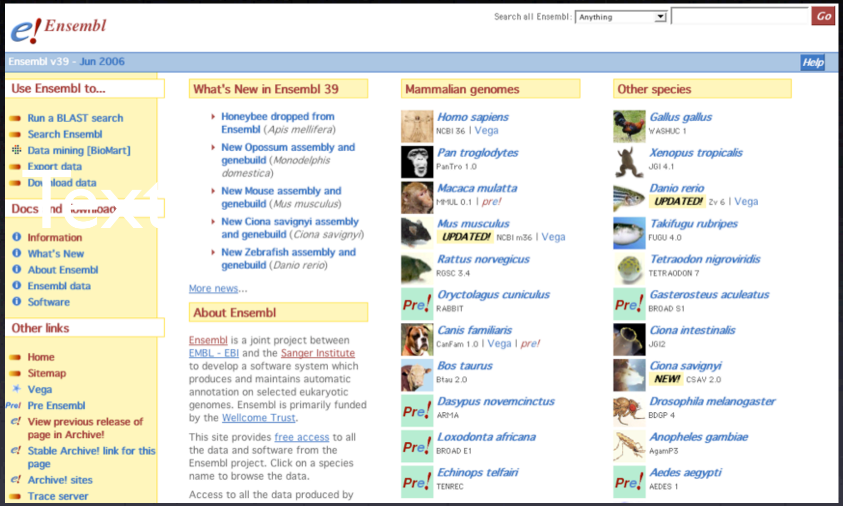
Data Integration using the Distributed Annotation System (DAS)

Andreas Prlić, Ewan Birney, Tony Cox, Thomas A. Down, Rob Finn, Stefan Gräf, David Jackson, Andreas Kähäri, Eugene Kulesha, Roger Pettett, James Smith, Jim Stalker, Tim J. P. Hubbard

wellcome trust sanger institute

EBI

- what is DAS
- what do we do with it
- DAS registration server
- latest developments



e! Ensembl

Search all Ensembl: Anything ▾ Go

Ensembl v39 - Jun 2006

Help

Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Export data
- Download data

Docs and download

- Information
- What's New
- About Ensembl
- Ensembl data
- Software

Other links

- Home
- Sitemap
- Vega
- Pre Ensembl
- View previous release of page in Archive!
- Stable Archive! link for this page
- Archive! sites
- Trace server

What's New in Ensembl 39

- Honeybee dropped from Ensembl (*Apis mellifera*)
- New Opossum assembly and genebuild (*Monodelphis domestica*)
- New Mouse assembly and genebuild (*Mus musculus*)
- New Ciona savignyi assembly and genebuild (*Ciona savignyi*)
- New Zebrafish assembly and genebuild (*Danio rerio*)

More news...

About Ensembl

Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. Ensembl is primarily funded by the Wellcome Trust.

This site provides free access to all the data and software from the Ensembl project. Click on a species name to browse the data.

Access to all the data produced by

Mammalian genomes

Homo sapiens
NCBI 36 | Vega

Pan troglodytes
PanTro 1.0

Macaca mulatta
MMUL 0.1 | *pre!*

Mus musculus
UPDATED! NCBI m36 | Vega

Rattus norvegicus
RGSC 3.4

Pre! *Oryctolagus cuniculus*
RABBIT

Canis familiaris
CanFam 1.0 | Vega | *pre!*

Bos taurus
Btau 2.0

Pre! *Dasypus novemcinctus*
ARMA

Pre! *Loxodonta africana*
BROAD E1

Pre! *Echinops telfairi*
TENREC

Other species

Gallus gallus
WASHUC 1

Xenopus tropicalis
JGI 4.1

Danio rerio
UPDATED! Zv 6 | Vega

Takifugu rubripes
FUGU 4.0

Tetraodon nigroviridis
TETRAODON 7

Pre! *Gasterosteus aculeatus*
BROAD S1

Ciona intestinalis
JGI2

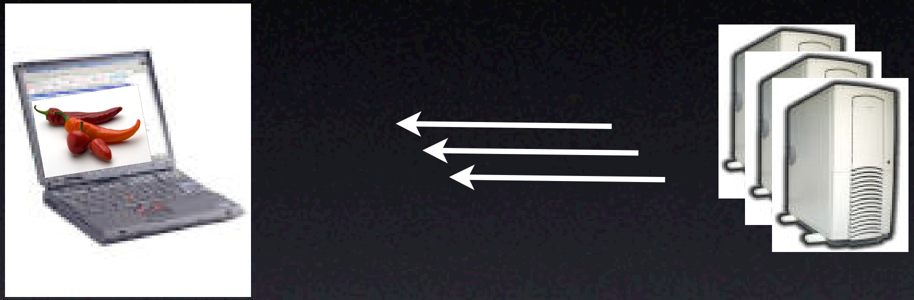
Ciona savignyi
NEW! CSAV 2.0

Drosophila melanogaster
BDGP 4

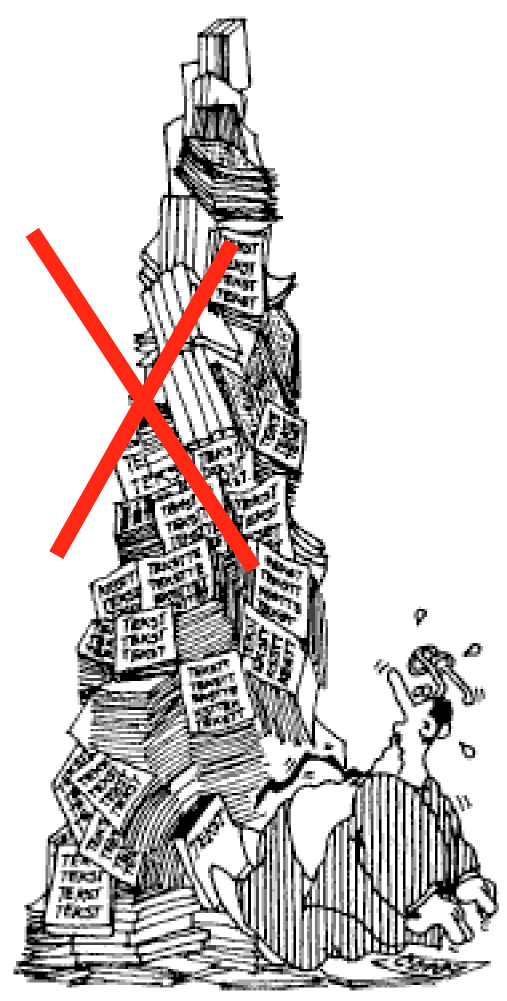
Anopheles gambiae
AgamP3

Pre! *Aedes aegypti*
AEDES 1

Integration of personal data into bioinf. resources



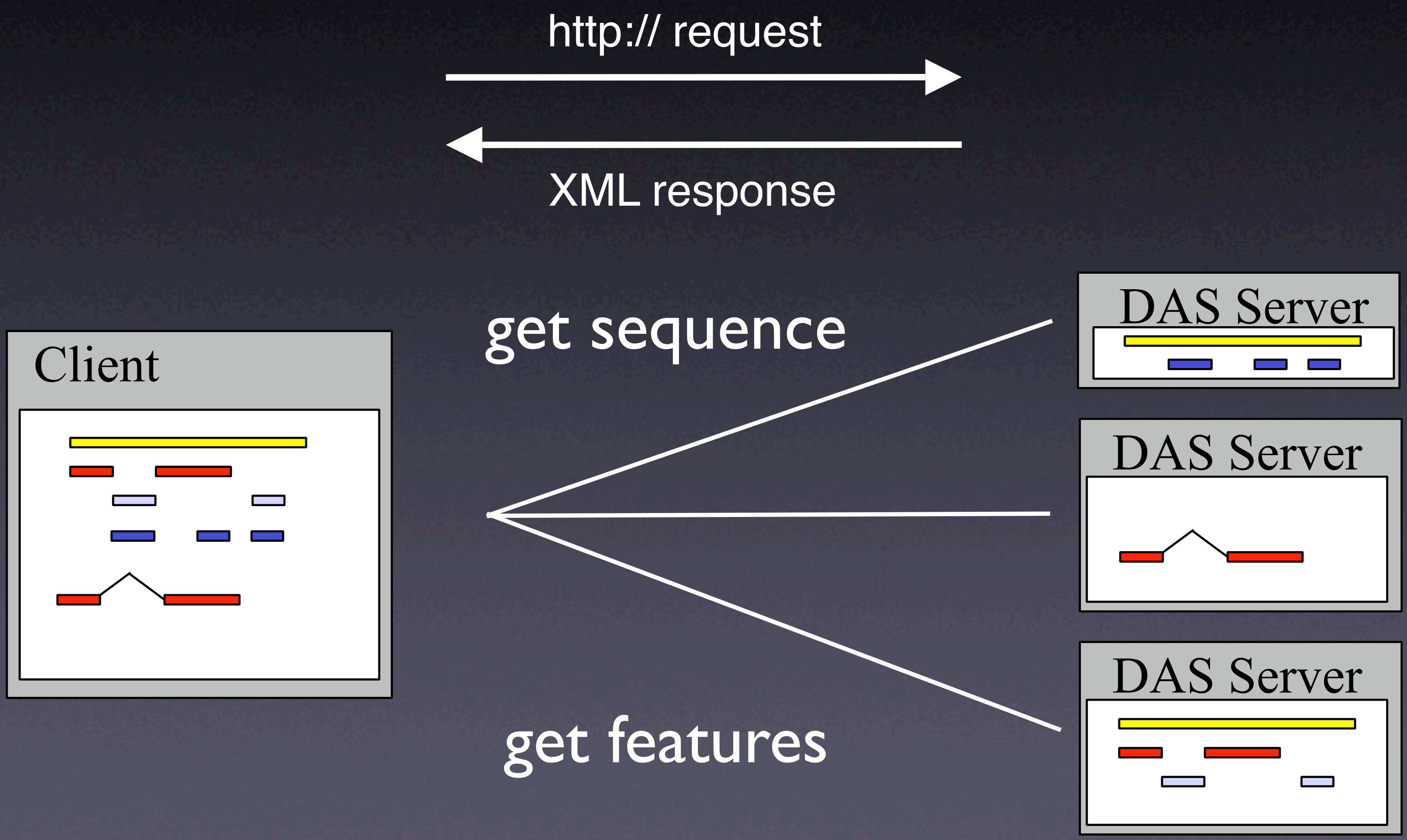
- Integration of annotations from external sources into local applications



- online access to most recent data versions
  - no need for local installations

DAS, how it works

Dowell, Jokerst, Allen, Eddy, Stein
BMC Bioninformatics 2001



http:// request

XML response

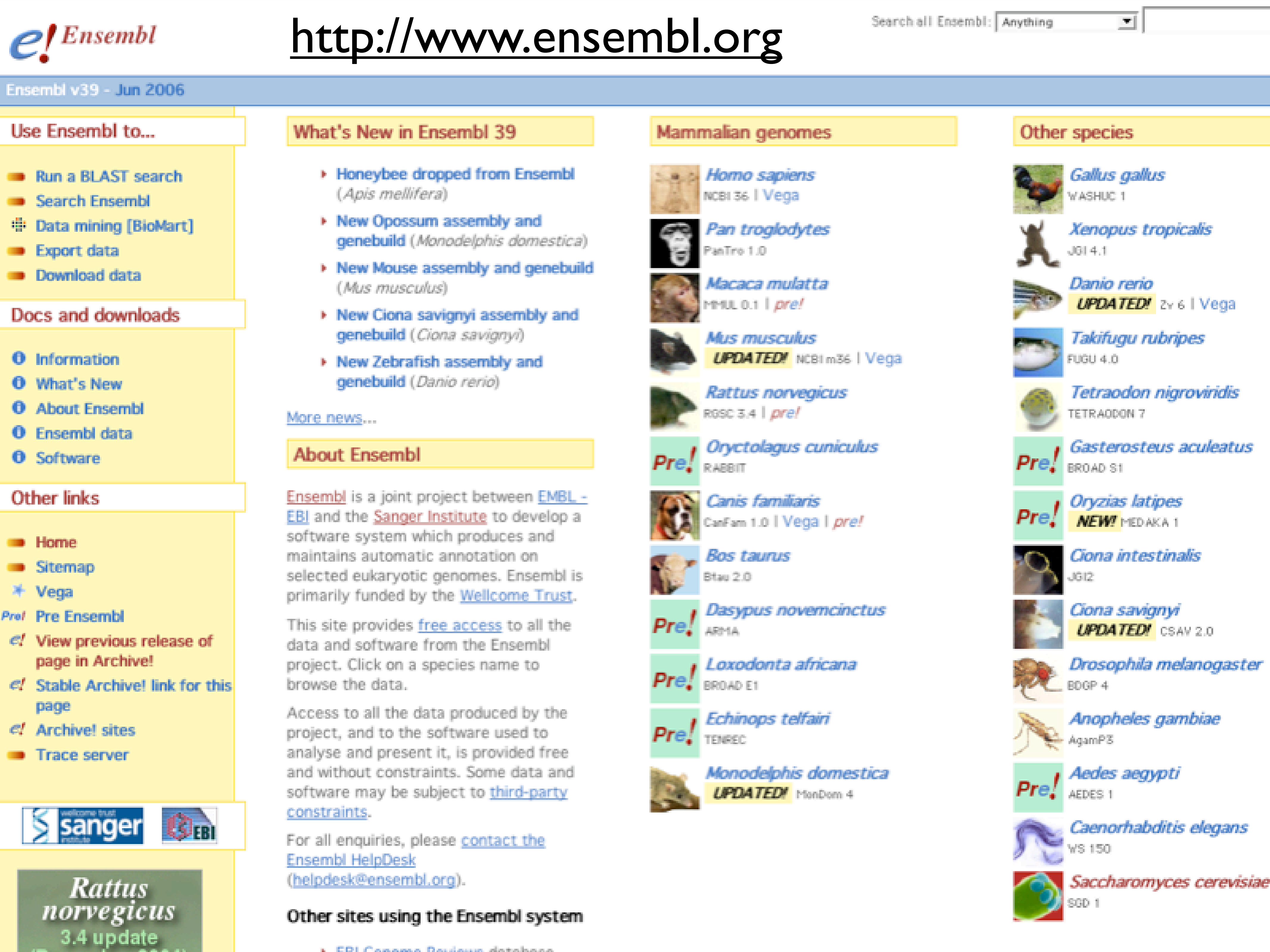
get sequence

get features

a few principles...

- Clients are “intelligent” (few)
- Servers are simple and easy to set up (many)
- (most of) data is precalculated
- libraries for server and client
- multiple programming languages

http://www.ensembl.org



Ensembl v39 - Jun 2006

Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Export data
- Download data

Docs and downloads

- Information
- What's New
- About Ensembl
- Ensembl data
- Software

Other links

- Home
- Sitemap
- Vega
- Pre Ensembl
- View previous release of page in Archive!
- Stable Archive! link for this page
- Archive! sites
- Trace server

What's New in Ensembl 39

- Honeybee dropped from Ensembl (*Apis mellifera*)
- New Opossum assembly and genebuild (*Monodelphis domestica*)
- New Mouse assembly and genebuild (*Mus musculus*)
- New Ciona savignyi assembly and genebuild (*Ciona savignyi*)
- New Zebrafish assembly and genebuild (*Danio rerio*)

More news...

About Ensembl

Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. Ensembl is primarily funded by the Wellcome Trust.

This site provides free access to all the data and software from the Ensembl project. Click on a species name to browse the data.

Access to all the data produced by the project, and to the software used to analyse and present it, is provided free and without constraints. Some data and software may be subject to third-party constraints.

For all enquiries, please contact the Ensembl HelpDesk (helpdesk@ensembl.org).

Other sites using the Ensembl system

- EBI Genome Reviews database

Mammalian genomes

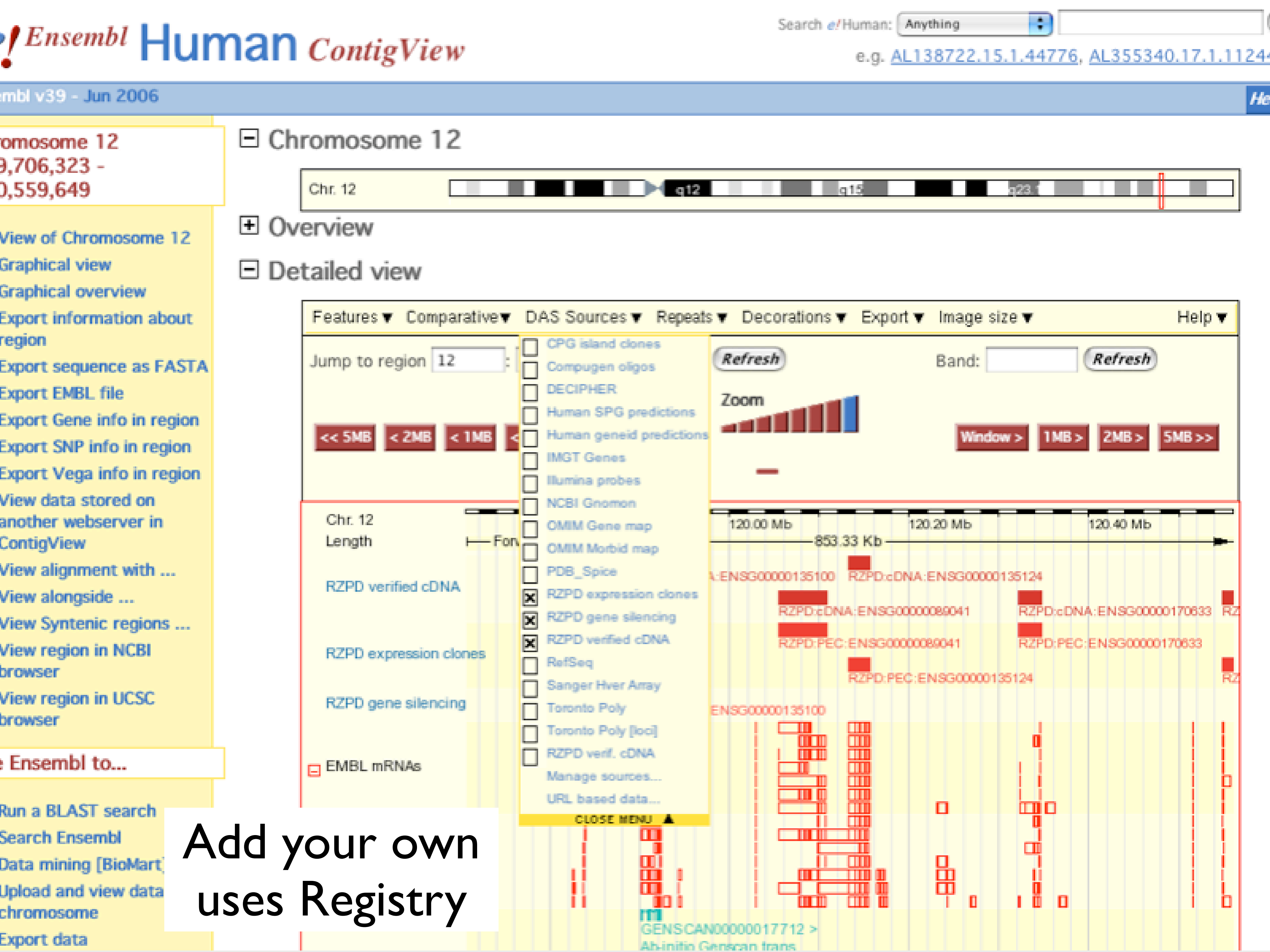
- *Homo sapiens* — NCBI 36 | Vega
- *Pan troglodytes* — PanTro 1.0
- *Macaca mulatta* — MMUL 0.1 | pre!
- *Mus musculus* — UPDATED! NCBI m36 | Vega
- *Rattus norvegicus* — RGSC 3.4 | pre!
- Pre! *Oryctolagus cuniculus* — RABBIT
- *Canis familiaris* — CanFam 1.0 | Vega | pre!
- *Bos taurus* — Btau 2.0
- Pre! *Dasypus novemcinctus* — ARMA
- Pre! *Loxodonta africana* — BROAD E1
- Pre! *Echinops telfairi* — TENREC
- *Monodelphis domestica* — UPDATED! MonDom 4

Other species

- *Gallus gallus* — WASHUC 1
- *Xenopus tropicalis* — JGI 4.1
- *Danio rerio* — UPDATED! Zv 6 | Vega
- *Takifugu rubripes* — FUGU 4.0
- *Tetraodon nigroviridis* — TETRAODON 7
- Pre! *Gasterosteus aculeatus* — BROAD S1
- Pre! *Oryzias latipes* — NEW! MEDAKA 1
- *Ciona intestinalis* — JGI2
- *Ciona savignyi* — UPDATED! CSAV 2.0
- *Drosophila melanogaster* — BDGP 4
- *Anopheles gambiae* — AgamP3
- Pre! *Aedes aegypti* — AEDES 1
- *Caenorhabditis elegans* — WS 150
- *Saccharomyces cerevisiae* — SGD 1

- > 20 vertebrates / model organism

- 5 mill. page impressions / week
- 100 mirrors/internal installations worldwide
- open source
- used for other species as well
- MySQL
- 5-10 G / species + 100 G multi species data



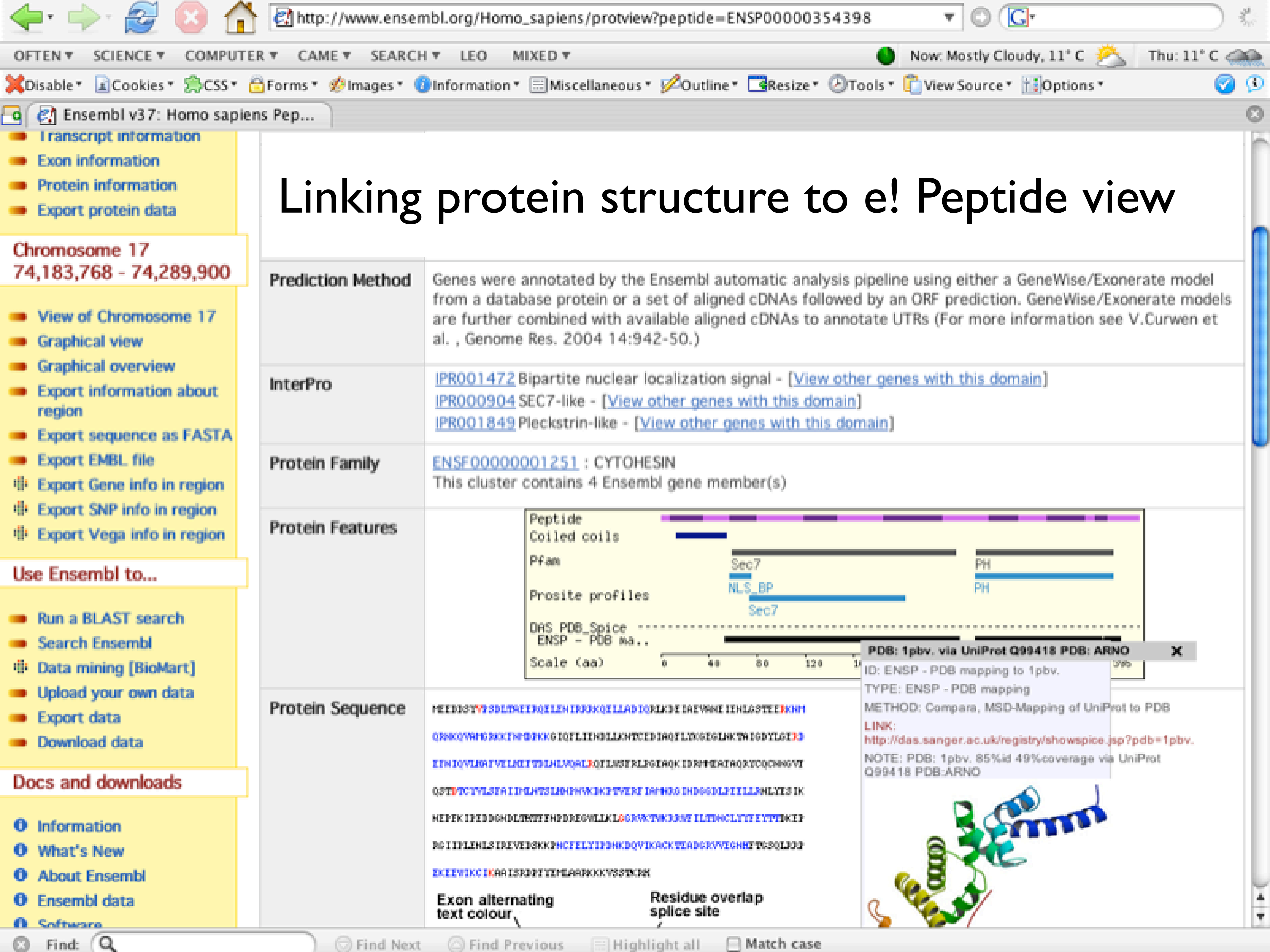
□ CPG island clones
□ Compugen oligos
□ DECIPHER
□ Human SPG predictions
□ Human geneid predictions
□ IMGT Genes
□ Illumina probes
□ NCBI Gnomon
□ OMIM Gene map
□ OMIM Morbid map
□ PDB_Spice
☒ RZPD expression clones
☒ RZPD gene silencing
☒ RZPD verified cDNA
□ RefSeq
□ Sanger Hver Array
□ Toronto Poly
□ Toronto Poly [loci]
□ RZPD verif. cDNA

Manage sources...

URL based data...

Add your own uses Registry

Linking protein structure to e! Peptide view



Prediction Method	Genes were annotated by the Ensembl automatic analysis pipeline using either a GeneWise/Exonerate model from a database protein or a set of aligned cDNAs followed by an ORF prediction. GeneWise/Exonerate models are further combined with available aligned cDNAs to annotate UTRs (For more information see V.Curwen et al. , Genome Res. 2004 14:942-50.)
InterPro	IPR001472 Bipartite nuclear localization signal - [View other genes with this domain] IPR000904 SEC7-like - [View other genes with this domain] IPR001849 Pleckstrin-like - [View other genes with this domain]
Protein Family	ENSF00000001251 : CYTOHESIN This cluster contains 4 Ensembl gene member(s)
Protein Features	
Protein Sequence	MEEDDSYVPSDLTAEEIQQILENIRRRKQILLADIQRLKDEIAEVANEIENLGSTEERKNM QRNKQVAMGRKKFNMDPKKGIQFLIENDLLKNTCEDIAQFLYKGEGLNKTAIGDYLGERD EFNIQVLHAFVELHEFTDLNLVQALRQFLWSFRLPGEAQKIDRMMEAIAQRYCQCNNGVF QSTDTCYVLSFAIIMLNTSLHNPNVKDKPTVERFIAMNRGINDGGDLPEELLRNLYESIK NEPFKIPEDDGNDLTHTFFNPDREGWLLKLGGRVKTWKRRWFILTDNCLYYFEYTTDKEP RGIIPLENLSIREVEDSKKPNCFELYIPNNKDQVIKACKTEADGRVVEGNHTVYRISAPT EKEEWIKCIKAAISRKPFYEMLAARKKKVSSTKRH

Exon alternating text colour

Residue overlap splice site

PDB: 1pbv. via UniProt Q99418 PDB: ARNO X

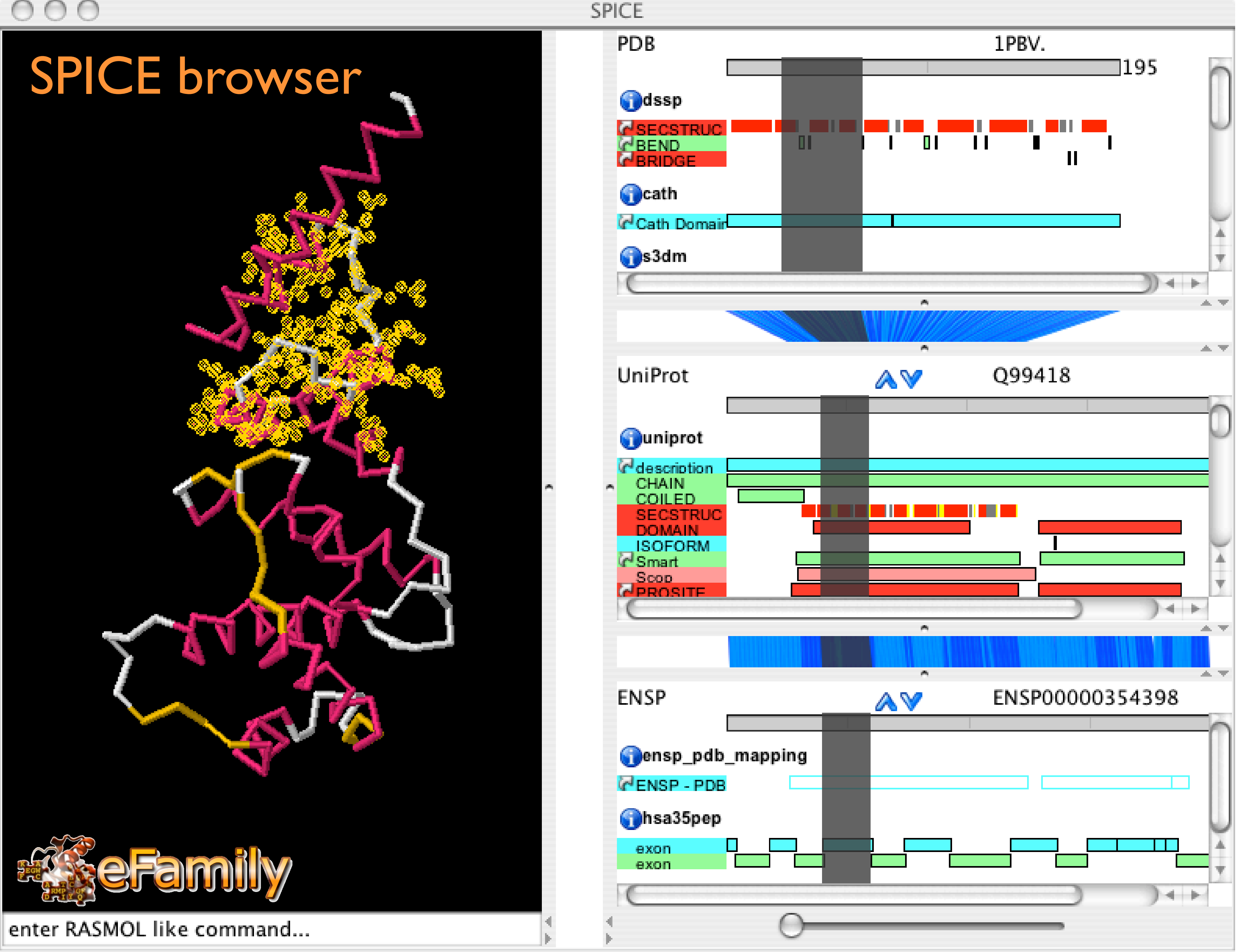
ID: ENSP - PDB mapping to 1pbv.

TYPE: ENSP - PDB mapping

METHOD: Compara, MSD-Mapping of UniProt to PDB

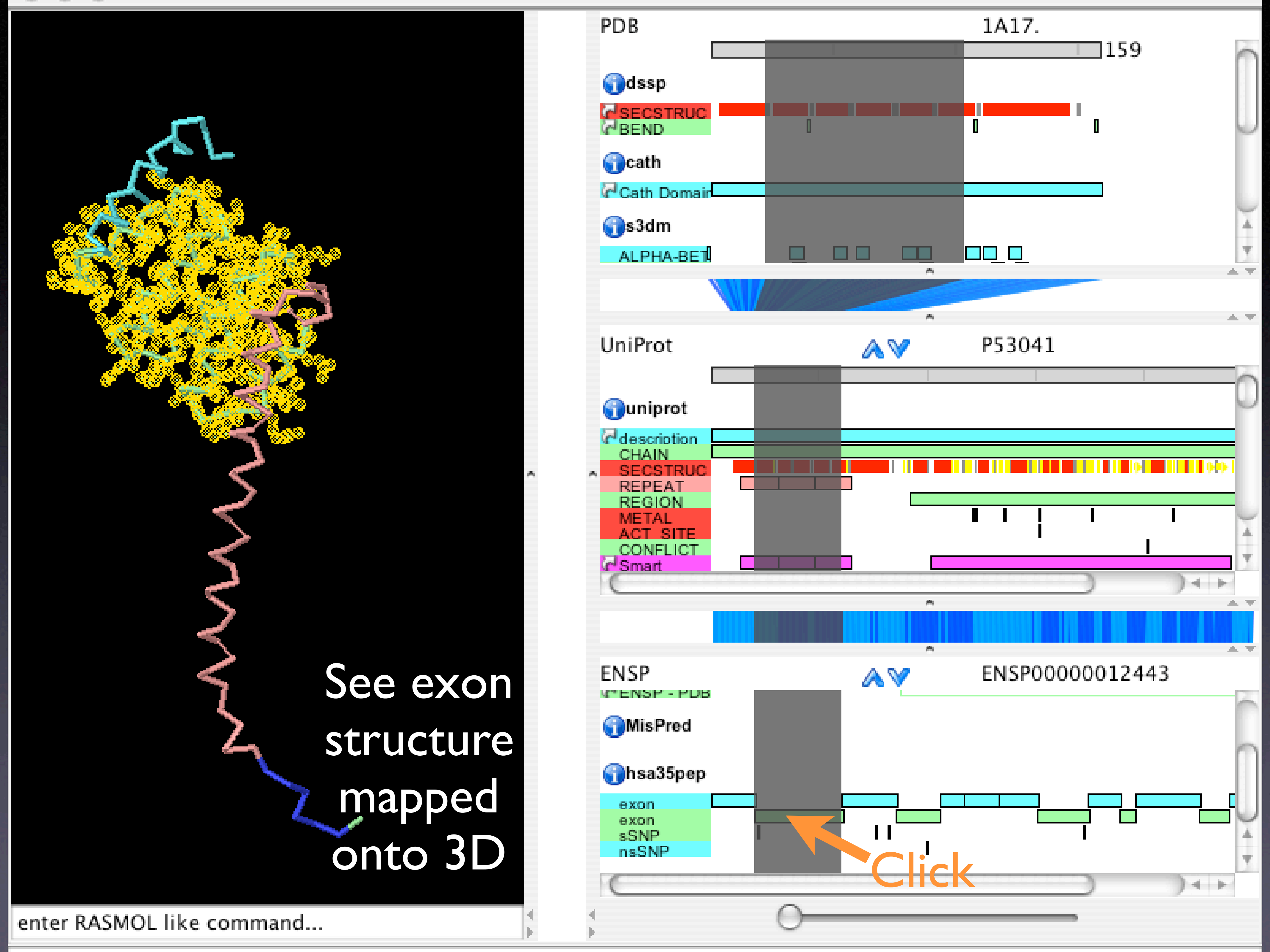
LINK: http://das.sanger.ac.uk/registry/showspice.jsp?pdb=1pbv.

NOTE: PDB: 1pbv. 85%id 49%coverage via UniProt Q99418 PDB:ARNO



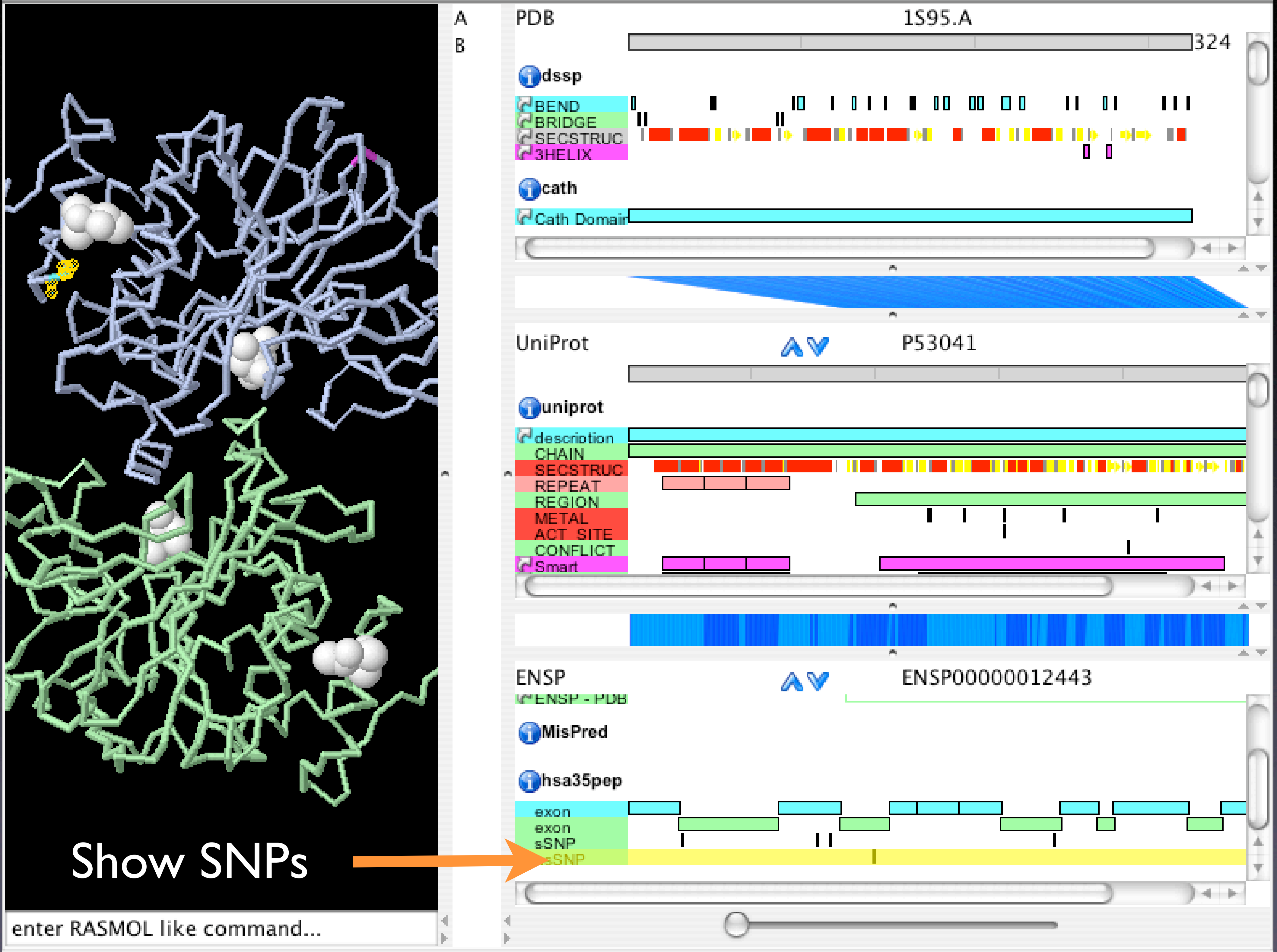
SPICE browser

http://www.efamily.org.uk/software/dasclients/spice



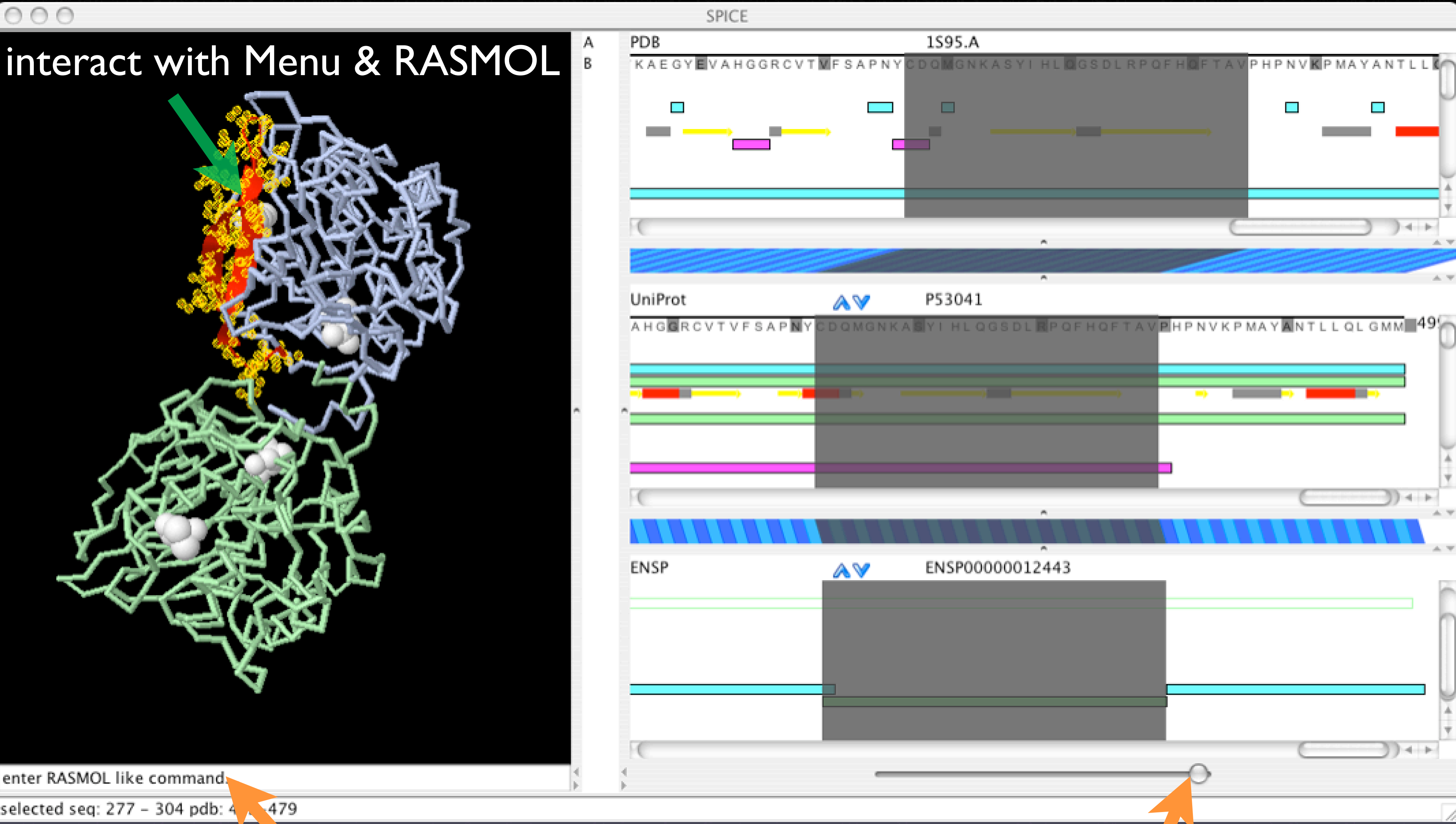
See exon structure mapped onto 3D

Click


A
B
PDB
1S95.A
324
dssp
BEND
BRIDGE
SECSTRUC
3HELIX
cath
Cath Domain
UniProt
P53041
uniprot
description
CHAIN
SECSTRUC
REPEAT
REGION
METAL
ACT_SITE
CONFLICT
Smart
ENSP
ENSP00000012443
ENSP - PDB
MisPred
hsa35pep
exon
exon
sSNP
nsSNP
enter RASMOL like command...

Show SNPs

interact with Menu & RASMOL

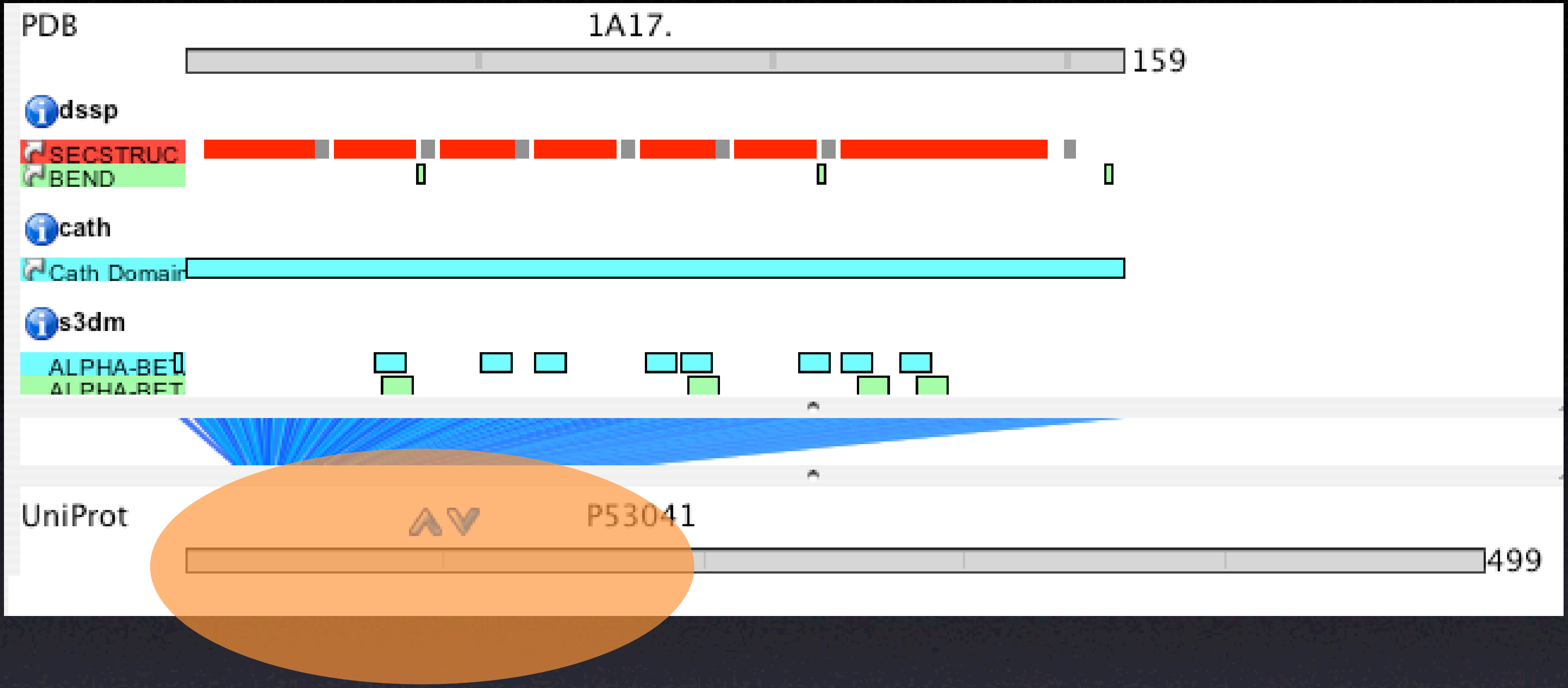

SPICE
A
B
PDB
1S95.A
KAEGYEVAHGGRCVTVFSAPNYCDQMGNKASYIHLQGSDLRPQFHQFTAVPHPNVKPMAYANTLLQ
UniProt
P53041
AHGGRCVTVFSAPNYCDQMGNKASYIHLQGSDLRPQFHQFTAVPHPNVKPMAYANTLLQLGMM
49
ENSP
ENSP00000012443

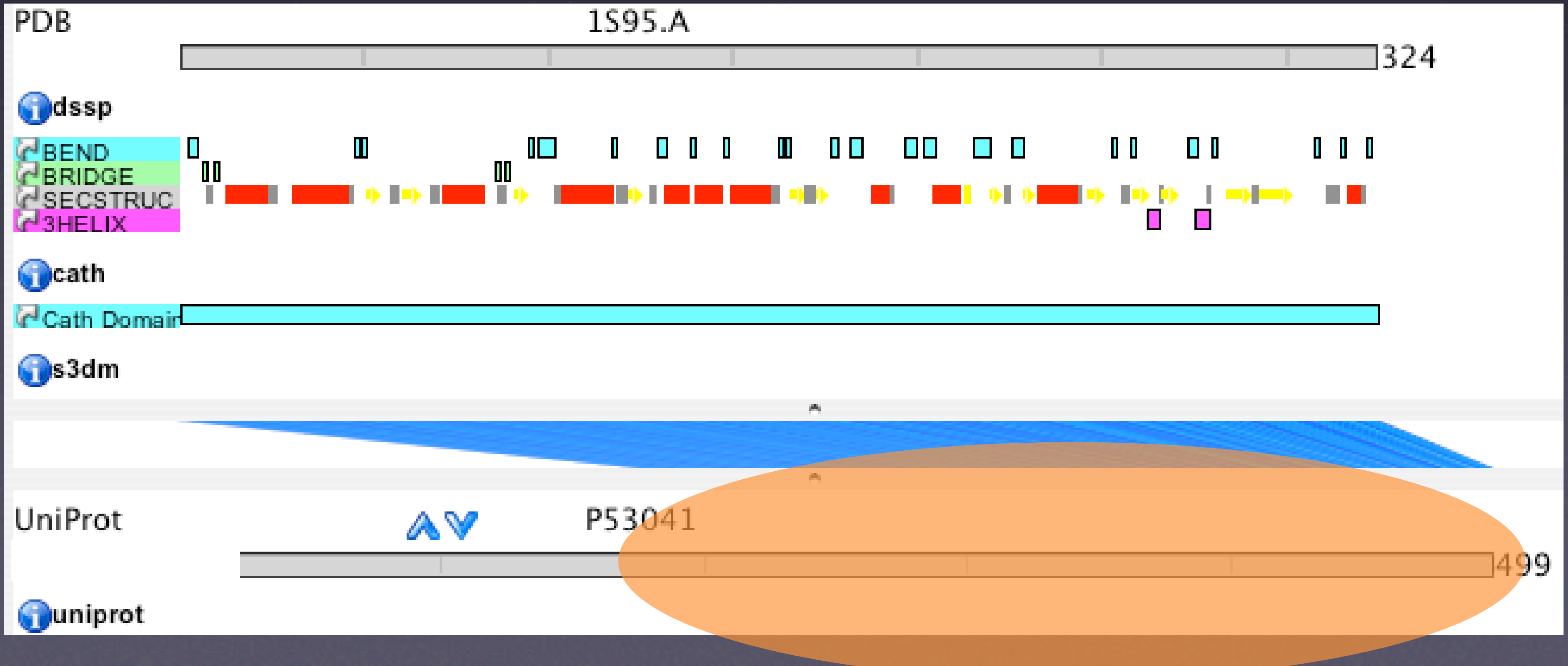
enter RASMOL like command

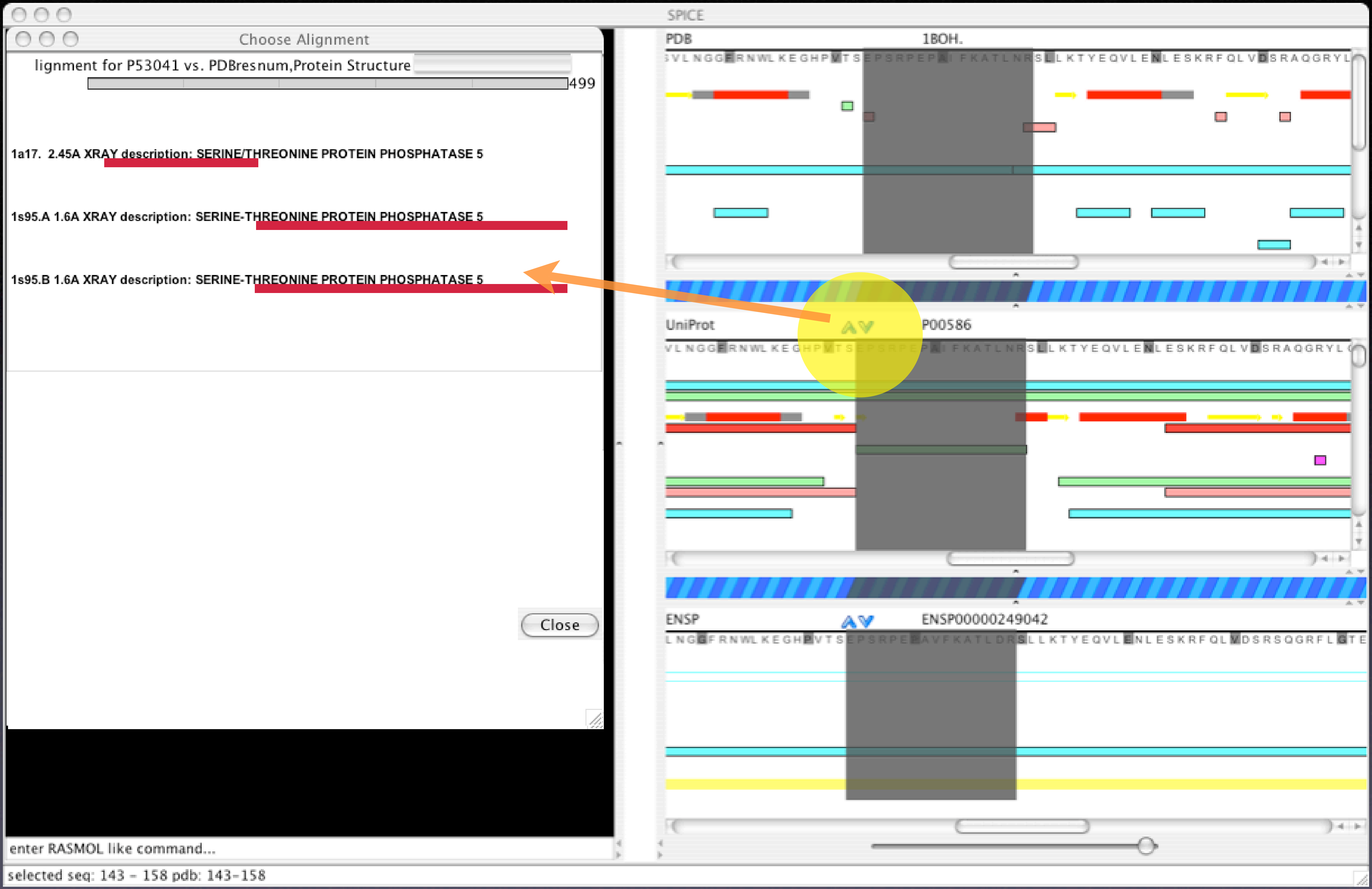
selected seq: 277 – 304 pdb: 4 -479

RASMOL commands

Zoom


PDB
1A17.
159
dssp
SECSTRUC
BEND
cath
Cath Domain
s3dm
ALPHA-BET
ALPHA-BET
UniProt
P53041
499


PDB
1S95.A
324
dssp
BEND
BRIDGE
SECSTRUC
3HELIX
cath
Cath Domain
s3dm
UniProt
P53041
499
uniprot



SPICE

Choose Alignment

lignment for P53041 vs. PDBresnum,Protein Structure

1a17. 2.45A XRAY description: SERINE/THREONINE PROTEIN PHOSPHATASE 5

1s95.A 1.6A XRAY description: SERINE-THREONINE PROTEIN PHOSPHATASE 5

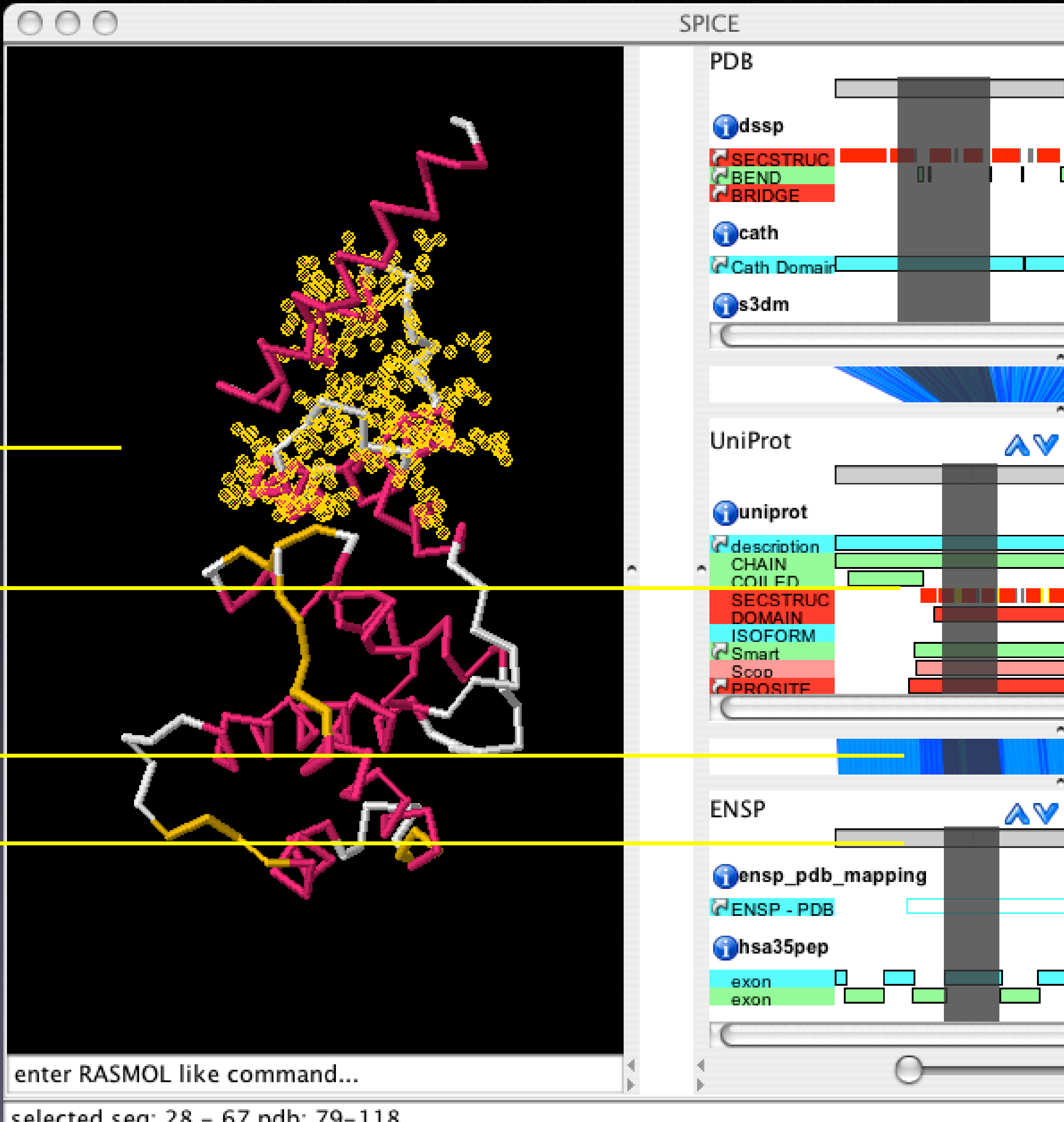
1s95.B 1.6A XRAY description: SERINE-THREONINE PROTEIN PHOSPHATASE 5

enter RASMOL like command...

selected seq: 143 - 158 pdb: 143-158

499
Close
PDB
1BOH.
SVLNGGERNWLKEGHPVTSEPSRPEPAIFKATLNRSLLKTYEQVLENLESKRFQLVDSRAQGRYL
UniProt
P00586
VLNGGERNWLKEGHPVTSEPSRPEPAIFKATLNRSLLKTYEQVLENLESKRFQLVDSRAQGRYL
ENSP
ENSP00000249042
LNGGFRNWLKEGHPVTSEPSRPEPAVFKATLDRSLLKTYEQVLENLESKRFQLVDSRSQGRFLGTE

DAS commands

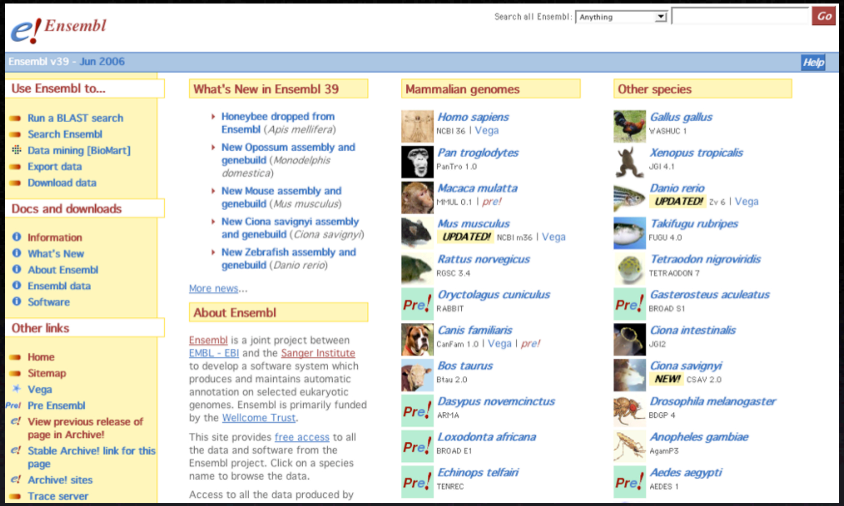


Structure

Features

Alignment

Sequence


e! Ensembl

Search all Ensembl: Anything ▾ Go

Ensembl v39 - Jun 2006 Help

Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Export data
- Download data

Docs and downloads

- Information
- What's New
- About Ensembl
- Ensembl data
- Software

Other links

- Home
- Sitemap
- Vega
- Pre Ensembl
- View previous release of page in Archive!
- Stable Archive! link for this page
- Archive! sites
- Trace server

What's New in Ensembl 39

- Honeybee dropped from Ensembl (*Apis mellifera*)
- New Opossum assembly and genebuild (*Monodelphis domestica*)
- New Mouse assembly and genebuild (*Mus musculus*)
- New Ciona savignyi assembly and genebuild (*Ciona savignyi*)
- New Zebrafish assembly and genebuild (*Danio rerio*)

More news...

About Ensembl

Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. Ensembl is primarily funded by the Wellcome Trust.

This site provides free access to all the data and software from the Ensembl project. Click on a species name to browse the data.

Access to all the data produced by

Mammalian genomes

Homo sapiens
NCBI 36 | Vega

Pan troglodytes
PanTro 1.0

Macaca mulatta
MMUL 0.1 | *pre!*

Mus musculus
UPDATED! NCBI m36 | Vega

Rattus norvegicus
RGSC 3.4

Pre! *Oryctolagus cuniculus*
RABBIT

Canis familiaris
CanFam 1.0 | Vega | *pre!*

Bos taurus
Btau 2.0

Pre! *Dasypus novemcinctus*
ARMA

Pre! *Loxodonta africana*
BROAD E1

Pre! *Echinops telfairi*
TENREC

Other species

Gallus gallus
WASHUC 1

Xenopus tropicalis
JGI 4.1

Danio rerio
UPDATED! Zv 6 | Vega

Takifugu rubripes
FUGU 4.0

Tetraodon nigroviridis
TETRAODON 7

Pre! *Gasterosteus aculeatus*
BROAD S1

Ciona intestinalis
JGI2

Ciona savignyi
NEW! CSAV 2.0

Drosophila melanogaster
BDGP 4

Anopheles gambiae
AgamP3

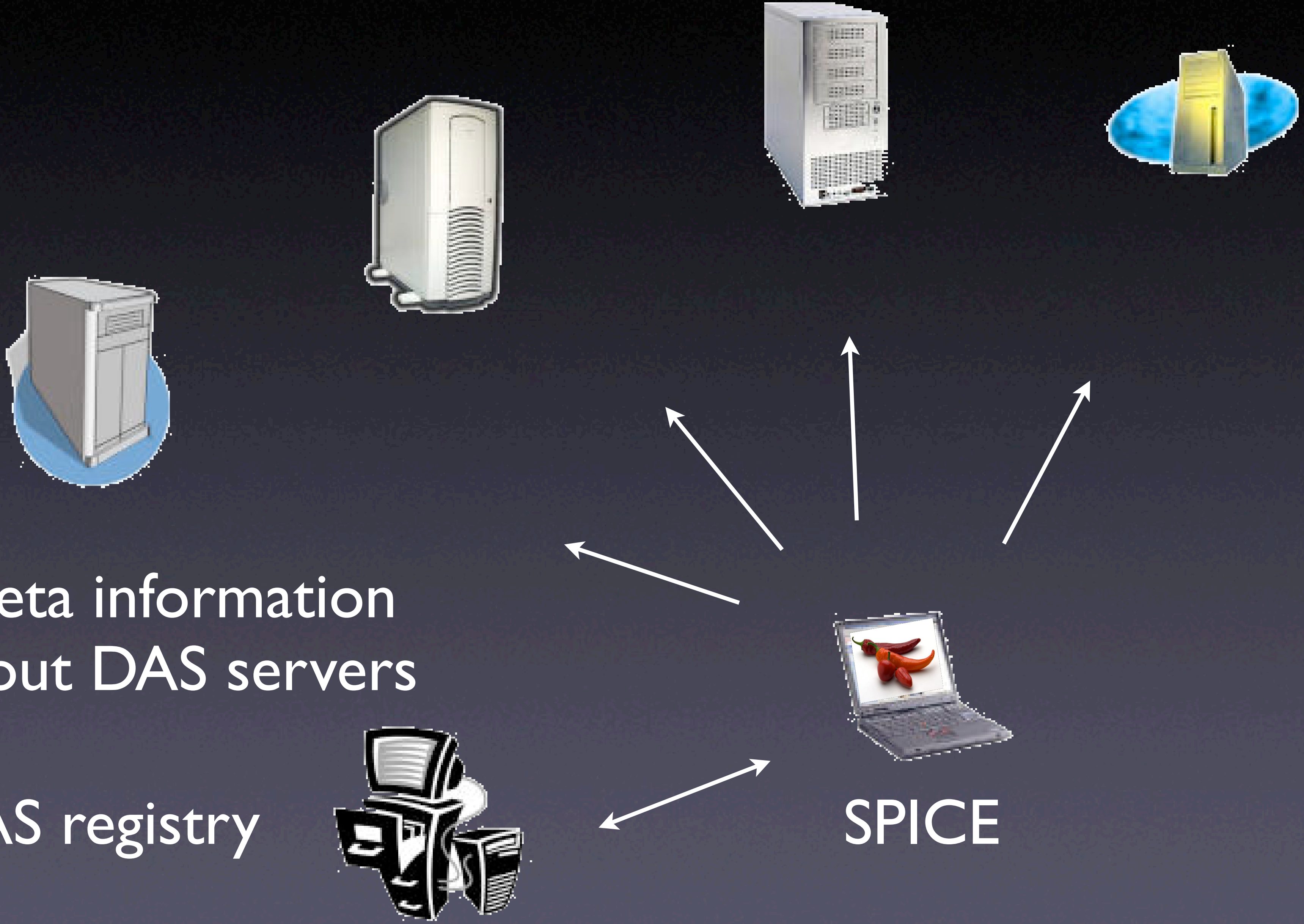
Pre! *Aedes aegypti*
AEDES 1

Java Web Start

auto install
latest version
send arguments

SPICE

DAS registry

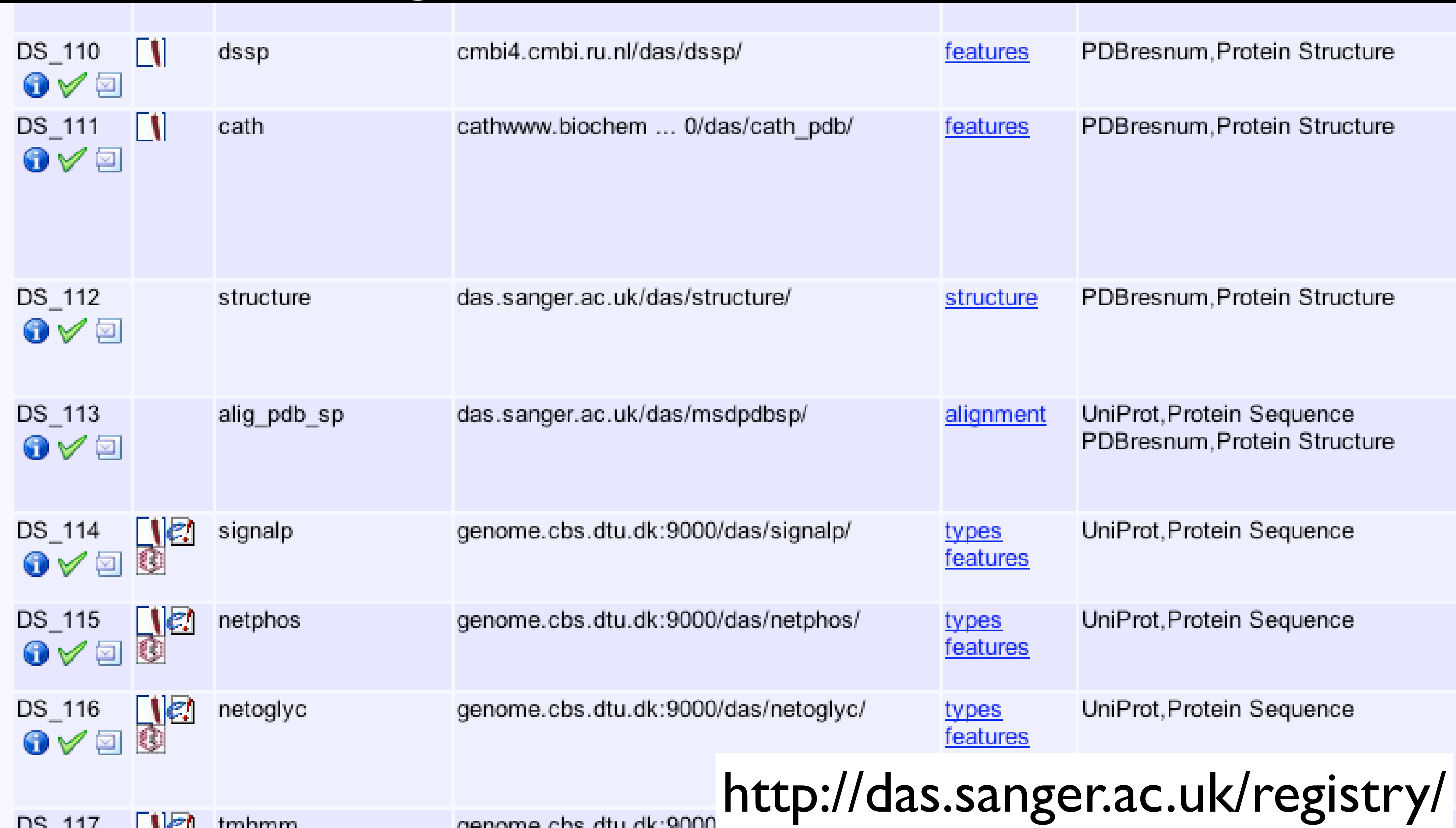


Meta information about DAS servers

DAS registry

SPICE

The DAS registration server



DS_110 ℹ ✓ ✉		dssp	cmbi4.cmbi.ru.nl/das/dssp/	features	PDBresnum,Protein Structure
DS_111 ℹ ✓ ✉		cath	cathwww.biochem ... 0/das/cath_pdb/	features	PDBresnum,Protein Structure
DS_112 ℹ ✓ ✉		structure	das.sanger.ac.uk/das/structure/	structure	PDBresnum,Protein Structure
DS_113 ℹ ✓ ✉		alig_pdb_sp	das.sanger.ac.uk/das/msdpdbsp/	alignment	UniProt,Protein Sequence PDBresnum,Protein Structure
DS_114 ℹ ✓ ✉		signalp	genome.cbs.dtu.dk:9000/das/signalp/	types features	UniProt,Protein Sequence
DS_115 ℹ ✓ ✉		netphos	genome.cbs.dtu.dk:9000/das/netphos/	types features	UniProt,Protein Sequence
DS_116 ℹ ✓ ✉		netoglyc	genome.cbs.dtu.dk:9000/das/netoglyc/	types features	UniProt,Protein Sequence
DS_117		tmhmm	genome.cbs.dtu.dk:9000		

http://das.sanger.ac.uk/registry/

DAS registration server

- allows to “publish” DAS servers & share with community
- communicates with clients
- regularly checks servers, sends notification

What is the glue?

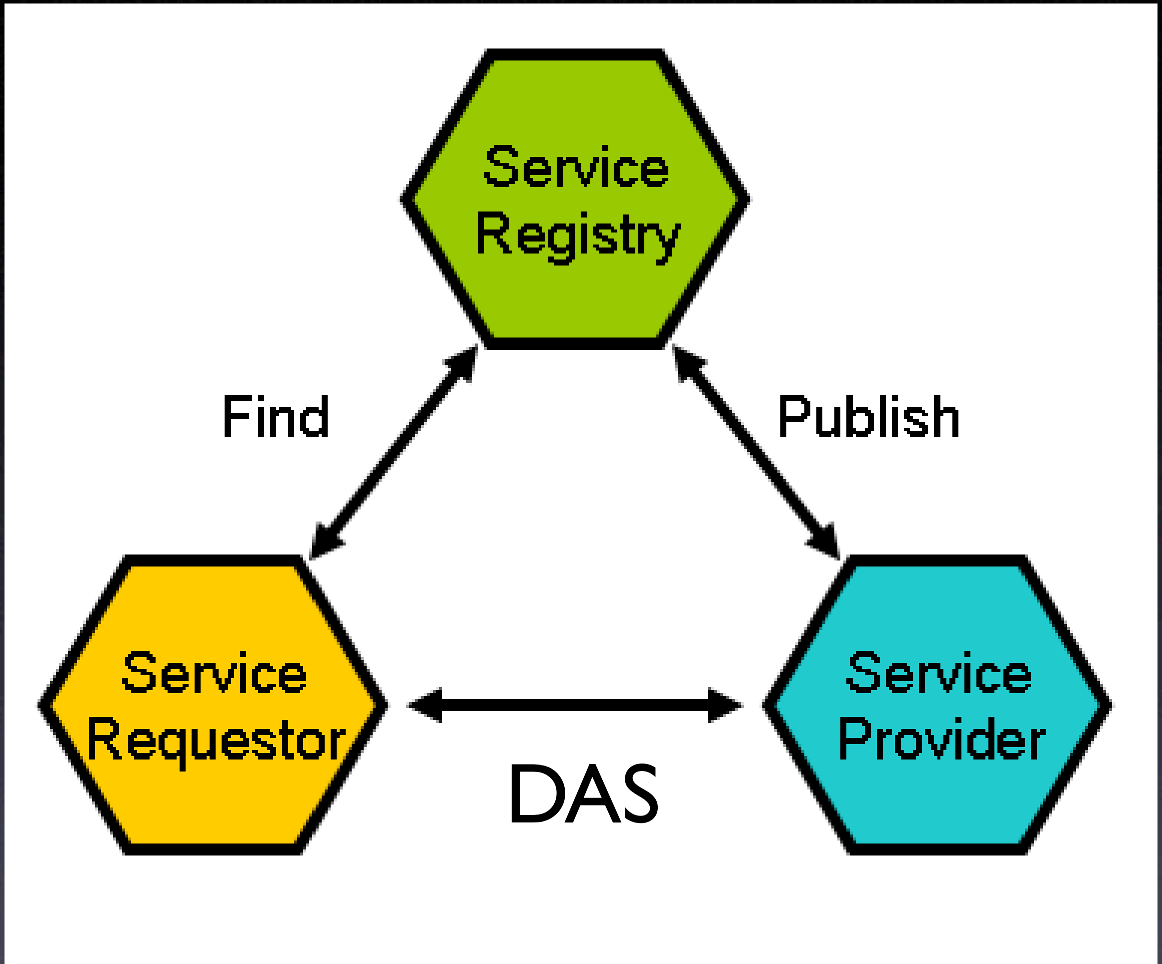
- “Coordinate Systems”
  - Authority
  - Type of data
  - Version
  - Organism (optional)

Clients and Coordinate Systems

- Ensembl - most of the views can display DAS sources from multiple CS
- SPICE - PDB, UniProt, Ensp
- Dasty - UniProt

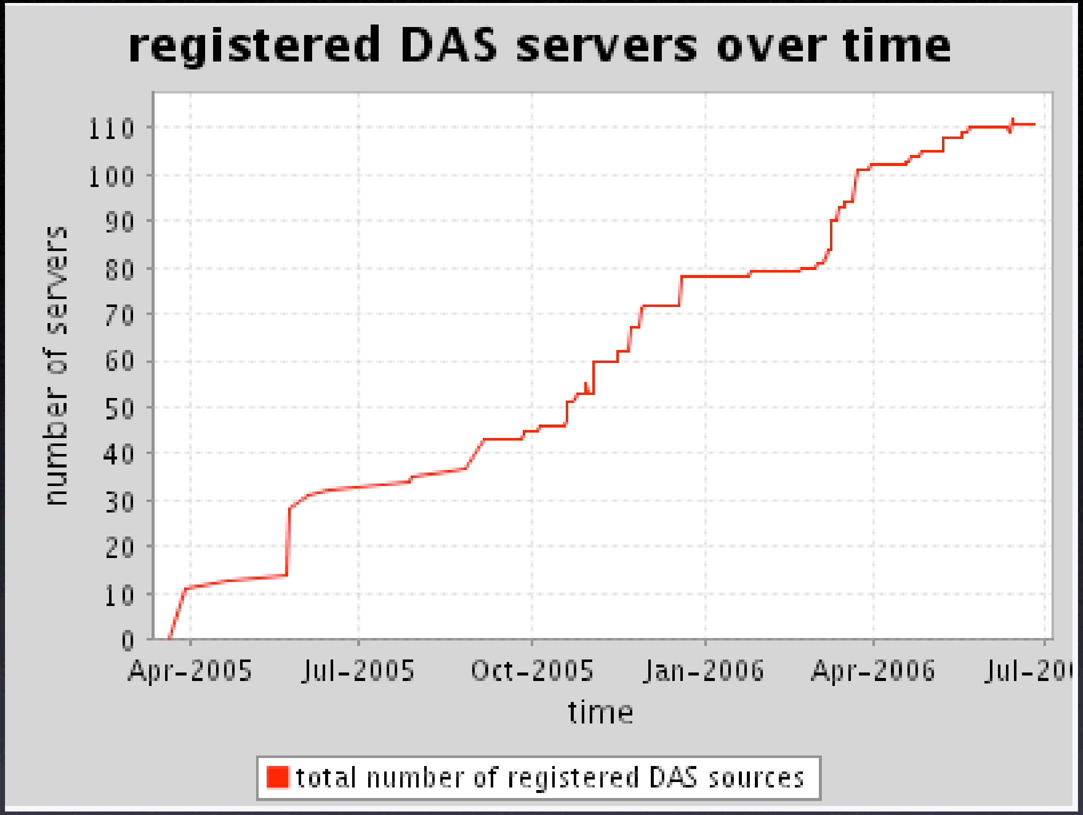
DAS registration server

e.g. Ensembl, SPICE



a DAS source

the DAS - SOA



111 DAS sources
26 institutions
12 countries

eFamily

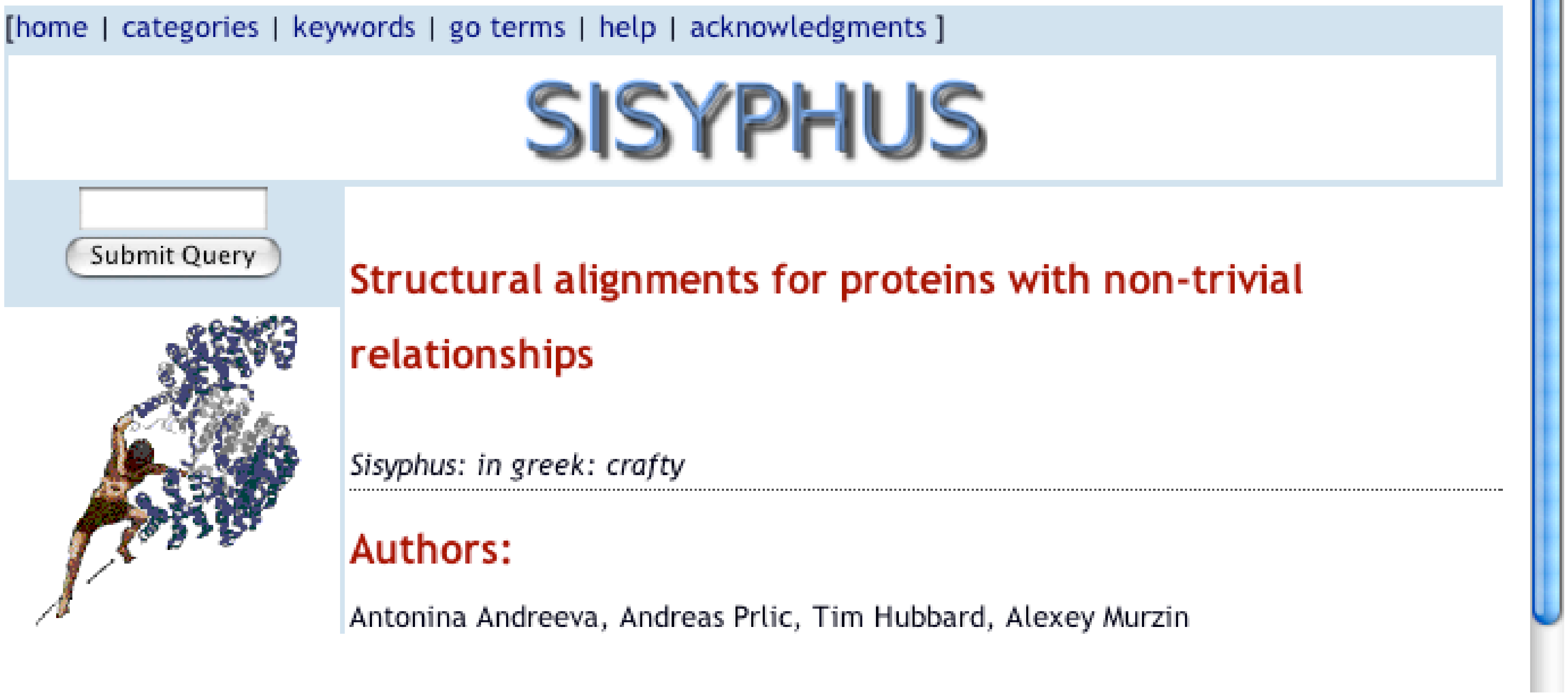
BIOSAPIENS NETWORK

+ others

DAS - issues

- inconsistent implementations
- no consistent use annotation types
- error handling
- searches not possible - in DAS/1
- open sharing of data - low security

http://sisyphus.mrc-cpe.cam.ac.uk


[home | categories | keywords | go terms | help | acknowledgments]
Submit Query

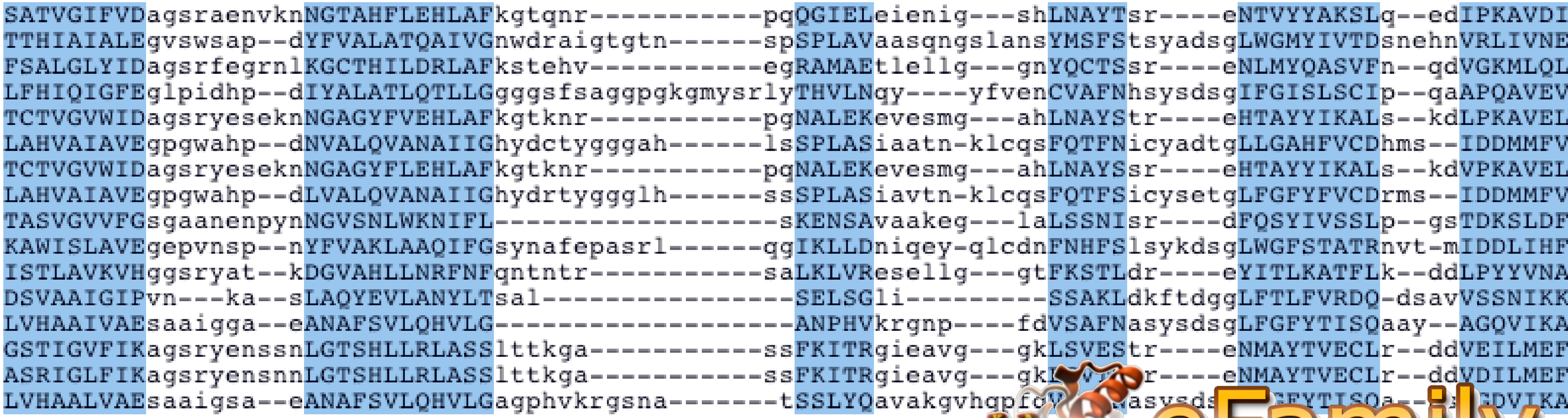
SISYPHUS

Structural alignments for proteins with non-trivial relationships

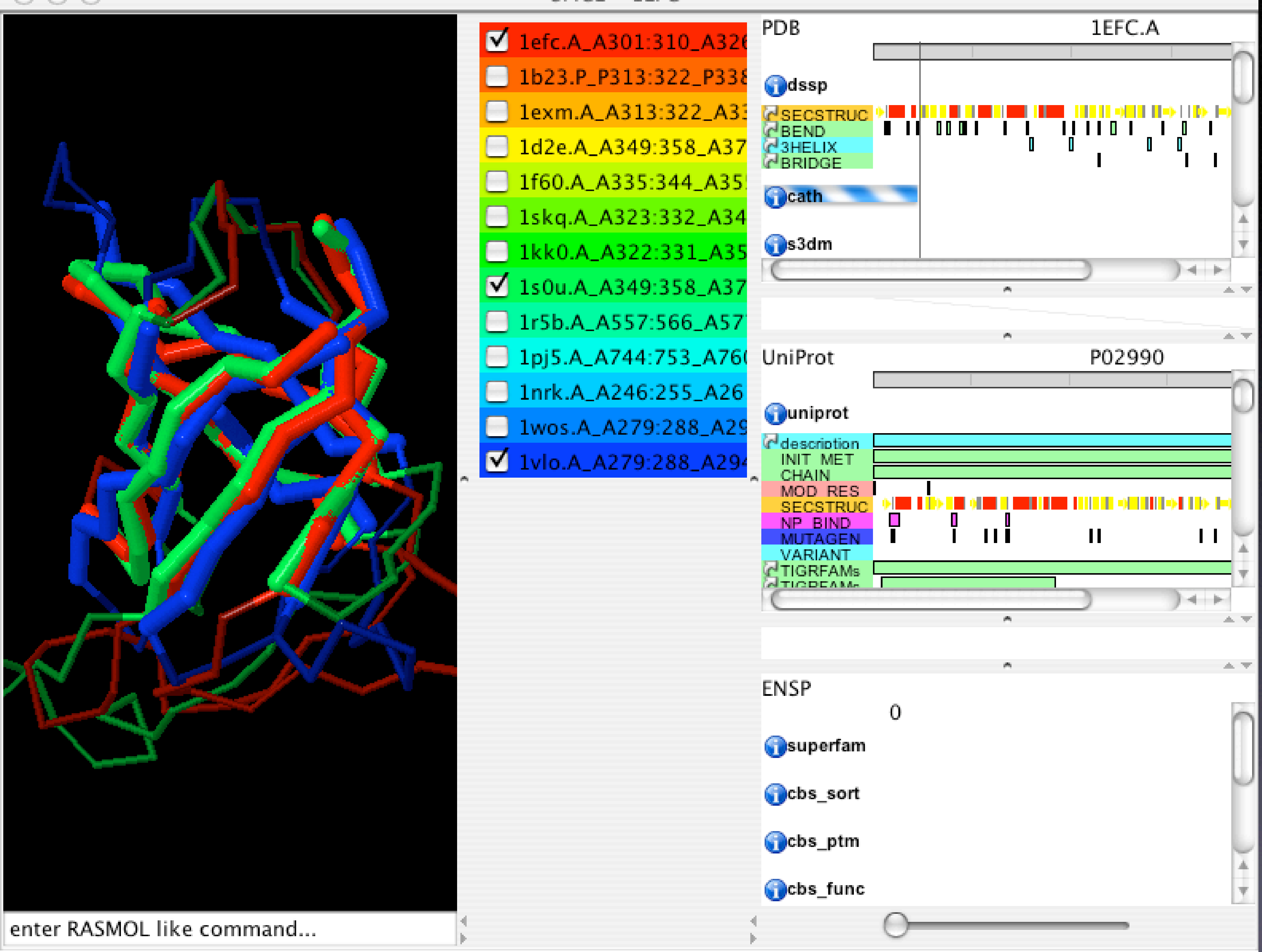
Sisyphus: in greek: crafty

Authors:

Antonina Andreeva, Andreas Prlic, Tim Hubbard, Alexey Murzin

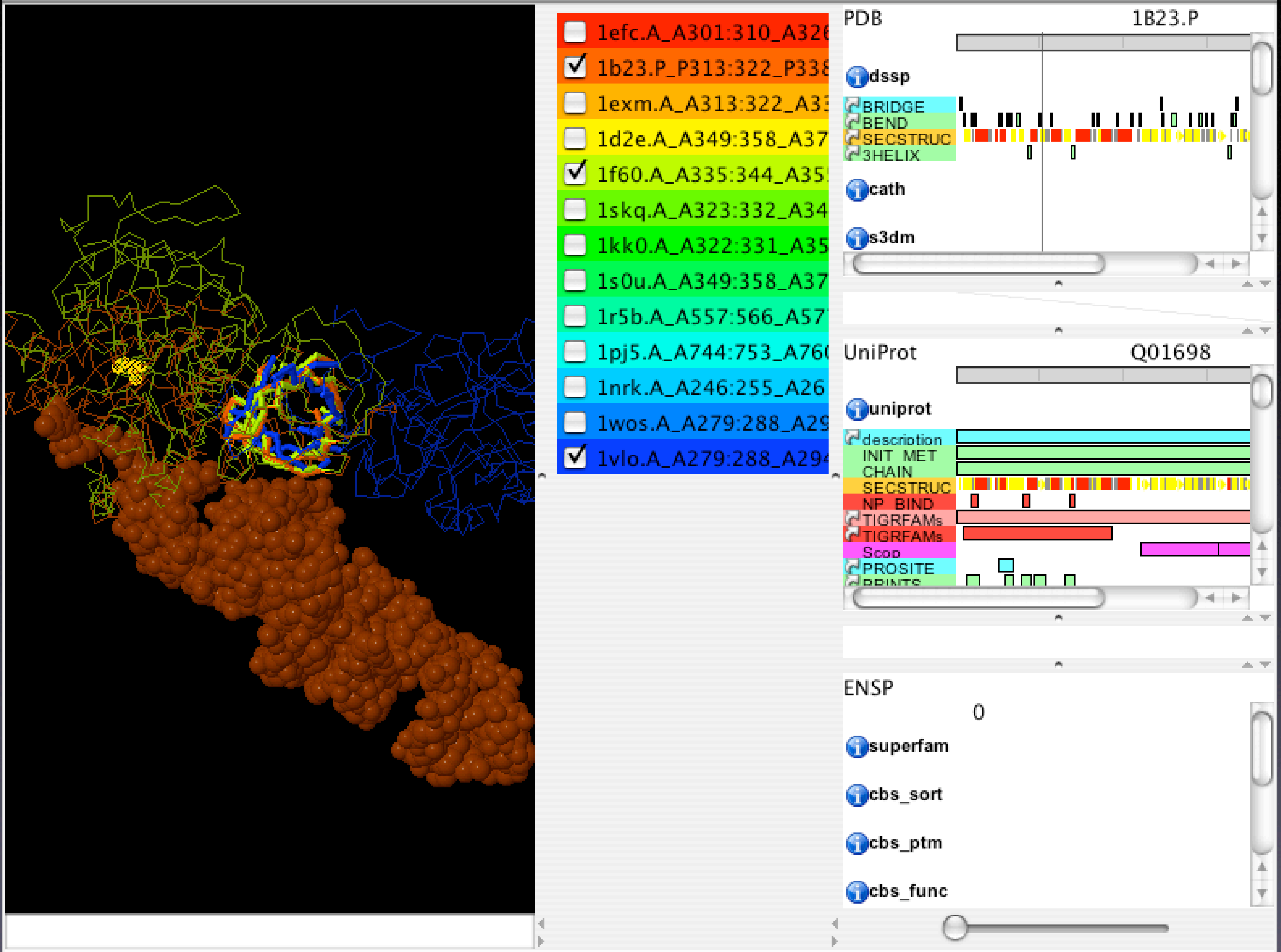


```
SATVGIFVDagsraenvknNGTAHFLEHLAFkgtqnr-----------pqQGIELeienig---shLNAYTsr----eNTVYYAKSLq--edIPKAVDI
TTHIAIALEgvswsap--dYFVALATQAIVGnwdraigtgtn------spSPLAVaasgngslansYMSFStsyadsgLWGMYIVTDsnehnVRLIVNE
FSALGLYIDagsrfegrnlKGCTHILDRLAFkstehv-----------egRAMAEtlellg---gnYQCTSsr----eNLMYQASVFn--qdVGKMLQL
LFHIQIGFEglpidhp--dIYALATLQTLLGgggsfsaggpgkgmysrlyTHVLNqy----yfvenCVAFNhsysdsgIFGISLSCIp--qaAPQAVEV
TCTVGVWIDagsryeseknNGAGYFVEHLAFkgtknr-----------pgNALEKevesmg---ahLNAYStr----eHTAYYIKALs--kdLPKAVEL
LAHVAIAVEgpgwahp--dNVALQVANAIIGhydctygggah------lsSPLASiaatn-klcqsFQTFNicyadtgLLGAHFVCDhms--IDDMMFV
TCTVGVWIDagsryeseknNGAGYFLEHLAFkgtknr-----------pqNALEKevesmg---ahLNAYSsr----eHTAYYIKALs--kdVPKAVEL
LAHVAIAVEgpgwahp--dLVALQVANAIIGhydrtygglh-------ssSPLASiavtn-klcqsFQTFSicysetgLFGFYFVCDrms--IDDMMFV
TASVGVVFGsgaanenpynNGVSNLWKNIFL------------------sKENSAvaakeg---laLSSNIsr----dFQSYIVSSLp--gsTDKSLDF
KAWISLAVEgepvnsp--nYFVAKLAAQIFGsynafepasrl------qgIKLLDniqey-qlcdnFNHFSlsykdsgLWGFSTATRnvt-mIDDLIHF
ISTLAVKVHggsryat--kDGVAHLLNRFNFqntntr-----------saLKLVResellg---gtFKSTLdr----eYITLKATFLk--ddLPYYVNA
DSVAAIGIPvn---ka--sLAQYEVLANYLTsal----------------SELSGli---------SSAKLdkftdggLFTLFVRDQ-dsavVSSNIKK
LVHAAIVAEsaaigga--eANAFSVLQHVLG-------------------ANPHVkrgnp----fdVSAFNasysdsgLFGFYTISQaay--AGQVIKA
GSTIGVFIKagsryenssnLGTSHLLRLASSlttkga-----------ssFKITRgieavg---gkLSVEStr----eNMAYTVECLr--ddVEILMEF
ASRIGLFIKagsryensnnLGTSHLLRLASSlttkga-----------ssFKITRgieavg---gk[illegible]r----eNMAYTVECLr--ddVDILMEF
LVHAALVAEsaaigsa--eANAFSVLQHVLGagphvkrgsna-------tSSLYQavakgvhqpfd[illegible]asysds[illegible]GFYTISQa--[illegible]GDVIKA
```



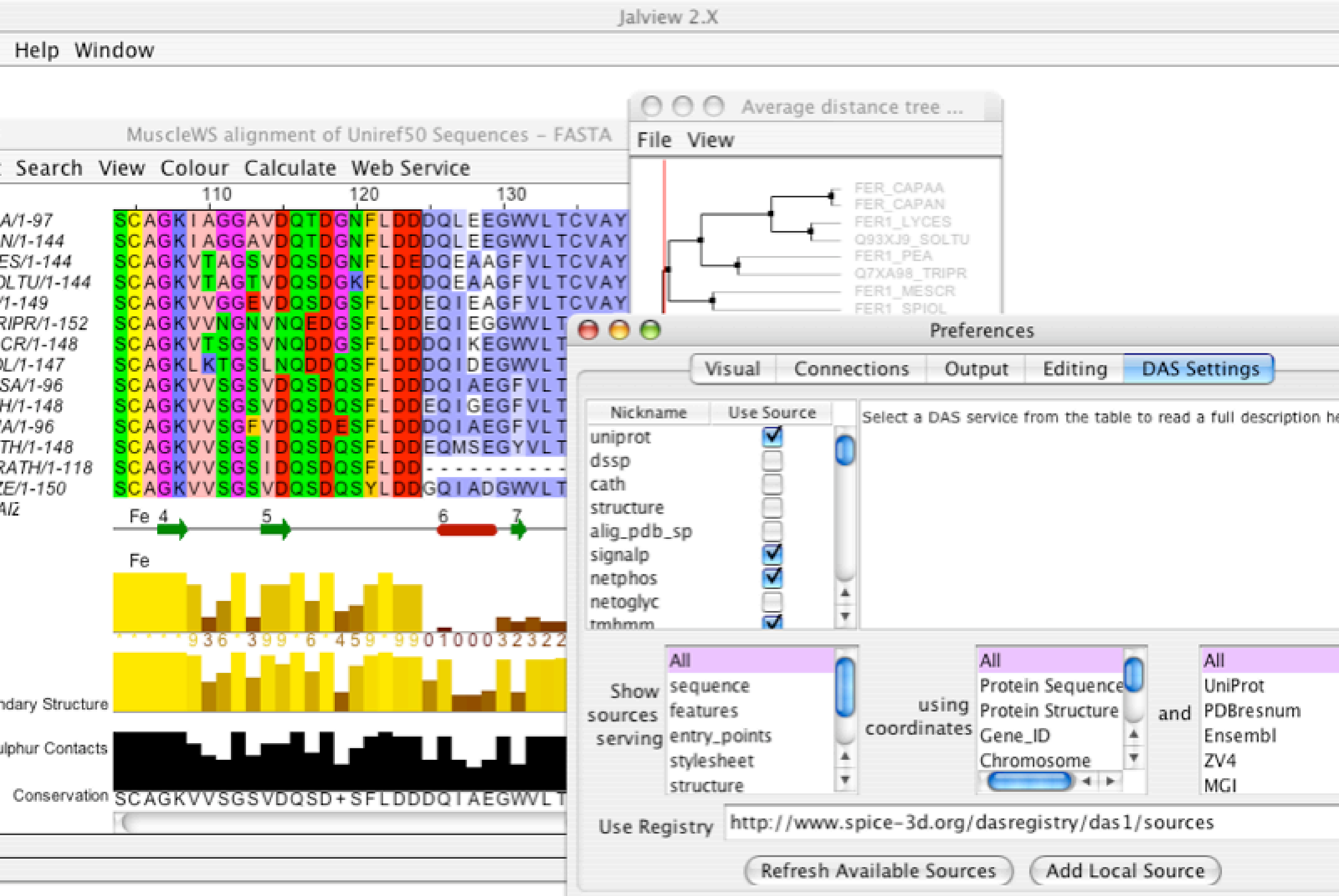
☑ 1efc.A_A301:310_A32
☐ 1b23.P_P313:322_P33
☐ 1exm.A_A313:322_A3
☐ 1d2e.A_A349:358_A37
☐ 1f60.A_A335:344_A35
☐ 1skq.A_A323:332_A34
☐ 1kk0.A_A322:331_A35
☑ 1s0u.A_A349:358_A37
☐ 1r5b.A_A557:566_A57
☐ 1pj5.A_A744:753_A76
☐ 1nrk.A_A246:255_A26
☐ 1wos.A_A279:288_A29
☑ 1vlo.A_A279:288_A29

enter RASMOL like command...



- ☐ 1efc.A_A301:310_A326
- ☑ 1b23.P_P313:322_P338
- ☐ 1exm.A_A313:322_A33
- ☐ 1d2e.A_A349:358_A37
- ☑ 1f60.A_A335:344_A35
- ☐ 1skq.A_A323:332_A34
- ☐ 1kk0.A_A322:331_A35
- ☐ 1s0u.A_A349:358_A37
- ☐ 1r5b.A_A557:566_A57
- ☐ 1pj5.A_A744:753_A76
- ☐ 1nrk.A_A246:255_A26
- ☐ 1wos.A_A279:288_A29
- ☑ 1vlo.A_A279:288_A29

- Alignment DAS:
- rotation matrices, shift vectors
- range information (optional)



```
A/1-97       SCAGKIAGGAVDQTDGNFLDDDQLEEGWVLTCVAY
N/1-144      SCAGKIAGGAVDQTDGNFLDDDQLEEGWVLTCVAY
ES/1-144     SCAGKVTAGSVDQSDGNFLDEDQEAAGFVLTCVAY
OLTU/1-144   SCAGKVTAGTVDQSDGKFLDDDQEAAGFVLTCVAY
/1-149       SCAGKVVGGEVDQSDGSFLDDEQIEAGFVLTCVAY
RIPR/1-152   SCAGKVVNGNVNQEDGSFLDDEQIEGGWVLT
CR/1-148     SCAGKVTSGSVNQDDGSFLDDDQIKEGWVLT
OL/1-147     SCAGKLKTGSLNQDDQSFLDDDQIDEGWVLT
SA/1-96      SCAGKVVSGSVDQSDQSFLDDDQIAEGFVLT
H/1-148      SCAGKVVSGSVDQSDQSFLDDEQIGEGFVLT
A/1-96       SCAGKVVSGFVDQSDESFLDDDQIAEGFVLT
TH/1-148     SCAGKVVSGSIDQSDQSFLDDEQMSEGYVLT
RATH/1-118   SCAGKVVSGSIDQSDQSFLDD----------
ZE/1-150     SCAGKVVSGSVDQSDQSYLDDGQIADGWVLT
```

*****936*399*6*459*990100032322

Conservation SCAGKVVSGSVDQSD+SFLDDDQIAEGWVLT

Nickname	Use Source
uniprot	☑
dssp	☐
cath	☐
structure	☐
alig_pdb_sp	☐
signalp	☑
netphos	☑
netoglyc	☐
tmhmm	☑

Select a DAS service from the table to read a full description he

Show sources serving: All, sequence, features, entry_points, stylesheet, structure

using coordinates: All, Protein Sequence, Protein Structure, Gene_ID, Chromosome

and: All, UniProt, PDBresnum, Ensembl, ZV4, MGI

Use Registry http://www.spice-3d.org/dasregistry/das1/sources

http://www.jalview.org A. Waterhouse, J. Procter, G. Barton

Acknowledgments

- T. Down, T. Hubbard
- Web Team, E. Kulesha, R. Pettett, T. Cox
- eFamily Project
- S. Gräf, A. Kahari, BioSapiens
- A. Murzin, A. Andreeva
- R. Finn, H.Hotz, A.Ahmed
- Jmol, Biojava, MSD, everybody who sets up DAS servers