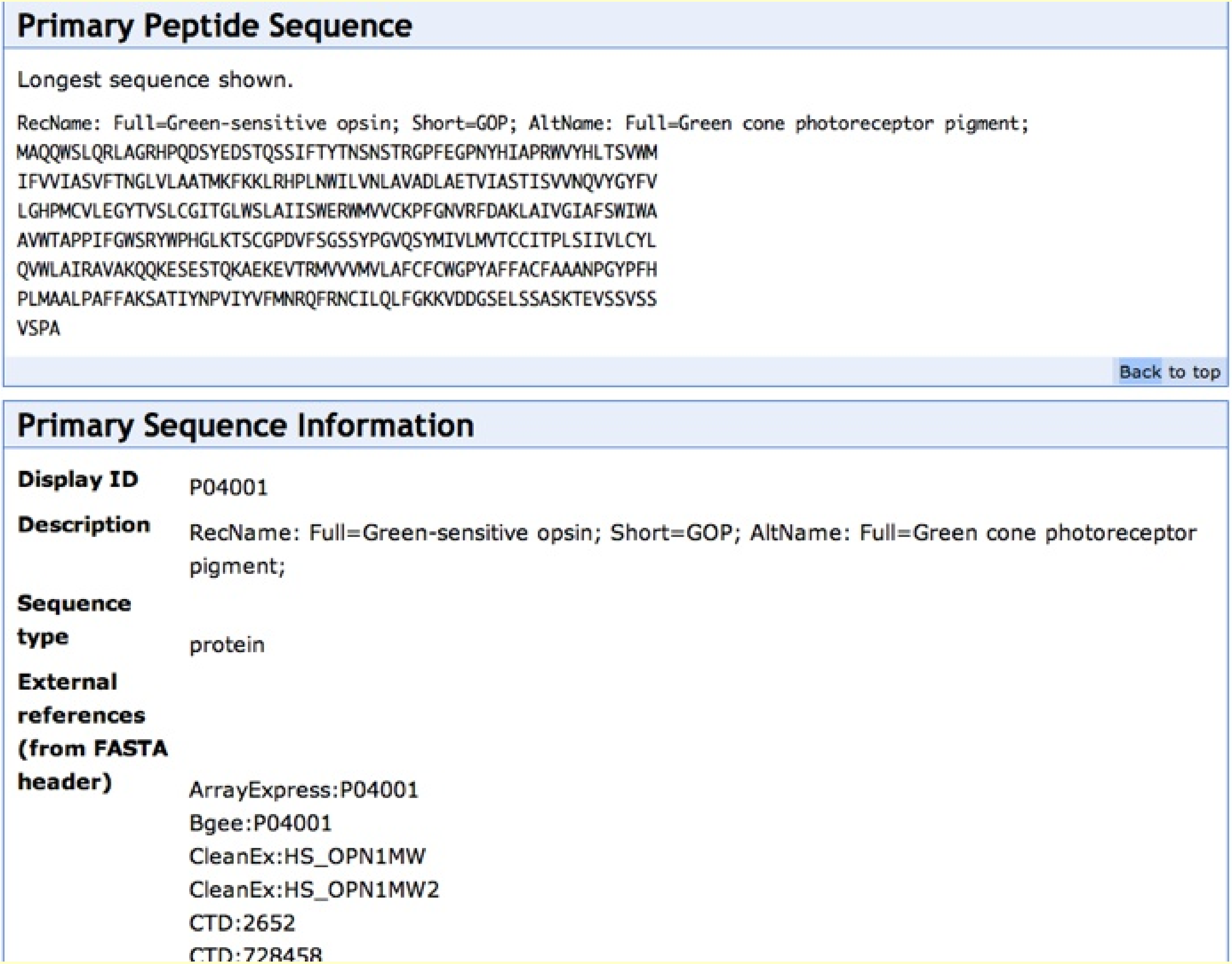
OPN1MW

Information

Symbol	OPN1MW
Name(s)	Green-sensitive opsin
Type	**protein**
Species	*Homo sapiens (human)*
Synonyms	GCP IPI00021725 OPN1MW OPN1MW2 OPSG_HUMAN
Database	UniProtKB, UniProtKB:P04001
Sequence	View sequence; use as BLAST query sequence

GO: Sequence Information for OPN1MW

http://www.geneontology.org/



Primary Peptide Sequence

Longest sequence shown.

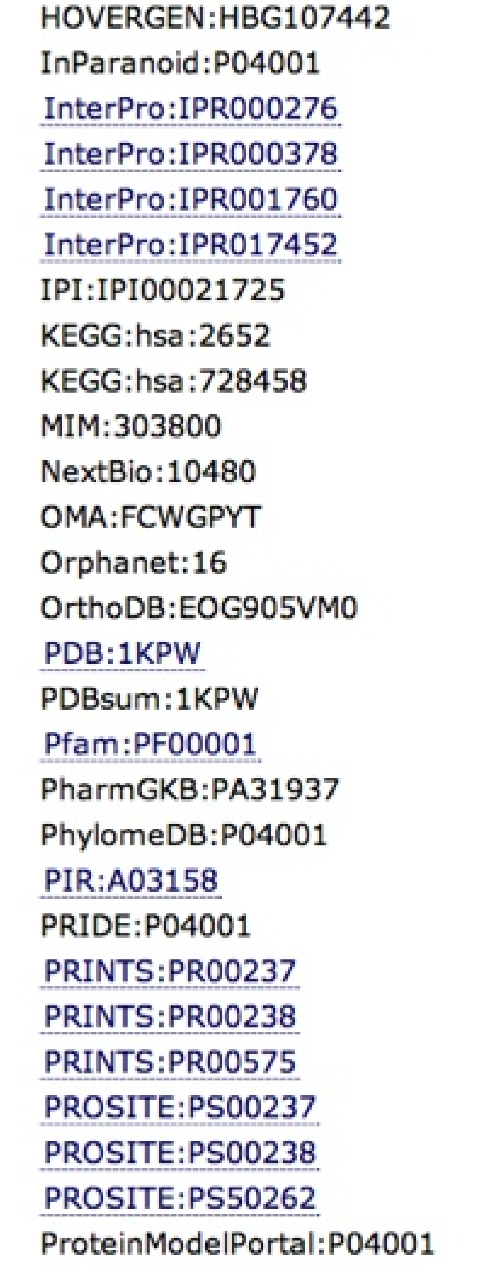
```
RecName: Full=Green-sensitive opsin; Short=GOP; AltName: Full=Green cone photoreceptor pigment;
MAQQWSLQRLAGRHPQDSYEDSTQSSIFTYTNSNSTRGPFEGPNYHIAPRWVYHLTSVWM
IFVVIASVFTNGLVLAATMKFKKLRHPLNWILVNLAVADLAETVIASTISVVNQVYGYFV
LGHPMCVLEGYTVSLCGITGLWSLAIISWERWMVVCKPFGNVRFDAKLAIVGIAFSWIWA
AVWTAPPIFGWSRYWPHGLKTSCGPDVFSGSSYPGVQSYMIVLMVTCCITPLSIIVLCYL
QVWLAIRAVAKQQKESESTQKAEKEVTRMVVVMVLAFCFCWGPYAFFACFAAANPGYPFH
PLMAALPAFFAKSATIYNPVIYVFMNRQFRNCILQLFGKKVDDGSELSSASKTEVSSVSS
VSPA
```

Primary Sequence Information

Display ID	P04001
Description	RecName: Full=Green-sensitive opsin; Short=GOP; AltName: Full=Green cone photoreceptor pigment;
Sequence type	protein
External references (from FASTA header)	ArrayExpress:P04001 Bgee:P04001 CleanEx:HS_OPN1MW CleanEx:HS_OPN1MW2 CTD:2652 CTD:728458

GO: Annotations for OPN1MW

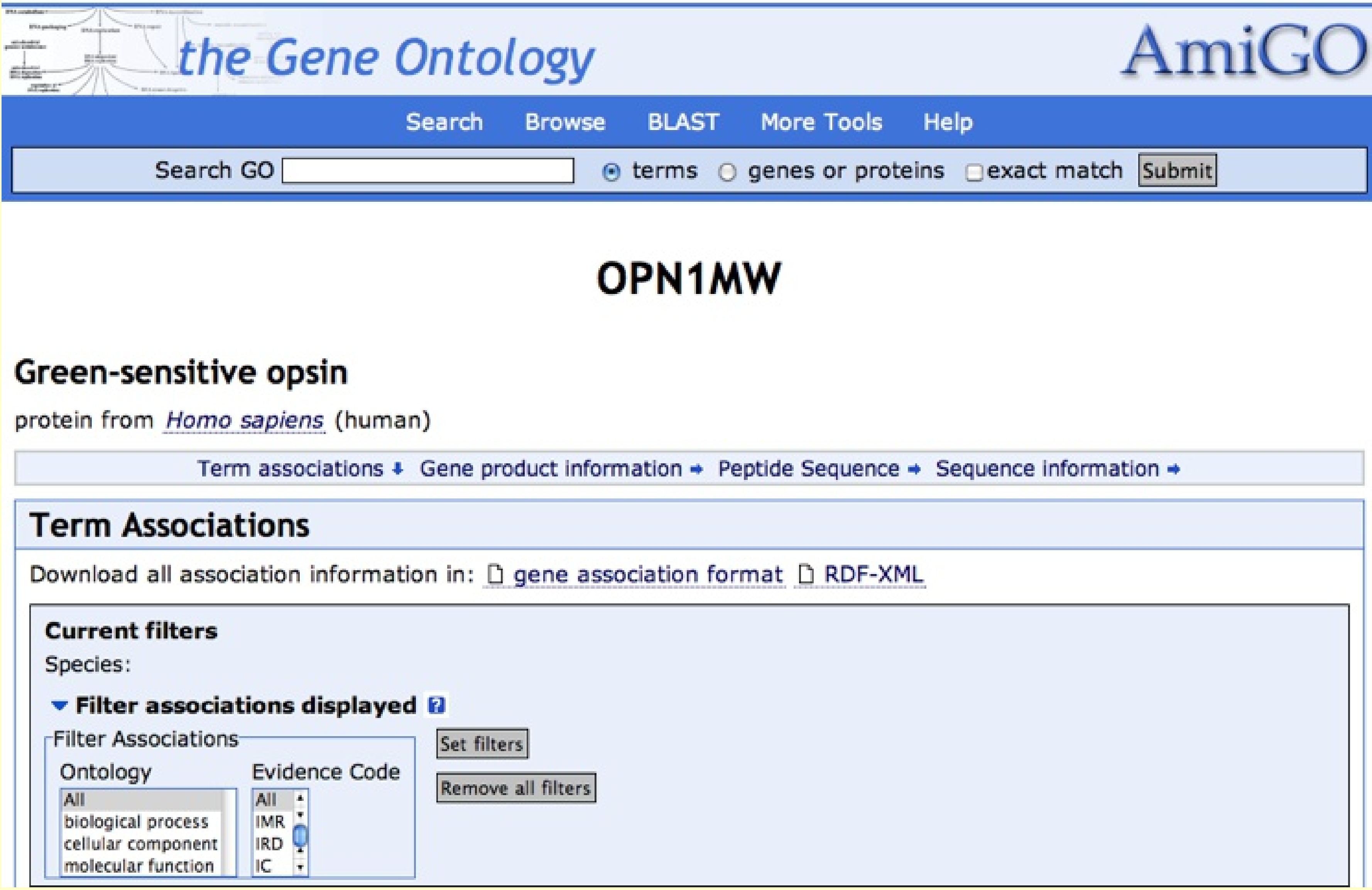
http://www.geneontology.org/



HOVERGEN:HBG107442
InParanoid:P04001
InterPro:IPR000276
InterPro:IPR000378
InterPro:IPR001760
InterPro:IPR017452
IPI:IPI00021725
KEGG:hsa:2652
KEGG:hsa:728458
MIM:303800
NextBio:10480
OMA:FCWGPYT
Orphanet:16
OrthoDB:EOG905VM0
PDB:1KPW
PDBsum:1KPW
Pfam:PF00001
PharmGKB:PA31937
PhylomeDB:P04001
PIR:A03158
PRIDE:P04001
PRINTS:PR00237
PRINTS:PR00238
PRINTS:PR00575
PROSITE:PS00237
PROSITE:PS00238
PROSITE:PS50262
ProteinModelPortal:P04001

GO: Gene Ontology Database

http://www.geneontology.org/



OPN1MW

Green-sensitive opsin

protein from *Homo sapiens* (human)

Term Associations

Download all association information in: gene association format RDF-XML

Current filters

Species:

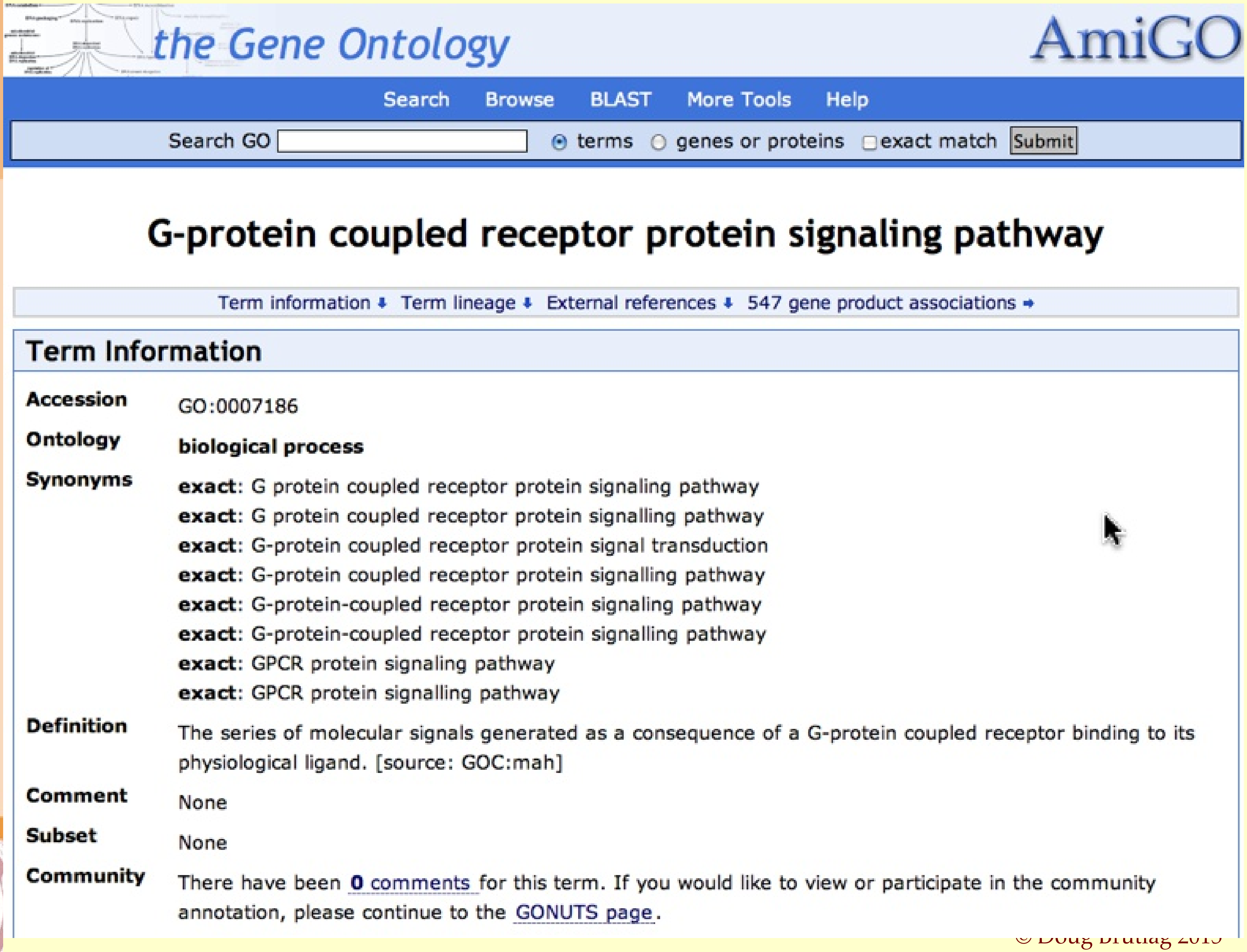
Filter associations displayed

Filter Associations

Ontology	Evidence Code
All	All
biological process	IMR
cellular component	IRD
molecular function	IC

GO: Gene Ontology Terms for OPN1MW

http://www.geneontology.org/



G-protein coupled receptor protein signaling pathway

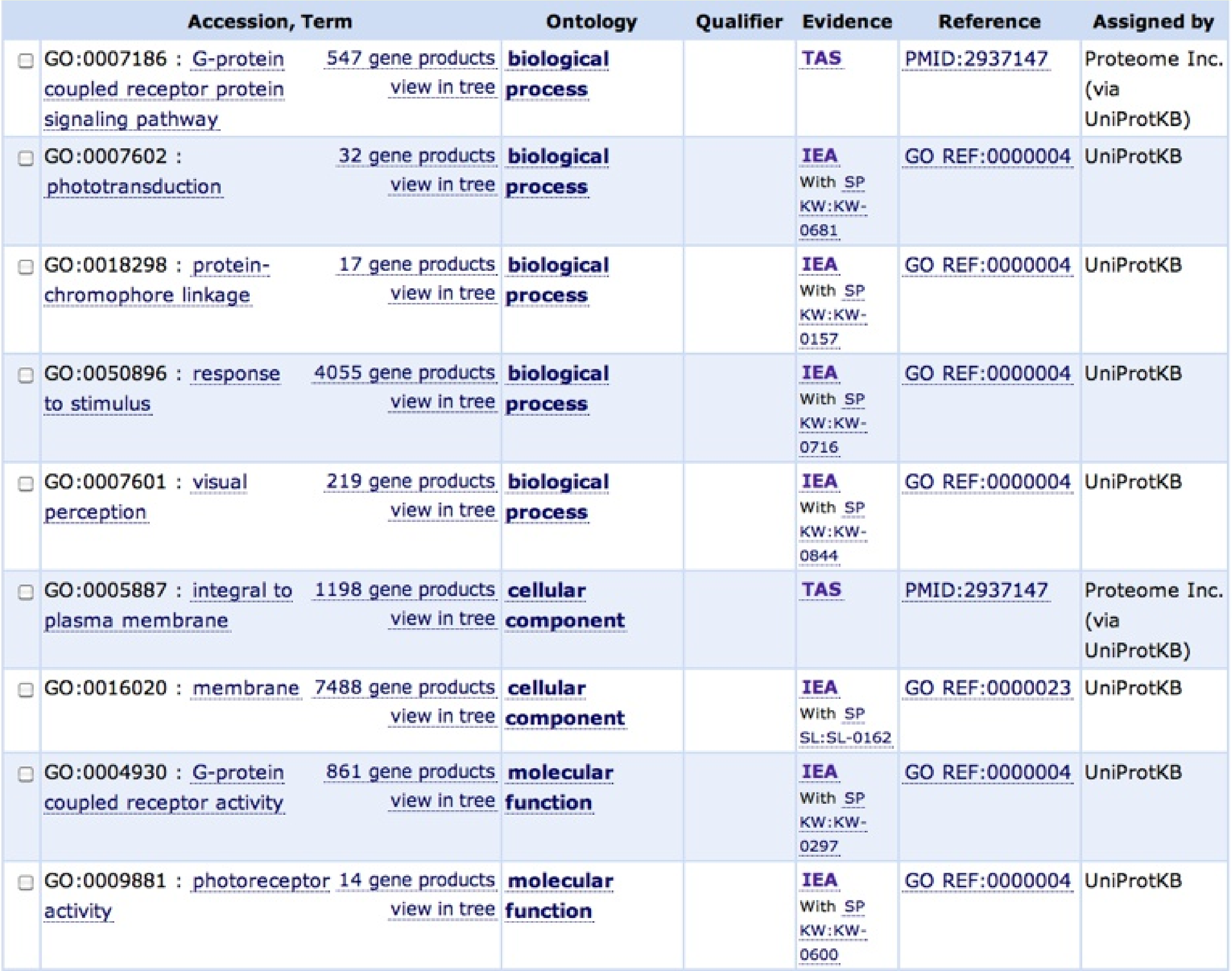
Term Information

Accession	GO:0007186
Ontology	**biological process**
Synonyms	**exact**: G protein coupled receptor protein signaling pathway **exact**: G protein coupled receptor protein signalling pathway **exact**: G-protein coupled receptor protein signal transduction **exact**: G-protein coupled receptor protein signalling pathway **exact**: G-protein-coupled receptor protein signaling pathway **exact**: G-protein-coupled receptor protein signalling pathway **exact**: GPCR protein signaling pathway **exact**: GPCR protein signalling pathway
Definition	The series of molecular signals generated as a consequence of a G-protein coupled receptor binding to its physiological ligand. [source: GOC:mah]
Comment	None
Subset	None
Community	There have been **0** comments for this term. If you would like to view or participate in the community annotation, please continue to the GONUTS page.

the Gene Ontology

GO: Gene Ontology Term GCRP

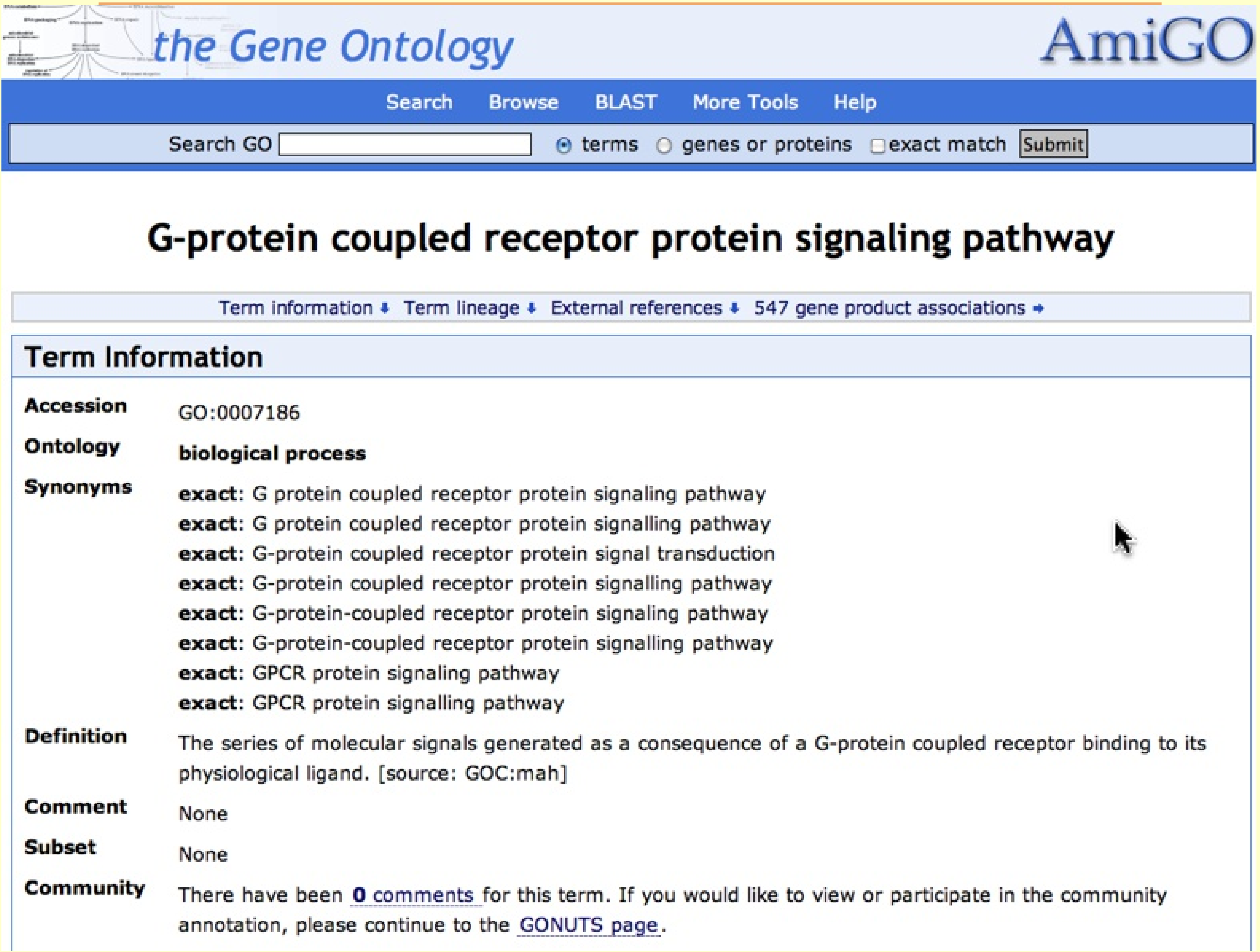
http://www.geneontology.org/



Accession, Term		Ontology	Qualifier	Evidence	Reference	Assigned by
☐ GO:0007186 : G-protein coupled receptor protein signaling pathway	547 gene products view in tree	biological process		TAS	PMID:2937147	Proteome Inc. (via UniProtKB)
☐ GO:0007602 : phototransduction	32 gene products view in tree	biological process		IEA With SP KW:KW-0681	GO REF:0000004	UniProtKB
☐ GO:0018298 : protein-chromophore linkage	17 gene products view in tree	biological process		IEA With SP KW:KW-0157	GO REF:0000004	UniProtKB
☐ GO:0050896 : response to stimulus	4055 gene products view in tree	biological process		IEA With SP KW:KW-0716	GO REF:0000004	UniProtKB
☐ GO:0007601 : visual perception	219 gene products view in tree	biological process		IEA With SP KW:KW-0844	GO REF:0000004	UniProtKB
☐ GO:0005887 : integral to plasma membrane	1198 gene products view in tree	cellular component		TAS	PMID:2937147	Proteome Inc. (via UniProtKB)
☐ GO:0016020 : membrane	7488 gene products view in tree	cellular component		IEA With SP SL:SL-0162	GO REF:0000023	UniProtKB
☐ GO:0004930 : G-protein coupled receptor activity	861 gene products view in tree	molecular function		IEA With SP KW:KW-0297	GO REF:0000004	UniProtKB
☐ GO:0009881 : photoreceptor activity	14 gene products view in tree	molecular function		IEA With SP KW:KW-0600	GO REF:0000004	UniProtKB

GO: Gene Ontology GCPR Term

http://www.geneontology.org/



G-protein coupled receptor protein signaling pathway

Term Information

Accession	GO:0007186
Ontology	**biological process**
Synonyms	**exact**: G protein coupled receptor protein signaling pathway **exact**: G protein coupled receptor protein signalling pathway **exact**: G-protein coupled receptor protein signal transduction **exact**: G-protein coupled receptor protein signalling pathway **exact**: G-protein-coupled receptor protein signaling pathway **exact**: G-protein-coupled receptor protein signalling pathway **exact**: GPCR protein signaling pathway **exact**: GPCR protein signalling pathway
Definition	The series of molecular signals generated as a consequence of a G-protein coupled receptor binding to its physiological ligand. [source: GOC:mah]
Comment	None
Subset	None
Community	There have been 0 comments for this term. If you would like to view or participate in the community annotation, please continue to the GONUTS page.

GO: Gene Ontology GCPR Term

http://www.geneontology.org/

Term Lineage

Switch to viewing term parents, siblings and children

Current filters

Species:

Filter tree view

- all : all [18294 gene products]
  - GO:0008150 : biological_process [14700 gene products]
    - GO:0023052 : signaling [3619 gene products]
      - GO:0023033 : signaling pathway [2580 gene products]
        - GO:0007166 : cell surface receptor linked signaling pathway [1519 gene products]
          - **GO:0007186 : G-protein coupled receptor protein signaling pathway** **[547 gene products]**

Actions...
Last action: Reset the tree

External References

- InterPro (268)
- MIPS_funcat (2)
  - 30.01.05.05
  - 30.05.02.24
- Pfam (12)
- PIRSF (1)
  - PIRSF037995
- PRINTS (258)
- PROSITE (9)

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.