Current Protocols in Bioinformatics

Exploring Short Linear Motifs using the ELM Database and Tools

Marc Gouw[1], *Hugo Samano*[1], *Kim Van Roey*[1], *Francesca Diella*[1], *Toby J. Gibson*[1], *Holger Dinkel*[1,2]

[1] *Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany*
[2] *Leibniz-Institute on Aging – Fritz Lipmann Institute (FLI), Beutenbergstrasse 11, 07745 Jena, Germany*

Keywords

Linear motifs, Bioinformatics, Protein-Protein Interaction, Molecular switches, Cell regulation

Significance Statement

Marc: TODO: still 23 words too long :(

Short Linear motifs (SLiMs) are small protein-interaction modules that play essential roles in signalling, cell cycle progression, trafficking and disease. SLiMs are short, often highly degenerate sequences with no stable secondary structure, and have proven difficult to study both experimentally and computationally. It has been suggested that the number of different motifs in the order of millions. Yet, despite their abundance and importance, we have a very limited understanding of which motifs exist and what they do. The Eukaryotic Linear Motif (ELM) database is an online resource containing over 3000 manually curated SLiMs from experimental literature grouped into 262 motif classes. This database can in turn be used to detect SLiMs in novel protein sequences. ELM is a valuable resource to the scientific community studying protein function, and will also play an important role in unravelling the complex nature of the biology.

Abstract

Marc: TODO: this abstract is still pretty rough :(

The Eukaryotic Linear Motif (ELM) resource (elm.eu.org) is a manually curated database of short linear motifs (SLiMs) as well as a tool to detect and visualize motifs in protein sequences.

In this unit we will explore how to browse and search through the collection of manually annotated data, and explain all of the different types of knowledge and data integrated into the database.

We also cover how this resource can be used to search for SLiM's in novel sequences, as well as giving examples of how to use the output of the ELM prediction pipeline to evaluate which detections may be biologically relevant.

Lastly we also give examples of how the ELM database can be searched as well queried programatically.

By the end of this unit the reader should have a much better understanding of how to navigate the different content types contained in the database, as well as how to use the ELM prediction pipeline for their own research.

Introduction

The activity and function of a protein is tightly regulated by its cellular environment. To interact with their surroundings, proteins use various types of binding modules that each display distinct binding properties (Wright and Dyson 1999). One prominent type of binding module consists of short linear motifs (SLiMs) (Diella 2008). These compact binding sites are generally located in intrinsically disordered regions (IDR) of the proteome and commonly bind to surfaces of a globular domain of a protein (Davey et al. 2012). SLiMs mediate different types of interactions that regulate protein functionality, and hence are important regulators of the dynamic processes involved in cell signalling (Van Roey et al. 2012) (Van Roey et al. 2014). The number of SLiM instances in the human proteome is currently suggested to be over one million (Tompa et al. 2014). Identifying SLiMs and elucidating their functionality is an essential step in understanding cell regulation. The Eukaryotic Linear Motif (ELM) resource contributes to this process by providing the necessary tools to researchers working on motifs. It consists of a database and a prediction tool. The database provides a categorised repository of experimentally validated linear motif classes and instances that were manually annotated from the literature. The ELM prediction tool in turn relies on annotated data, both from the ELM database and other resources, to accurately analyse unknown sequences for candidate motifs and assist researchers in selecting the most plausible ones for experimental validation and discard likely false positive hits, saving them valuable time and resources (Dinkel et al. 2012). The following protocols will guide users through the different ELM applications, explaining how to browse the curated data available in ELM, how to analyse a protein sequence for putative motifs, and how to interpret these data and avoid common pitfalls in SLiM discovery.

Protocol 1 Exploring the Content of the ELM Database

The core of the ELM database is a repository of manually annotated motifs and instances. As of January 2017, ELM contains over 260 motif classes and over 3000 experimentally validated and manually curated instances. The motif classes and motif instances have been uploaded by a large group of annotators from around the globe. The complete catalogue of manually curated data can be searched, browsed and explored on the ELM website.

Each motif class describes a short linear motif, a short sequence of amino acids with a dedicated function. Since the motifs are often degenerate, each motif class is represented using "regular expressions": a symbolic representation expressing a complex pattern of letters (or amino acids). For example, the regular expression "[FY].L.P" is to be read as: The first amino acid is a Phenylalaline 'F' or a Tyrosine 'Y' followed by any one amino acid '.', a Leucine 'L', one more arbitrary amino acid '.' and finally a Proline 'P'.

In all cases the annotator condenses information from the primary and secondary literature into manageable abstracts, accompanies the motif definition with a list of experimental instances, as well as links to external resources including biological pathways, diseases, Gene Ontology and other protein resources. In this protocol we explore the several data-types stored in the database, as well as links to external resources.

Necessary Resources

Software & Hardware

A modern browser such as Firefox, Chrome, or Safari. ELM is best viewed on a laptop or desktop computer, although tablets and smartphones will also work.

Database content overview

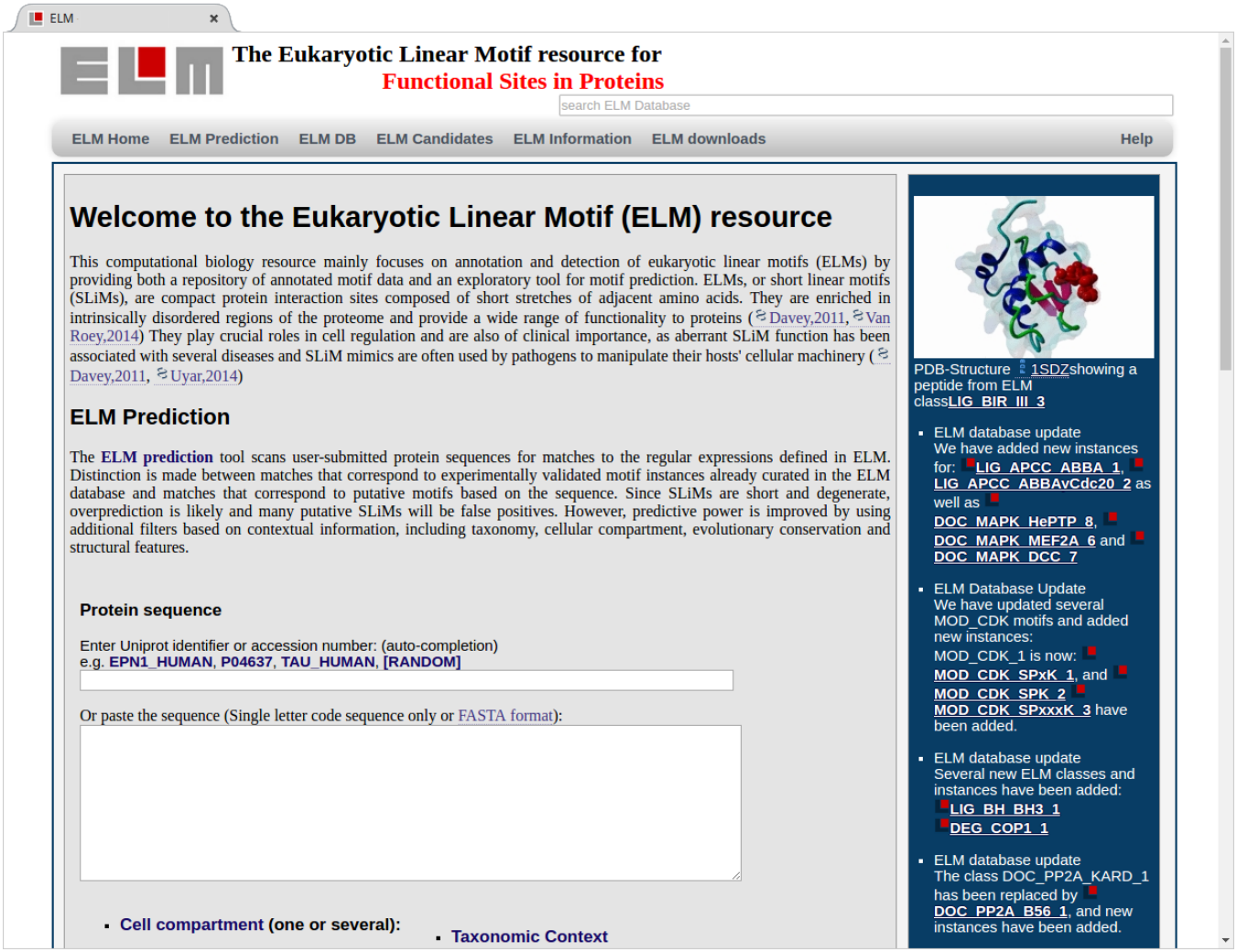


Figure 1: The homepage of the ELM database (elm.eu.org).

1. The ELM database is an online web resource. Open a browser and navigate to elm.eu.org to visit the homepage (Figure 1). This page shows a brief explanation of the ELM resource, and a form to search for SLiMs (which we cover in further detail in Protocol 3 and Protocol 4). The column to the right is continually updated with the latest news about changes and additions to the database.

2. On the ELM homepage click on the menu link **ELM DB** for an overview of the database statistics (Figure 2). This page displays the types and amounts of annotations contained in the database and

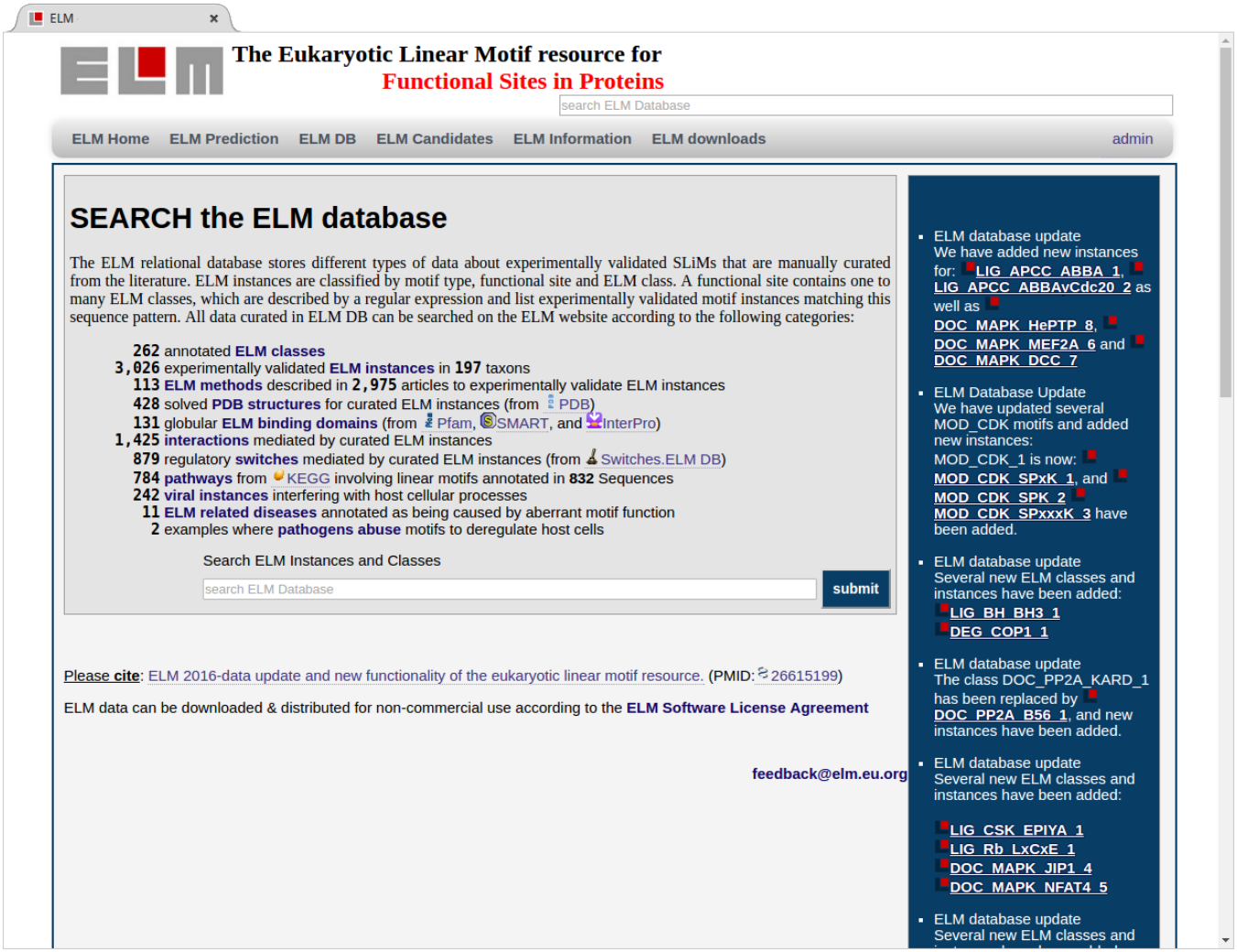


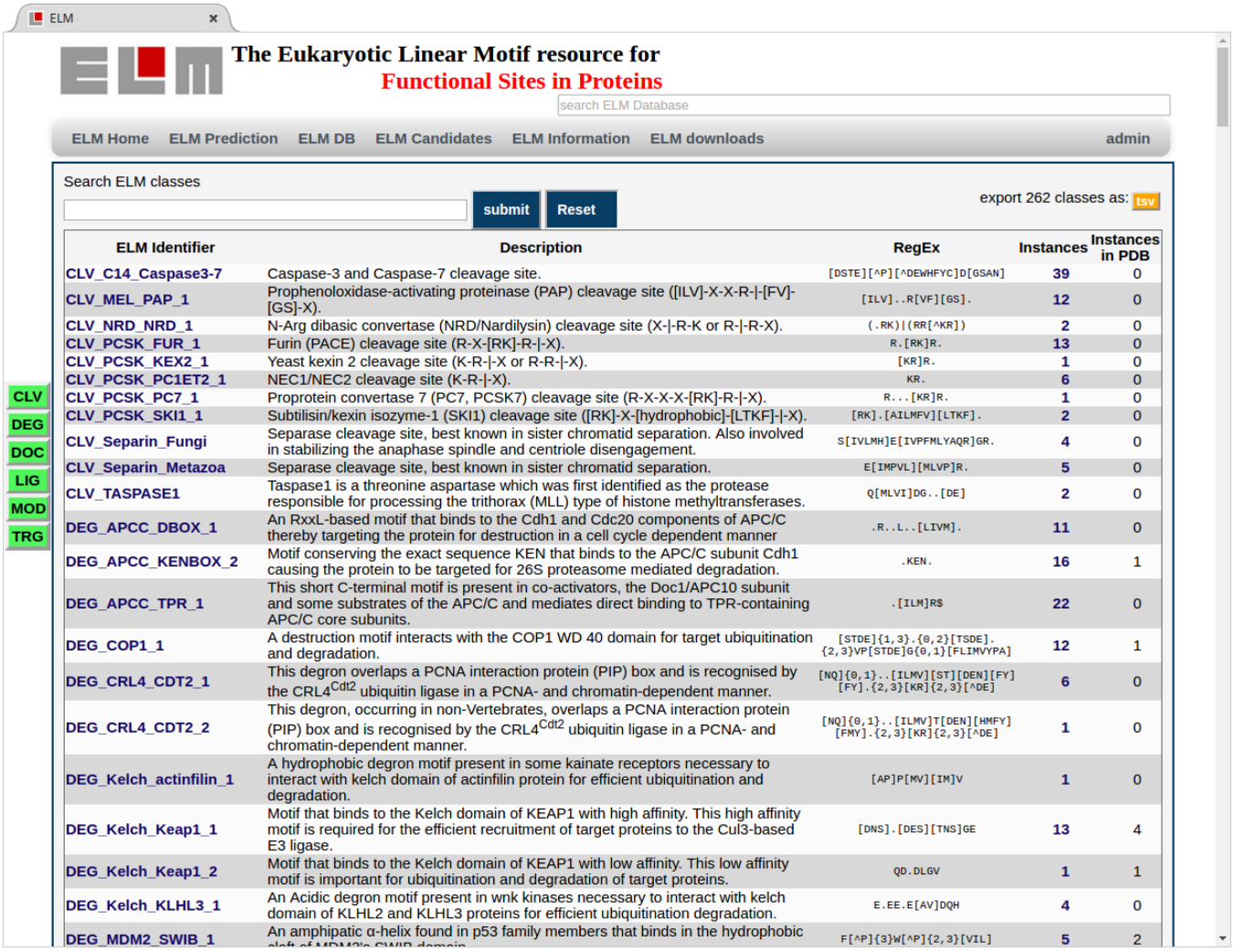
Figure 2: The ELM database statistics overview page shows the most up-to-date database statistics. As of January 2017, ELM has just over 3000 annotated instances in 262 different motif classes.

a few links to third-part databases. Each line contains at least one link that will take you to the corresponding contents page. For example: Clicking on **ELM classes** will take you to the page showing all classes annotated in ELM. We will be exploring these content overview pages in this protocol.

Browsing motif classes and annotated instances

3. Click on the sub-menu **ELM classes** under **ELM DB** to visit the page listing all of the ELM classes (Figure 3). For each class, the following information is provided: ELM identifier, short description, regular expression, number of instances annotated for each class, and number of structures available. For details on each class, click on the ELM identifier; to get a list of annotated instances for an individual class, click on the number of instances.

   Use the search bar at the top of the page to filter for certain motif classes. For example, typing "MAPK" and hitting ***submit*** *will perform a full-text search on all motif classes in the ELM*



ELM Identifier	Description	RegEx	Instances	Instances in PDB
CLV_C14_Caspase3-7	Caspase-3 and Caspase-7 cleavage site.	[DSTE][^P][^DEWHFYC]D[GSAN]	39	0
CLV_MEL_PAP_1	Prophenoloxidase-activating proteinase (PAP) cleavage site ([ILV]-X-X-R-|-[FV]-[GS]-X).	[ILV]..R[VF][GS].	12	0
CLV_NRD_NRD_1	N-Arg dibasic convertase (NRD/Nardilysin) cleavage site (X-|-R-K or R-|-R-X).	(.RK)|(RR[^KR])	2	0
CLV_PCSK_FUR_1	Furin (PACE) cleavage site (R-X-[RK]-R-|-X).	R.[RK]R.	13	0
CLV_PCSK_KEX2_1	Yeast kexin 2 cleavage site (K-R-|-X or R-R-|-X).	[KR]R.	1	0
CLV_PCSK_PC1ET2_1	NEC1/NEC2 cleavage site (K-R-|-X).	KR.	6	0
CLV_PCSK_PC7_1	Proprotein convertase 7 (PC7, PCSK7) cleavage site (R-X-X-X-[RK]-R-|-X).	R...[KR]R.	1	0
CLV_PCSK_SKI1_1	Subtilisin/kexin isozyme-1 (SKI1) cleavage site ([RK]-X-[hydrophobic]-[LTKF]-|-X).	[RK].[AILMFV][LTKF].	2	0
CLV_Separin_Fungi	Separase cleavage site, best known in sister chromatid separation. Also involved in stabilizing the anaphase spindle and centriole disengagement.	S[IVLMH]E[IVPFMLYAQR]GR.	4	0
CLV_Separin_Metazoa	Separase cleavage site, best known in sister chromatid separation.	E[IMPVL][MLVP]R.	5	0
CLV_TASPASE1	Taspase1 is a threonine aspartase which was first identified as the protease responsible for processing the trithorax (MLL) type of histone methyltransferases.	Q[MLVI]DG..[DE]	2	0
DEG_APCC_DBOX_1	An RxxL-based motif that binds to the Cdh1 and Cdc20 components of APC/C thereby targeting the protein for destruction in a cell cycle dependent manner	.R..L..[LIVM].	11	0
DEG_APCC_KENBOX_2	Motif conserving the exact sequence KEN that binds to the APC/C subunit Cdh1 causing the protein to be targeted for 26S proteasome mediated degradation.	.KEN.	16	1
DEG_APCC_TPR_1	This short C-terminal motif is present in co-activators, the Doc1/APC10 subunit and some substrates of the APC/C and mediates direct binding to TPR-containing APC/C core subunits.	.[ILM]R$	22	0
DEG_COP1_1	A destruction motif interacts with the COP1 WD 40 domain for target ubiquitination and degradation.	[STDE]{1,3}.{0,2}[TSDE].{2,3}VP[STDE]G{0,1}[FLIMVYPA]	12	1
DEG_CRL4_CDT2_1	This degron overlaps a PCNA interaction protein (PIP) box and is recognised by the $CRL4^{Cdt2}$ ubiquitin ligase in a PCNA- and chromatin-dependent manner.	[NQ]{0,1}..[ILMV][ST][DEN][FY][FY].{2,3}[KR]{2,3}[^DE]	6	0
DEG_CRL4_CDT2_2	This degron, occurring in non-Vertebrates, overlaps a PCNA interaction protein (PIP) box and is recognised by the $CRL4^{Cdt2}$ ubiquitin ligase in a PCNA- and chromatin-dependent manner.	[NQ]{0,1}..[ILMV]T[DEN][HMFY][FMY].{2,3}[KR]{2,3}[^DE]	1	0
DEG_Kelch_actinfilin_1	A hydrophobic degron motif present in some kainate receptors necessary to interact with kelch domain of actinfilin protein for efficient ubiquitination and degradation.	[AP]P[MV][IM]V	1	0
DEG_Kelch_Keap1_1	Motif that binds to the Kelch domain of KEAP1 with high affinity. This high affinity motif is required for the efficient recruitment of target proteins to the Cul3-based E3 ligase.	[DNS].[DES][TNS]GE	13	4
DEG_Kelch_Keap1_2	Motif that binds to the Kelch domain of KEAP1 with low affinity. This low affinity motif is important for ubiquitination and degradation of target proteins.	QD.DLGV	1	1
DEG_Kelch_KLHL3_1	An Acidic degron motif present in wnk kinases necessary to interact with kelch domain of KLHL2 and KLHL3 proteins for efficient ubiquitination degradation.	E.EE.E[AV]DQH	4	0
DEG_MDM2_SWIB_1	An amphipatic α-helix found in p53 family members that binds in the hydrophobic	F[^P]{3}W[^P]{2,3}[VIL]	5	2

Figure 3: The list of all motif classes annotated in the ELM database.

database containing the term "MAPK". The green buttons on the left can also be used to filter this table. For example, toggling the "DOC" button will remove all DOC classes from the table (and clicking it again will bring them back). Lastly, the yellow ***tsv*** *link can be used to export all motif classes as a "tab separated values" file.*

4. Search the table for the term DOC_CYCLIN_1 and click on **DOC_CYCLIN_1** in the left column to navigate to the page with details about the DOC_CYCLIN_1 motif class (Figure 4). This page contains a description of the functional site class (a Cyclin recognition site), and a short description of the ELM and its regular expression, as well as a probability score, the taxonomic distribution of the motif and which domain (if any) is responsible for the interaction.

   The probability score is the probability that the regular expression represents a random selection of amino acids (similar to an information content score). A lower score indicates that the motif pattern is more difficult to find by chance in a random sequence.

5. Scroll further down the DOC_CYCLIN_1 page (Figure 4) to view more details about this motif

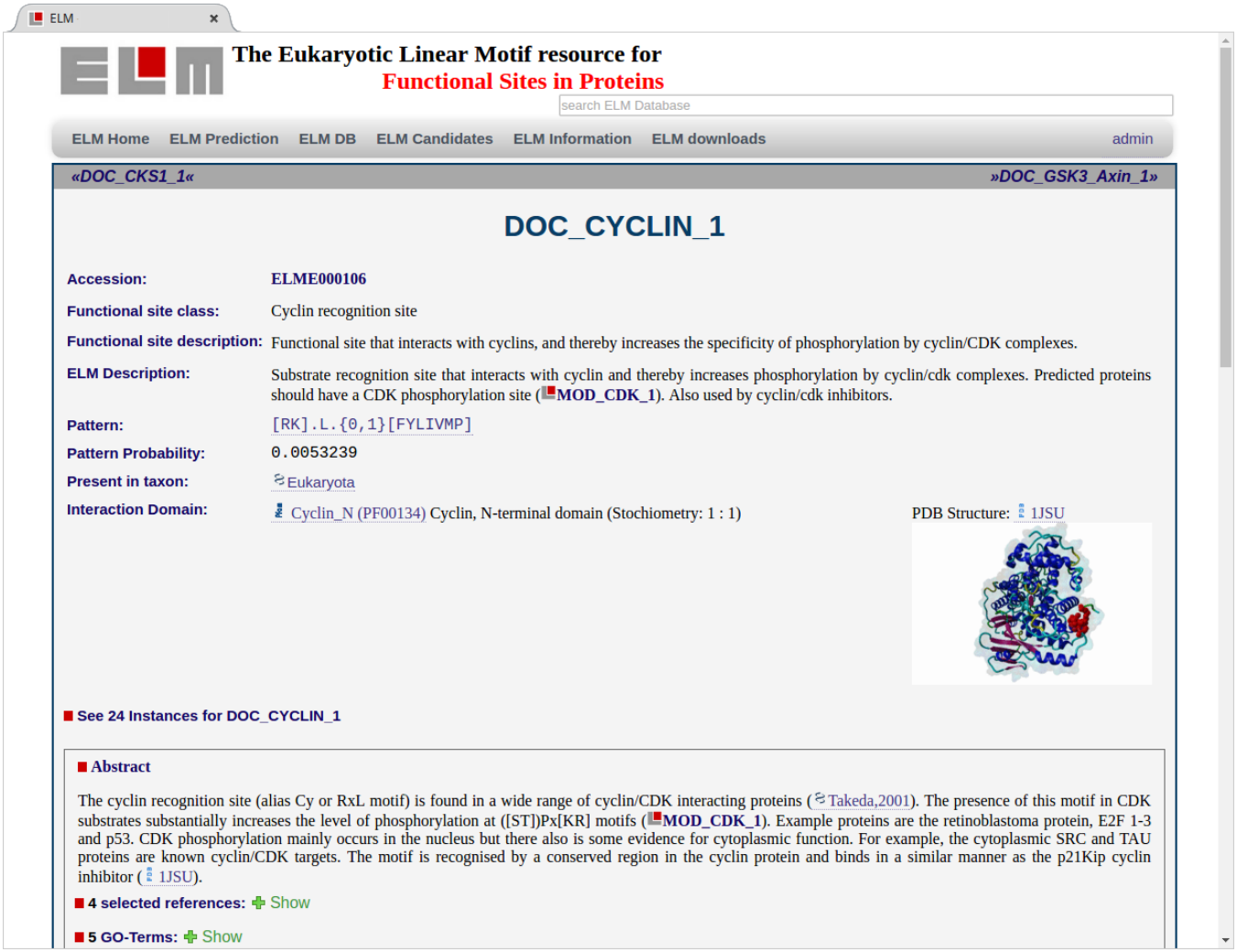


Figure 4: The motif details page for DOC_CYCLIN_1. This page contains all of the manual annotation details for the DOC_CYCLIN_1 motif, the biological background summarized from the scientific literature including links to the primary literature and to external resources (Pubmed (NCBI Resource Coordinators 2017), the Gene Ontology (Gene Ontology Consortium 2017), PDB (Berman et al. 2002) and more).

(Figure 5) The "abstract" contains a description of the biological relevance of the motif (for example its involvement in cellular processes and pathways). annotation. Click on the **show** button next to the "selected references" header for a list of publications relevant to this motif. Click on **show** next to "GO terms" for a complete list of all Gene Ontology (GO) terms annotated for this motif.

6. Scroll further down the DOC_CYCLIN_1 page to view the "Instances" header (Figure 5) This table contains the list of all annotated DOC_CYCLIN_1 instances in the database of this motif. This includes the protein identifier, the start and end positions of the instance, the specific sequence matching the regular expression representing the motif and the "logic" of the instance. The "# Ev." indicates the number of experimental evidences associated with the annotation. "Organism" indicates in which species in which the protein is found. Lastly the "Notes" column contains links to any "interactions" or "switches" present in the database, as well as links to PDB, if the structure

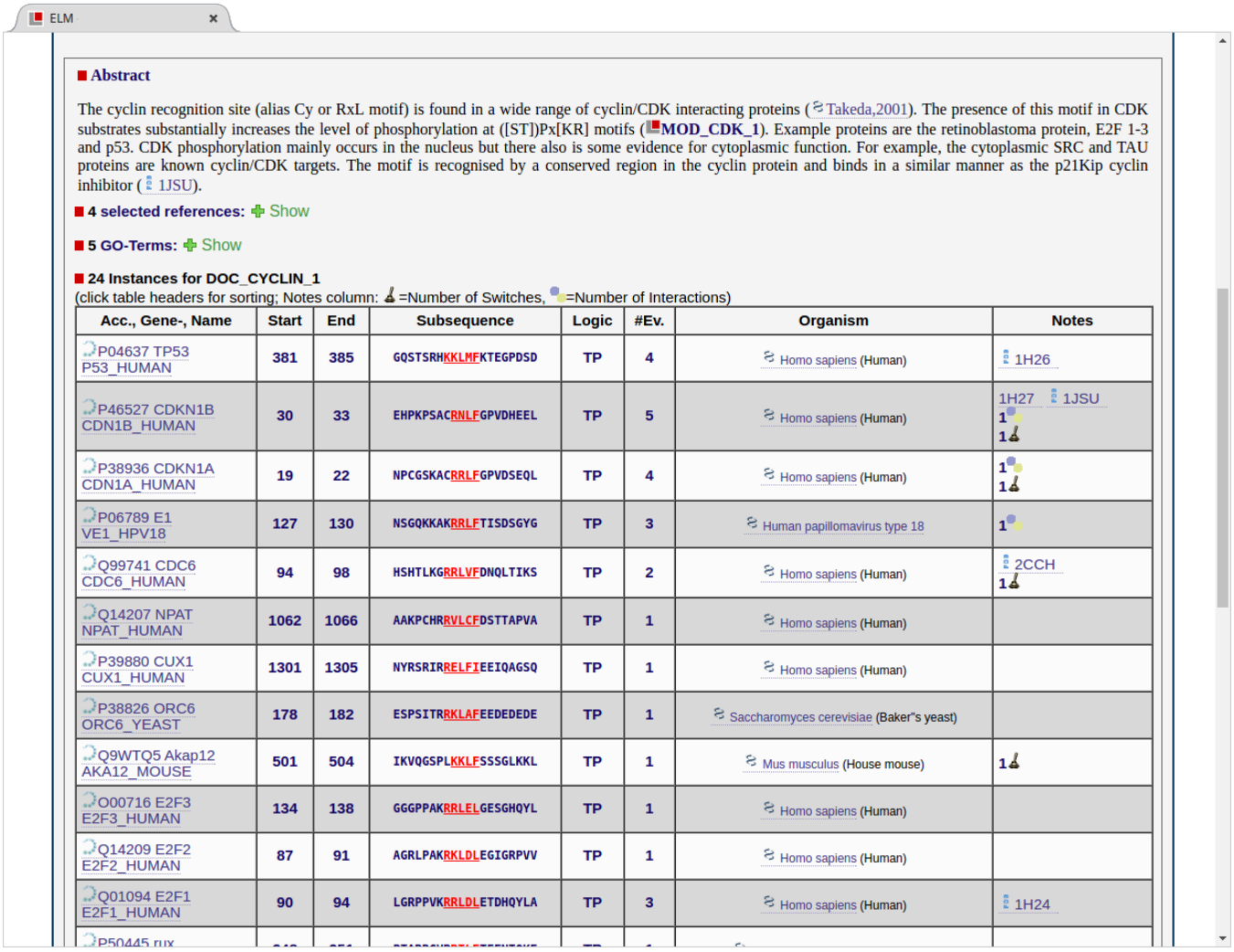


Figure 5: The second part of the DOC_CYCLIN_1 motif details page shows the motif abstract GO terms, and the list of annotated instances.

exists in PDB.

The instance "logic" is an annotation of whether this is a bona-fide instance, or whether it is a non-functional instance. TP (True positive) indicates the instance is annotated with experimental evidence showing, that it is functional. FP (False Positive) instances have experimental evidence suggesting function, but are believed to be non-functional, after careful examination by our annotators. TN (True Negative) instances have been experimentally determined to be non-functional, and U (Unknown) instances do not have enough evidence to determine whether it is functional or not. The overwhelming majority of instances in ELM are TPs.

7. Click on the sub-menu **ELM instances** under **ELM DB** to visit the page where you can search and browse the instances annotated in ELM (Figure 6). Note that only the first hundred instances matching the search criteria are shown. The search form can be used to filter results by a full text search, by instance logic, or by organism.

   This table can be filtered by motif class using the green toggle filters on the left hand side. Lastly,

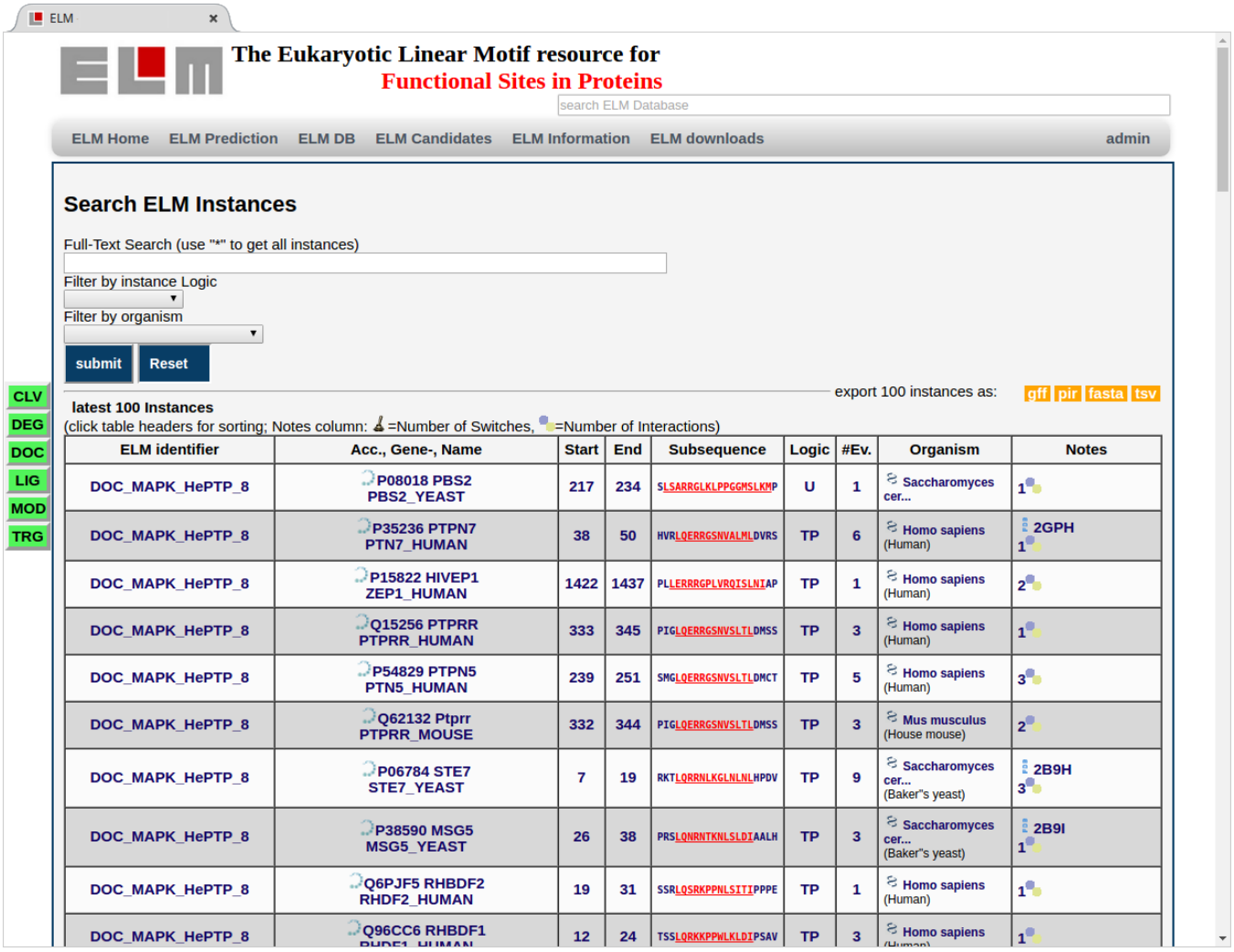


Figure 6: The "instances" page can be used to search for instances in the ELM database.

the yellow buttons at the top of the page can be used to download the instances in the following formats: GFF, PIR, FASTA or TSV.

8. Type "p53_human" in the search box to search for ELM Instances in this protein. Find the row for the ELM class DOC_CYCLIN_1 and click on the instance sub-sequence (highlighted in red) to go to the instance details page of this instance (Figure 7). The top part of the page contains details about the instance and the protein it was identified in, as well as a link to the UniProt entry for the protein (UniProt Consortium 2015).

9. Scroll down to the "Instance Evidence" header to view details on the experimental evidence used to annotate this instance. The each experimental method is annotated using the Proteomics Standards Initiative Method Identifier (PSI-MI) (Kerrien et al. 2007), as well as the references in which the experiments were published.

   The "biosource" indicates whether method is in vivo, in vitro, in silico *or a combination of these. The "logic" column indicates whether this experiment supports or contradicts this instance being*

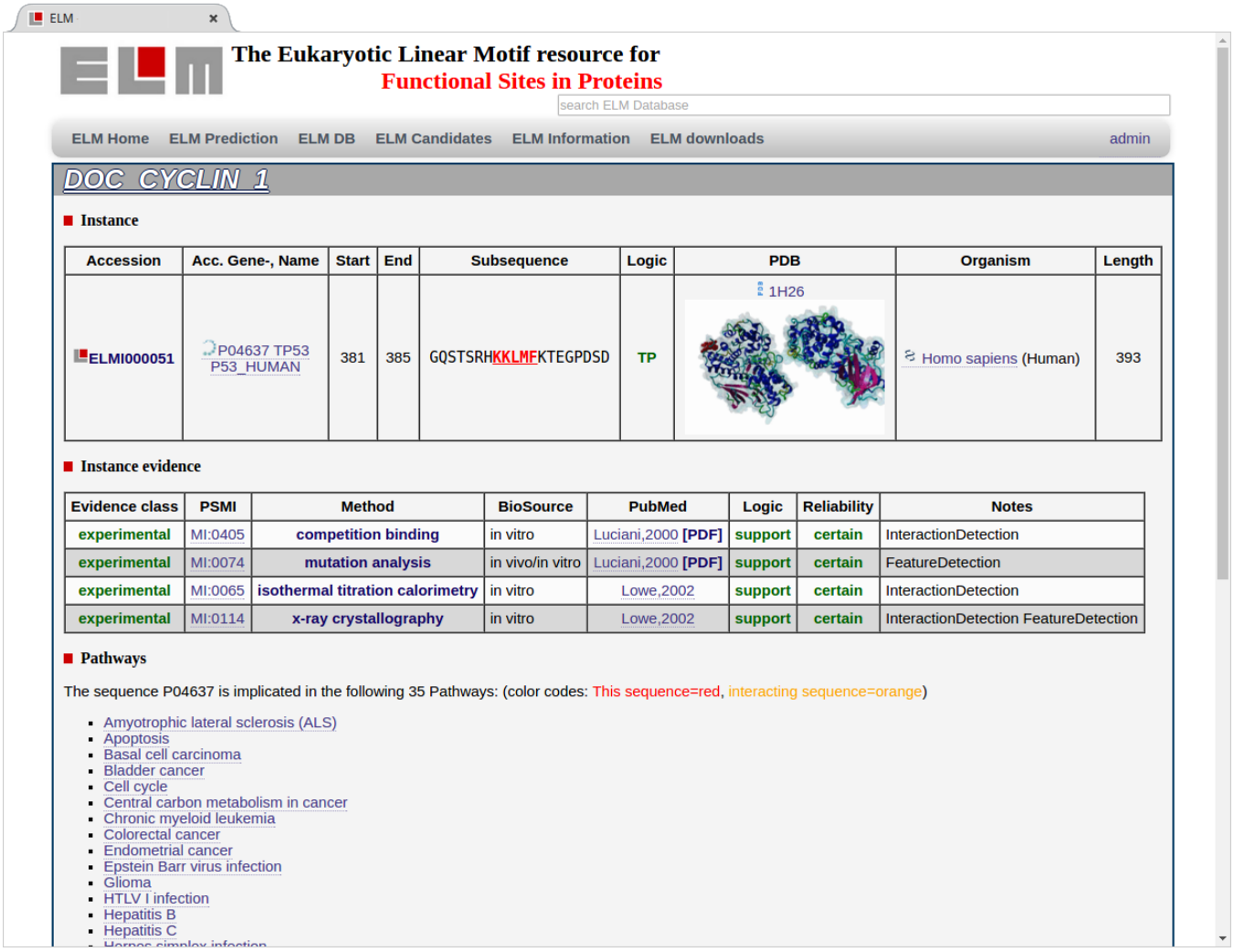


Figure 7: The instance details page for the DOC_CYCLIN_1 instance annotated for protein P53_HUMAN with start/end position "381–385". This page also contains links to many external databases including UniProt (UniProt Consortium 2015), PDB (Berman et al. 2002), NCBI taxonomy and Pubmed (NCBI Resource Coordinators 2017), and KEGG pathways (Kanehisa et al. 2016), as well as the PSI-MI controlled vocabulary (Kerrien et al. 2007).

functional. Each method is also annotated with a reliability assessment, which can be any of certain, likely, unlikely or unspecified.

Finding Switches and molecular interactions

10. Repeat the previous search by clicking on the sub-menu **ELM instances** under **ELM DB** and type "p53_human" in the search box. This time, find the ELM instance of the motif DOC_WW_PIN1_4 with the start/end position "30–35". (You can sort the table by clicking on the header lines: click on "Start" to sort by start position). Click on the start/end position or the sub-sequence that will take you to the details page (Figure 8). This page is similar to that described for the p53 instance DOC_CYCLIN_1 (Figure 7). Additionally, for this instance, there is information available about its

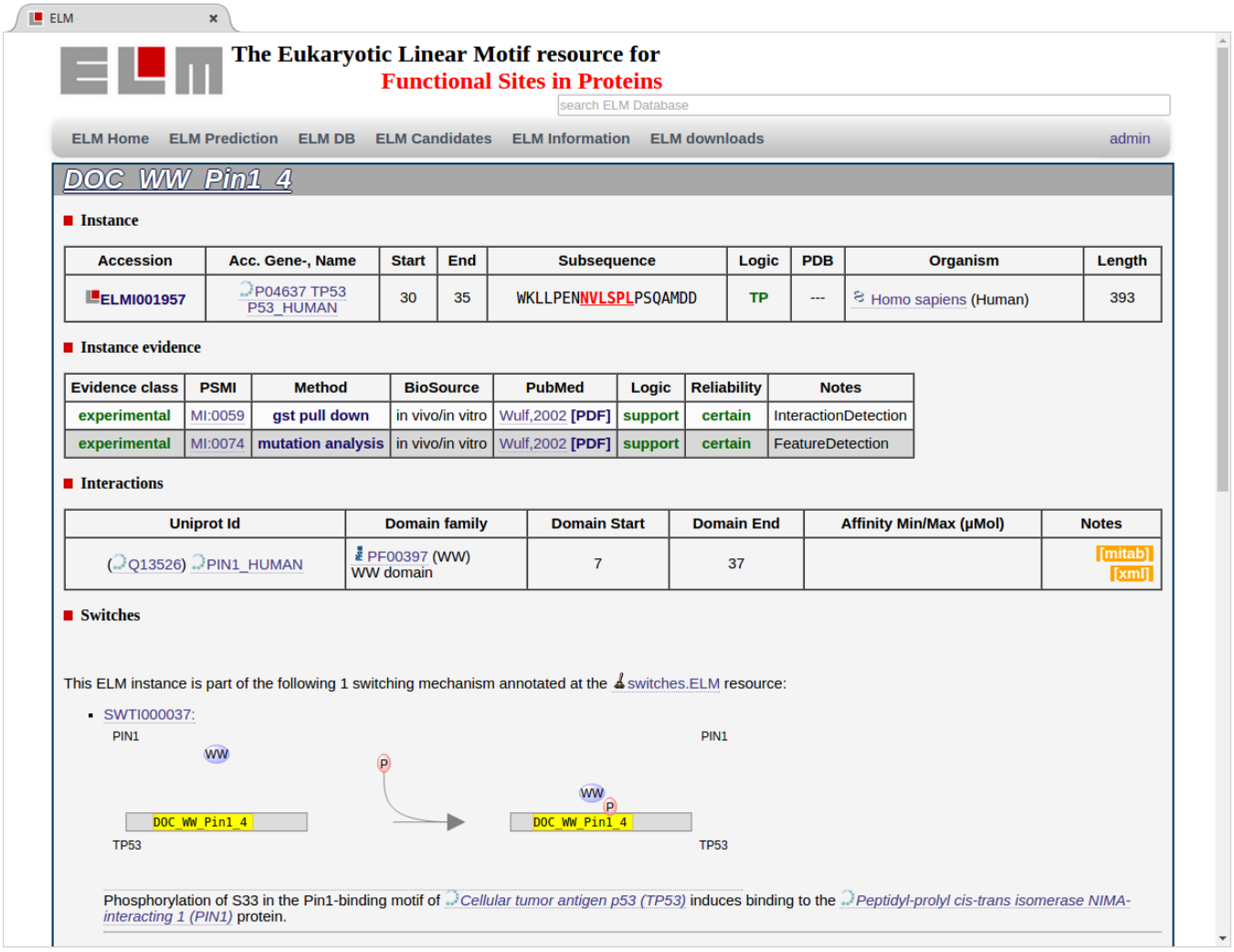


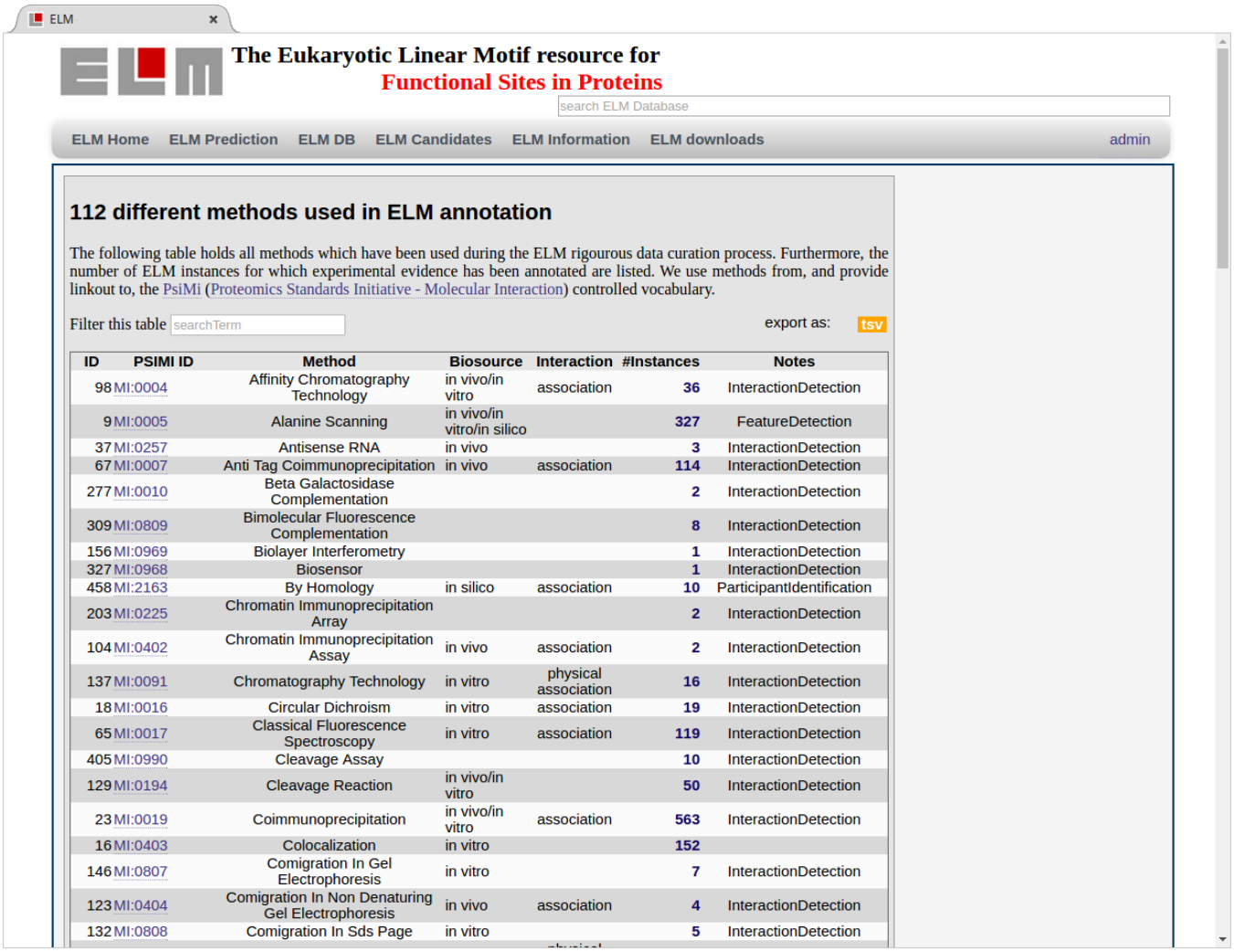
Figure 8: The instance details page for the DOC_WW_PIN1_4 instance found in human p53 (P53_HUMAN) with start/end position "30–35".

interaction partner and a molecular switch, which is mediated by this motif instance.

11. Scroll down to the "Interactions" header to view information about this instance's interactions (Fig.ure 8). This instance interacts with PIN1_HUMAN via the "WW" domain (PFAM identifier PF00397; found on position 7–37 in PIN1_HUMAN. If available, binding affinities are also shown here. Interaction data is made available in *MiTab* and *XML* format (Kerrien et al. 2007), and can be downloaded by clicking on the yellow buttons in the right column.

12. Scroll further down to the "Switches" section for a brief overview of the switches details of this instance obtained from "switches.ELM" (Van Roey et al. 2013) (Figure 8). This particular instance is part of a phosphorylation-dependant molecular switch – only if p53 is phosphorylated on residue Serine-33 can it bind to the protein "Peptidyl-prolyl cis-trans isomerase NIMA-interacting 1 (PIN1)". Clicking on the diagram will open an external link to the `switches.elm.eu.org` website where more detail can be found.

Exploring Links to External Protein Resources



ID	PSIMI ID	Method	Biosource	Interaction	#Instances	Notes
98	MI:0004	Affinity Chromatography Technology	in vivo/in vitro	association	36	InteractionDetection
9	MI:0005	Alanine Scanning	in vivo/in vitro/in silico		327	FeatureDetection
37	MI:0257	Antisense RNA	in vivo		3	InteractionDetection
67	MI:0007	Anti Tag Coimmunoprecipitation	in vivo	association	114	InteractionDetection
277	MI:0010	Beta Galactosidase Complementation			2	InteractionDetection
309	MI:0809	Bimolecular Fluorescence Complementation			8	InteractionDetection
156	MI:0969	Biolayer Interferometry			1	InteractionDetection
327	MI:0968	Biosensor			1	InteractionDetection
458	MI:2163	By Homology	in silico	association	10	ParticipantIdentification
203	MI:0225	Chromatin Immunoprecipitation Array			2	InteractionDetection
104	MI:0402	Chromatin Immunoprecipitation Assay	in vivo	association	2	InteractionDetection
137	MI:0091	Chromatography Technology	in vitro	physical association	16	InteractionDetection
18	MI:0016	Circular Dichroism	in vitro	association	19	InteractionDetection
65	MI:0017	Classical Fluorescence Spectroscopy	in vitro	association	119	InteractionDetection
405	MI:0990	Cleavage Assay			10	InteractionDetection
129	MI:0194	Cleavage Reaction	in vivo/in vitro		50	InteractionDetection
23	MI:0019	Coimmunoprecipitation	in vivo/in vitro	association	563	InteractionDetection
16	MI:0403	Colocalization	in vitro		152	
146	MI:0807	Comigration In Gel Electrophoresis	in vitro		7	InteractionDetection
123	MI:0404	Comigration In Non Denaturing Gel Electrophoresis	in vivo	association	4	InteractionDetection
132	MI:0808	Comigration In Sds Page	in vitro		5	InteractionDetection

Figure 9: The list of all experimental methods used in the ELM database, along with their PSI-MI identifiers.

13. Click on the sub-menu **ELM methods** under **ELM DB** to see a list of all experimental methods, which have been used to identify motifs and instances (Figure 9). This table shows the internal method identifier in the first column, a link to the corresponding entry in the PSI-MI database (Kerrien et al. 2007), and the method name as annotated by the PSI-MI controlled vocabulary, as well as the type of experiment (*in vitro*, *in vivo*, *in silico*, or a combination of these). Clicking on the link in the "instances" column will list all instances annotated using that method.

    The filter bar on the top page can be used to filter the list of methods. The TSV link creates a downloadable file in "tab separated values" format.

14. Click on the sub-menu **ELM PDB structures** under **ELM DB** to see a list of all macromolecular structures in the ELM database (Figure 10). Structures annotated in ELM ideally (but not always)

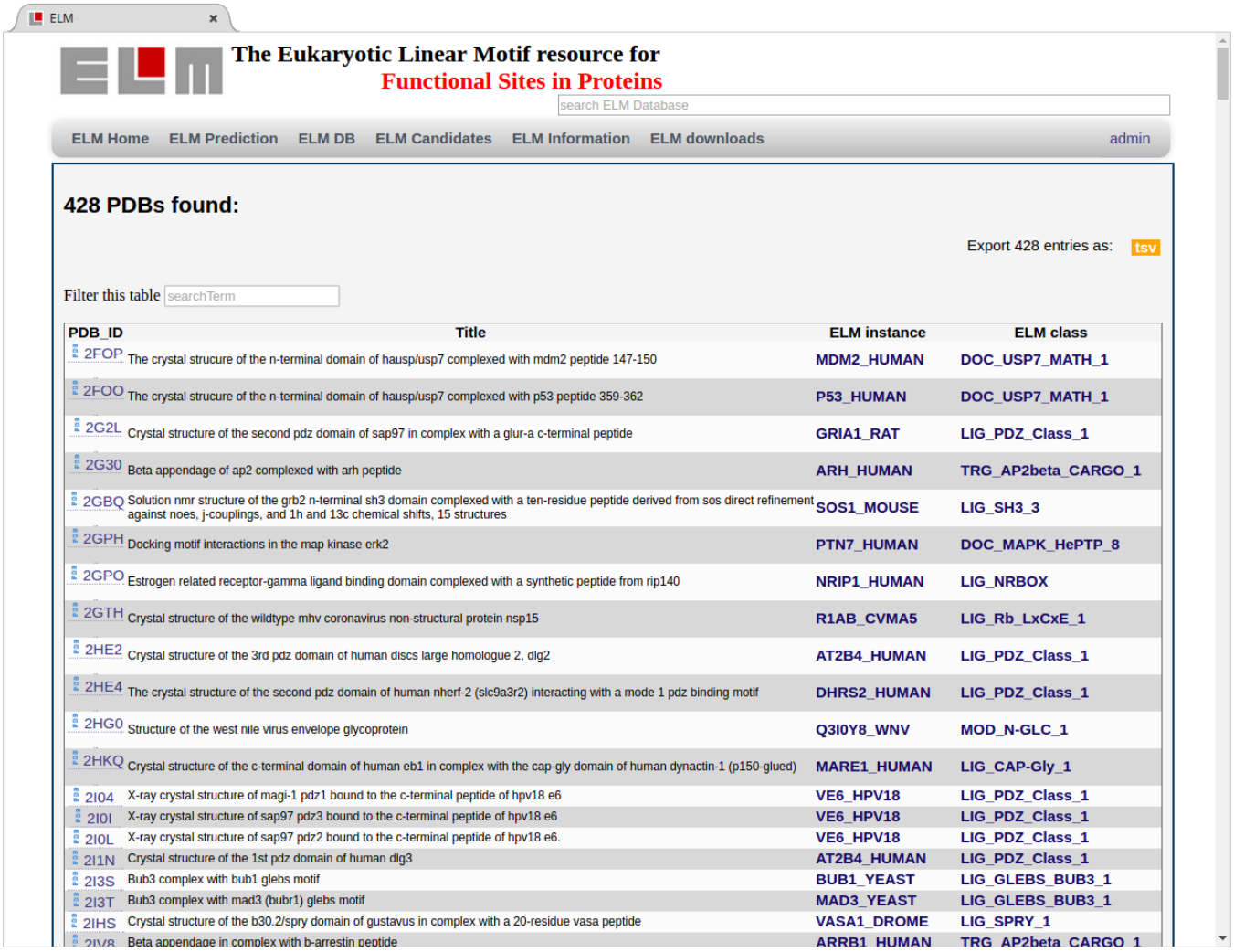
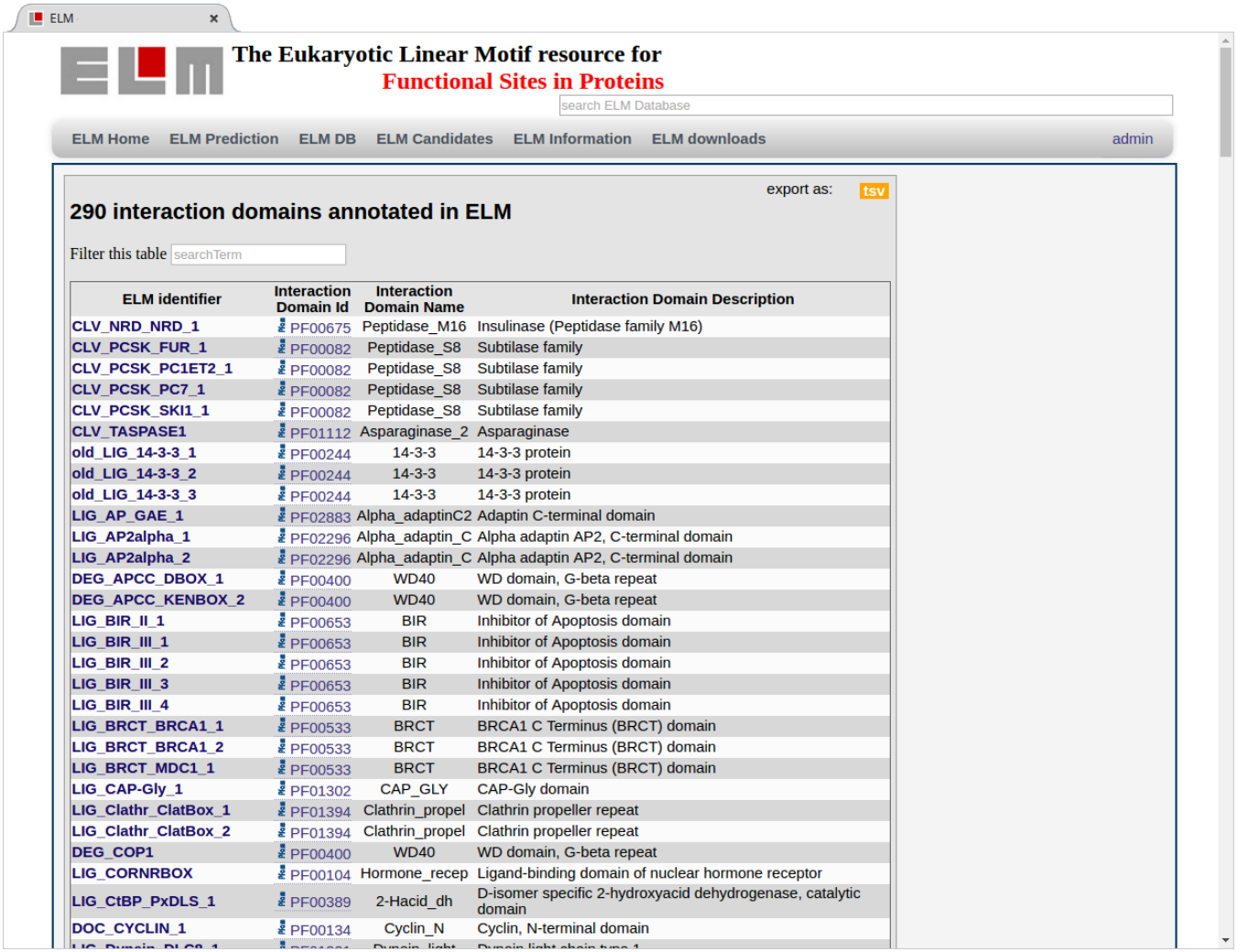


Figure 10: The list of all known PDB structures for which annotated motif instances exist in the ELM database.

show both interaction partners (motif and domain). This page also contains links to RCSB/PDB (Berman et al. 2002), the individual instance and the motif class of that instance.

15. Click on the sub-menu **ELM binding domains** under **ELM DB** to see a complete list of all the interaction domains in ELM (Figure 11). This table shows the ELM classes that have been annotated with a corresponding interaction domain divided by the ELM class, a link to the PFAM (Finn et al. 2016), SMART (Letunic et al. 2015) or InterPro (Finn et al. 2017) domain, as well as the name of the interacting domain followed by a brief description.

16. Click on the sub-menu **ELM switches** under **ELM DB** to see a complete list of all the molecular switches annotated in ELM (Figure 12). This table shows the motif class, contains a link to UniProt, as well as the start and stop positions of the motif mediating the switch. The last two columns show links to the `switches.elm.eu.org` website, and a brief description of the switch (taken from the switches.ELM database, see Van Roey et al. (2013)).



ELM identifier	Interaction Domain Id	Interaction Domain Name	Interaction Domain Description
CLV_NRD_NRD_1	PF00675	Peptidase_M16	Insulinase (Peptidase family M16)
CLV_PCSK_FUR_1	PF00082	Peptidase_S8	Subtilase family
CLV_PCSK_PC1ET2_1	PF00082	Peptidase_S8	Subtilase family
CLV_PCSK_PC7_1	PF00082	Peptidase_S8	Subtilase family
CLV_PCSK_SKI1_1	PF00082	Peptidase_S8	Subtilase family
CLV_TASPASE1	PF01112	Asparaginase_2	Asparaginase
old_LIG_14-3-3_1	PF00244	14-3-3	14-3-3 protein
old_LIG_14-3-3_2	PF00244	14-3-3	14-3-3 protein
old_LIG_14-3-3_3	PF00244	14-3-3	14-3-3 protein
LIG_AP_GAE_1	PF02883	Alpha_adaptinC2	Adaptin C-terminal domain
LIG_AP2alpha_1	PF02296	Alpha_adaptin_C	Alpha adaptin AP2, C-terminal domain
LIG_AP2alpha_2	PF02296	Alpha_adaptin_C	Alpha adaptin AP2, C-terminal domain
DEG_APCC_DBOX_1	PF00400	WD40	WD domain, G-beta repeat
DEG_APCC_KENBOX_2	PF00400	WD40	WD domain, G-beta repeat
LIG_BIR_II_1	PF00653	BIR	Inhibitor of Apoptosis domain
LIG_BIR_III_1	PF00653	BIR	Inhibitor of Apoptosis domain
LIG_BIR_III_2	PF00653	BIR	Inhibitor of Apoptosis domain
LIG_BIR_III_3	PF00653	BIR	Inhibitor of Apoptosis domain
LIG_BIR_III_4	PF00653	BIR	Inhibitor of Apoptosis domain
LIG_BRCT_BRCA1_1	PF00533	BRCT	BRCA1 C Terminus (BRCT) domain
LIG_BRCT_BRCA1_2	PF00533	BRCT	BRCA1 C Terminus (BRCT) domain
LIG_BRCT_MDC1_1	PF00533	BRCT	BRCA1 C Terminus (BRCT) domain
LIG_CAP-Gly_1	PF01302	CAP_GLY	CAP-Gly domain
LIG_Clathr_ClatBox_1	PF01394	Clathrin_propel	Clathrin propeller repeat
LIG_Clathr_ClatBox_2	PF01394	Clathrin_propel	Clathrin propeller repeat
DEG_COP1	PF00400	WD40	WD domain, G-beta repeat
LIG_CORNRBOX	PF00104	Hormone_recep	Ligand-binding domain of nuclear hormone receptor
LIG_CtBP_PxDLS_1	PF00389	2-Hacid_dh	D-isomer specific 2-hydroxyacid dehydrogenase, catalytic domain
DOC_CYCLIN_1	PF00134	Cyclin_N	Cyclin, N-terminal domain

Figure 11: A list of all interactions annotated in the database.

Visualizing KEGG pathways from ELM

17. Click on the sub-menu **ELM pathways** under **ELM DB** to see a list of all KEGG pathways contained in ELM (Figure 13). Pathways are taken from and mapped onto the "Kyoto Encyclopedia of Genes and Genomes" (KEGG (Kanehisa et al. 2016)).

18. On the "ELM pathways" page (Figure 14) click on the link **Gallus gallus** to navigate to the page containing all pathways annotated for chicken.

19. On the page with pathways annotated for chicken (Figure 14), click on **Adherens junction** to the KEGG entry for this pathway, with each protein's color corresponding to ELM classes (see the color legend right side of Figure 15).

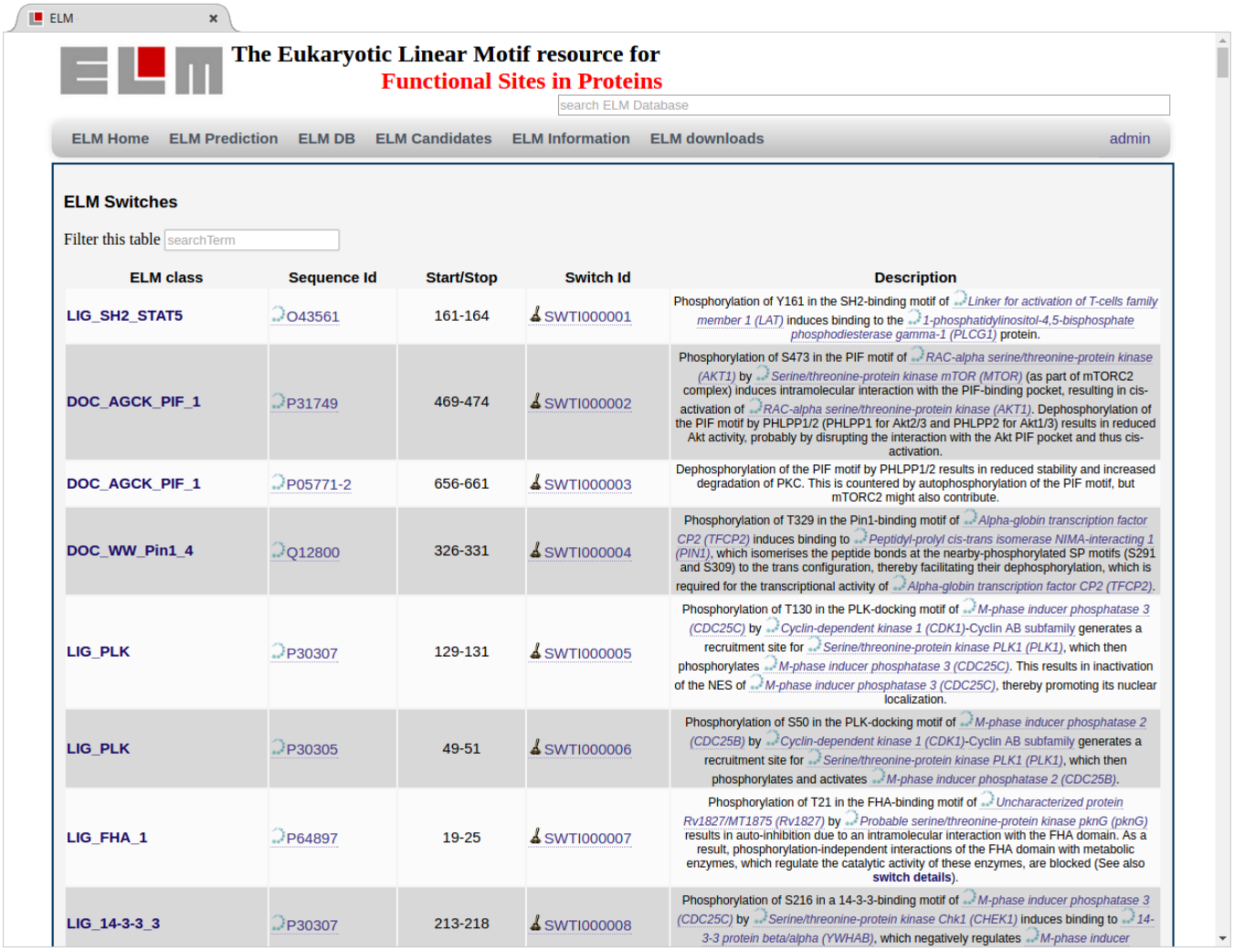


Figure 12: A list of all molecular switches that are based on instances from ELM and which are annotated at the `switches.elm.eu.org` database.

Browsing Infections and Diseases

20. Click on the sub-menu **ELM virus instances** under **ELM DB** to see a list of all instances in ELM that have been annotated as being abused by viruses (Figure 16). (The columns are identical to those listed in step 7, see Figure 6).

    The green buttons on the left can be used to filter this table by motif class. Click on the yellow links on the top right of the page to download the (complete) table in GFF, PIR, FASTA or TSV format.

21. Click on the sub-menu **ELM diseases** under **ELM DB** to see a list of diseases which are mediated by short linear motifs accompanied by a short description of the disease as well as the role of the motif. (Figure 17). Disease information is taken from the "Online Mendelian Inheritance in Man" (OMIM) database (McKusick 2007).

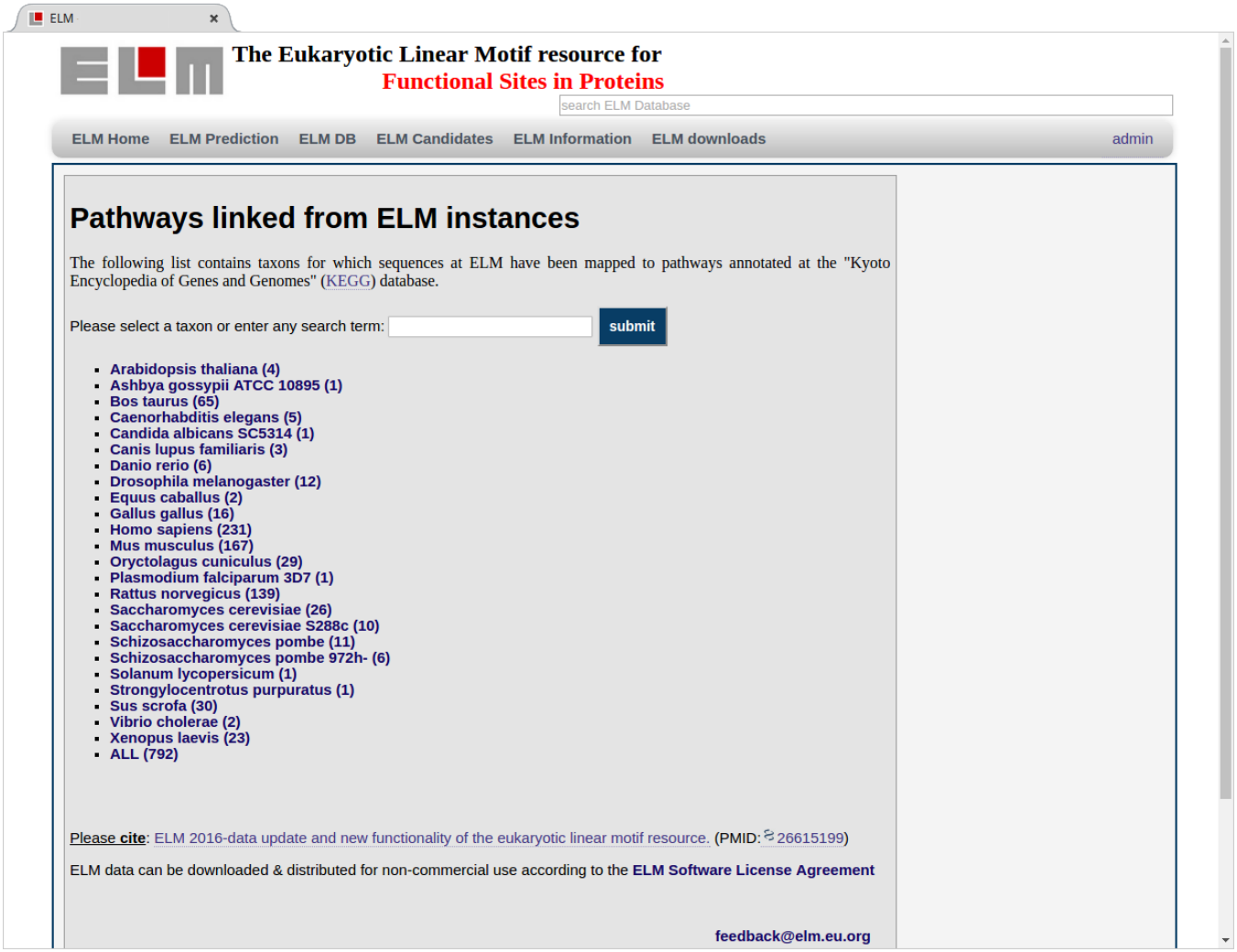


Figure 13: A list of organisms for which pathways from KEGG have been mapped to protein sequences of instances in ELM. The number in brackets denotes the number of different pathways per organism.

Finding Help and Frequently Asked Questions

22. Click on the **Help** button on the right of the top navigation menu to visit the ELM Help page. This page has answers to the most Frequently asked questions, which you can see by clicking on a particular question. For example: Click on “Regular expressions” for a detailed description of the symbols used to build regular expressions to define motif classes.

Protocol 2 Exploring the Content of the ELM Database Using the General Search

A general search text box is available to query the entire collection of manually curated information in the ELM database. This search field can be found in the header of the ELM database website (see for example

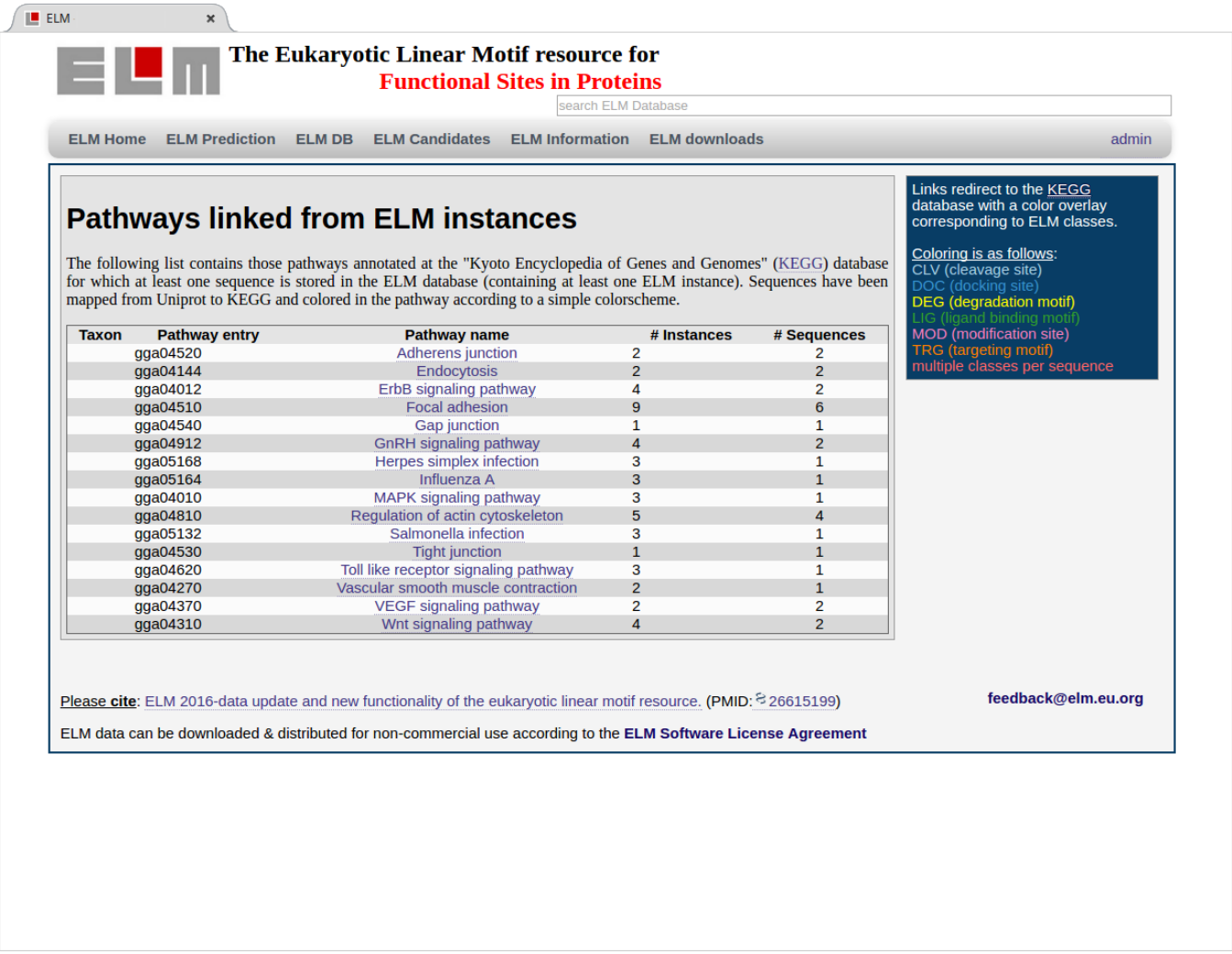


Figure 14: A list of all KEGG pathways for the taxon *Gallus gallus* involving proteins annotated in ELM. Note that multiple instances from a single sequence can be annotated for one pathway. The color scheme on the right is used for coloring instances on the KEGG website (see Figure 15).

Fig. 21). This search peforms a full-text query across multiple selected data sources in the ELM database, including ELM classes, instances, candidates, and switches. Using this general search is helpful in getting information about a particular protein and its annotation status in the ELM database (eg. full instance vs. candidate).

Necessary Resources

Software & Hardware

A modern browser such as Firefox, Chrome, or Safari. ELM is best viewed on a laptop or desktop computer, although tablets and smartphones will also work.

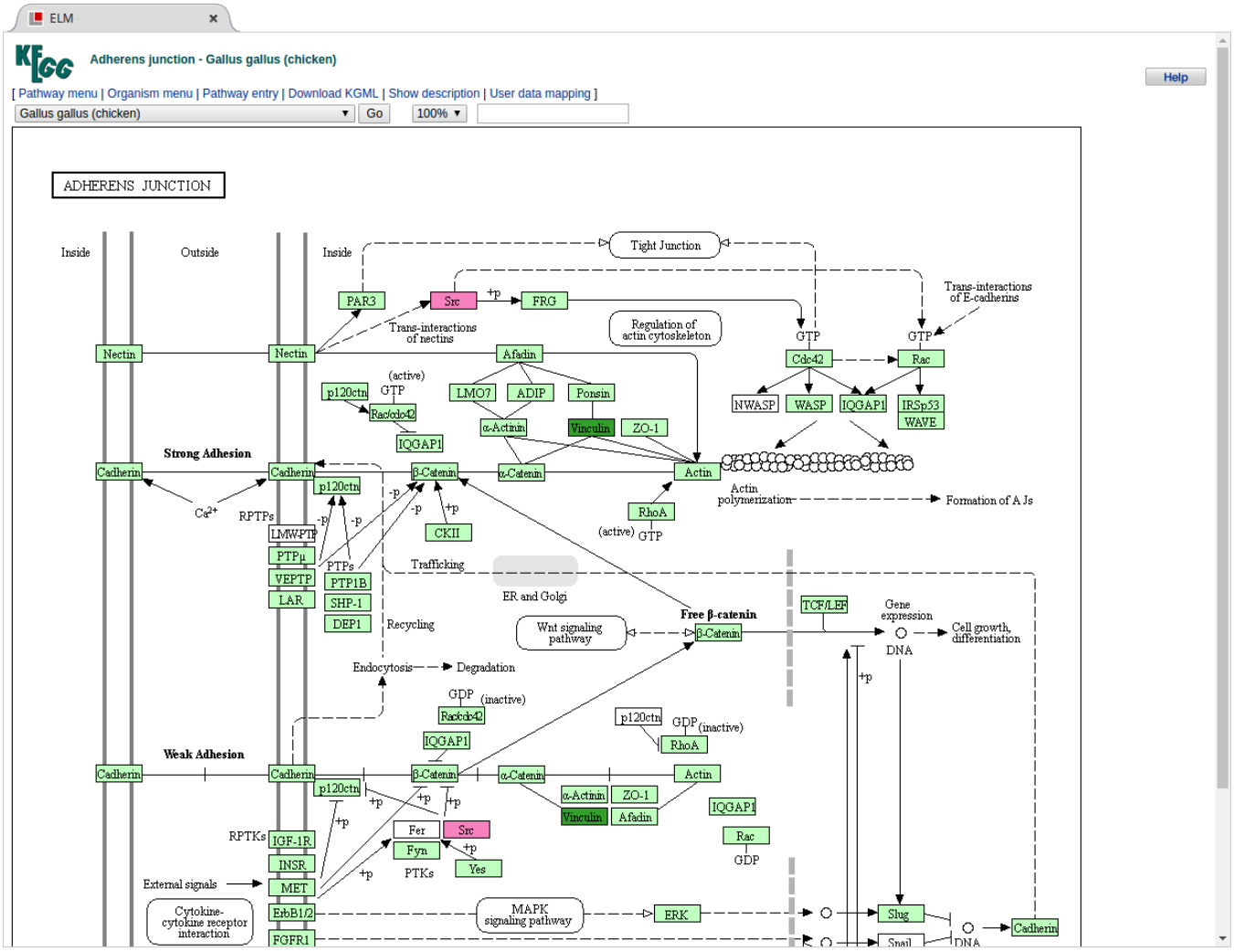


Figure 15: An overlay of ELM annotations of proteins in the Adherens junction pathway in *Gallus gallus*. The coloring of genes/proteins is as follows: light blue=CLV (cleavage site), dark blue=DOC (docking site), yellow=DEG (degradation motif), green=LIG (ligand binding motif), pink=MOD (modification site), orange=TRG (targeting motif), red=multiple classes per sequence. Light green boxes are colored by KEGG and represent the KEGG hyperlinks to GENES entries (see KEGG help http://www.kegg.jp/kegg/document/help_pathway.html).

Using the General Search

A modern browser such as Firefox, Chrome, or Safari. ELM is best viewed on a laptop or desktop computer, although tablets and smartphones will also work.

1. Use the general search field (on the top of the page) to do a general search for p53 using its UniProt identifier by typing “P04637” in the search field and hitting “Enter”. This will perform a query across multiple tables of the ELM database to find any matches to the search query “P04637”. In this case, the results are grouped into matching instances (Figure 19) candidate classes and switches

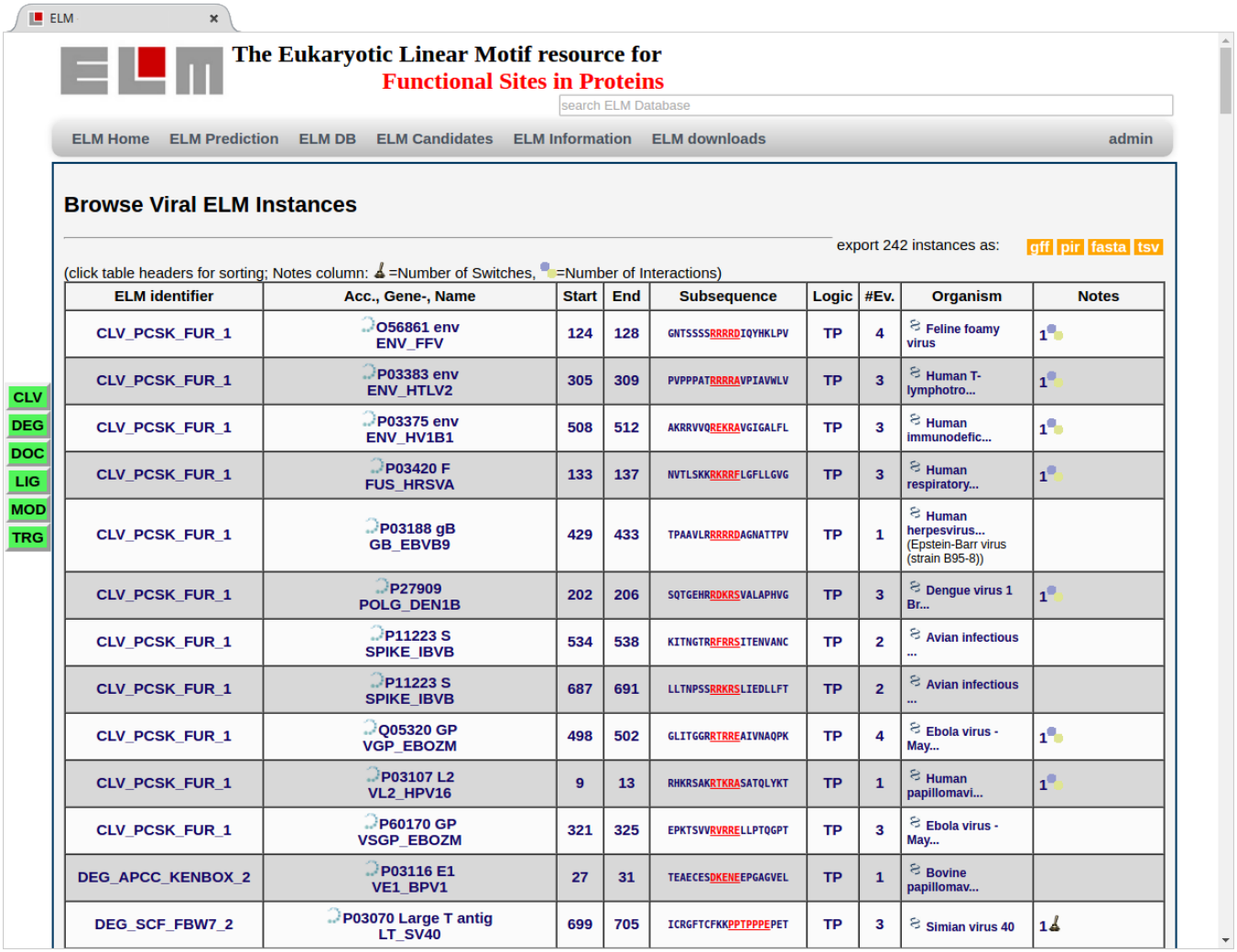


Figure 16: A table of the ELM instances abused by viruses.

(Figure 20). As there are no classes with "P04637" in the name or description, no classes are returned for this section of the query.

> *The "candidate classes" are a collection of putative future ELM classes, which are not yet fully annotated, often submitted by ELM users. Keep in mind that these are a first draft on ELM classes and are still pending curation.*

2. Perform a search using the keyword "p53" in the general search field instead of its UniProt identifier "P04637". The set of results retrieved using this term as the search query (Figure 21) are different, returning 31 instances and 44 switches (instead of 14 and 11). The reason for this is that the phrase "p53" also matches the UniProt identifier of CDH1_YEAST (P53197). This is probably not what you had in mind when using this search term, so it is important to keep in mind to check for such false positive search results, when searching the database.

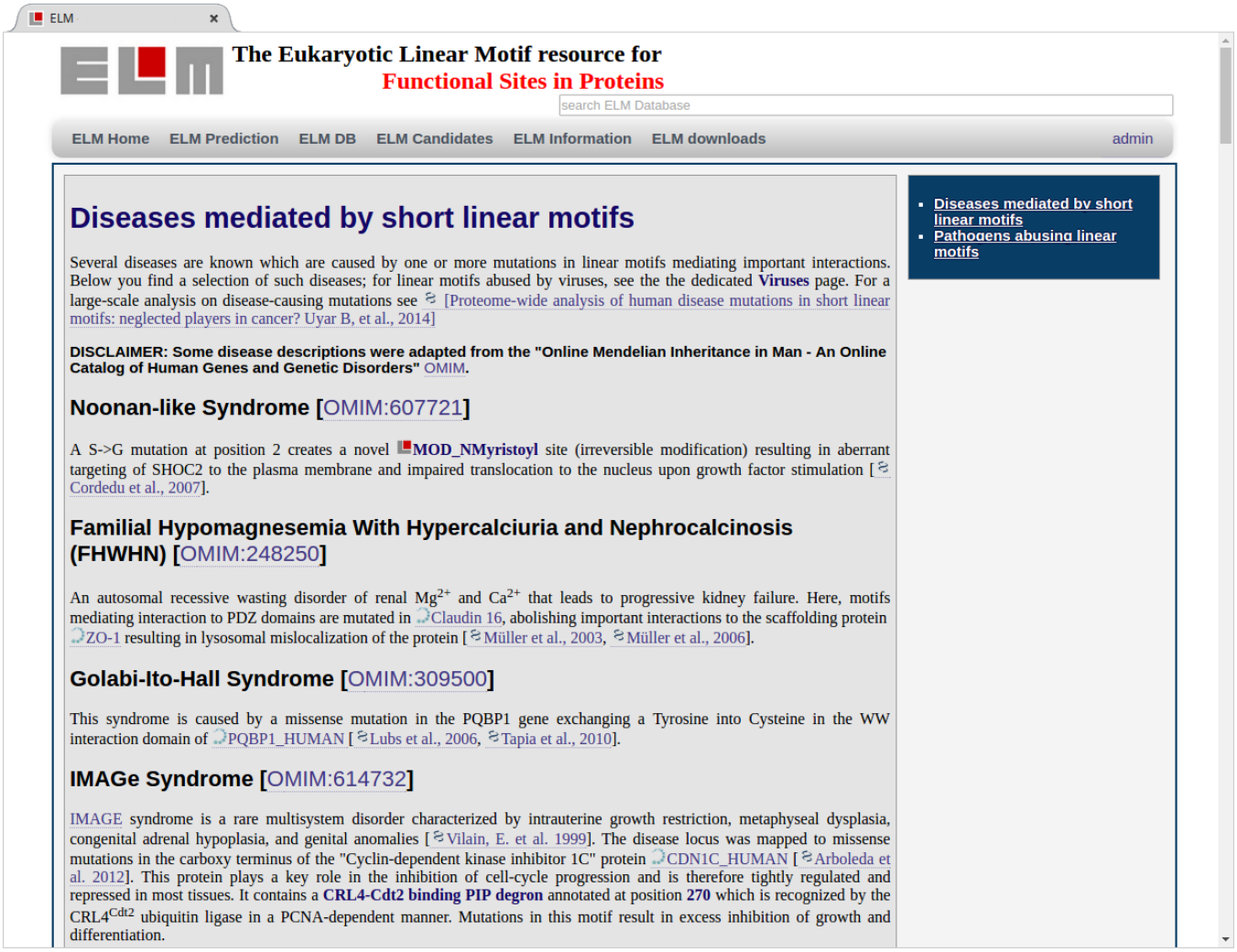


Figure 17: A list of diseases that are mediated by short linear motifs. Disease description has been adapted from the “Online Mendelian Inheritance in Man“ (OMIM) database (McKusick 2007) and enriched with information about the role of the motif.

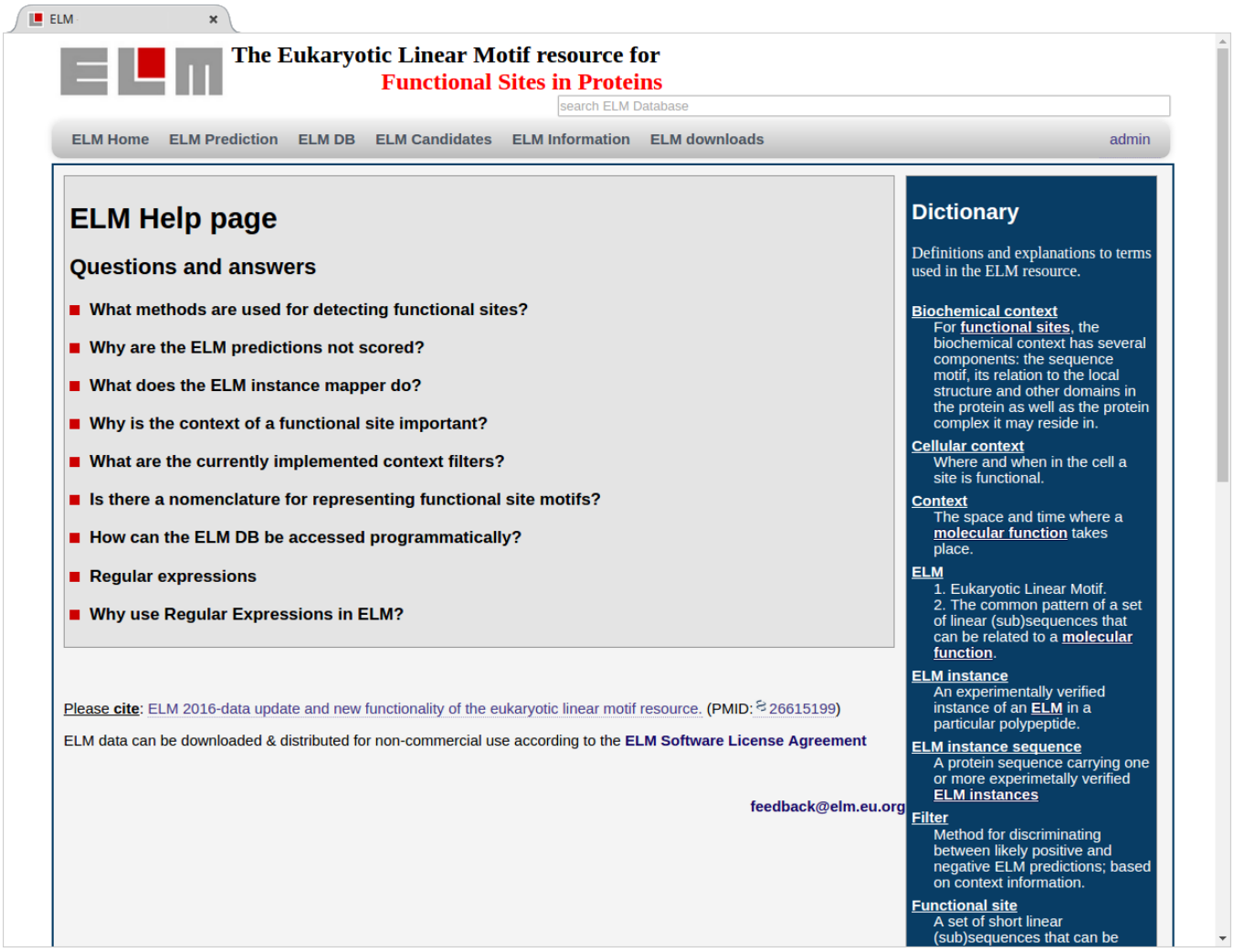


Figure 18: The ELM "Help" and "Questions & Answers" page. Users are encouraged to contact the authors via email if their questions regarding the ELM database can not be answered by this page.

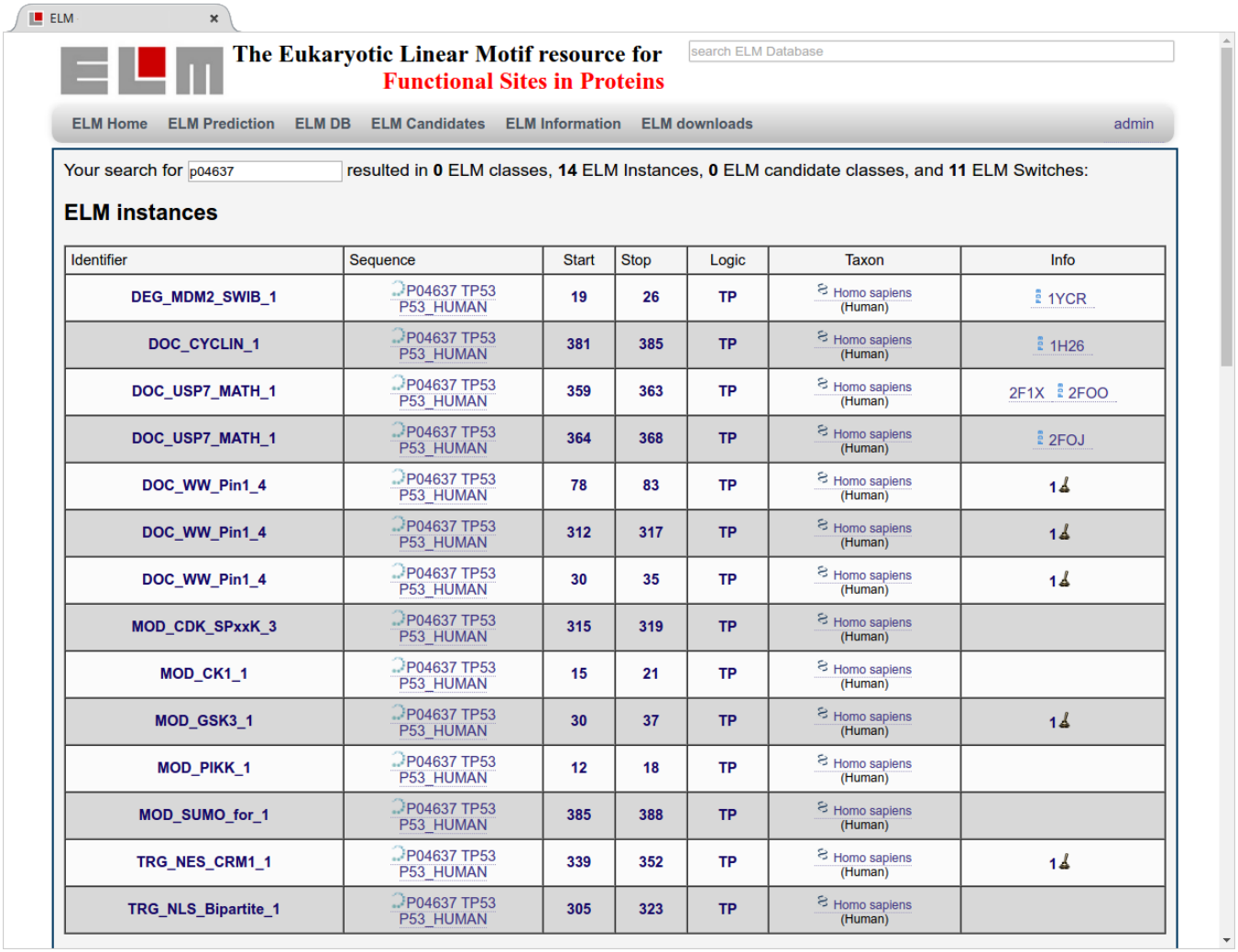


Figure 19: The instances retrieved when performing a general search for P53_HUMAN using its UniProt identifier "P04637".

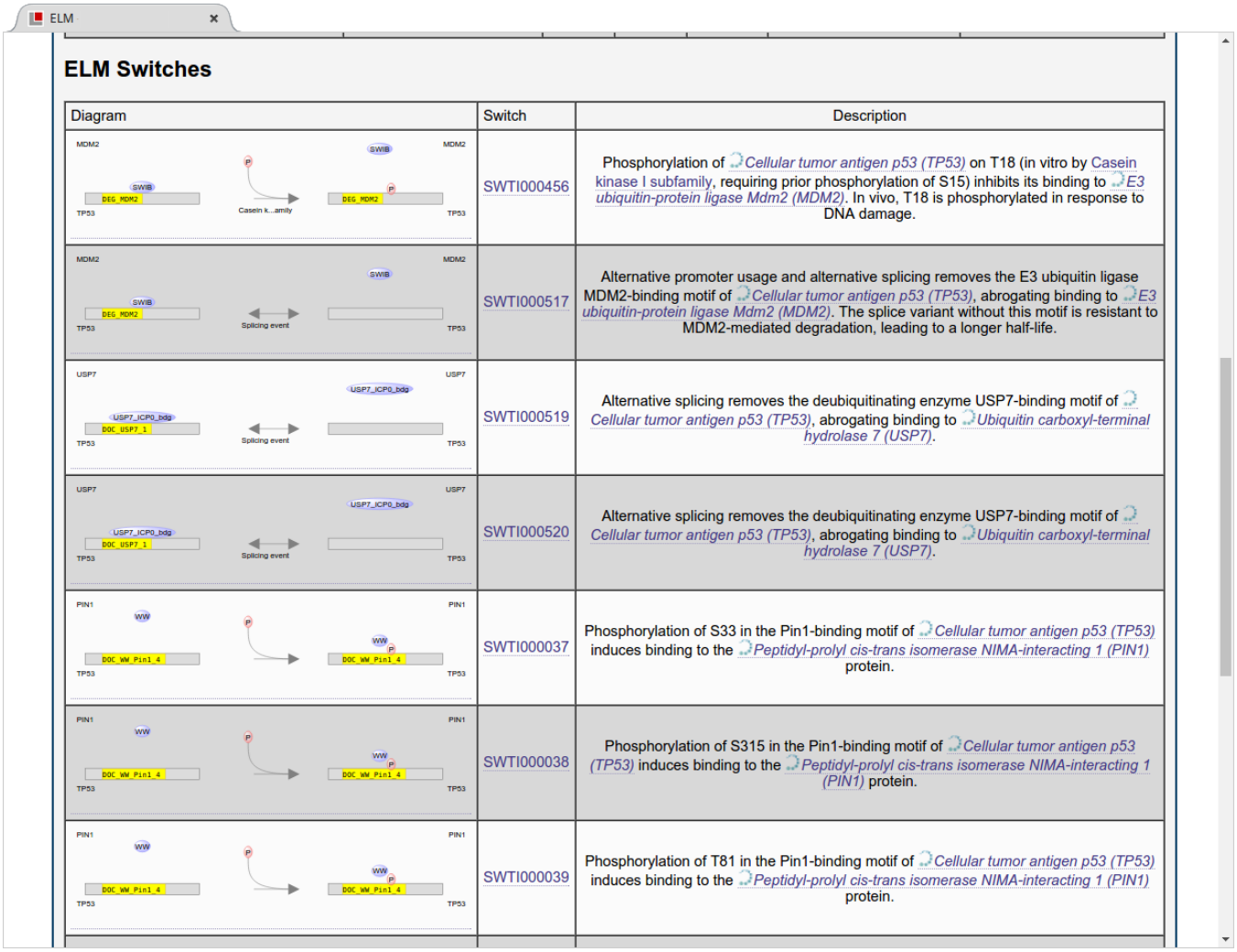


Figure 20: The switches found when performing a general search for P53_HUMAN using its UniProt identifier “P04637”,

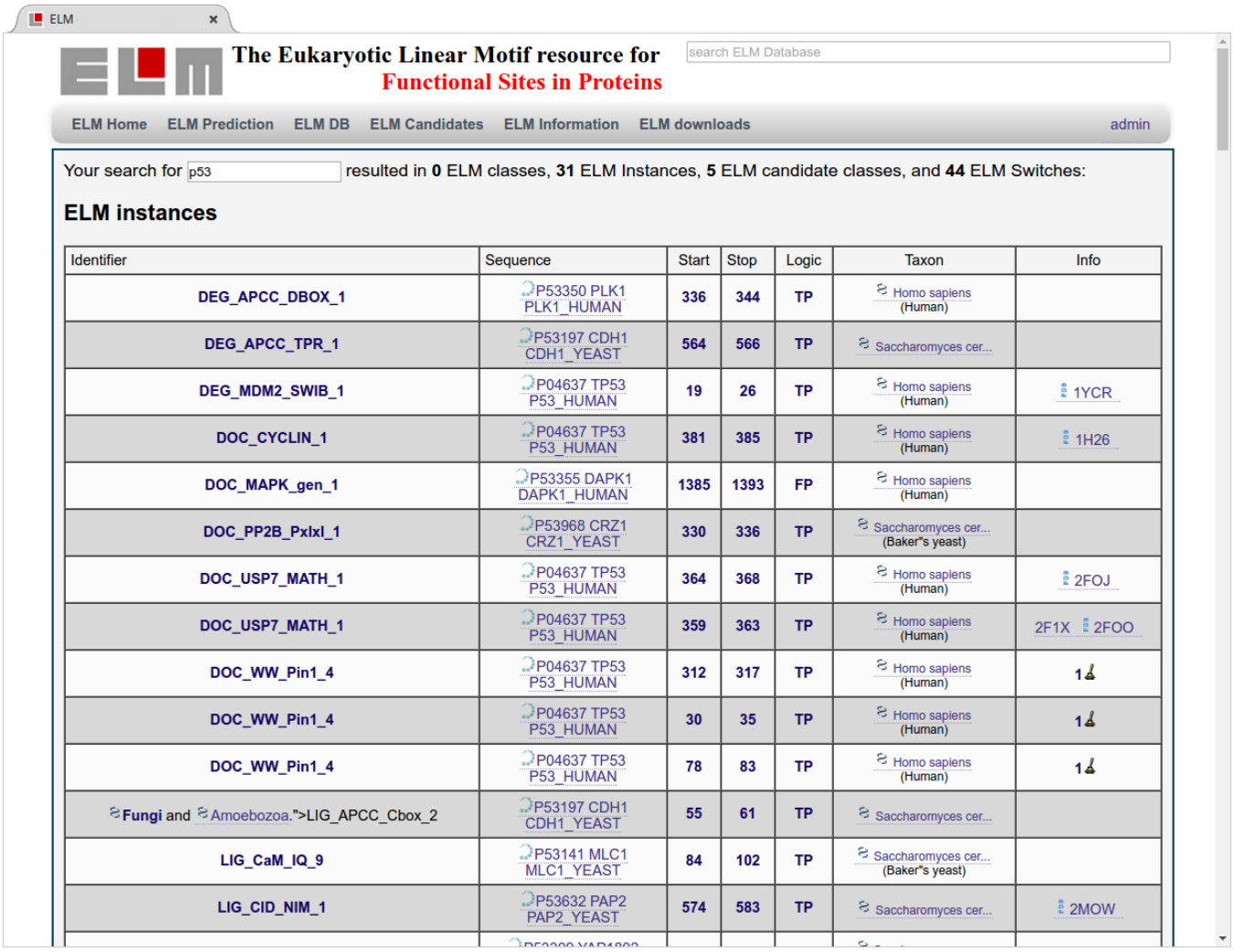


Figure 21: The results retrieved when performing a general search for P53_HUMAN using the query "p53".

Protocol 3 Detecting Short Linear Motifs in Protein Sequences

One of the most useful (and used) features in ELM is the ability to detect motifs in proteins and sequences. Given a protein's amino acid sequence, the "ELM Predictions" pipeline searches for occurrences of each motif class using regular expressions, applies a set of filters to remove false positives and creates a diagram to visualize resulting set of putative motifs.

In this protocol we will be viewing the manually annotated data of a typical protein, using p53 (UniProt ID: P53_HUMAN/P04637) as an example. We will cover how to find the manually annotated motifs and instances, the references used to annotate each instance, the experimental protocols used, and additional information including relationships to biological pathways (KEGG), diseases (OMIM) and molecular switches (switches.ELM).

Necessary Resources

Software & Hardware

A modern browser such as Firefox, Chrome, or Safari. ELM is best viewed on a laptop or desktop computer, although tablets and smartphones will also work.

Predicting ELM instances: Input form

1. Open a browser, and navigate to the ELM homepage: http://elm.eu.org. Enter the UniProt ID P53_HUMAN in the search field labelled "Enter a UniProt identifier or accession number". While typing, the page should autocomplete your input "P53_HUMAN / P04637 (*Homo sapiens*)" and already pre-fill other fields of the input form (Figure 22). Click on this entry to confirm that you want to search for motifs in this protein. Click on **Submit** to send the query to the server.

   The autocompletion mechanism queries UniProt for the given protein identifier; if it succeeds, then additional information from UniProt will be used to auto-populate the input boxes. In this example, P53_HUMAN *is recognized as a human protein, and so "Homo sapiens" is automatically filled in the "Taxonomic Context" field. Also, p53 has been annotated (by UniProt) to be localized to nucleus, cytosol, endoplasmic reticulum and mitochondrion, so these are also automatically applied as search criteria. The motif cutoff of "100" is a sufficiently high (lenient) threshold to ensure no motif class is filtered out based on motif probability.*

2. Select the search criteria (optional). To restrict the search to include motifs that are active only in certain cellular compartments, select one or more from the "Cell compartment" list (use the "control" key to select more than one option). It is also possible to select a "Taxonomic Context" to restrict the search to motifs from certain species. Start typing a species name in the "Taxonomic Context" input field to get an auto-completed list of species to select from. Additionally, a "Motif probability cutoff" can be used to only retain ELM classes whose pattern probability is below the given value. These filters are implemented later in **??**). For now, leave all all filters at the default values that were auto-populated for p53.

   TODO: Repeat search using stringent filters (homo sapiens, nucleus, 0.01) Do we want to do this? - Marc

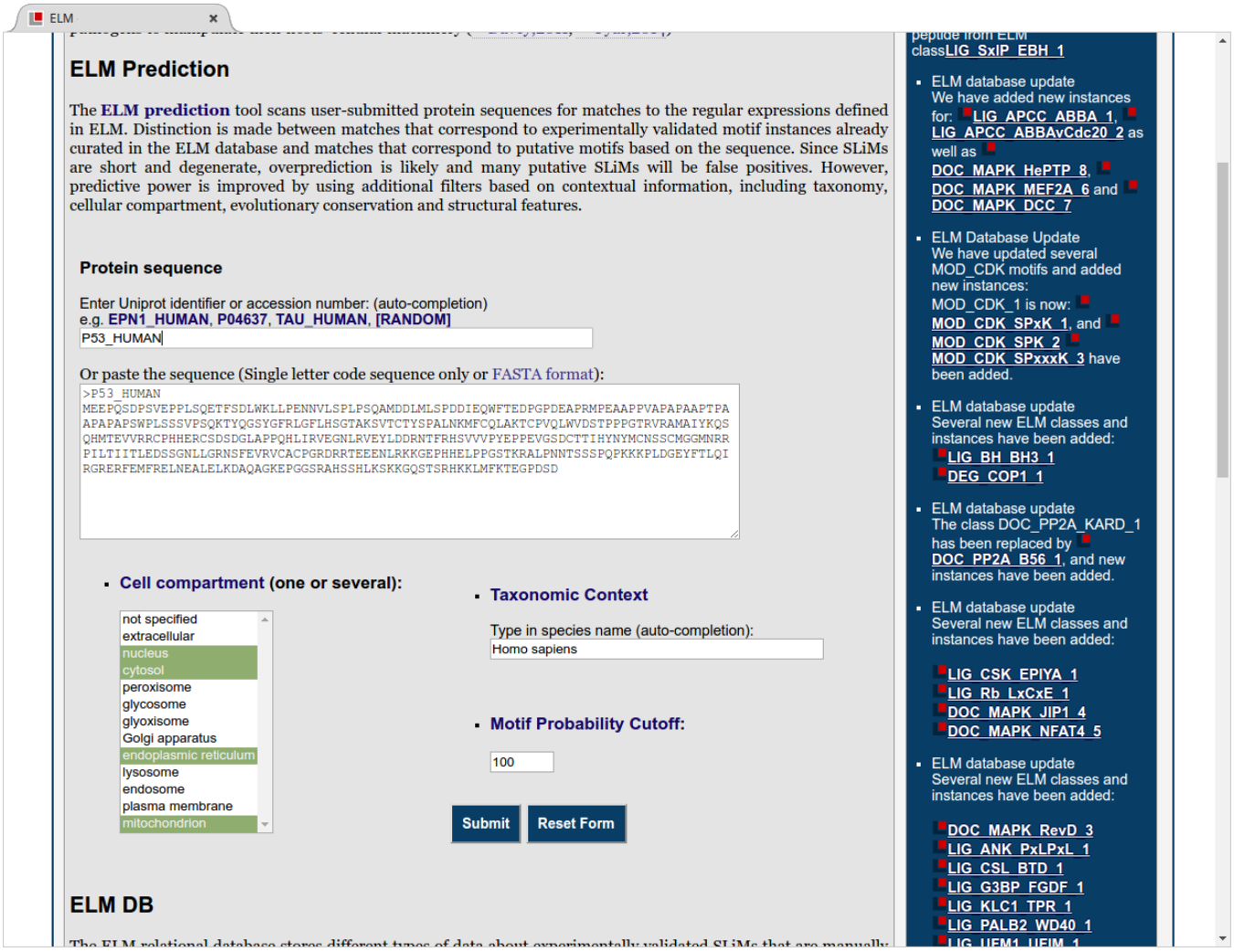


Figure 22: The ELM input page for predicting motifs in a protein. Here, all fields have been filled in; strictly only the protein ID field is necessary to perform a search.

YES, we do. HD

Interpreting the prediction results: Graphical Summary

3. Click **submit** to start searching for motifs. You will be brought to an intermediate page indicating that your results are being processed, and should be redirected to the final results page within a minute. You can bookmark this page: The results are stored for a week.

   The results are summarized in the first figure on the results page (see Figure 23). The graphical summary shows the results generated by the ELM prediction pipeline, combined with additional filters and information from external resources. The visualization should help you to interpret the results and to assess whether or not a motif is present in a sequence, as well as how likely it is to be functional, based on its structural context and evolutionary conservation. Motif instances that are manually annotated in the database appear as red (TP) or yellow (FP) ovals in the graphic. Blue/gray squares represent predicted motif occurrences.

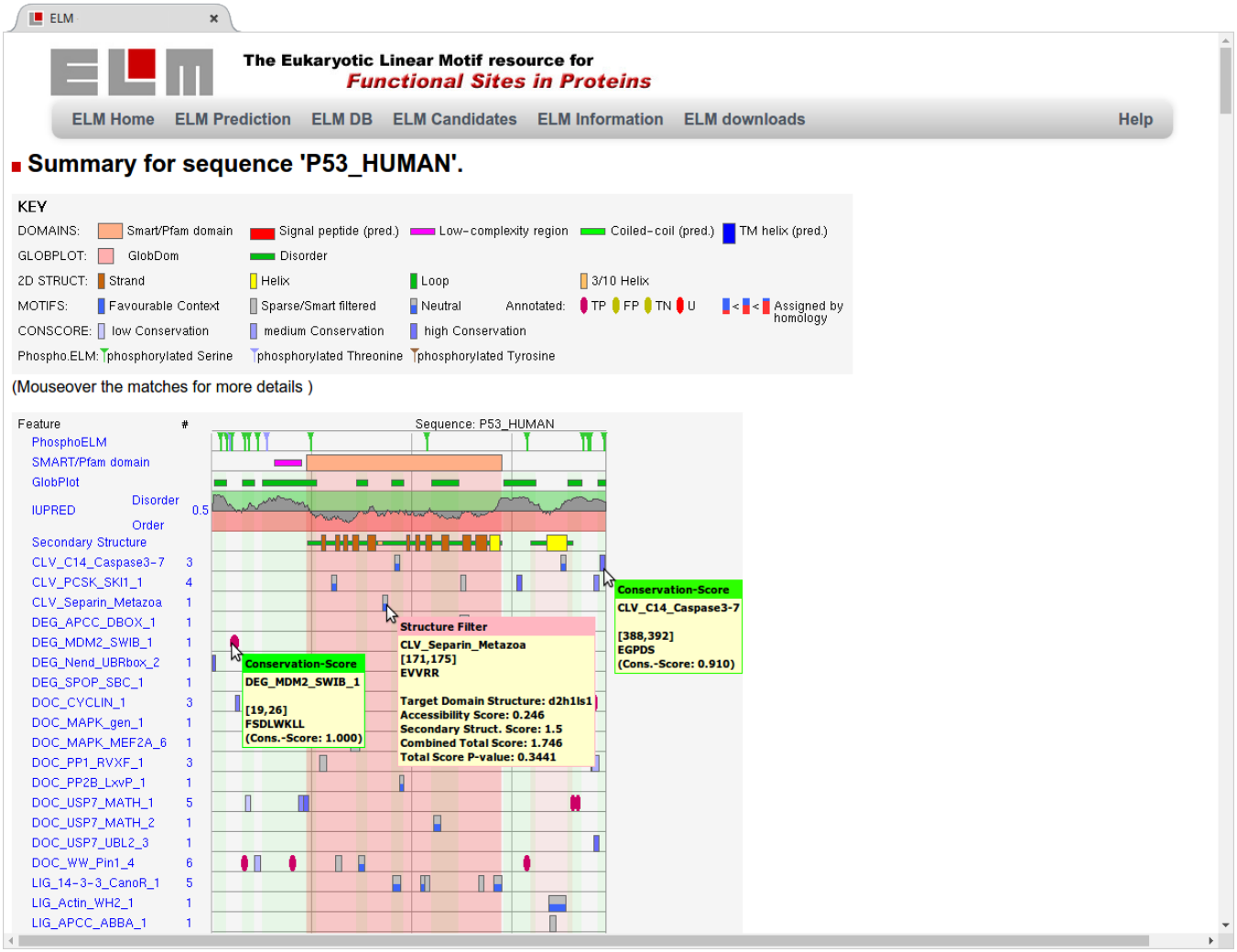


Figure 23: The graphical results summary of the ELM Prediction pipeline for P53_HUMAN. Note that not all motif detections are shown (the image is truncated at the bottom). The top five rows show a set of structural features. Annotated and predicted motifs are shown as differently colored ovals/boxes. The information screens for three motifs are shown: CLV_C14_CASPASE3-7, DEG_MDM2_SWIB_1 and CLV_SEPARIN_METAZOA.

4. The first row contains phosphorylation sites, as retrieved from PhosphoELM (Dinkel et al. 2011), and shows whether the phosphorylated amino acid is a serine, threonine or tyrosine. PhosphoELM is a database of manually annotated phosphorylation sites, obtained from scientific publications from low and high-throughput experiments. You can follow the link to PhosphoELM by clicking on the phosphorylation site in the image to get more information on individual phosphorylation sites.

   Phosphorylation sites are only available when the search is performed with a protein accession (eg. not *with a FASTA sequence alone) in step 1 and there is relevant information annotated in the PhosphoELM database. Phosphorylation sites are relevant to interpret ELM motif predictions, when the predicted motifs require to be phosphorylated (as in several docking and ligand binding motifs) and for predicting phosphorylation motifs.*

5. The second row shows SMART and PFAM domains detected by the SMART database (Schultz et al. 1998; Letunic et al. 2015; Schultz et al. 1998) (Figure 23). Hover the mouse over these domains to see their names and exact start and end positions.

   In order to be functional, motifs need to be accessible, and therefore they are usually not found within globular domains or structured regions (Davey et al. 2012). Any motif detected by the ELM prediction pipeline inside of a SMART domain are less likely to be functional, and therefore are shown with a gray box background (see commentary section "Structure Filter" at page 46).

6. The third row shows globular and disordered regions in the sequence, as predicted by GlobPlot (Linding et al. 2003). The fourth and fifth rows contain results from IUPred (Dosztányi et al. 2005), another predictor of disordered protein regions. Protein segments with an IUPred score above 0.5 are considered to be disordered (see commentary section "Disorder Filter" at page 46).

   Motifs are typically only functional when found in intrinsically disordered regions. Any motif occurrence detected by the ELM prediction pipeline that falls within disordered regions are more likely to be functional.

7. The 5th row (Figure 23) contains information on the protein's structure (see commentary section "Structure Filter" at page 46). The secondary structure is predicted by mapping the motif occurrence onto high quality reference domain structures (Via et al. 2009). Check the graphical representation, and whether the output of the secondary structure filter and the disorder predictors agree with each other with respect to which parts of the sequence are considered structured or disordered.

8. The remainder of the figure (below "secondary structure" output) displays predicted and annotated motif instances, overlayed with the structural context from rows 2 and 3 (SMART domains and GlobPlot). A blue square indicates a single motif occurrence, and intensity of the color indicates the conservation of this sequence across a group of homologous proteins. Boxes in gray are motif occurrences that have been filtered out by the structure filter. Boxes that are blue & gray are neutral (residing in structural context, but the secondary structure detected a loop region). If the sequence is already present in the ELM database, any motif instances that have already been annotated are shown as ovals. Lastly, any motifs detected that are annotated to be functional in homologous sequences, are shown as red & blue rectangles (see commentary section "Instance Mapper" at page 47).

   In the case that not enough homologous sequences were detected to build an alignment, no conservation score can be calculated. Therefore all of the motif occurrences will be shown in a uniform shade of blue.

9. Place the cursor over the blue box for motif occurrence CLV_C14_CASPASE3-7 at the end of the sequence (position 388–392). This will trigger the green and yellow information screen shown on the top right in Fig. 23. This motif is in a disordered region, and has not been filtered out by the structural filter. Also, its conservation score of 0.910 is very high, indicating that this motif is highly conserved.

   The conservation score is based on how conserved the sequence is across a set of homologous proteins (see commentary section "Conservation Filter" at page 46).

10. Place the cursor over the blue & gray rectangle for motif CLV_SEPARIN_METAZOA at position 171–175, a motif which was flagged as "neutral" by the ELM prediction pipeline. This will trigger the

information screen (with the pink header) shown in Figure 23 to appear. This motif resides inside of the p53 PFAM domain, and thus has been subjected to "structural filtering". However, the secondary structure prediction suggests that this motif occurs within the looped region of this domain, so may be accessible.

The information screen pop-up shows scores for all of the individual criteria used by the secondary structure filter: The name of the domain, the accessibility score, secondary structure score, a combined total score, and the associated total score P-value (Via et al. 2009).

11. Place the cursor over the red oval for DEG_MDM2_SWIB_1 at position 19–26. This motif is an annotated instances in the ELM database, and is therefore a bona-fide experimentally validated instance.

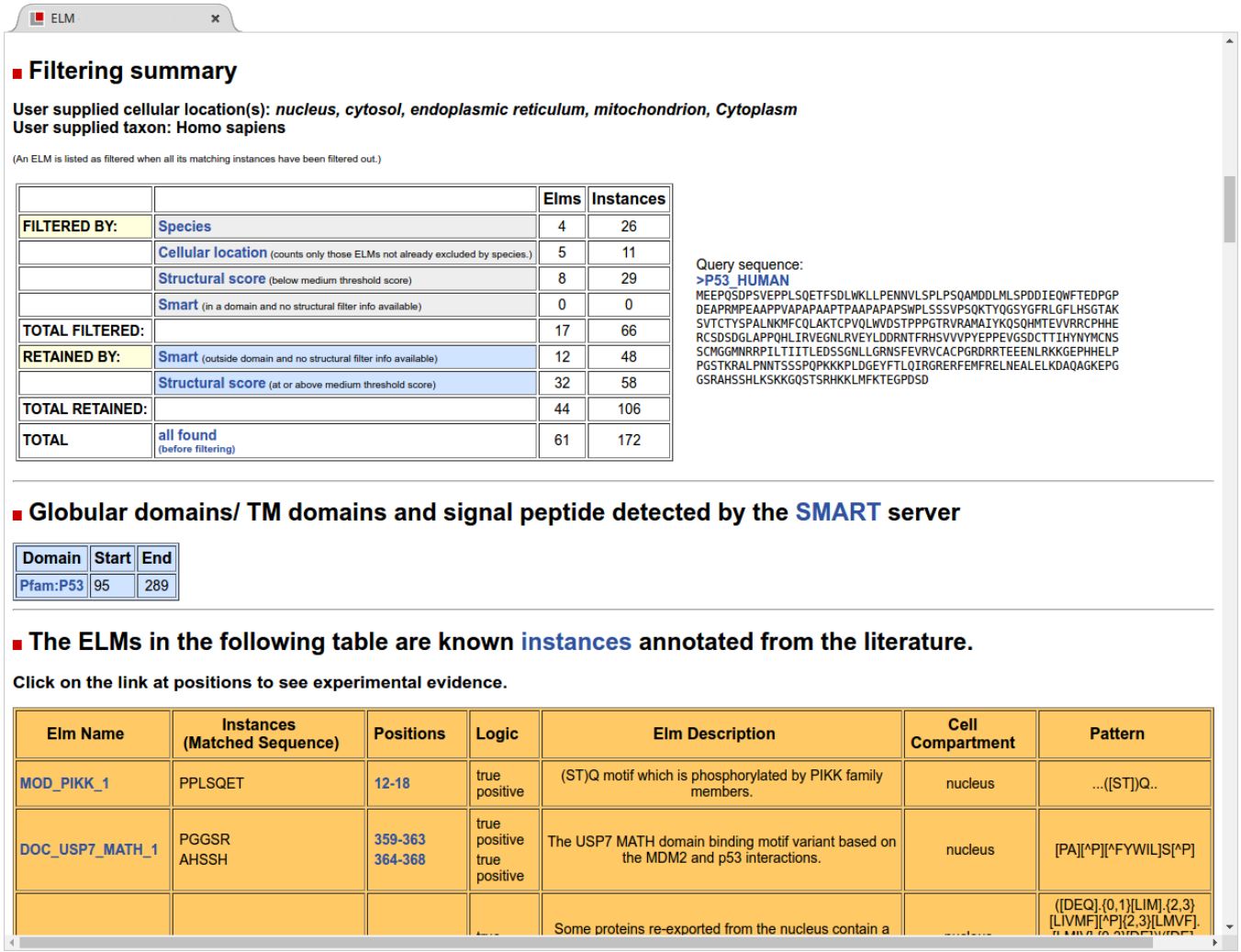


Figure 24: This section of the results contains additional details on the homologue alignments used to calculate the conservation score, filtering results and globular domains.

12. Scroll down to below the results graphic to find additional information on the ELM prediction pipeline's results (Figure 24). The first section contains links to download or view the multiple

sequence alignments of homologous proteins used to calculate the conservation score. Click on the link "Click here to enable the multiple sequence alignment viewer" to open the alignment in Jalview (note: this requires the Java browser plugin, which might not be available on some browsers). Alternatively you can also download the "alignment", "conservation features" and "phosphosite features" files separately to view on a desktop (non-browser) installation of Jalview (Waterhouse et al. 2009).

The search for possible homologues is performed against the UniRef90 database, a dataset of protein sequences with less than 90 percent identity between any two of them (Suzek et al. 2007). It may occur that the BLAST results are not finished when the results page is shown: We suggest to refresh the page if you see the message "Either not enough data available to calculate a sequence alignment or the calculations haven't finished yet". In some cases it is also possible that no homologues will be detected. If you have refreshed the page after waiting for more than 3 minutes, this is most likely the case.

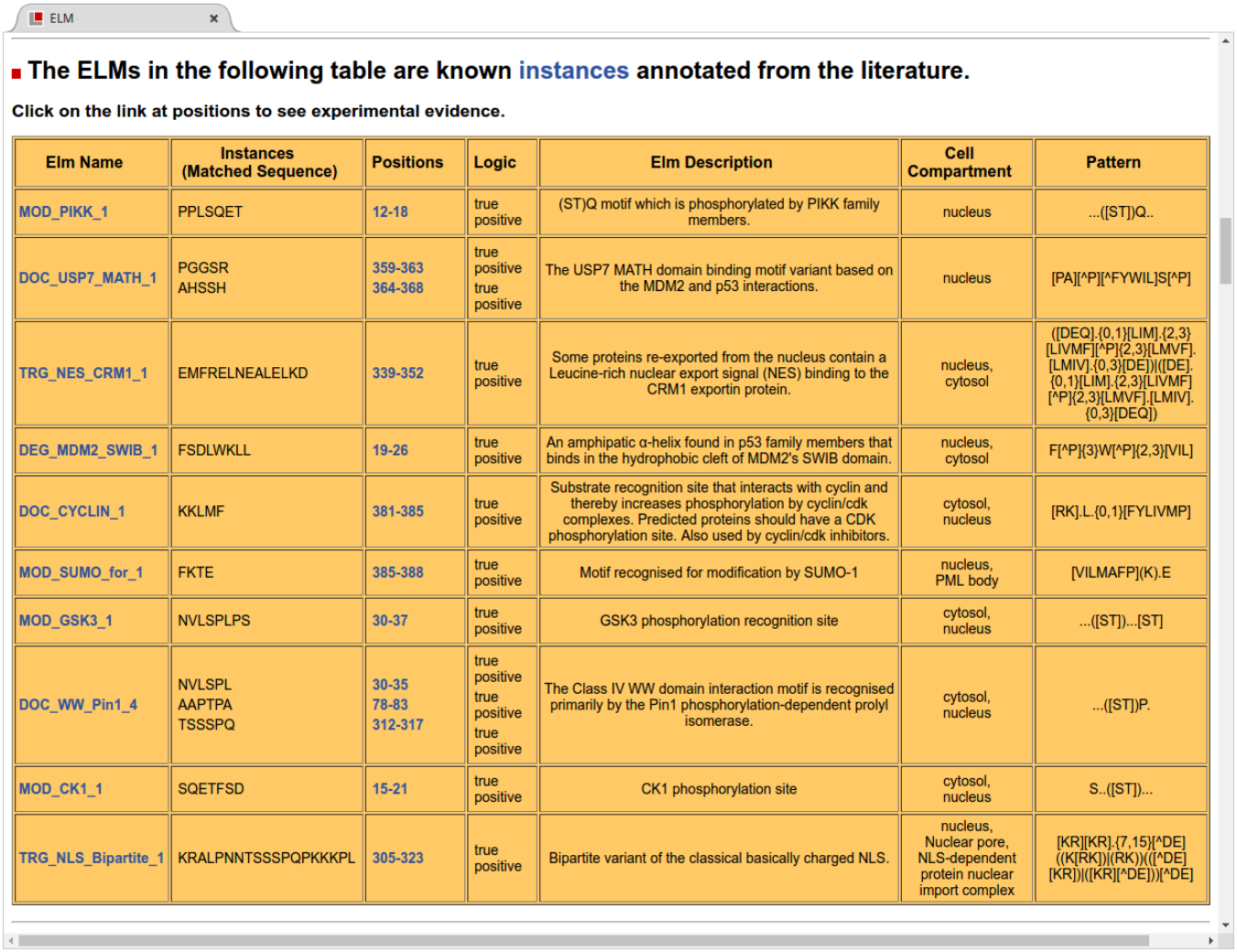
13. Scroll down to the section titled "Filtering Summary" to view some statistics about how many motifs and instances were filtered out (Fig. 24). The first two lines contain information on which filters were applied in step 1 of this protocol. In this case 4 motifs representing 26 instances were filtered out as they did not occur in *Homo sapiens*. An additional 5 motifs (representing 11 instances) were filtered out because they are not annotated to the cell compartments automatically filled in on the search page (step 1). The next three lines ("SMART" & "Structural score") show how many motifs and instances were not removed by the SMART and Secondary structure filters. A total of 42 motifs (representing 106 instances) passed the structural filter.

    Note that the graphical summary above does not contain sequences filtered out by the "cell compartment" and "taxonomic context" filters. However those filtered out by the SMART and Structural scores are shown in the graphic above (as gray rectangles).

14. Scroll down to the section with the header "Globular domains/ TM domains and signal peptide detected by the SMART server" (Figure 24). This section contains information on which domains were detected by the SMART server, and their positions. Clicking on their names will bring you to the entry for that domain on the SMART or PFAM homepage. In this case the only domains detected is the "p53" PFAM domain.

15. On the results page, scroll down to the heading: "The ELMs in the following table are known instances annotated from the literature" (25). This table has details of the motifs and instances which have been manually annotated in the ELM database. The columns show each motif name, the sequence(s) that matched the motif as well as their starting and ending positions and the logic of the annotation followed by a short description of each motif, to which cell compartments its has been associated, and finally the regular expression of the motif.

16. Scroll further down to the section title "Results of ELM motif search after globular domain filtering, structural filtering and context filtering" to obtain an overview of all of the motifs and motif instances detected (26) Each of the rows is a "predicted" motif: A sequence matching a motif's regular expression has been detected that has also passed the "structural filter". Each row displays the motif identified, the matching peptide sequence and its position. Additional information is shown about the motif, its cell compartment and its regular expression. If the motif was detected in a homologue, the column "PHI-Blast Instance mapping" contains a link to the multiple sequence alignment of the



The ELMs in the following table are known instances annotated from the literature.

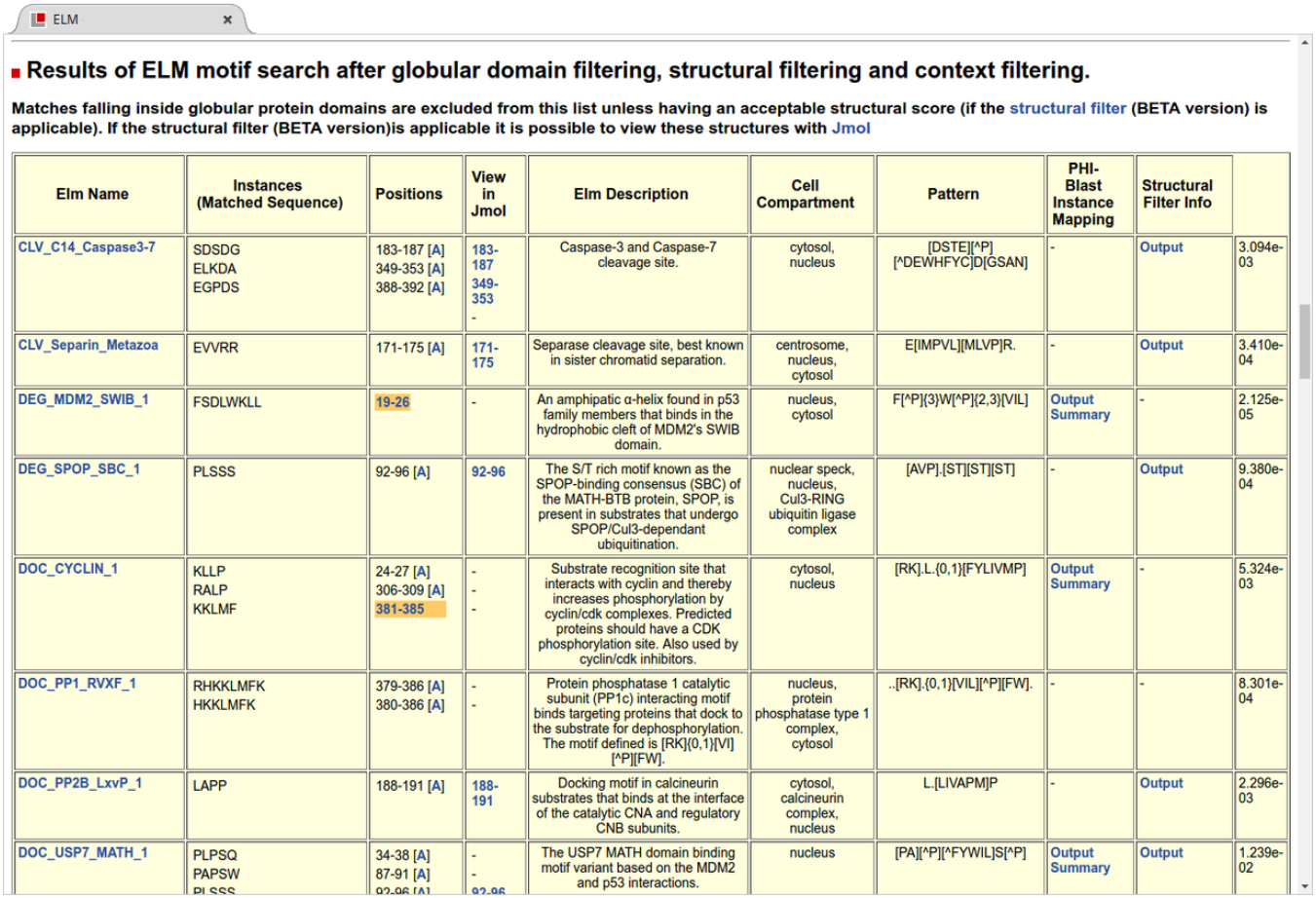
Click on the link at positions to see experimental evidence.

Elm Name	Instances (Matched Sequence)	Positions	Logic	Elm Description	Cell Compartment	Pattern
MOD_PIKK_1	PPLSQET	12-18	true positive	(ST)Q motif which is phosphorylated by PIKK family members.	nucleus	...([ST])Q..
DOC_USP7_MATH_1	PGGSR AHSSH	359-363 364-368	true positive true positive	The USP7 MATH domain binding motif variant based on the MDM2 and p53 interactions.	nucleus	[PA][^P][^FYWIL]S[^P]
TRG_NES_CRM1_1	EMFRELNEALELKD	339-352	true positive	Some proteins re-exported from the nucleus contain a Leucine-rich nuclear export signal (NES) binding to the CRM1 exportin protein.	nucleus, cytosol	([DEQ].{0,1}[LIM].{2,3}[LIVMF][^P]{2,3}[LMVF].[LMIV].{0,3}[DE])|([DE].{0,1}[LIM].{2,3}[LIVMF][^P]{2,3}[LMVF].[LMIV].{0,3}[DEQ])
DEG_MDM2_SWIB_1	FSDLWKLL	19-26	true positive	An amphipatic α-helix found in p53 family members that binds in the hydrophobic cleft of MDM2's SWIB domain.	nucleus, cytosol	F[^P]{3}W[^P]{2,3}[VIL]
DOC_CYCLIN_1	KKLMF	381-385	true positive	Substrate recognition site that interacts with cyclin and thereby increases phosphorylation by cyclin/cdk complexes. Predicted proteins should have a CDK phosphorylation site. Also used by cyclin/cdk inhibitors.	cytosol, nucleus	[RK].L.{0,1}[FYLIVMP]
MOD_SUMO_for_1	FKTE	385-388	true positive	Motif recognised for modification by SUMO-1	nucleus, PML body	[VILMAFP](K).E
MOD_GSK3_1	NVLSPLPS	30-37	true positive	GSK3 phosphorylation recognition site	cytosol, nucleus	...([ST])...[ST]
DOC_WW_Pin1_4	NVLSPL AAPTPA TSSSPQ	30-35 78-83 312-317	true positive true positive true positive	The Class IV WW domain interaction motif is recognised primarily by the Pin1 phosphorylation-dependent prolyl isomerase.	cytosol, nucleus	...([ST])P.
MOD_CK1_1	SQETFSD	15-21	true positive	CK1 phosphorylation site	cytosol, nucleus	S..([ST])...
TRG_NLS_Bipartite_1	KRALPNNTSSSPQPKKKPL	305-323	true positive	Bipartite variant of the classical basically charged NLS.	nucleus, Nuclear pore, NLS-dependent protein nuclear import complex	[KR][KR].{7,15}[^DE]((K[RK])|(RK))(([^DE][KR])|([KR][^DE]))[^DE]

Figure 25: The ELM prediction pipeline section displaying the p53 motifs that are "known", and have been annotated in the ELM database.

homologous proteins. If a motif instance has been filtered out by the "structural filter", the "Structural filter info" column contains a link to a page with details on why. The last column contains information on the Probability filter: the probability reflects the chance to observe this motif in any random amino acid sequence (see section Protocol 1)

17. Scroll further down to the heading "List of excluded ELMs falling inside SMART/PFAM domains and/or scoring poorly with the structural filter (if applicable)." (Figure 27) This table is similar to the one described above, but shows motif matches which were rejected by the structural filter.



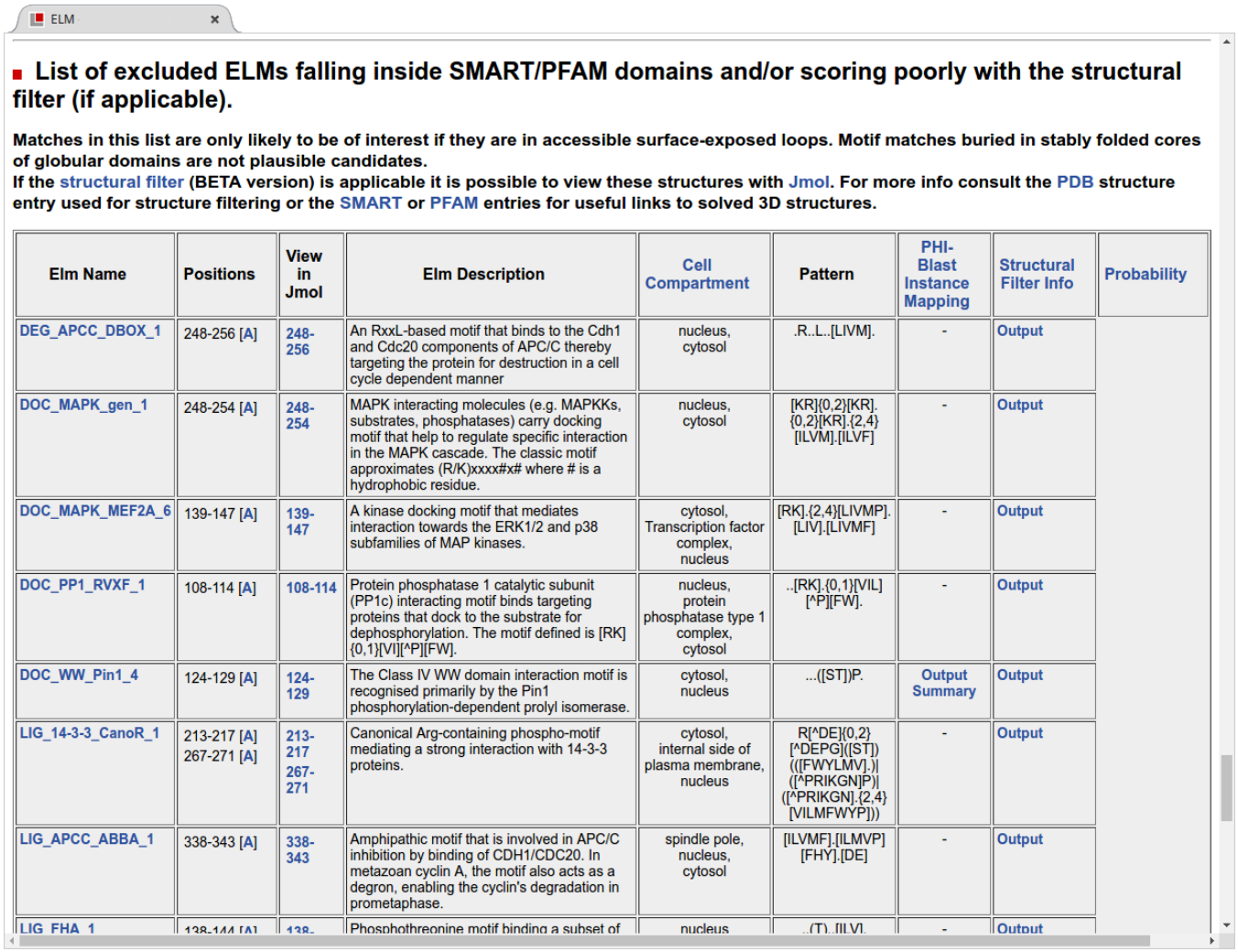
ELM

Results of ELM motif search after globular domain filtering, structural filtering and context filtering.

Matches falling inside globular protein domains are excluded from this list unless having an acceptable structural score (if the structural filter (BETA version) is applicable). If the structural filter (BETA version)is applicable it is possible to view these structures with Jmol

Elm Name	Instances (Matched Sequence)	Positions	View in Jmol	Elm Description	Cell Compartment	Pattern	PHI-Blast Instance Mapping	Structural Filter Info	
CLV_C14_Caspase3-7	SDSDG ELKDA EGPDS	183-187 [A] 349-353 [A] 388-392 [A]	183-187 349-353 -	Caspase-3 and Caspase-7 cleavage site.	cytosol, nucleus	[DSTE][^P][^DEWHFYC]D[GSAN]	-	Output	3.094e-03
CLV_Separin_Metazoa	EVVRR	171-175 [A]	171-175	Separase cleavage site, best known in sister chromatid separation.	centrosome, nucleus, cytosol	E[IMPVL][MLVP]R.	-	Output	3.410e-04
DEG_MDM2_SWIB_1	FSDLWKLL	19-26	-	An amphipatic α-helix found in p53 family members that binds in the hydrophobic cleft of MDM2's SWIB domain.	nucleus, cytosol	F[^P]{3}W[^P]{2,3}[VIL]	Output Summary	-	2.125e-05
DEG_SPOP_SBC_1	PLSSS	92-96 [A]	92-96	The S/T rich motif known as the SPOP-binding consensus (SBC) of the MATH-BTB protein, SPOP, is present in substrates that undergo SPOP/Cul3-dependant ubiquitination.	nuclear speck, nucleus, Cul3-RING ubiquitin ligase complex	[AVP].[ST][ST][ST]	-	Output	9.380e-04
DOC_CYCLIN_1	KLLP RALP KKLMF	24-27 [A] 306-309 [A] 381-385	- - -	Substrate recognition site that interacts with cyclin and thereby increases phosphorylation by cyclin/cdk complexes. Predicted proteins should have a CDK phosphorylation site. Also used by cyclin/cdk inhibitors.	cytosol, nucleus	[RK].L.{0,1}[FYLIVMP]	Output Summary	-	5.324e-03
DOC_PP1_RVXF_1	RHKKLMFK HKKLMFK	379-386 [A] 380-386 [A]	- -	Protein phosphatase 1 catalytic subunit (PP1c) interacting motif binds targeting proteins that dock to the substrate for dephosphorylation. The motif defined is [RK]{0,1}[VI][^P][FW].	nucleus, protein phosphatase type 1 complex, cytosol	..[RK].{0,1}[VIL][^P][FW].	-	-	8.301e-04
DOC_PP2B_LxvP_1	LAPP	188-191 [A]	188-191	Docking motif in calcineurin substrates that binds at the interface of the catalytic CNA and regulatory CNB subunits.	cytosol, calcineurin complex, nucleus	L.[LIVAPM]P	-	Output	2.296e-03
DOC_USP7_MATH_1	PLPSQ PAPSW PLSSS	34-38 [A] 87-91 [A] 92-96 [A]	- - 92-96	The USP7 MATH domain binding motif variant based on the MDM2 and p53 interactions.	nucleus	[PA][^P][^FYWIL]S[^P]	Output Summary	Output	1.239e-02

Figure 26: This table contains the list of putative motifs detected in the query sequence (only the top part of the table is shown). These are "predictions" in the sense that each of these motifs could be found in the sequence after applying (structure/context) filtering, however no experimental evidence has been annotated (yet) to determine if they are biologically functional.



List of excluded ELMs falling inside SMART/PFAM domains and/or scoring poorly with the structural filter (if applicable).

Matches in this list are only likely to be of interest if they are in accessible surface-exposed loops. Motif matches buried in stably folded cores of globular domains are not plausible candidates.
If the structural filter (BETA version) is applicable it is possible to view these structures with Jmol. For more info consult the PDB structure entry used for structure filtering or the SMART or PFAM entries for useful links to solved 3D structures.

Elm Name	Positions	View in Jmol	Elm Description	Cell Compartment	Pattern	PHI-Blast Instance Mapping	Structural Filter Info	Probability
DEG_APCC_DBOX_1	248-256 [A]	248-256	An RxxL-based motif that binds to the Cdh1 and Cdc20 components of APC/C thereby targeting the protein for destruction in a cell cycle dependent manner	nucleus, cytosol	.R..L..[LIVM].	-	Output	
DOC_MAPK_gen_1	248-254 [A]	248-254	MAPK interacting molecules (e.g. MAPKKs, substrates, phosphatases) carry docking motif that help to regulate specific interaction in the MAPK cascade. The classic motif approximates (R/K)xxxx#x# where # is a hydrophobic residue.	nucleus, cytosol	[KR]{0,2}[KR].{0,2}[KR].{2,4}[ILVM].[ILVF]	-	Output	
DOC_MAPK_MEF2A_6	139-147 [A]	139-147	A kinase docking motif that mediates interaction towards the ERK1/2 and p38 subfamilies of MAP kinases.	cytosol, Transcription factor complex, nucleus	[RK].{2,4}[LIVMP].[LIV].[LIVMF]	-	Output	
DOC_PP1_RVXF_1	108-114 [A]	108-114	Protein phosphatase 1 catalytic subunit (PP1c) interacting motif binds targeting proteins that dock to the substrate for dephosphorylation. The motif defined is [RK]{0,1}[VI][^P][FW].	nucleus, protein phosphatase type 1 complex, cytosol	..[RK].{0,1}[VIL][^P][FW].	-	Output	
DOC_WW_Pin1_4	124-129 [A]	124-129	The Class IV WW domain interaction motif is recognised primarily by the Pin1 phosphorylation-dependent prolyl isomerase.	cytosol, nucleus	...([ST])P.	Output Summary	Output	
LIG_14-3-3_CanoR_1	213-217 [A] 267-271 [A]	213-217 267-271	Canonical Arg-containing phospho-motif mediating a strong interaction with 14-3-3 proteins.	cytosol, internal side of plasma membrane, nucleus	R[^DE]{0,2}[^DEPG]([ST])(([FWYLMV].)|([^PRIKGN]P)|([^PRIKGN].{2,4}[VILMFWYP]))	-	Output	
LIG_APCC_ABBA_1	338-343 [A]	338-343	Amphipathic motif that is involved in APC/C inhibition by binding of CDH1/CDC20. In metazoan cyclin A, the motif also acts as a degron, enabling the cyclin's degradation in prometaphase.	spindle pole, nucleus, cytosol	[ILVMF].[ILMVP][FHY].[DE]	-	Output	
LIG_FHA_1	138-144 [A]	138-	Phosphothreonine motif binding a subset of	nucleus	..(T)..[ILV]	-	Output	

Figure 27: This table contains the list of putative motifs detected in the query sequence (only the top part of the table is shown) which were excluded by the structural or context filter.

Protocol 4 Detecting Short Linear Motifs in Unknown Protein Sequences

The ELM motif detection pipeline is a very powerful way to obtain a lot of information on which motifs are present, in which structural context they are. However, determining which motifs are actually true positive detections requires interpreting all of these results, as well incorporating as much biological knowledge as possible. In this protocol we will be following a typical example of how one might use the ELM pipeline to search for motifs in novel sequences.

Some pathogens have evolved short linear motifs in effector proteins to modify the intracellular signalling of their host cell for their own convenience (Via et al. 2015). The Gram-negative bacteria *Chromobacterium violaceum* is a opportunistic pathogen of humans whose mechanism of pathogenicity remain poorly understood. Its genome encodes a type three secretion system (T3SS) that is used by different pathogens to translocate bacterial proteins into the infected cells. Interestingly, the genes encoding this T3SS as well as other genes located in the same genomic location are very similar to the ones in *Salmonella spp.* except for a couple of genes including the modular protein SptP (de Brito et al. 2004). SptP in *Salmonella spp.* is a secreted protein tyrosine phosphatase (Kaniga et al. 1996) whose closest homolog in *C. violaceum* is the protein CV_0974. To further understand the possible biological function of the protein CV_0974 we will use it as an example application of the ELM server motif detection pipeline.

Necessary Resources

Software & Hardware

A modern browser such as Firefox, Chrome, or Safari. ELM is best viewed on a laptop or desktop computer, although tablets and smartphones will also work.

Files

You need to download the following file from UniProt containing the FASTA sequence of CV_0974/Q7NZE8: http://www.uniprot.org/uniprot/Q7NZE8.fasta

Submitting a query to ELM

1. Click on the "ELM Predictions" button in the menu to access the search query page (Fig. 28). Here you should provide the amino acid sequence of CV_0974 from UniProt (http://www.uniprot.org/uniprot/Q7NZE8.fasta), and enter it as FASTA format into the "sequence input text box".

2. In the "Taxonomic Context" field, enter the text "Homo sapiens". This will limit the search to motif classes which have been annotated for human proteins.

   Certain human diseases occur when motifs are hijacked by opportunistic pathogens (see Step 20 in Protocol 1 and Figure 16). By limiting the search to human motifs, we will identify motifs which are known to exist in humans and thus may be the target of motif hijacking.

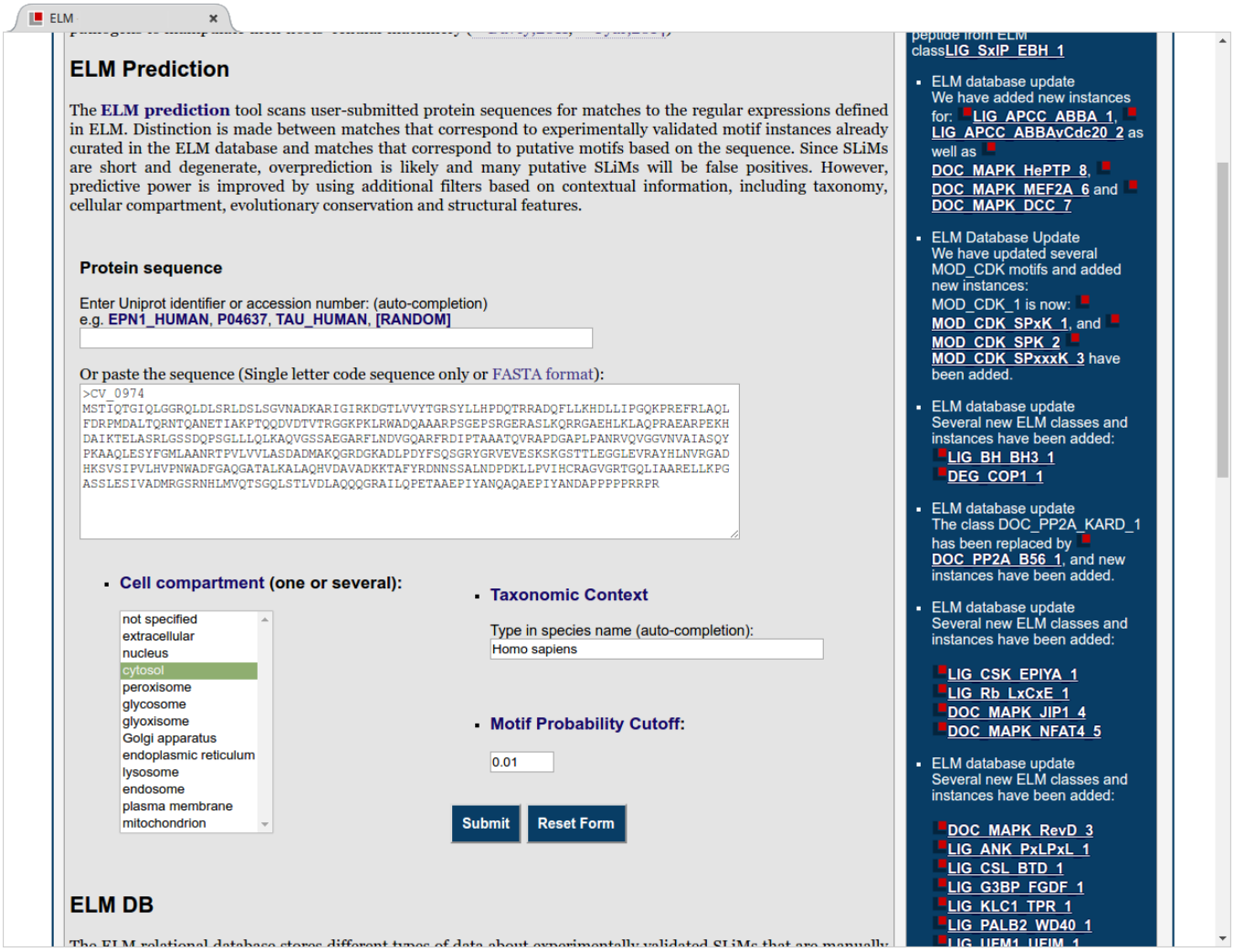


Figure 28: The input query page for finding motifs in ELM. The sequence for *C. violaceum* protein CV_0974 is used as an example for this protocol.

3. The bacterial protein CV_0974 is likely to be an effector protein, similarly to its homologue SptP. As SptP localizes to the cytosol, and we assume the same for CV_0974 so you should select "cytosol" in the "Cell compartment" field.

4. For now, you should set the "Motif probability cutoff" to the same value you used in Protocol 3 and enter "100" into this box.

Interpreting the results

Interpreting the prediction results

5. Hit **Submit** to send the query to the ELM prediction pipeline. The results are summarized in the first figure on the results page (see Fig. 29) See steps 3 – 10 of Protocol 3 for a description of the graphical summary output. In this case, there are a lot of putative motif hits and it will be difficult

Figure 29: The graphical results summary of the ELM prediction pipeline for the protein "Probable Tyrosine phosphate" (CV_0974). Note that not all motif detections are shown as the image is truncated at the bottom.

to investigate all of these. So you should try to limit your search to most promising candidates first, by trying to remove false positive hits.

6. Go back to the "ELM Predictions" page (it's probably still open in another browser tab) and enter the same values as used before in steps 1-4. Now however, in order to remove false positive hits, you should set the "Motif probability cutoff value" to a more stringent threshold. Enter "0.001" for this value. This will exclude all motif with a probability score higher than this, limiting the results to motifs which are less likely to be found by random chance (False Positives).

7. Hit **Submit** to send the query to the ELM prediction pipeline. The results are summarized in the first figure on the results page (see Fig. **??**). You notice the amount of motif classes detected has been reduced drastically (from 54 to 11).

8. By focussing on motif instances that reside in disordered context (in the amino- or carboxy-terminus

of this protein) you have successfully limited the list of ELM motif classes down to 11. (Now would be a good point in time to read up on the details of the remaining motif classes to get a better idea, about which motifs might be functional in your protein. Click on the motif identifiers on the left side of the graphical summary to go to the details page for the motif classes). For this protocol, however, just click on the motif identifier LIG_CSK_EPIYA_1 on the left side of the graphical summary to go to the details page for this particular motif class.

9. Find the entry for LIG_CSK_EPIYA_1 in the graphical summary and check the structural context for the two LIG_CSK_EPIYA_1 motifs. Both of them fall outside of the SMART domain PTPC, and reside in a region with a protein disorder (IUPred) score higher than 0.5. In the "Functional site description" it is stated that 'bacterial proteins usually have repeats of EPIYA motifs.' The ELM prediction results did indeed also detect two EPIYA motifs in a 20 amino acid range, lending further support to the likelihood that there are indeed two functional EPIYA motifs in CV_0974, which in turn suggests that these motifs may be involved in *C. violaceum's* pathogenicity.

   When this motif is tyrosine-phosphorylated it is recognized by C-terminal kinase. Different effector proteins from human pathogens like Bartonella henselae, Helicobacter pylori and Haemophilus ducreyi have been reported to use this motif to interfere with the host signalling network to induce proliferation or to avoid phagocytosis (Selbach et al. 2009; Tsutsumi 2002; Dodd et al. 2014).

Protocol 5 (Alternative Protocol) Searching the ELM database using the REST API

Many researchers are interested in large-scale analyses rather than information about individual protein sequences. To this end, individual queries to the ELM webserver with a single protein id at a time, are not practical.

For this reason, as much information as possible is made available via a REST interface (Fielding and Taylor 2002). This allows the user to interact with the ELM database and ELM webserver via scriptable URL requests. Each request can easily be tested in the browser before it is being automated in a script.

In this section we will explore the various ways in which data can downloaded both in using the browser as well as via the commandline.

Necessary Resources

Software

A modern browser such as Firefox, Chrome, or Safari. ELM is best viewed on a laptop or desktop computer, although tablets and smartphones will also work. There exist several REST client plugins for different browsers, however these are not needed for this protocol. Ideally use a commandline tool such as `curl` (`http://curl.haxx.se/`) in a terminal window. This program is available in any of the major operating systems: OSX, Windows and Linux. Of course, `curl` is only one of many different ways to access web content programatically, and we suggest to use whichever program you feel is better suited for your tasks.

Downloading all ELM classes

1. Direct your browser to the URL '`elm.eu.org/downloads`' or select **ELM Downloads** from the main menu (Figure 30). This page contains links and descriptions on how to download ELM data in text format. The datasets are split into several smaller collections (for example "Classes", "Instances", etc). Each table contains links (in orange) to download the data in appropriate formats.

   Each table also shows the 'last modified date' indicating when the data was last updated. This information is useful if you want to know when to update your local data with the most up to date ELM data as it allows you to determine whether you need to update or not.

2. Click on the first orange **html** link in the table "Classes" to navigate to the following URL: '`elm.eu.org/elms/elm_index.html`'. This page shows all of the annotated ELM classes in the database. This page is the same one as shown in Figure 3.

3. Navigate to the following URL: '`elm.eu.org/elms.html?q=CSK`' specifying `q=CSK` to limit the list of ELMs to those matching the search query "CSK". This page is again similar to the one shown in Figure 3, but with less classes.

   *This search result is identical to the result you would obtain by doing a "manual" search on the ELM Classes page (eg. typing 'CSK' in the search box and clicking **submit**) as described in step 3 of Protocol 1 (see Figure 3).*

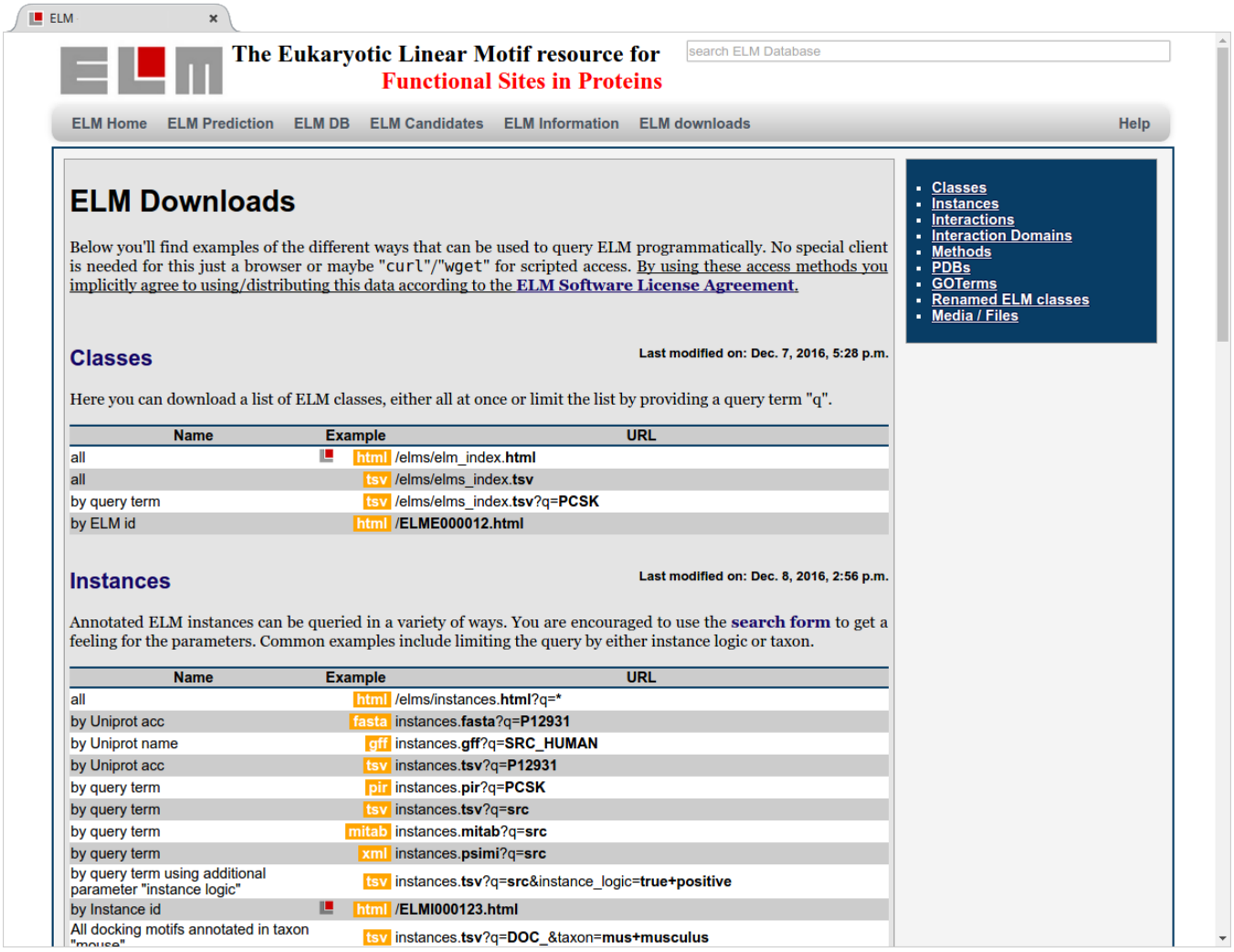


Figure 30: The ELM downloads page, which holds information about the different types of data (such as "Classes", "Instances", etc; see menu to the right) that can be obtained from the server. The orange boxes are clickable links, the URL following them are used to highlight the URL scheme used by the server (bold font denotes specifics used in the examples such as query terms, or formats).

4. Open the following URL: '`elm.eu.org/elms.tsv?q=CSK`' to download a list of classes that match the search query "CSK" (as in the previous step) in the "tab separated values" format. Note that this time we used the file extension '.tsv' instead of '.html' as before. By exchanging the '*.html*' part of the URL with '*.tsv*', we ask the webserver to give us the data in "tab-separated values" format.

   Depending on which browser you are using, the file may open directly in your browser, or you may be prompted to download the file or save it to a separate location. In the latter two cases you can open the downloaded file using a (plain) text file viewer, or possible a spreadsheet viewer (such as Microsoft Excel or LibreOffice Calc).

5. Type the following command into a command line terminal to download the same data from the previous step directly into the terminal: `curl 'http://elm.eu.org/elms/elms_index.tsv?q=CSK'`. The output should look similar to Figure 31. The column names are the same as shown in Figure 3.

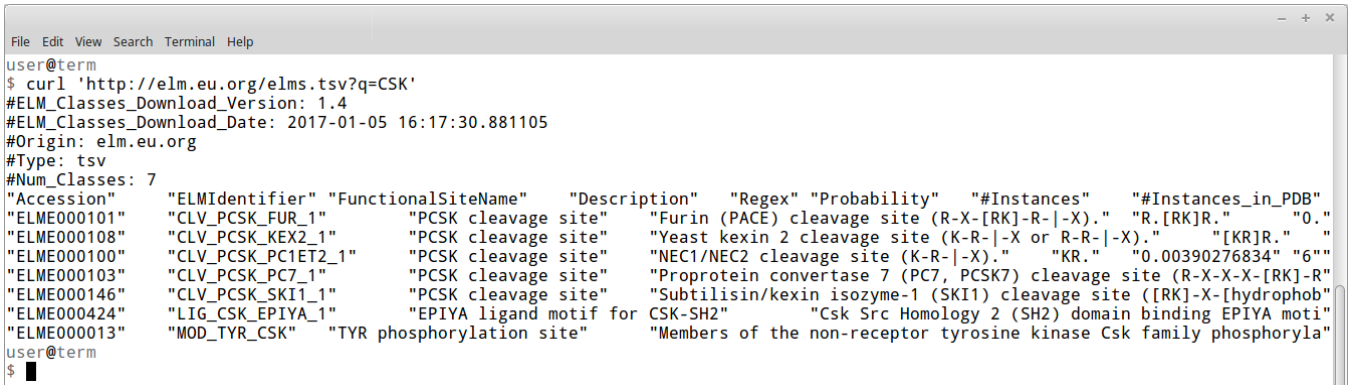


Figure 31: A screenshot of a terminal window using `curl` to download all ELM classes matching the term 'CSK'.

> *Use the curl option "`-o`" to save the results directly to a file. For example:* `curl -o classes.tsv 'http://elm.eu.org/elms/elms_index.tsv?q=CSK'` *will save the data to a file called* classes.tsv.

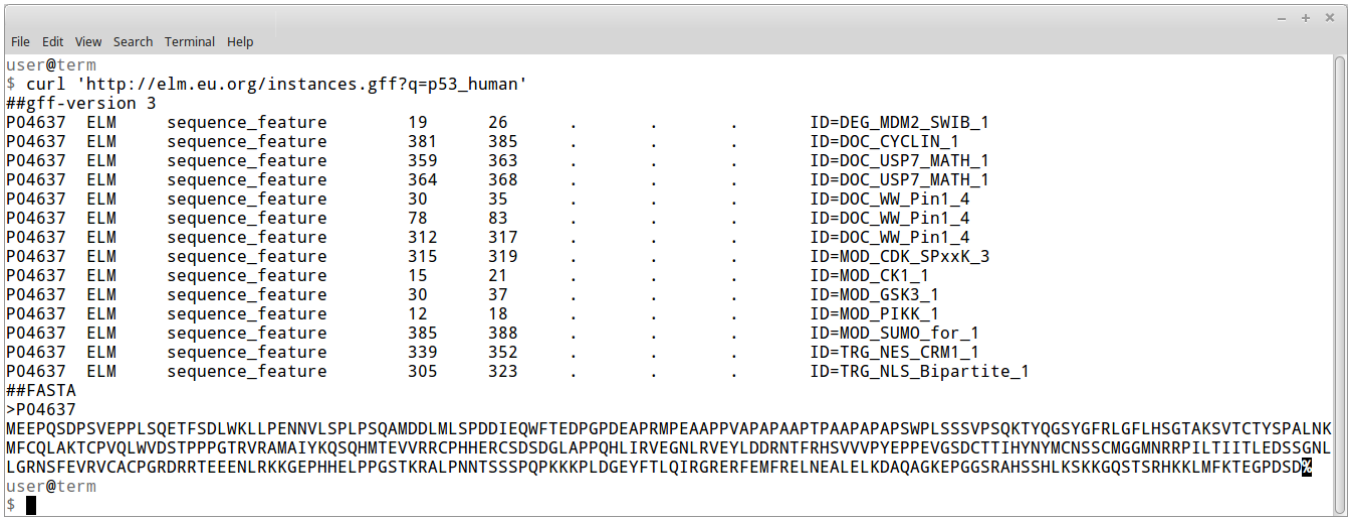


Figure 32: Screenshot of a terminal window using `curl` to download all ELM instances annotated for sequence P53_HUMAN.

6. To download a list of all motif instances detected in the protein sequence of human p53, type the following command into a terminal: `curl 'http://elm.eu.org/instances.gff?q=p53_human'`. The output should look similar to that shown in Figure 33. The output is in the "General Feature Format" (see www.ensembl.org/info/website/upload/gff.html#moreinfo), with the FASTA formatted sequence appended to the end of the output.

   > *Many other file formats are available for downloading instances annotations; see the downloads page for available options including the FASTA, GFF, PIR, or PSI-MI format (either XML or MiTab).*

7. To download a list of all instances matching the search query "CLV" annotated for the taxon "yellow fever mosquito (*Aedes aegypti*)", enter the following command into a terminal: `curl`

`'http://elm.eu.org/instances.tsv?q=CLV&taxon=aedes+aegypti'` (In general, any species name can be used, however remember to replace all "spaces" with "+"). This should return a single instance, the only one matching CLV in *A. aegypti*.

8. More data (interactions, domains, methods, etc.) can be downloaded from ELM in analogous fashion as shown in the preceding steps. Take a look at the ELM Downloads page (elm.eu.org/downloads, figure 30) for an overview of which datasets can be downloaded, and what different filters and formats are available for each dataset.

Protocol 6 (Alternate Protocol) Detecting Short Linear Motifs in Sequences using the REST API

Querying ELM for motifs in a given sequence (as described in Protocol 3 and Protocol 4), gives you a nice overview of putative and possibly annotated motifs in your query protein with a graphical representation using colors to highlight different regions of the protein sequence (eg. disordered vs. globular). It is however difficult to analyse a large set of protein sequences in this manner. Therefore, the ELM server provides an interface which you can use to submit your sequence in a programmatic way. Of course, this way, you won't receive the graphical output representation, but are limited to textual data representation.

Currently, there exists a single URL (`elm.eu.org/start_search/`) to accept such queries. You can choose to either submit a UniProt name or accession (eg. '`elm.eu.org/start_search/P53_HUMAN.tsv`') or submit your raw sequence (e.g. '`elm.eu.org/start_search/MAPRGFSCLLLLTSEIDLPVKRRA`'.) If the URL ends in '.tsv' then the server assumes you are using a UniProt id or accession; if it doesn't, then it assumes you are using raw sequence. See below for details.

Necessary Resources

Software

A modern browser such as Firefox, Chrome, or Safari. ELM is best viewed on a laptop or desktop computer, although tablets and smartphones will also work. There exist several REST client plugins for different browsers, however these are not needed for this protocol. Ideally use a commandline tool such as `curl` (`http://curl.haxx.se/`) in a terminal window. This program is available in any of the major operating systems: OSX, Windows and Linux. Of course, `curl` is only one of many different ways to access web content programatically, and we suggest to use whichever program you feel is better suited for your tasks. (Note that in some of the following screenshots, we append the following command to the commandline "`| column -t`" to make the output more readable. All this does is to properly align all columns at tabstops.)

Submitting a query to ELM via the REST API

1. Use `curl` to query ELM for all motifs predicted to occur in human p53 by typing the following into a terminal: '`curl 'http://elm.eu.org/start_search/P53_HUMAN.tsv`'. See Figure 33 for example output. Each resulting row represents a motif detection, and the first column "elm_identifier" indicates which ELM class was identified, multiple matches to the same class are represented in multiple lines. The columns "start" and "stop" show that first and last amino acid positions that matched the motif. The column "is_annotated" is True if this motif has been annotated in the ELM database as an (experimentally validated) motif instance. The column "is_phiblastmatch" is True if a match was found by the ELM Instance mapper indicating that an experimentally validated instance in a homologous sequence was found (see commentary section "Instance Mapper" at page 47). The column "is_filtered" shows whether or not this motif was rejected by any of the ELM prediction filters (structure, topodom, taxon), whereby "topodomfilter" uses information from UniProt to determine the protein's "topology" with respect to trans-membrane domains or extracellular re-

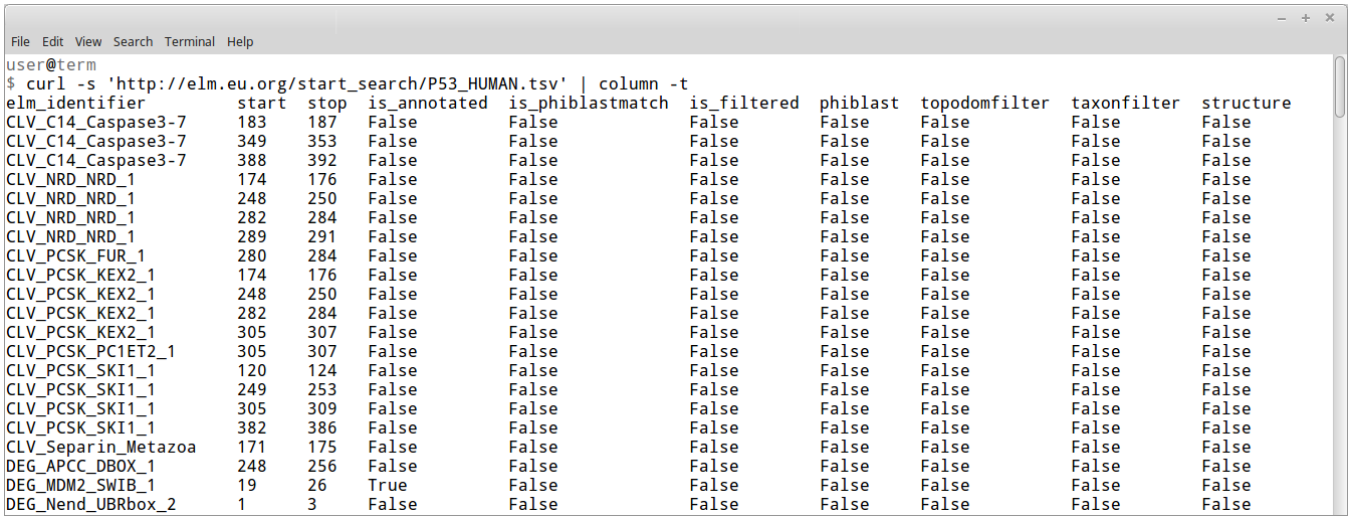


Figure 33: The commandline output when `curl` is used to download all motifs predicted in human p53.

gions. The columns “taxonfilter” and “structure” indicate that an instance has been filtered by the taxonomy or secondary structure filter, respectively (see commentary sections "Taxon Filter" and "Structure Filter").

In FigureF33 we use a slightly more advanced command to get the output to look nice in the terminal. We specified the `-s` *option to silence all* `curl` *output other than the downloaded file, and piped the output directly to the* `column` *command (this command exists on most Liux and OSX machines).*

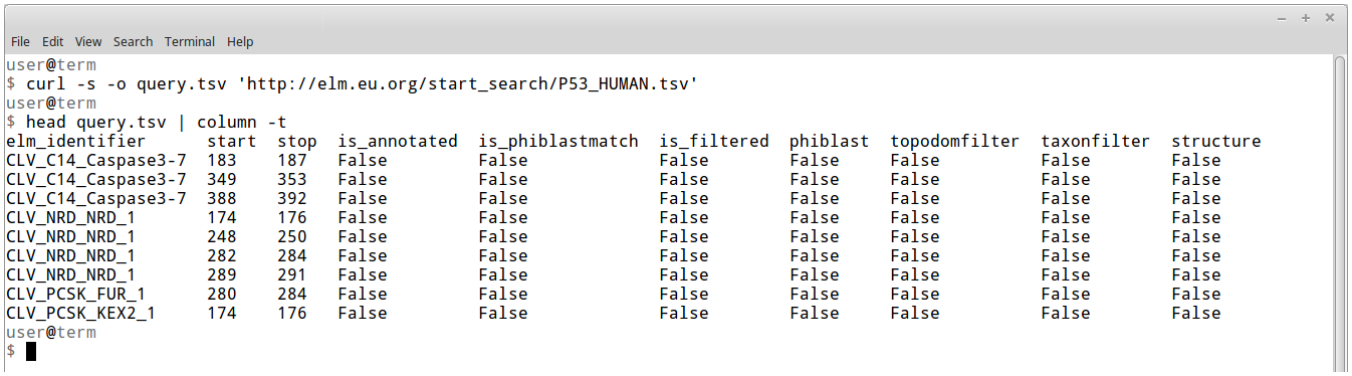


Figure 34: It is possible to send amino acid sequences to the ELM Prediction pipeline. In this case we have used the curl option “`-o`” to download directly to the file `query.tsv`, and use a combination of the `head` and `column` commands to display the first 10 rows to the terminal.

2. Use `curl` to query ELM via protein sequence by using the URL ’elm.eu.org/start_search/MAPRGFSCLLLLTSEIDLPVKRRA’ (Figure 34). In this case the query is an arbitrary short peptide sequence, but this can (of course) contain any sequence you are interested in analysing. The output format is exactly the same as in the previous step.

   This way of querying ELM is unfortunately not stable for long protein sequences. Different browsers and computers have different maximum lengths for URLs, and the excess text is often

simply ignored. We recommend not using this method for sequences longer than 2000 amino acids.

Guidelines for Understanding Results

Marc: TODO: Still working on this bit.

The annotations in the ELM database have all been selected and In most cases you can safely assume that the annotations present in the database are all of high confidence. These annotators have a lot of exprience in reading the scientific literature, and know how to distinguish high confidence from suggestive expriments Gibson et al. (2015). You are also encouraged to dig deeper into each annotation. In all cases each entry contains descriptions of the expriments performed and links to the original reasearch.

Understanding the results generated by the ELM prediction pipeline can sometimes require some extra work. In Protocol 3 and Protocol 4 we gave a few examples of how to read and interpret the results from the ELM prediction pipleline. These are bioinformatics predictions, and therefore will rely on a heuristic which will make mistakes. In general the prediction pipeline attempts to make as many predictions as possible, at the risk of making some False Positive predictions as well.

In cases all of the intermediarte results generated by the ELM prediction pipeline are made available to aid you in deciding which preditions are worth further investigation. Looking at the multiple sequence alignments used to generate the conservation score can (for example) help determine why a seemingly likely motif may have a (falsely) low confidence score. On the contrary you may have reason to belive that a predicted motif that was rejected by ELM's structure filter may actually exposed in a different structural conformation.

instructions: A brief discussion of the theory and applications of your

notes: Maybe mention how findings are relevant to the lab? For example: Manually annotated content should be reliable, although one should look at the 'confidence' in the instance annotation. Predictions are probably trustworthy, but you need to take into account the 'confidence score', and other features like whether its in a domain, etc...

Commentary:

instructions: A brief discussion of the theory and applications of your

Background Information

In order to interpret the data contained in ELM and the results produced by the ELM prediction tool, it is important to have a basic understanding of SLiM's and how they are affected by their structural and biological context. This background information summarises the different functionalities of SLiMs, describes the degenerate nature of motif sequences, and emphasises the need for contextual data for confident SLiM prediction.

ELM categorises SLiMs depending on their functionality

SLiMs mediate different types of interactions, and based on this functionality, the ELM classes annotated in the ELM database are grouped into six main ELM types (Figure 35, Dinkel et al. (2014)). They can function as ligand binding sites or as sites for post-translational modification (PTM). Some ligand SLiMs are recognised by components of the cellular transport machinery and function as localisation signals that target proteins to specific sub-cellular compartments (TRG type). Other ligand SLiMs are abundantly present in interfaces that mediate the assembly of large macromolecular complexes and in highly modular scaffold proteins that act as multivalent platforms for protein complex assembly (LIG type). Docking motifs are ligand SLiMs that recruit modification enzymes to their substrates by binding to a site on the enzyme that is distinct from the active site (DOC type). A subset of these, known as degrons, recruit ubiquitin ligases, which subsequently polyubiquitylate their substrates and hence target them for proteasomal degradation (DEG type). SLiMs that act as sites for PTM can be targeted by specific enzymes for the addition or removal of a small chemical group (e.g. phosphorylation), a sugar molecule (e.g. glycosylation), a protein (e.g. ubiquitylation), or another moiety (e.g. lipidation) (MOD type). Other PTM SLiMs mediate proteolytic cleavage by acting as target site for proteolytic enzymes (CLV type), or are recognised for structural modification by isomerases that catalyse cis-trans isomerisation of the peptide backbone (DOC type), see Van Roey et al. (2014) and Lee et al. (2015).

ELM regular expressions reflect the degenerate nature of SLiMs

As their name suggests, SLiMs are compact, being composed of a limited number of adjacent amino acids. Most of a motif's binding specificity however is conferred by only a subset of these amino acids. Those few residues that directly interact with the binding partner are evolutionary conserved, although in many cases a subset of amino acids that share certain properties (such as similar charge, size or hydrophobicity) are allowed in these hotspot positions. In the motif positions that contribute little to the interaction, there are even less constraints, i.e. a broader range of amino acids is allowed in these positions (Davey et al. 2012). This sequence flexibility is captured in the regular expressions that are defined for each motif class. A first consequence of this degeneracy is that SLiMs co-operatively engage in interactions of relatively low affinity. Hence these binding events are transient and reversible, and can be readily modulated, for instance by PTM. These characteristics make SLiM-based interactions ideal mediators of the dynamic processes involved in cell signalling (Van Roey et al. 2012). Another consequence is that it might take only a few or even a single point mutation to generate or disrupt a functional motif in a protein. The associated ability to evolve convergently might underlie the proliferation of SLiMs and the rewiring of interactomes (Davey et al. 2015; Kim et al. 2012). Conversely, several SLiM-associated diseases have been characterised to date, for instance Liddle syndrome (Furuhashi et al. 2005).

ELM integrates data to increase the confidence of SLiM prediction

Due to their degenerate nature, motif sequences contain only very little information, and many short sequences in a proteome will match motif patterns. However, most of these matches will not represent functional motifs, and hence, when scanning a proteome for putative motifs using only the motif sequence patterns will yield a large number of false positive instances, far exceeding the number of true motifs. Therefore, reliable motif detection cannot go without experimental validation of candidate motifs, using different types of experiments and techniques (Gibson et al. 2015). This however does not mean that

bioinformatics analysis cannot guide researchers towards a subset of candidate motifs that have a higher probability to be functional and help rule out those candidate motifs that are likely to be false positives. Taking into account additional information, besides a match to a sequence pattern defining a SLiM, can greatly narrow the selection of putative motifs for experimental validation. Additional data for in silico analysis include conservation of the motif sequence, the location of the motif within the protein's structure and its accessibility for its binding partner, validated interaction with the binding partner, and in-cell co-localisation with the binding partner. The availability and usefulness of these additional data for SLiM discovery depends on their extensive and correct biocuration. A vast and increasing amount of biological data is available in a wide variety of sources, including the literature and large-scale datasets. In order to facilitate integration of data, they need to be collected, annotated and formatted in central data and knowledge repositories. The ELM database provides such a repository for experimentally validated linear motif classes and instances. The ELM prediction tool in turn relies on annotated data, both from the ELM database and other resources, to accurately analyse unknown sequences for candidate motifs and assist researchers in selecting the most plausible ones for experimental validation and discard likely false positive hits, saving them valuable time and assets (Dinkel et al. 2012).

ELM Filters

Disorder Filter ELM uses two different predictors of globularity/disorder: Firstly, GlobPlot developed by Linding et al. (2003), uses amino acid propensities derived from a set of proteins to detect regions of globularity/disorder in any given protein sequence. Secondly, IUPred by Dosztányi et al. (2005), which, unlike GlobPlot, has not been trained on any dataset, but rather uses a position-specific scoring scheme assessing the tendency of any given amino acid to reside in either an ordered or disordered region. IUPred assigns a score between 0.0 to 1.0 to each amino acid of a protein sequence, whereby protein segments with an IUPred score above 0.5 are considered to be disordered. ELM displays IUPred scores as a colored line on either green (disordered) or red (globular) background, see rows four and five of Figure 23.

Structure Filter The structural filter, especially developed for ELM by Via et al. (2009), assesses accessibility and secondary structural context derived from experimentally solved protein structures. It maps putatively functional motif occurrences onto a representative domain structure and scores these motifs for solvent-accessibility and secondary structure context. ELM displays this information as overlay boxes in the graphical output, whereby the user needs to hover over individual instance entries within structural context (see the CLV_PCSK_SKI1_1 example in fig. 23).

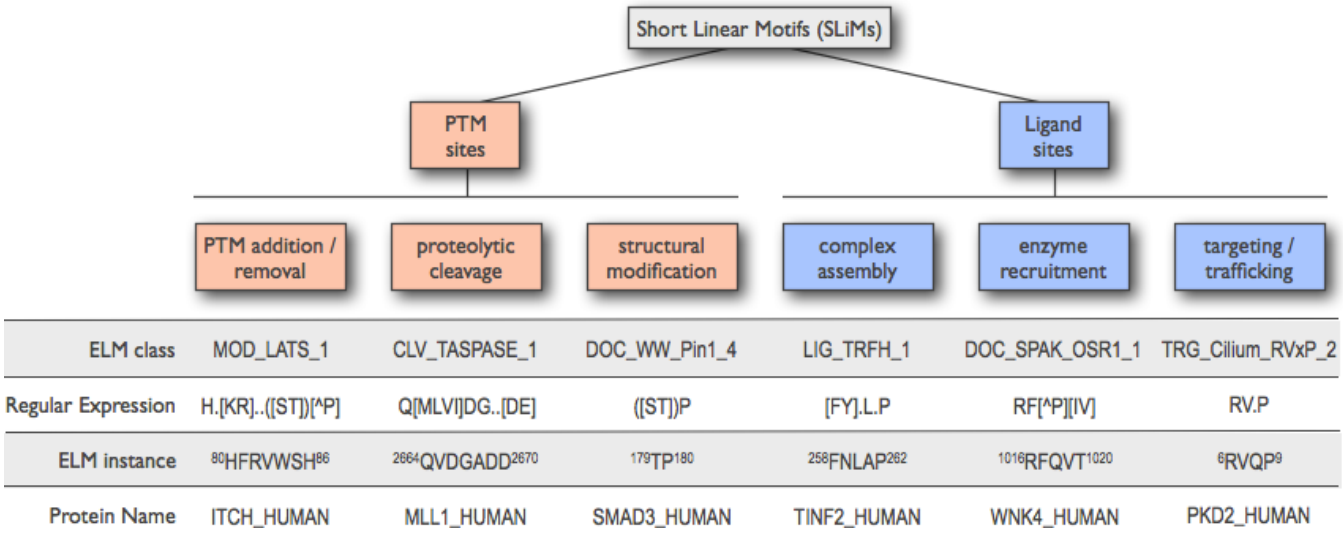
Conservation Filter This filter method for scoring the conservation of linear motif instances was developed by Chica et al. (2008) and subsequently implemented into the ELM pipeline. It requires only primary sequence-derived information (e.g. a multiple alignment and the sequence tree) and implicitly takes into account the degenerate nature of short linear motif patterns. By auto-generating multiple sequence alignments from a non-redundant database, generating distance-trees and taking into account motif degeneracy, it assesses for each ELM motif class found in any given sequence its conservation. The conservation score ranges from 0.0 (the predicted instance is present only in the query sequence) to 1.0 (full conservation of the motif regular expression in all the informative sequences). In the graphical representation, motif conservation is indicated by coloring instances in different shades of blue, whereby darker shades of blue represent higher conservation. In the case that not enough homologous sequences were detected to build an alignment, no conservation score can be calculated. Therefore all of the motif occurrences will be

shown in a uniform shade of blue.

Since conserved motifs in structural regions are most likely conserved for structural integrity rather than motif function, one always has to assess the context when inspecting conservation score. Generally, best motif candidates are those with high conservation scores in regions of unstructured, unconserved regions.

Taxon Filter Each ELM class is annotated with one or more taxonomic ranges, for which experimental evidence has been found for the particular class (see "Present in Taxon: Eukaryota" in figure 4). This information is then used to filter taxons outside the annotated range whenever a user submits a query sequence to the ELM database (see the **taxonomic context** field in the ELM search input form in fig. 22).

Instance Mapper The ELM instance mapper takes all annotated instances from the ELM database, generates a BLAST database from it and uses `phiblast` to detect sequence stretches in the query sequence which are similar to sequences in this database. This allows the instance mapper to effectively map known instances (for which experimental evidence exists) onto homologous sequences of unknown function.



ELM class	MOD_LATS_1	CLV_TASPASE_1	DOC_WW_Pin1_4	LIG_TRFH_1	DOC_SPAK_OSR1_1	TRG_Cilium_RVxP_2
Regular Expression	H.[KR]..([ST])[^P]	Q[MLVI]DG..[DE]	([ST])P	[FY].L.P	RF[^P][IV]	RV.P
ELM instance	^{80}HFRVWSH^{86}	^{2664}QVDGADD^{2670}	^{179}TP^{180}	^{258}FNLAP^{262}	^{1016}RFQVT^{1020}	^{6}RVQP^{9}
Protein Name	ITCH_HUMAN	MLL1_HUMAN	SMAD3_HUMAN	TINF2_HUMAN	WNK4_HUMAN	PKD2_HUMAN

Figure 35: For each ELM class, the functional category to which it belongs is indicated by a three-letter prefix. Each ELM class is defined by a regular expression. Peptide sequences in proteins that match the regular expression of a specific ELM class and that were experimentally validated to be functional motifs are captured as ELM instances of that class. Degrons are a specific subtype of enzyme-recruiting docking motifs (see text for a detailed description).

Critical Parameters and Troubleshooting

Factors that influence the protocol and to which special attention should be paid. Common problems with the protocols, their causes, and potential solutions. The information may be presented in tabular form or it may be combined with Critical Parameters.

Internet Resources with Annotations

http://www.clustal.org/omega Clustal Omega (Sievers et al. 2011) is a tool for the alignment of multiple nucleic acid and protein sequences.

http://www.jalview.org Jalview (Waterhouse et al. 2009) is a Java desktop application (and browser applet) that employs web services for sequence alignment and visualization.

http://proviz.ucd.ie ProViz (Jehl et al. 2016) is an interactive protein exploration tool, which searches several databases for information about a given query protein. Data relevant to the protein like an alignment of homologues, linear motifs, post translational modifications, domains, secondary structure, sequence variations and others are graphically represented relative to their position in the protein.

References

Berman, H. M., T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook, and C. Zardecki. 2002. The protein data bank. Acta crystallographica. Section D, Biological crystallography 58:899–907.

Chica, C., A. Labarga, C. M. Gould, R. López, and T. J. Gibson. 2008. A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. BMC bioinformatics 9:229.

Davey, N. E., M. S. Cyert, and A. M. Moses. 2015. Short linear motifs âĂŞ ex nihilo evolution of protein regulation. Cell Communication and Signaling 13:43.

Davey, N. E., K. Van Roey, R. J. Weatheritt, G. Toedt, B. Uyar, B. Altenberg, A. Budd, F. Diella, H. Dinkel, and T. J. Gibson. 2012. Attributes of short linear motifs. Molecular bioSystems 8:268–81.

de Brito, C. F. A., C. B. Carvalho, F. Santos, R. T. Gazzinelli, S. C. Oliveira, V. Azevedo, and S. M. R. Teixeira. 2004. Chromobacterium violaceum genome: molecular mechanisms associated with pathogenicity. Genetics and molecular research : GMR 3:148–61.

Diella, F. 2008. Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Frontiers in Bioscience Volume:6580.

Dinkel, H., C. Chica, A. Via, C. M. Gould, L. J. Jensen, T. J. Gibson, and F. Diella. 2011. Phospho.elm: a database of phosphorylation sites–update 2011. Nucleic acids research 39:D261–7.

Dinkel, H., S. Michael, R. J. Weatheritt, N. E. Davey, K. Van Roey, B. Altenberg, G. Toedt, B. Uyar, M. Seiler, A. Budd, L. Jödicke, M. A. Dammert, C. Schroeter, M. Hammer, T. Schmidt, P. Jehl, C. McGuigan, M. Dymecka, C. Chica, K. Luck, A. Via, A. Chatr-Aryamontri, N. Haslam, G. Grebnev, R. J. Edwards, M. O. Steinmetz, H. Meiselbach, F. Diella, and T. J. Gibson. 2012. Elm–the database of eukaryotic linear motifs. Nucleic acids research 40:D242–51.

Dinkel, H., K. Van Roey, S. Michael, N. E. Davey, R. J. Weatheritt, D. Born, T. Speck, D. Krüger, G. Grebnev, M. Kubań, M. Strumillo, B. Uyar, A. Budd, B. Altenberg, M. Seiler, L. B. Chemes, J. Glavina, I. E.

Sánchez, F. Diella, and T. J. Gibson. 2014. The eukaryotic linear motif resource ELM: 10 years and counting. Nucleic Acids Research 42:D259–D266.

Dodd, D. A., R. G. Worth, M. K. Rosen, S. Grinstein, N. S. C. van Oers, and E. J. Hansen. 2014. The haemophilus ducreyi LspA1 protein inhibits phagocytosis by using a new mechanism involving activation of c-terminal src kinase. mBio 5:e01178–14–e01178–14.

Dosztányi, Z., V. Csizmok, P. Tompa, and I. Simon. 2005. Iupred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics (Oxford, England) 21:3433–4.

Fielding, R. T. and R. N. Taylor. 2002. Principled design of the modern web architecture. ACM Transactions on Internet Technology 2:115–150.

Finn, R. D., T. K. Attwood, P. C. Babbitt, A. Bateman, P. Bork, A. J. Bridge, H.-Y. Chang, Z. Dosztányi, S. El-Gebali, M. Fraser, J. Gough, D. Haft, G. L. Holliday, H. Huang, X. Huang, I. Letunic, R. Lopez, S. Lu, A. Marchler-Bauer, H. Mi, J. Mistry, D. A. Natale, M. Necci, G. Nuka, C. A. Orengo, Y. Park, S. Pesseat, D. Piovesan, S. C. Potter, N. D. Rawlings, N. Redaschi, L. Richardson, C. Rivoire, A. Sangrador-Vegas, C. Sigrist, I. Sillitoe, B. Smithers, S. Squizzato, G. Sutton, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, I. Xenarios, L.-S. Yeh, S.-Y. Young, and A. L. Mitchell. 2017. Interpro in 2017-beyond protein family and domain annotations. Nucleic acids research 45:D190–D199.

Finn, R. D., P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman. 2016. The pfam protein families database: towards a more sustainable future. Nucleic acids research 44:D279–85.

Furuhashi, M., K. Kitamura, M. Adachi, T. Miyoshi, N. Wakida, N. Ura, Y. Shikano, Y. Shinshi, K.-i. Sakamoto, M. Hayashi, N. Satoh, T. Nishitani, K. Tomita, and K. Shimamoto. 2005. Liddle's Syndrome Caused by a Novel Mutation in the Proline-Rich PY Motif of the Epithelial Sodium Channel β-Subunit. The Journal of Clinical Endocrinology & Metabolism 90:340–344.

Gene Ontology Consortium. 2017. Expansion of the gene ontology knowledgebase and resources. Nucleic acids research 45:D331–D338.

Gibson, T. J., H. Dinkel, K. Van Roey, and F. Diella. 2015. Experimental detection of short regulatory motifs in eukaryotic proteins: tips for good practice as well as for bad. Cell Communication and Signaling 13:42.

Jehl, P., J. Manguy, D. C. Shields, D. G. Higgins, and N. E. Davey. 2016. Proviz-a web-based visualization tool to investigate the functional and evolutionary features of protein sequences. Nucleic acids research 44:W11–5.

Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. 2016. Kegg as a reference resource for gene and protein annotation. Nucleic acids research 44:D457–62.

Kaniga, K., J. Uralil, J. B. Bliska, and J. E. Galán. 1996. A secreted protein tyrosine phosphatase with modular effector domains in the bacterial pathogen salmonella typhimurlum. Molecular Microbiology 21:633–641.

Kerrien, S., S. Orchard, L. Montecchi-Palazzi, B. Aranda, A. F. Quinn, N. Vinod, G. D. Bader, I. Xenarios, J. Wojcik, D. Sherman, M. Tyers, J. J. Salama, S. Moore, A. Ceol, A. Chatr-Aryamontri, M. Oesterheld, V. Stümpflen, L. Salwinski, J. Nerothin, E. Cerami, M. E. Cusick, M. Vidal, M. Gilson, J. Armstrong, P. Woollard, C. Hogue, D. Eisenberg, G. Cesareni, R. Apweiler, and H. Hermjakob. 2007. Broadening the horizon–level 2.5 of the hupo-psi format for molecular interactions. BMC biology 5:44.

Kim, J., I. Kim, J.-S. Yang, Y.-E. Shin, J. Hwang, S. Park, Y. S. Choi, and S. Kim. 2012. Rewiring of PDZ Domain-Ligand Interaction Network Contributed to Eukaryotic Evolution. PLoS Genetics 8:e1002510.

Lee, R. V. D., M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, P. M. Kim, R. W. Kriwacki, C. J. Old, R. V. Pappu, P. Tompa, V. N. Uversky, P. E. Wright, and M. M. Babu. 2015. Classi fi cation of Intrinsically Disordered Regions and Proteins. Prog Biophys Mol Biol .

Letunic, I., T. Doerks, and P. Bork. 2015. Smart: recent updates, new developments and status in 2015. Nucleic acids research 43:D257–60.

Linding, R., R. B. Russell, V. Neduva, and T. J. Gibson. 2003. Globplot: Exploring protein sequences for globularity and disorder. Nucleic acids research 31:3701–8.

McKusick, V. A. 2007. Mendelian inheritance in man and its online version, omim. American journal of human genetics 80:588–604.

NCBI Resource Coordinators. 2017. Database resources of the national center for biotechnology information. Nucleic acids research 45:D12–D17.

Schultz, J., F. Milpetz, P. Bork, and C. P. Ponting. 1998. Smart, a simple modular architecture research tool: identification of signaling domains. Proceedings of the National Academy of Sciences of the United States of America 95:5857–64.

Selbach, M., F. E. Paul, S. Brandt, P. Guye, O. Daumke, S. Backert, C. Dehio, and M. Mann. 2009. Host cell interactome of tyrosine-phosphorylated bacterial proteins. Cell Host & Microbe 5:397–403.

Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Molecular systems biology 7:539.

Suzek, B. E., H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu. 2007. Uniref: comprehensive and non-redundant uniprot reference clusters. Bioinformatics (Oxford, England) 23:1282–8.

Tompa, P., N. E. Davey, T. J. Gibson, and M. M. Babu. 2014. A million peptide motifs for the molecular biologist. Molecular cell 55:161–9.

Tsutsumi, R. 2002. Attenuation of helicobacter pylori CagAmiddle dotSHP-2 signaling by interaction between CagA and c-terminal src kinase. Journal of Biological Chemistry 278:3664–3670.

UniProt Consortium. 2015. Uniprot: a hub for protein information. Nucleic acids research 43:D204–12.

Van Roey, K., H. Dinkel, R. J. Weatheritt, T. J. Gibson, and N. E. Davey. 2013. The switches.elm resource: a compendium of conditional regulatory interaction interfaces. Science signaling 6:rs7.

Van Roey, K., T. J. Gibson, and N. E. Davey. 2012. Motif switches: decision-making in cell regulation. Current opinion in structural biology 22:378–85.

Van Roey, K., B. Uyar, R. J. Weatheritt, H. Dinkel, M. Seiler, A. Budd, T. J. Gibson, and N. E. Davey. 2014. Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. Chemical reviews 114:6733–78.

Via, A., C. M. Gould, C. Gemünd, T. J. Gibson, and M. Helmer-Citterich. 2009. A structure filter for the eukaryotic linear motif resource. BMC bioinformatics 10:351.

Via, A., B. Uyar, C. Brun, and A. Zanzoni. 2015. How pathogens use linear motifs to perturb host cell networks. Trends in Biochemical Sciences 40:36–48.

Waterhouse, A. M., J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton. 2009. Jalview version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics (Oxford, England) 25:1189–91.

Wright, P. E. and H. J. Dyson. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. Journal of molecular biology 293:321–31.