
InterPro

InterPro Team

Apr 03, 2024

ABOUT INTERPRO

1	About InterPro	1
2	Citing InterPro	3
2.1	Latest publications	3
2.2	All previous publications	3
3	InterPro tutorials & Webinars	7
3.1	Tutorials	7
3.2	Webinars	7
4	Upcoming courses and webinars	9
5	Previous courses	11
5.1	Structural bioinformatics course (in person)	11
5.2	Bioinformatics resources for protein biology (virtual)	11
5.3	Structural bioinformatics course (virtual)	11
5.4	Introduction to InterPro workshop (virtual)	11
5.5	Bioinformatics resources for protein biology (virtual)	12
5.6	Structural bioinformatics course (virtual)	12
5.7	Structural bioinformatics course (virtual)	12
5.8	Bioinformatics Resources for Protein Biology	12
6	InterPro Entries : essential information	13
6.1	InterPro entry types	13
6.2	Other entry and page types	13
6.3	Entry relationships	14
6.4	Overlapping entries	14
6.5	Ontologies	14
7	InterPro website banner	15
7.1	Navigation banner and menu	15
8	InterPro homepage	17
8.1	InterPro resource overview	18
8.2	Search box	18
8.3	Data	19
8.4	News and information	22
9	How to search the InterPro website?	23
9.1	Quick search	23
9.2	Sequence search	23

9.3	Text search	27
9.4	Domain architecture search	28
9.5	Using Browse feature to search and filter InterPro	28
10	Protein sequence viewer	37
11	Browsing entries in the InterPro website	43
11.1	InterPro entry page	44
11.2	Member database page	49
11.3	Protein entry page	55
11.4	Structure entry page	57
11.5	Taxonomy entry page	57
11.6	Proteome entry page	60
11.7	Set/Clan entry page	61
12	How to download InterPro data?	63
12.1	Download page	63
12.2	Export button	63
12.3	Your downloads	64
12.4	InterPro Application Programming Interface (API)	65
13	Release notes	67
13.1	General information	67
13.2	Other statistics	67
14	Settings page	71
14.1	Navigation settings	71
14.2	Notification settings	72
14.3	User interface settings	73
14.4	Cache settings	74
14.5	Server settings	74
14.6	Developer Information	75
15	Frequently Asked Questions (FAQs)	77
15.1	General Questions	77
15.2	Sequence searches (InterProScan)	79
15.3	Web Interface	80
15.4	Application Programming Interface (API)	81
15.5	Troubleshooting	82
15.6	Additional help	82
16	InterProScan	83
16.1	Documentation	83
16.2	Web services	83
16.3	Web based tools	83
16.4	Source code	84
16.5	Previous releases	84
16.6	License	84
16.7	Follow us & reporting bugs	84
17	InterPro consortium member databases	85
17.1	CATH-Gene3D	85
17.2	CDD	85
17.3	HAMAP	86
17.4	MobiDB Lite	86

17.5	NCBIfam	86
17.6	PANTHER	86
17.7	Pfam	87
17.8	PIRSF	87
17.9	PRINTS	87
17.10	PROSITE profiles	87
17.11	SFLD	87
17.12	SMART	88
17.13	SUPERFAMILY	88
18	About Pfam	89
19	About PRINTS	91
20	About SFLD	93
21	About AntiFam	95
21.1	How to use AntiFam	95
21.2	Superkingdom-specific sets	96
21.3	Acknowledgements	96
21.4	How to cite AntiFam	96
22	InterPro team	97
22.1	Team members	97
22.2	Previous contributors	97
23	Funding	99
24	Privacy	101
25	License	103
26	Literature references	105
27	Protein families card game	107
27.1	Protein families game	107
27.2	Understanding the biology	107
27.3	Ask questions or give feedback	113

ABOUT INTERPRO

InterPro is a resource that provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites. To classify proteins in this way, InterPro uses predictive models, known as signatures, provided by several collaborating databases (referred to as member databases) that collectively make up the InterPro consortium. A key value of InterPro is that it combines protein signatures from these member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool. We add further value to InterPro entries by providing detailed functional annotation as well as adding relevant GO terms that enable automatic annotation of millions of GO terms across the protein sequence databases.

InterPro integrates signatures from the following 13 member databases:

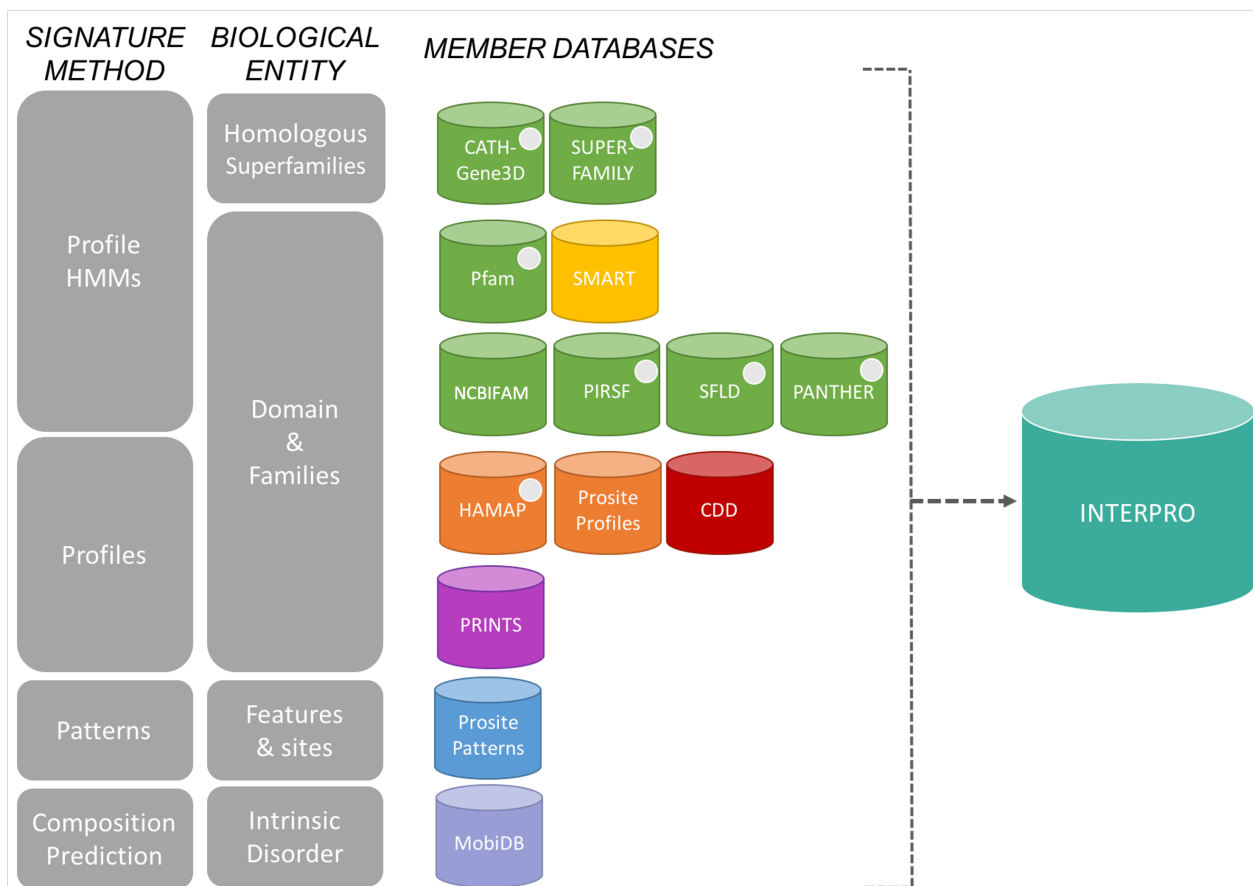
CATH, CDD, HAMAP, MobiDB Lite, Panther, Pfam, PIRSF, PRINTS, Prosite, SFLD, SMART, SUPERFAMILY AND NCBIfam (the [InterPro consortium](#) section gives further information about the individual databases).

The member databases use a variety of different methods to classify proteins. Each of the databases has a particular focus (e.g. protein domains defined from structure, or full length protein families with shared function). We strive to integrate the signatures from the member databases into InterPro entries and to identify where different member database entries are the same entity.

You can use the [InterPro website](#) to obtain information about individual protein families, domains, important sites, perform a sequence search or browse through InterPro annotations. We have designed the website to be intuitive for new users meaning it is not essential to read this documentation. However, in the following sections you will find a wealth of specialised and powerful features that can be easily overlooked. You may also want to check out our list of [training materials and webinars](#).

InterPro is updated approximately every 8 weeks. The [release notes page](#) contains information about what has changed in each release.

All information in InterPro is freely available. You can download InterPro data for local analyses from the [Download](#) page, or use the [InterPro API](#). Find out more about the project by exploring the [latest papers](#).



CITING INTERPRO

2.1 Latest publications

If you find InterPro useful for your research, please cite the following publications:

2.1.1 InterPro

[InterPro in 2022](#) Typhaine Paysan-Lafosse, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, Julian Gough, Daniel H Haft, Ivica Letunić, Aron Marchler-Bauer, Huaiyu Mi, Darren A Natale, Christine A Orengo, Arun P Pandurangan, Catherine Rivoire, Christian J A Sigrist, Ian Sillitoe, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, Alex Bateman, *Nucleic Acids Research* (2022), gkac993, PMID: [36350672](#)

2.1.2 InterProScan

[InterProScan 5: genome-scale protein function classification](#) Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, Sarah Hunter Bioinformatics (2014), PMID: [24451626](#)

2.2 All previous publications

[The InterPro protein families and domains database: 20 years on](#) Matthias Blum, Hsin-Yu Chang, Sara Chuguransky, Tiago Grego, Swaathi Kandasamy, Alex Mitchell, Gift Nuka, Typhaine Paysan-Lafosse, Matloob Qureshi, Shriya Raj, Lorna Richardson, Gustavo A Salazar, Lowri Williams, Peer Bork, Alan Bridge, Julian Gough, Daniel H Haft, Ivica Letunic, Aron Marchler-Bauer, Huaiyu Mi, Darren A Natale, Marco Necci, Christine A Orengo, Arun P Pandurangan, Catherine Rivoire, Christian J A Sigrist, Ian Sillitoe, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, Alex Bateman, Robert D Finn *Nucleic Acids Research* (2020), gkaa977, PMID: [33156333](#)

[InterPro in 2019: improving coverage, classification and access to protein sequence annotations](#) Alex L Mitchell, Teresa K Attwood, Patricia C Babbitt, Matthias Blum, Peer Bork, Alan Bridge, Shoshana D Brown, Hsin-Yu Chang, Sara El-Gebali, Matthew I Fraser, Julian Gough, David R Haft, Hongzhan Huang, Ivica Letunic, Rodrigo Lopez, Aurélien Luciani, Fabio Madeira, Aron Marchler-Bauer, Huaiyu Mi, Darren A Natale, Marco Necci, Gift Nuka, Christine Orengo, Arun P Pandurangan, Typhaine Paysan-Lafosse, Sebastien Pesseat, Simon C Potter, Matloob A Qureshi, Neil D Rawlings, Nicole Redaschi, Lorna J Richardson, Catherine Rivoire, Gustavo A Salazar, Amaia Sangrador-Vegas, Christian J A Sigrist, Ian Sillitoe, Granger G Sutton, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Siew-Yit Yong, Robert D Finn *Nucleic Acids Research* (2019) Database Issue 47:D351–D360, PMID: [30398656](#)

[InterPro in 2017 — beyond protein family and domain annotations](#) Robert D. Finn, Teresa K. Attwood, Patricia C. Babbitt, Alex Bateman, Peer Bork, Alan J. Bridge, Hsin-Yu Chang, Zsuzsanna Dosztányi, Sara El-Gebali, Matthew Fraser, Julian Gough, David Haft, Gemma L. Holliday, Hongzhan Huang, Xiaosong Huang, Ivica Letunic, Rodrigo Lopez, Shennan Lu, Aron Marchler-Bauer, Huaiyu Mi, Jaina Mistry, Darren A. Natale, Marco Necci, Gift Nuka, Christine A. Orengo, Youngmi Park, Sebastien Pesseat, Damiano Piovesan, Simon C. Potter, Neil D. Rawlings, Nicole Redaschi, Lorna Richardson, Catherine Rivoire, Amaia Sangrador-Vegas, Christian Sigrist, Ian Sillitoe, Ben Smithers, Silvano Squizzato, Granger Sutton, Narmada Thanki, Paul D Thomas, Silvio C. E. Tosatto, Cathy H. Wu, Ioannis Xenarios, Lai-Su Yeh, Siew-Yit Young, Alex L. Mitchell *Nucleic Acids Research* (2017), Database Issue 45:D190–D199, PMID: 27899635

[GO annotation in InterPro: why stability does not indicate accuracy in a sea of changing annotation](#) Sangrador-Vegas A, Mitchell AL, Chang HY, Yong SY, Finn RD *Database: the Journal of Biological Databases and Curation* (2016), 1–8, PMID: 26994912

[The InterPro protein families database: the classification resource after 15 years](#) Alex Mitchell, Hsin-Yu Chang, Louise Daugherty, Matthew Fraser, Sarah Hunter, Rodrigo Lopez, Craig McAnulla, Conor McMenamin, Gift Nuka, Sebastien Pesseat, Amaia Sangrador-Vegas, Maxim Scheremetjew, Claudia Rato, Siew-Yit Yong, Alex Bateman, Marco Punta, Teresa K. Attwood, Christian J.A. Sigrist, Nicole Redaschi, Catherine Rivoire, Ioannis Xenarios, Daniel Kahn, Dominique Guyot, Peer Bork, Ivica Letunic, Julian Gough, Matt Oates, Daniel Haft, Hongzhan Huang, Darren A. Natale, Cathy H. Wu, Christine Orengo, Ian Sillitoe, Huaiyu Mi, Paul D. Thomas, Robert D. Finn *Nucleic Acids Research* (2015), Database issue 43:D213–21, PMID: 25428371

[InterPro in 2011: new developments in the family and domain prediction database](#) Sarah Hunter; Philip Jones; Alex Mitchell; Rolf Apweiler; Teresa K. Attwood; Alex Bateman; Thomas Bernard; David Binns; Peer Bork; Sarah Burge; Edouard de Castro; Penny Coggill; Matthew Corbett; Ujjwal Das; Louise Daugherty; Lauranne Duquenne; Robert D. Finn; Matthew Fraser; Julian Gough; Daniel Haft; Nicolas Hulo; Daniel Kahn; Elizabeth Kelly; Ivica Letunic; David Lonsdale; Rodrigo Lopez; Martin Madera; John Maslen; Craig McAnulla; Jennifer McDowall; Conor McMenamin; Huaiyu Mi; Prudence Mutowo-Muellenet; Nicola Mulder; Darren Natale; Christine Orengo; Sebastien Pesseat; Marco Punta; Antony F. Quinn; Catherine Rivoire; Amaia Sangrador-Vegas; Jeremy D. Selengut; Christian J. A. Sigrist; Maxim Scheremetjew; John Tate; Manjulapramila Thimmajananathan; Paul D. Thomas; Cathy H. Wu; Corin Yeats; Siew-Yit Yong *Nucleic Acids Research* (2012), Database issue 40:D306–D312, PMID: 22096229

[Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation](#) Burge, S., Kelly, E., Lonsdale, D., Mutowo-Muellenet, P., McAnulla, C., Mitchell, A., Sangrador-Vegas, A., Yong, S., Mulder, N., Hunter, S. *Database: the Journal of Biological Databases and Curation* (2012), PMID: 22301074

[The InterPro BioMart: federated query and web service access to the InterPro Resource](#) Jones P., Binns D., McMenamin C., McAnulla C., Hunter S. *Database: the Journal of Biological Databases and Curation* (2011), PMID: 21785143

[InterPro protein classification](#) McDowall J, Hunter S. *Methods Mol Biol.* (2011) Database issue 694:37–47, PMID: 21082426

[InterPro: the integrative protein signature database](#) Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C. *Nucleic Acids Res.* (2009), Database issue 37:D211–5, PMID: 18940856

[The InterPro database and tools for protein domain analysis](#) Mulder NJ, Apweiler R. *Curr Protoc Bioinformatics* (2008), Chapter 2:Unit 2.7, PMID: 18428686

[InterPro and InterProScan: tools for protein sequence classification and comparison](#) Mulder N, Apweiler R. *Methods Mol Biol* (2007), Database issue 396:59–70, PMID: 18025686

[InterProScan: protein domains identifier](#) Quevillon E., Silventoinen V., Pillai S., Harte N., Mulder N., Apweiler R., Lopez R. *Nucleic Acids Research* (2005), Vol. 33, Issue suppl 2, PMID: 15980438

[New developments in the InterPro database](#) Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJ, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C. *Nucleic Acids Research* (2005), Database issue 35:D224-8, PMID: [17202162](#)

[InterPro, progress and status in 2005](#) Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Ponting CP, Quevillon E, Selengut J, Sigrist CJ, Silventoinen V, Studholme DJ, Vaughan R, Wu CH. *Nucleic Acids Res, Database issue* 33:D201-5, PMID: [15608177](#)

[The InterPro Database, 2003 brings increased coverage and new features](#) Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJ, Vaughan R, Zdobnov EM. *Nucleic Acids Res* (2003), 1;31(1):315-8, PMID: [12520011](#)

[HMM-based databases in InterPro](#) Bateman A, Haft DH. *Brief Bioinform* (2002), 3(3):236-45, PMID: [12230032](#)

[InterPro: an integrated documentation resource for protein families, domains and functional sites](#) Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley R, Courcelle E, Durbin R, Falquet L, Fleischmann W, Gouzy J, Griffiths-Jones S, Haft D, Hermjakob H, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Orchard S, Pagni M, Peyruc D, Ponting CP, Servant F, Sigrist CJ; InterPro Consortium. *Brief Bioinform* (2002), 3(3):225-35, PMID: [12230031](#)

[Interactive InterPro-based comparisons of proteins in whole genomes](#) Kanapin A, Apweiler R, Biswas M, Fleischmann W, Karavidopoulou Y, Kersey P, Kriventseva EV, Mittard V, Mulder N, Oinn T, Phan I, Servant F, Zdobnov E. *Bioinformatics* (2002), 18(2):374-5, PMID: [11847096](#)

[InterProScan — an integration platform for the signature-recognition methods in InterPro](#) Zdobnov EM, Apweiler R. *Bioinformatics* (2001), 17(9):847-8, PMID: [11590104](#)

[InterPro — an integrated documentation resource for protein families, domains and functional sites](#) Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJ, Zdobnov EM; InterPro Consortium. *Bioinformatics* (2000), 16(12):1145-50, PMID: [11159333](#)

[The InterPro database, an integrated documentation resource for protein families, domains and functional sites](#) Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJ, Zdobnov EM. *Nucleic Acids Res* (2001), 1;29(1):37-40, PMID: [11125043](#)

INTERPRO TUTORIALS & WEBINARS

3.1 Tutorials

Tutorials related to InterPro are available online:

- [InterPro quick tour](#)
- [Functional and structural analysis of protein sequences](#)
- [Protein classification: An introduction to EMBL-EBI resources](#)
- [A case study of 3 protein family building methodologies](#)
- [Genome3D annotations in InterPro](#)

3.2 Webinars

Recorded webinars related to InterPro are available online:

3.2.1 2023

[Exploring protein families and domains using InterPro](#)

3.2.2 2022

[Finding Pfam's protein families data in the InterPro website](#)

3.2.3 2021

[A guide to proteomics data analysis using UniProt and InterPro](#)

3.2.4 2020

- Exploring protein families and domains using InterPro
- Finding Pfam's protein families data in the InterPro website
- A guide to proteomics data analysis using UniProt and InterPro

3.2.5 2019

Genome3D annotations in InterPro webinar

UPCOMING COURSES AND WEBINARS

If you would like us to train employees/students from your company/institution, [contact us](#).

PREVIOUS COURSES

5.1 Structural bioinformatics course (in person)

Date: 2 - 6 October 2023

Venue: Virtual - EMBL-EBI

[Programme details](#)

5.2 Bioinformatics resources for protein biology (virtual)

Date: 25 - 27 April 2023

Venue: Virtual - EMBL-EBI

[Programme details](#)

5.3 Structural bioinformatics course (virtual)

Date: Monday 17 - Friday 21 October 2022

Venue: Virtual - EMBL-EBI

[Programme details](#)

5.4 Introduction to InterPro workshop (virtual)

Date: 21 September 2022 15:30 (GMT-3)

Speakers: Typhaine Paysan-Lafosse and Sara Chuguransky

Venue: Virtual - 3rd Women in Bioinformatics & Data Science LA Conference

[Programme details](#)

5.5 Bioinformatics resources for protein biology (virtual)

Date: 21 February - 2 March 2022

Venue: Virtual - EMBL-EBI

[Programme details](#)

5.6 Structural bioinformatics course (virtual)

Date: Monday 11 - Friday 15 October 2021

Venue: Virtual - EMBL-EBI

[Programme details](#)

5.7 Structural bioinformatics course (virtual)

Date: Monday 23 - Friday 27 November 2020

Venue: Virtual - EMBL-EBI

[Programme details](#)

5.8 Bioinformatics Resources for Protein Biology

Date: Tuesday 10 - Thursday 12 March 2020

Venue: European Bioinformatics Institute (EMBL-EBI) - Training Room 1 - Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

[Programme details](#)

INTERPRO ENTRIES : ESSENTIAL INFORMATION

An InterPro entry is created for each protein family, domain or important site signature that is integrated into InterPro from one or more of its *13 member databases*. Where signatures from two or more member databases describe the same family, domain or site, the member database signatures are brought together under one InterPro entry.

An InterPro entry provides a written description of the family, domain or site and lists the contributing member database signatures. Each entry has a name, a unique InterPro identifier and an *entry type*. *Go terms* associated with the entry are also displayed. For each InterPro entry further information is provided showing, for example, the proteins, structures and pathways matching this entry along with taxonomic distribution. This information can be easily viewed by *Browsing entries in the InterPro website*.

6.1 InterPro entry types

InterPro entries are created for protein families, domains, sites, repeats and homologous superfamilies, defined as follows:

F Family - a group of proteins that share a common evolutionary origin reflected by their related functions, sequence homology or similarities in their structure.

D Domain - a distinct functional, structural or sequence unit often found associated with other types of domains.

S Site - a short sequence containing one or more conserved residues, including: active sites, binding sites, conserved sites and sites of post-translational modification.

R Repeat - A short sequence (usually <50 amino acids) typically repeated many times within a protein.

H Homologous Superfamily - a group of proteins that share a common evolutionary origin, reflected by similarity in their structure, even if sequence similarity is low. This entry type contains signatures from the CATH-Gene3D and SUPERFAMILY member databases exclusively.

6.2 Other entry and page types

In addition to the main *InterPro Entries*, which bring together protein signatures from the member databases consortium, InterPro also provides entry pages for the individual *member database signatures* and for *proteins*, *structures*, *taxons*, *proteomes* and *sets/clans* integrated or used by InterPro. These entry pages also have further information available that can be viewed by *Browsing entries in the InterPro website*. More information is available in the corresponding *train online section*.

6.3 Entry relationships

InterPro entries that represent a subset of proteins from another InterPro entry are identified as “children” of the “parent” entry. InterPro displays these connections between entries in the “Family Relationships” or “Domain Relationships” sections. Entries at the top of these hierarchies describe broad families or domains that share higher level structure and/or function, while those entries at the bottom describe more specific functional subfamilies or structural/functional subclasses of domains. More information is available in the corresponding [train online section](#).

6.4 Overlapping entries

Relationships between homologous superfamilies and either family or domain entries are generated automatically using the Jaccard and containment indexes. These relationships are shown in the Overlapping homologous superfamilies/Overlapping entries section on the InterPro entry pages. More information is available in the corresponding [train online section](#).

6.5 Ontologies

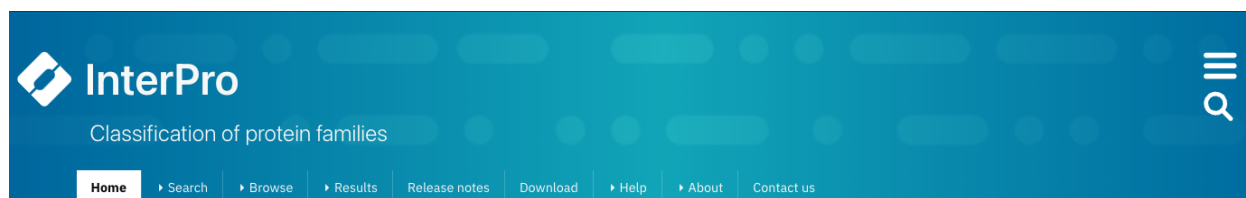
InterPro uses several standards and ontologies:

- the [NCBI Taxonomy](#) for taxa: the NCBI assigns unique taxonomic identifiers for all organisms (taxa) that are represented in UniProtKB. As these taxonomic identifiers are stable, InterPro uses them to let users search the resource by organism;
- the [Gene Ontology \(GO\)](#) for functions, processes, cellular components: InterPro2Go (<https://doi.org/10.1093/database/bar068>) is a manually created mapping between InterPro entries and GO terms. Where an InterPro entry hits a set of functionally similar proteins, GO terms describing the conserved function or location are associated with the InterPro entry.
- the Nomenclature Committee of the [International Union of Biochemistry and Molecular Biology](#) (NC-IUBMB) via [IntEnz](#): Enzyme Commission (EC) numbers describe enzyme-catalyzed reactions and are available in UniProtKB, e.g. [P17050](#). Where an InterPro entry hits reviewed/Swiss-Prot proteins annotated with EC numbers, the EC numbers are associated to the InterPro entry.
- [Reactome](#) and [MetaCyc](#) for pathways. Where an InterPro entry hits a reviewed/Swiss-Prot protein involved in a pathway described by Reactome, the pathway is associated to the InterPro entry. As reactions in MetaCyc include EC numbers, InterPro uses EC numbers assigned to an entry (as described above) and to a metabolic pathway to link InterPro entries and MetaCyc pathways.

INTERPRO WEBSITE BANNER

Every page in InterPro has an identical banner with some handy features described below.

7.1 Navigation banner and menu



The navigation banner contains:

7.1.1 Navigation menu tabs

Home, Search, Browse, Results, Release notes, Download, Help, About and Contact us.

7.1.2 Quick search box

The magnifying glass icon on the right side of the website banner can be clicked to show a text entry component and performs a *Quick search*.

7.1.3 Settings sidebar



The hamburger icon on the right opens the settings sidebar.

The settings sidebar provides another way to access different parts of the website and is the only way of accessing the *settings page*.

INTERPRO HOMEPAGE

The InterPro homepage can be split into in the following sections:

1. *InterPro resource overview*
2. *Search box*
3. *Data*
4. *News and information*

8.1 InterPro resource overview

This section (section 1 in the figure above) gives an overview of the InterPro resource and a link to the latest InterPro publication. The release version and date are displayed under the graphic, the user can click on it to access the [Release notes](#).

8.2 Search box

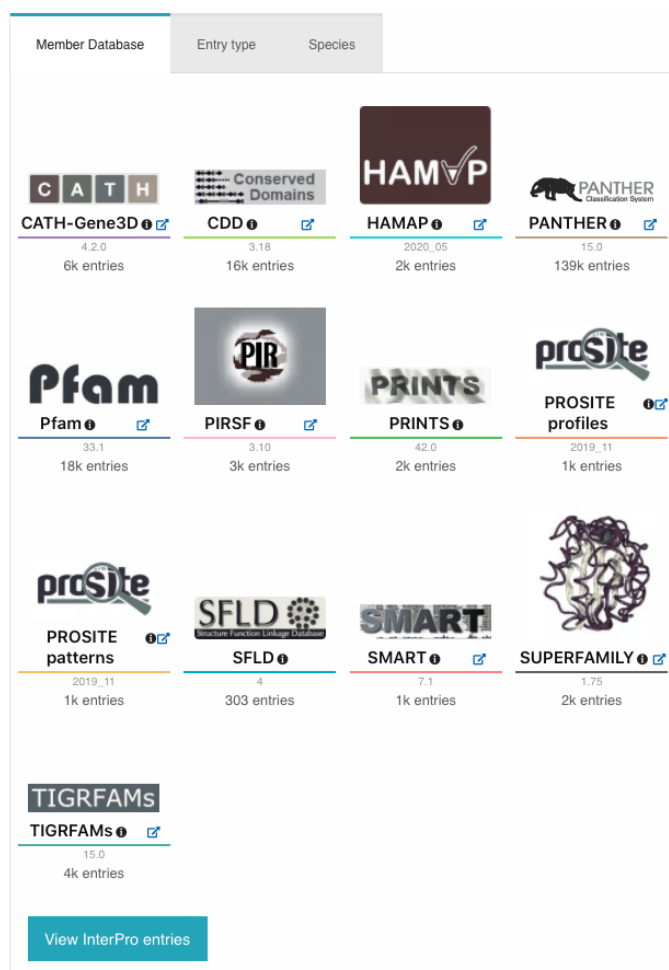
The search section (section 2 in the figure above) shows a multifunctional search component, which allows the selection of one of the five types of search supported by InterPro. More details about searching are available on the [How to search the InterPro website?](#) section.

The screenshot displays the InterPro search interface. At the top, there are three tabs: 'Search by sequence' (selected), 'Search by text', and 'Search by Domain Architecture'. Below the tabs, the 'Sequence, in FASTA format' section is visible. It contains a large text input field with the placeholder 'Enter your sequence'. Below the input field, there are two buttons: 'Choose file' and 'Example protein sequence'. Further down, there is a section titled 'Advanced options' with a right-pointing triangle icon. At the bottom of this section, there are two buttons: 'Search' and 'Clear'. In the bottom right corner of the search area, it says 'Powered by InterProScan'.

8.3 Data

The data section (section 3 in the figure above) gives an overview of InterPro data with shortcuts to different views of the data, and highlights the latest InterPro entries on the right hand side.

8.3.1 Member databases



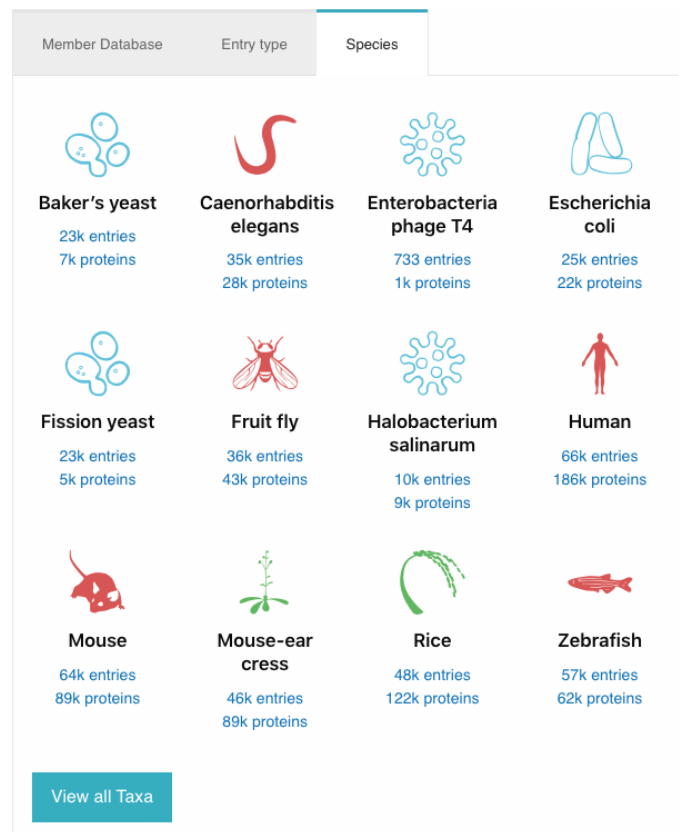
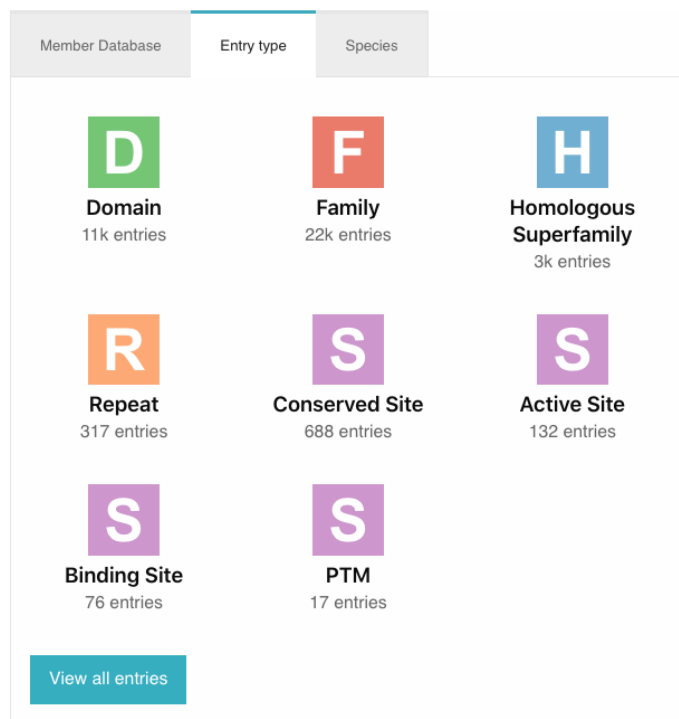
This section shows icons for the *InterPro consortium member databases*, along with information about the version of the member database and an estimate of the number of signatures from that resource which are in the current InterPro release. Each of the member database icons links to the *browse feature* showing data filtered to match the selected member database.

8.3.2 Entry type component

This section shows the icon and number of entries for each of the InterPro entry types. Clicking on an icon will display the browse feature showing InterPro data filtered by the selected entry type.

This component shows icons for *InterPro entry types*. An estimate of the number of entries corresponding to each type is shown under each icon. Clicking on an icon will display the *browse feature component* showing InterPro data filtered by the selected entry type.

8.3.3 Species component



The Species component shows a set of icons corresponding to several key species and an estimate of the number of entries and proteins associated with each species. Clicking on an icon will display the associated [Taxonomy entry page](#) for the selected organism. Clicking on the text below the icon will display the [Entries](#) or [Proteins](#) tabs, respectively.

8.3.4 Latest Entries component

Here we show a list of the latest InterPro entries with their entry type, followed by their name and accession number. The clickable icons beneath the text show the number of proteins, domain architectures, taxa, structures and member databases matching the entry. Each of the icons is clickable and provides a shortcut to the corresponding section of the *InterPro entry page*.

8.3.5 Favourites Entries component

This section provides a quick access to the list of favourite InterPro entries previously selected by clicking on the star icon in an InterPro entry page.

When a new version of InterPro has been released and one or more the Favourite entries have been updated, a button “**Check for updates**” is displayed.

When clicking on it, differences for each updated entry are displayed in a github diff style. The user can choose to apply the update or keep the previous annotation.

Latest entries

Favourite entries

Recent Search

F Paxillin

1k047826

IPR001904

H Asp/Glu racemase

55k024k50

IPR001920

F Bifunctional kinase-pyrophosphorylase

17k1112k2

F Profilin

6k612k61

IPR005455

D Amyloidogenic glycoprotein, extracellular

3k056913

IPR008154

F Cytochrome c4-like

12k05k7

IPR024167

Check for updates

[Home](#)
[Search](#)
[Browse](#)
[Results](#)
[Release notes](#)
[Download](#)
[Help](#)
[About](#)

Entry: IPR000001

Update

Saved	Latest
<div>Name:</div> <div>- Name: k ringle test</div> <div>- Short: test</div>	<div>Name:</div> <div>+ Name: Kringle</div> <div>+ Short: Kringle</div>
<div>Member_databases:</div> <div>Cdd:</div> <div>- Cd00108: Refu</div>	<div>Member_databases:</div> <div>Cdd:</div> <div>+ Cd00108: Kringle</div> <div>domain; Kringle domains are believed to play a role in binding mediat...</div>

8.3.6 Recent search component

Latest entries

Favourite entries

Recent Search

✕ KIAA1841

✕ DUF3342

✕ DUF1241

✕ PF00235

✕ PF00022

✕ PF02932

✕ albumin

✕ SARS-Cov-2

✕ SARS-Cov

✕ kinase

Clear History

When performing a Text search, the text is stored locally and accessible through this section, so the user can retrieve the data of interest easily the next time they visit the website. Unwanted saved Text searches can be removed by clicking on the cross icon, The “Clear History” button allows to clear the search history.

8.4 News and information

The final section of the homepage (section 4 in the *InterPro homepage* figure above) comprises components linking to the InterPro [feed](#), the articles from the [InterPro Blog](#) and technical aspects of the website.

The **Spotlight** section shows a selection of the latest articles from the [InterPro Blog](#). We publish a range of articles on the blog, from technical information about the resources run by the team to protein focus articles which deliver details about interesting entries from InterPro data.

The **Tools and libraries** section provides quick access to some of the tools and software used throughout the website.

HOW TO SEARCH THE INTERPRO WEBSITE?

A search can be performed on the *InterPro homepage* using the *Search box* component, by clicking on the Search tab in the *navigation menu*, or by clicking on the magnifying glass in the *navigation banner*. There are five different types of search available in InterPro:

- *Quick search*
- *Sequence search*
- *Text search*
- *Domain architecture search*
- *Using Browse feature to search and filter InterPro*

9.1 Quick search



The magnifying glass in the navigation banner allows a quick search for a specified keyword. A search can be triggered by entering some text and pressing the enter/return key or clicking the magnifying glass. If the keyword is text, the results will be displayed as described in the *Text search*. If the keyword entered is an accession, it automatically redirects to the corresponding *InterPro page* under the **Browse** tab in the *navigation menu*.

9.2 Sequence search

A sequence or a batch of sequences can be submitted in FASTA format in the dedicated text area or by uploading a fasta file. The “**Advanced options**” allows users to select the InterPro member databases of interest to search against (by default they are all selected). The sequence search is performed using the *InterProScan software*. While the sequence search is running, the user can continue to navigate through the website, other browser tabs or applications and will get a pop-up notification when the job has been completed (this requires the browser notifications to be allowed).

by sequence by text by domain architecture

Sequence, in FASTA format

This form allows you to scan your sequence for matches against the InterPro protein signature databases, using InterProScan tool. Enter or paste a protein sequence in FASTA format (complete or not - e.g. `EMPIGSKERPTFFEIFKTRCNKADLGPISLN`), with a maximum length of 40,000 amino acids.

Please note that can scan up 100 sequences at a time. Alternatively, read [more about InterProScan](#) for other ways of running sequences through InterProScan.

Enter your sequence

[Choose file](#) Example protein sequence

► Advanced options

[Search](#) [Clear](#)

Powered by InterProScan

9.2.1 Sequence search results

Results of a protein sequence search are available under the **Results** tab in the navigation menu under **Your InterProScan Searches** section. This page displays the protein sequence searches you have performed in the last seven days, with the most recent one being displayed at the top. The status column gives an indication of whether or not the search has completed (green tick symbol / searching), if the search has been saved locally (the results will still be available even after the seven days limit set up on InterPro servers), or if the results have been imported (file symbol). Clicking on the job id or on the text in the results column opens a page where the results are summarised in a protein sequence viewer (more detailed information is provided for the [Protein sequence viewer](#)).

Previously ran searches can be imported either by typing the job ID in the **Import** text box, for searches performed in the last seven days on our servers, or by uploading an [InterProScan](#) output file in JSON format, the job is added to the Results table. If the second option is chosen and InterProScan was run using nucleotide sequences, a job result is created for each Open Reading Frame (ORF) and ORFs from the same nucleotide sequence are grouped accordingly. This import feature can be used by users requiring to have InterProScan graphic output formats for publications and other uses.

When a search has been run using a previous version of InterProScan, it can be re-run using the latest version of the software. When a batch of sequences has been submitted, group actions allow to Delete All, Re-run All, and Download All the submitted sequences at once. If the search has been run in the last seven days, the results can be downloaded in JSON, XML and TSV formats, thereafter, if the search has been saved locally, the results are only available in JSON format.

On the search results page, some general information on the submitted sequence is provided, followed by the predicted InterPro protein family membership when available ([1] in the figure above). The search can be saved by clicking on the **Save in Browser** button. The status will be changed to “**Imported file**”. This means that the results will be available behind the usual seven days limit on the browser and machine the save has been done, and will only be deleted if the user deletes the job by clicking on the bin icon.

The sequence submitted is shown in its full length at the top of the protein sequence viewer (grey bar) [2]. This is followed by a summary of the representative domains composing the protein, when available [3]. InterPro entries and signatures matches are displayed in categories classified by [InterPro entry types](#). Each coloured bar represents a domain, protein family, or important site that has been matched to part or all the length of the submitted protein sequence.

- The top coloured bar represents the InterPro entry [4a, 5a].

InterPro Classification of protein families




Home Search Browse **Results** Release notes Download Help About Contact us


Result / InterProScan















Your InterProScan Search Results ⁱ

Your InterProScan search results are shown below. Searches may take varying times to complete. You can navigate to other pages and once the search is finished, you will receive a notification. The results will be available for 7 days.


Alternatively, you can import the results of an InterProScan run (in JSON format) into this page in order to view your search results interactively.

Submit a new search  Import:  

1 - 5 of 5 results 

RESULTS	CREATED	STATUS	ACTION
iprscan5-R20240115-162102-0037-52327852-np2 	34 seconds ago	 Searching	
EMBOSS_001 	2 days ago		
mol:protein subsequence:1-350 length:350 Results are being processed on the InterProScan server 	1 week ago		
20240112-164446-3			
seq2 	2 days ago		<ul style="list-style-type: none"> Delete All FASTA input TSV output JSON output XML output GFF output Resubmit All
seq3 	2 days ago		

Previous **1** Next

 This service is part of the ELIXIR infrastructure
InterPro is an ELIXIR Core Data Resource


This service is part of the GBC
InterPro is a Global Core Biodata Resource 

Fig. 1: Summary of sequence searches jobs.

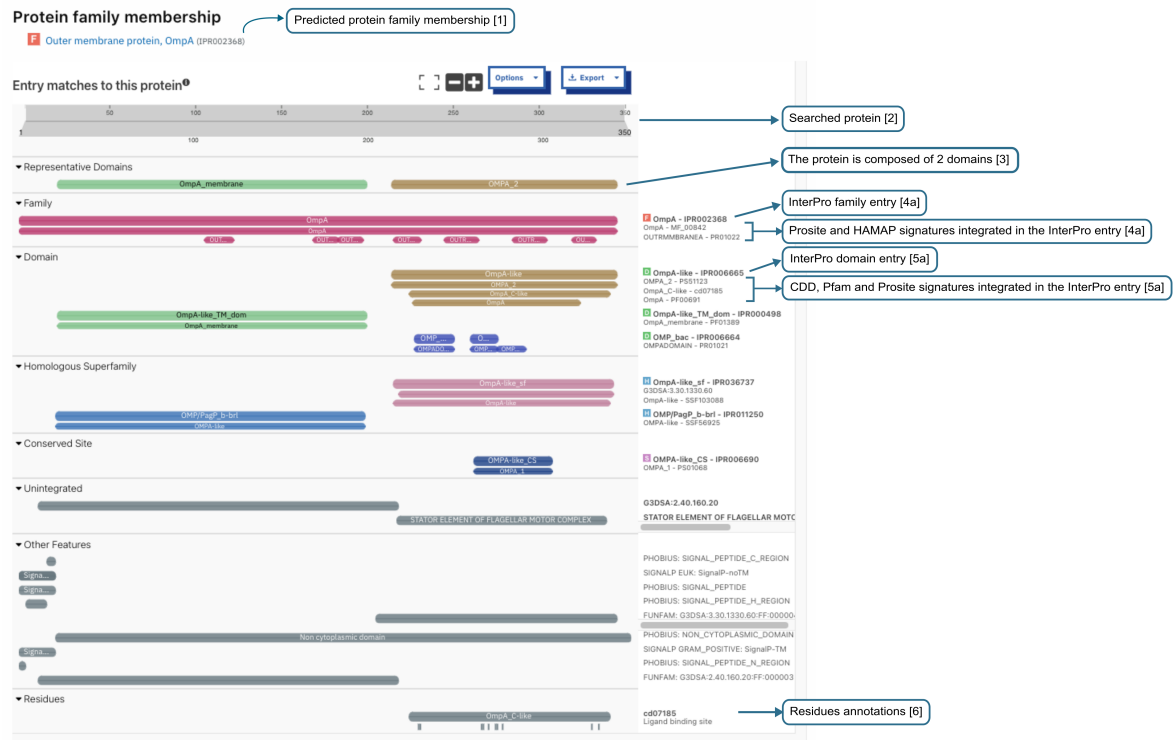


Fig. 2: Example of protein sequence viewer as displayed in the search result page.

- Directly below the InterPro entry, additional coloured bars display the member database signatures that contributed to that InterPro entry [4b, 5b].

In the example above, four InterPro entries (1 family and 3 domain entries) have been found matching the submitted sequence. The first InterPro entry is for a protein family [4a], containing two member database signatures, in this case from Prosite (PR01022) and HAMAP (MF_00842)[4b]. The following three InterPro matches are domains. The top InterPro domain entry [5a] contains signatures from 3 member databases (Pfam, CDD and Prosite) [5b] which all represent the same domain. The remaining two InterPro domains contain one member database signature.

Looking at the **Other features** section, we also learn that the protein has a signal peptide at its N-terminal end. CDD also provides per residue annotations which are displayed in the corresponding category at the bottom of the viewer [6].


Additionally to the InterPro matches, information about the GO terms associated to the InterPro entries and PANTHER signatures matching the protein are displayed below the sequence viewer when available. The GO terms are assigned manually to InterPro entries using on the [Gene Ontology](#) and reflect the Biological process, Molecular function or Cellular location the protein may have.

9.3 Text search

The text search is available by selecting the “By Text**” section under the **Search** tab in the website menu. The text search allows to search the following information in the database:

- Name or keyword (e.g. [Afadin](#))
- InterPro accession (e.g. [IPR000562](#))
- Member database signature accession (e.g. [PF00040](#))
- Protein accession (e.g. [P04937](#)) or identifier/short name (e.g. [FINC_RAT](#))
- PDB structure (e.g. [6AR9](#))
- Gene name (e.g. [BRCA2](#))
- GO terms (e.g. [GO:0005911](#))
- Proteome accession (e.g. [UP000000304](#))
- Taxonomy accession (e.g. [7240](#))
- Set/Clan accession (e.g. [CL0451](#))

Entering a **name**, or **keywords**, retrieves a list of all the InterPro entries and InterPro member database signatures that contain these searched words in their title or description. By default the term searched is highlighted in the results

list and the description is shortened, clicking on the  symbol located on the left hand side of the **Export** button removes the highlight and shows the full description text. The setting is saved and also applied to other text searches throughout the website.

Entering an **accession number** gives an exact match and a quick access to the corresponding InterPro page. It also displays the list of the InterPro entries and any member database signatures linked to that accession number/identifier.

Selecting the accession number or name of any entry in the list of entries opens the corresponding InterPro page (e.g. [member database signature](#), [InterPro entry](#)). An overview of the entry is provided and tabs on the left hand-side menu allow specific information for the entry to be viewed, for example the species in which a protein has been found, or structures matching an entry. More information on the [browsing an InterPro page](#) section.

9.4 Domain architecture search

The screenshot shows the 'by domain architecture' search tab. It includes a search bar, a list of domains to include (PF03165, PF03166), a list of domains to exclude, and options for 'Order of domain matters' and 'Exact match'. A 'Database' dropdown is set to 'Pfam'. An 'Export' button is visible. Numbered annotations point to various features: [1,2] points to the 'Add a domain' button; [3a] points to the domain list; [3] points to the 'Domains order' section; [4] points to the 'Exclude a domain' button; [5] points to the 'Exact match' checkbox; [6] points to the 'Database' dropdown; and [7] points to the 'Export' button.

This search option allows the retrieval of protein sequences that contain specific Pfam/InterPro domains in a particular arrangement referred to as a “domain architecture”. For example, protein sequences containing both a SH2 domain and SH3 domain can be retrieved. Domains that the proteins should or should not contain can be included or excluded from the domain architecture respectively. Selecting “**Order of domain matters**” offers the possibility to arrange the domains in a particular order. Selecting “**Exact match**” performs the search to find proteins containing the selected domains only (no extra domain in the proteins). Domains can be selected by entering a domain name, a Pfam accession, or an InterPro accession if a Pfam entry is integrated in it.

Once a search is performed the corresponding results are displayed below the search component and show the number of proteins followed by the corresponding domain architecture. For each domain architecture, the domain size is displayed based on the real length of the domain, using a protein of reference. When hovering over a domain, more details are available in a tooltip, including the domain’s position. Clicking on the number of proteins redirects to the **Browse** tab in the *navigation menu* under the protein section, showing the list of proteins which can be filtered to a specific member database, if required, as described in the *browse feature*.

By default, Pfam entries are shown in the results. This can be changed to show InterPro entries by toggling the Pfam checkbox to InterPro and vice versa.

The domain architectures can be downloaded in JSON and TSV formats through the **Export** button.

9.5 Using Browse feature to search and filter InterPro

The browse search page can be accessed by clicking on the Browse tab in the *navigation menu*. The browse search provides a powerful functionality to select subsets of data available in InterPro by selecting filters according to the results required. For example, this page can be used to browse all entries which have a contributing signature from a particular member database e.g. HAMAP, or to retrieve all proteins from a certain taxon, e.g. *Escherichia coli*, that contain a specific domain eg OmpA-like domain.

Below we describe how to use the browse search feature:

InterPro Classification of protein families

Home Search Browse Results Release notes Download Help About

/ Browse / By Entry / InterPro

Filter By Clear Collapse All

InterPro Type

- All 39k
- Family 23k
- Domain 11k
- Homologous Superfamily 3k
- Repeat 321
- Conserved Site 692
- Active Site 132
- Binding Site 76
- PTM 17

GO Terms

- All
- Molecular Function MF 12k
- Biological Process BP 11k
- Cellular Component CC 6k

1 - 20 of 39k entries in InterPro

Search entries Export

TYPE	ACCESSION	NAME	INTEGRATED SIGNATURE(S)	GO TERMS
D	IPR000001	Kringle	SM00130 PF00051 PS50070 cd00108	
F	IPR000003	Retinoid X receptor/HNF4	PR00545	DNA binding steroid hormone receptor activity zinc ion binding regulation of transcription, DNA-templated nucleus
F	IPR000006	Metallothionein, vertebrate	PR00860 PTHR23299	metal ion binding
D	IPR000007	Tubby, C-terminal	PF01167 PR01573	
D	IPR000008	C2 domain	PR00360 PF00168 SM00239 PS50004	
F	IPR000009	Protein phosphatase 2A regulatory subunit PR55	PR00600 PIRSF037309 PTHR11871	protein phosphatase regulator activity protein phosphatase type 2A complex
D	IPR000010	Cystatin domain	SM00043 PF00031 PF16845 cd00042	cysteine-type endopeptidase inhibitor activity

1. Select a data type

The browse page opens up with **7 data types** to allow browsing of InterPro entries, Member databases signatures, Proteins, Structures, Taxonomies, Proteomes or Sets.

InterPro Classification of protein families

Home Search Browse Results Release notes Download Help About

/ Search / Sequence

Search InterPro

by sequence

By InterPro

By Member DB

By Protein

By Structure

By Taxonomy

By Proteome

By Set

2. Select any additional filters

The filters options displayed for each data type will vary as appropriate.

9.5.1 Member database filter


① Select your database:

AntiFam	263
CATH-Gene3D	7k
CDD	19k
HAMAP	2k
NCBIfam	7k
PANTHER	16k
Pfam	21k
PIRSF	3k
PRINTS	2k
PROSITE profiles	1k

The “**Select your database**” option is available when Browsing by Member DB, Protein, Structure, Taxonomy and Set. It allows results to be retrieved from all or a selection of *InterPro member databases*. Only the databases that contain signatures for the chosen data type are displayed as options. By default all the member databases are selected, except when Browsing by Member DB, where Pfam is the default option selected.

9.5.2 Text filter

The “**Search entries**” box allows results to be filtered to match the text entered. For example, the text could be a keyword that might be found in entry names. It also allows specific protein names or taxa to be entered. By default the term searched is highlighted in yellow in the results list, this can be

disabled by clicking on the  symbol appearing between the text box and **Export** button once the search has started, the setting is saved and also applied to other text searches throughout the website.

9.5.3 Data-type specific filters

InterPro entry filters

When **Browse by InterPro** is selected, two filter types can be applied:

- **InterPro Type:** limits the data in the *data views* to the selected *InterPro entry types*.
- **Go Terms:** filters by selected Go terms from *InterPro2GO*.

Member database filters

Try / InterPro

Collapse All ▼

39k

23k

11k

3k

692

322

132

76

17

MF 11k

BP 11k

CC 6k

Protein filters

1 Select your database:

AntiFam	263
CATH-Gene3D	7k
CDD	19k
HAMAP	2k
NCBIfam	7k
PANTHER	16k
Pfam	21k
PIRSF	3k
PRINTS	2k
PROSITE profiles	1k

Filter By

Clear | Collapse All ▼

▼ Member Database Entry Type

<input checked="" type="radio"/> All	21k
<input type="radio"/> Family	12k
<input type="radio"/> Domain	8k
<input type="radio"/> Repeat	826
<input type="radio"/> Coiled Coil	189
<input type="radio"/> Disordered	121
<input type="radio"/> Conserved Site	119

▼ InterPro State

<input checked="" type="radio"/> Both	21k
<input type="radio"/> Integrated	20k
<input type="radio"/> Unintegrated	709

When **Browse by Member DB** is selected and a member database has been chosen, subsequent filters can be applied:

- **Member Database Entry Type:** select the types of signatures required. This is dependent on the database type selected. For example, if a database contains both domains and family signatures you can filter the results for a specific type.
- **InterPro state:** select all signatures from the selected database or only those signatures that have been integrated into InterPro.

Just as with the *Member DB* data type, **Protein** filters change based on the selection in the *member database filter* component. The basic filters are displayed irrespective of the selection made and an extra filter when the “**All Proteins**” option is selected.

Database selected

If a member database has been selected, the following filters are displayed:

- **UniProt Curation:** the [UniProtKB](#) is split into two sections. The reviewed set is manually curated (SwissProt) and the unreviewed set is derived from public databases automatically integrated into UniProt (TrEMBL).
- **Taxonomy:** this filter allows the displayed list of proteins to be limited to certain organisms.
- **Sequence Status:** this filter allows proteins to be limited to complete proteins or fragments.

All Proteins

Filter By Clear | Collapse All ▼

▼ UniProt Curation

☒ Both 156M

☐ Reviewed 534k

☐ Unreviewed 156M

▼ Taxonomy

☒ All

☐ Homo Sapiens 158k

☐ Arabidopsis Thaliana 106k

☐ Oryza Sativa Subsp. Japonica 81k

☐ Mus Musculus 68k

☐ Danio Rerio 55k

☐ Drosophila Melanogaster 30k

☐ Caenorhabditis Elegans 19k

☐ Escherichia Coli 10k

☐ Halobacterium Salinarum 7k

☐ Saccharomyces Cerevisiae 5k

☐ Schizosaccharomyces Pombe 4k

☐ Escherichia Virus T4 612

▼ Sequence Status

☒ Both/All 156M

☐ Complete Sequence 140M

☐ Fragment 16M

Additionally to the filters mentioned above, when the “**All Proteins**” option is selected in the *member database filter* component, the **Matching Entries** filter is displayed. This filter allows the selection of proteins which do or do not contain matches to entries in the InterPro dataset.

▼ Matching Entries

☐ With Matches 170M

☐ Without Matches 36M

☒ All proteins 206M

Structure filters

Structure filters do not vary depending on which option has been selected in the *member database filter* component.

- **Experiment Type:** this filter allows selection of structures based on the type of experimental data the structure is based on.
- **Resolution:** this filter allows structures to be selected based on the resolution of the structure.

9.5.4 Data Display Options

The data display is the main part of the results section in the browse page and shows the data selected in the *data type menu*. The actual details shown will also be dependent on the selected data type.



Tabular view



The tabular view is the default view and is available for all *InterPro data types*. The table view icon formats data into a tabular view composed of rows representing individual entities. The table header describes the contents of each column. Clicking on one of the rows redirects to the corresponding *InterPro page*.

TYPE	ACCESSION	NAME	INTEGRATED SIGNATURE(S)	GO TERMS
				BP MF CC
D	IPR000001	Kringle	SM00130 PF00051 PS50070 cd00108	
F	IPR000003	Retinoid X receptor/HNF4	PR00545	<ul style="list-style-type: none"> nucleus DNA binding steroid hormone receptor activity zinc ion binding regulation of transcription, DNA-templated
F	IPR000006	Metallothionein, vertebrate	PR00860 PTHR23299	<ul style="list-style-type: none"> metal ion binding

Fig. 3: Tabular view example for InterPro entry data type

Grid view

The grid view is available for all *InterPro data types*. It displays a series of cards summarising details of the entities being viewed. Clicking on one of the cards redirects to the corresponding *InterPro page*.

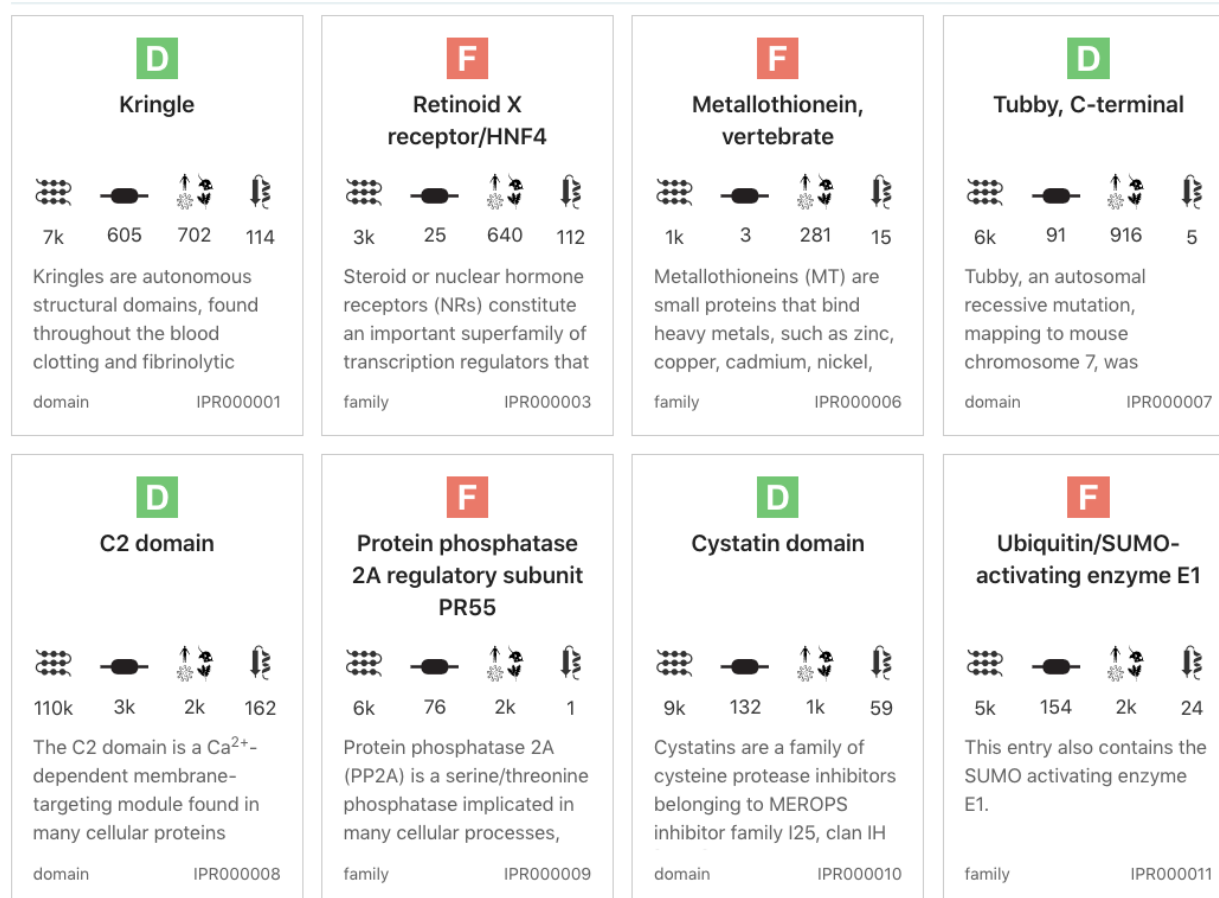
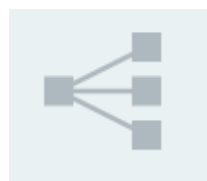
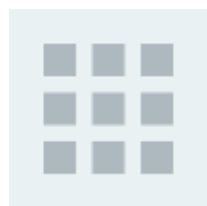


Fig. 4: Grid view example for InterPro entry data type

Tree view



The tree view is currently only enabled for taxonomy data. The tree view icon is only shown where a tree view is possible. The taxonomy tree viewer can be navigated by clicking on nodes or using keyboard arrow keys. This component is also used in the *Taxonomy entry page*.

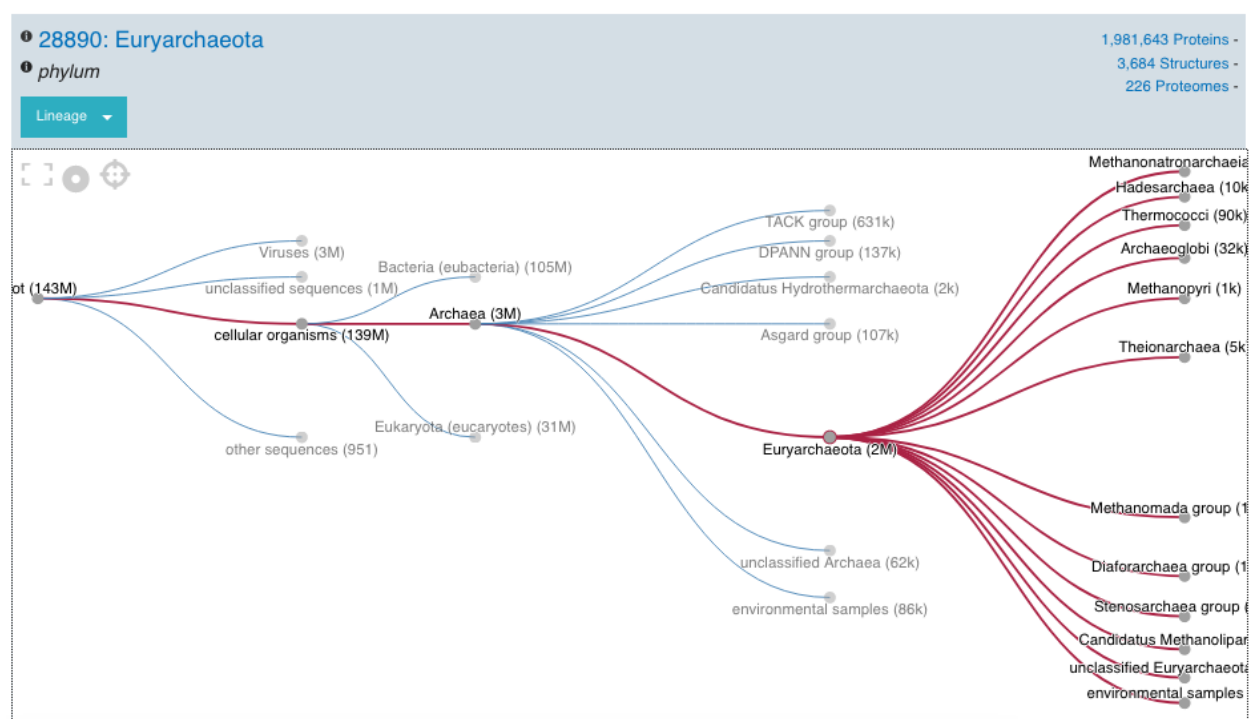


Fig. 5: Tree view example for Euryarchaeota phylum

PROTEIN SEQUENCE VIEWER

A common element on several InterPro website pages is the protein sequence viewer (in the [sequence search result](#), on the [protein](#) and [structure](#) pages). It summarises the InterPro entries (IPR) (top coloured bar) and member database signatures matches to the protein or structure being looked at, represented by the grey bar at the top of the viewer, categorised by *InterPro entry types*.

The *AlphaFold confidence* track is displayed in the protein sequence viewer in the [protein page](#) and in the [AlphaFold subpage](#) when a predicted structure is available.

The *Representative Domains* track is displayed in the protein sequence viewer in the [protein page](#). This representation is generated automatically using the type of the member databases models, which might differ from the InterPro entries types. When multiple models are overlapping, the representative domain is chosen by selecting the model covering the longest region of the protein. Be aware that in case of models made of multiple fragments, not all the fragments are necessarily chosen as representative, they are considered as individual entities for the selection.

Various options, make it easy to work with (as illustrated in the figure above):

1. Clicking on the Full screen button at the top of the viewer will switch to full screen view.
2. The viewer can be zoomed in and out by:
 - a. Clicking the two buttons (+ and -) at the top right corner.
 - b. Dragging the grey scale at the top to the desired positions on both left and right sides
 - c. Pressing the [Ctrl] key and scroll through the viewer
3. More options that customise the viewer are grouped under **Options** dropdown.
 - A. **Colour By** allows to change the colours in which the InterPro entries and signatures bars based on accession, member database or domain relationship.
 - B. The labels on the right side of the viewer can be customised. The **Accession** labels are shown by default. To see names and/or short names along with accession, the name/short name checkboxes should be ticked or if the user prefers to see the names/short names alone, the respective options should be selected.
 - C. **Save as image** allows to take a snapshot of the viewer and is saved as an image (.png).
 - D. **Collapse All** allows to collapse all the signatures bars displayed in the viewer at once to only display the InterPro entries bars.
 - E. The tooltips are shown when hovering over each bar. They can be disabled by unchecking the **Tooltip Active** option.
4. Residues annotations are provided by the CDD, SFLD and PIRSR databases.
5. Clicking on the header of a category (say Unintegrated) hides the bars for the entire category.

InterPro Classification of protein families

Home Search Browse Results Release notes Download Help About Contact us

Entry matches to this protein: Drag the gray scale to zoom in/out [2b]

Full screen view [1]

Customisation options [3]

Zoom buttons [2a]

Clicking on the header hides all the bars within the category [5]

Residues annotations [4]

AlphaFold Confidence

Representative Domains

Family

Domain

Homologous Superfamily

Binding Site

Unintegrated

Other Features

Residues

Other Residues

HisP_aminotrans - IPR005861
HisC - TIGR01141
HisC_aminotrans_2 - MF_01023
Pat - IPR024892
Phe_aminotrans_2 - MF_01513

Aminotransferase_I/II - IPR0048
Aminotran_3_2 - PF00155

PyrdxIP-dep_Trfase_major - IPR
G3DSA:3.40.640.10

PyrdxIP-dep_Trfase_small - IPR
G3DSA:3.90.1150.10

PyrdxIP-dep_Trfase - IPR015424
SSF53383

Aminotrans_II_pyridoxalP_BS - I
AA_TRANSFER_CLASS_2 - PS00599

PTHR43643
AAT_like - cd00609

Pfam-N: PF00155

cd00609
Catalytic residue
Homodimer interface
Pyridoxal 5'-phosphate binding site

MOD_RES: N6-(pyridoxal phosphi
MOD_RES: N6-(pyridoxal phosphi



Fig. 1: Tooltip example.

Options ▾

When zoomed in, panning can be achieved by either dragging the scale at the top or by dragging any bar in the desired direction (see figure below).

Entry matches to this protein ⓘ Move the scale left or right to pan the viewer

143 50 100 150 200 250 300 350 235

ITDRTRLIFVCNPNPTSTVVDALVRFVDVADPADILIAIDEAYVEYIRDGLLPNSLELALSRSNVVLRTFSKAYGLAGLRVGYAIGHPEL

▼ AlphaFold Confidence pLDDT ⓘ

▼ Representative Domains AAT_like AAT_like - cd00609

▼ Family

HisP_aminotrans HisP_aminotrans - IPR005861
hisC - TIGR01141
HisC_aminotrans_2 - MF_01023

Pat Pat - IPR024892
Phe_aminotrans_2 - MF_01513

▼ Domain Pan left or right by moving the bar accordingly

Aminotransferase_I/II Aminotran_1_2 - PF00155

▼ Homologous Superfamily

PyrdxIP-dep_Trfase_major PyrdxIP-dep_Trfase - IPR00048
G3DSA:3.40.640.10

PyrdxIP-dep_Trfase_small PyrdxIP-dep_Trfase - IPR00048
G3DSA:3.90.1150.10

PyrdxIP-dep_Trfase PyrdxIP-dep_Trfase - IPR015424
SSF53383

▼ Binding Site

Aminotran... AA_TRANSF...

▼ Unintegrated

AAT_like PTHR43643
AAT_like - cd00609

▼ Other Features

Pfam-N: PF00155

▼ Residues

AAT_like cd00609
Catalytic residue
Homodimer interface
Pyridoxal 5'-phosphate binding site

▼ Other Residues

MOD_RES: N6-(pyridoxal phosphé
MOD_RES: N6-(pyridoxal phosphé

For some proteins, additional information are provided by resources other than the member database consortium, they are displayed under the **Other features** category of the viewer. Available data include:

- Disordered regions from [MobiDB](#)
- Transmembrane regions from [Phobius](#) and/or [TMHMM](#)
- Coiled regions from [COILS](#)
- Cytoplasmic/non-cytoplasmic domains from [Phobius](#)

- Signal peptide regions from [SignalP](#) and/or [Phobius](#)
- Spurious protein from [AntiFam](#)
- [CATH-FunFams](#) is an automatically generated profile HMM database, with FunFams entries segregated by an entropy-based approach that distinguishes different patterns of conserved residues, corresponding to differences in functional determinants
- [Pfam-N](#) annotations result from a deep learning methodology developed by the Google Research team led by Dr Lucy Colwell to increase the Pfam coverage of protein sequences
- Eukaryotic linear motifs from [ELM](#)

For some proteins, we also have annotations that are fetched directly from the resource API. These annotations are displayed under the **External Sources** category of the viewer. Note: by default this category is collapsed. Available data include:

- 3D structure and domain predictions from the [Genome3D consortium](#)
- Intrinsically disordered proteins from [DisProt](#)
- Tandem repeat from [RepeatsDB](#)

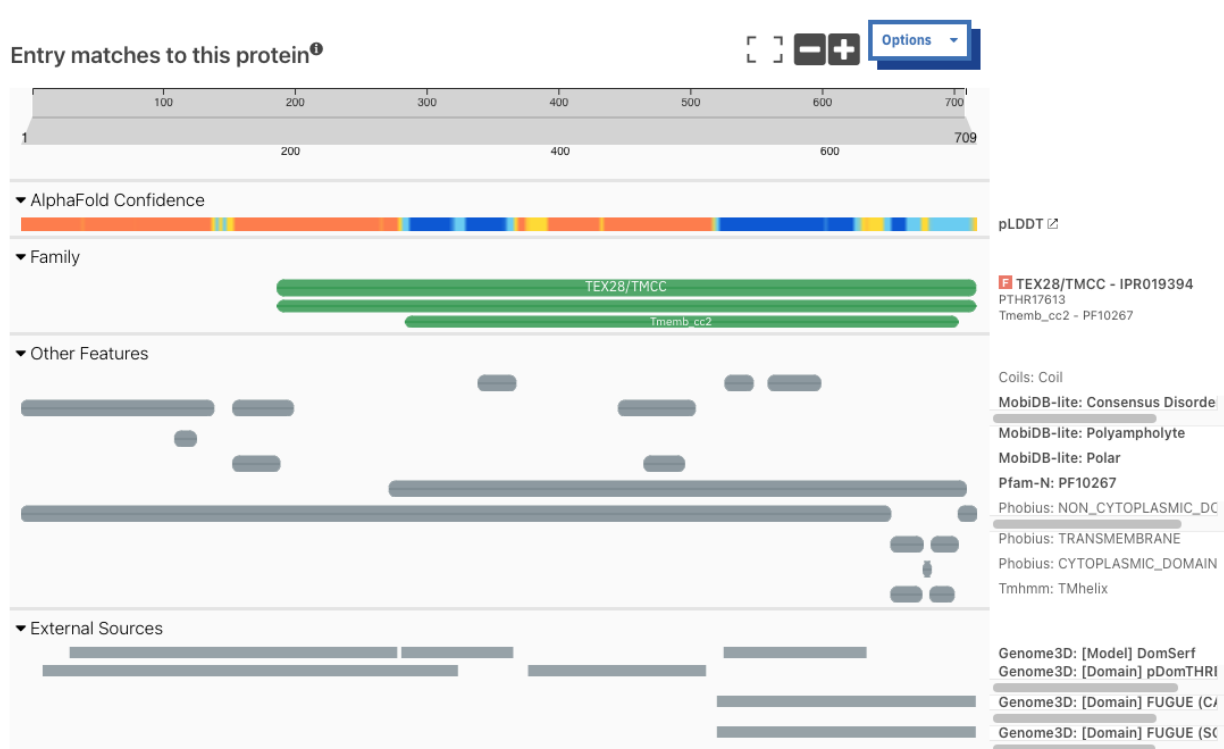


Fig. 2: Protein sequence viewer External Sources for [O75069](#)

BROWSING ENTRIES IN THE INTERPRO WEBSITE

You can get to entry pages in InterPro in lots of different ways. Commonly this will involve clicking on a link to an entry from one of the *search methods*. This section describes the different types of entries and what you will find for each of their pages.

There are 7 categories of entry pages in InterPro:

- *InterPro entry*
- *Member database signature*
- *Protein*
- *Structure*
- *Taxonomy*
- *Proteome*
- *Set/Clan*

The following entry data tabs are available when appropriate. We describe each in detail in the first entry page it appears in. Most entry data tabs will be described within the *InterPro entry page*.

- *Proteins*
- *Domain architectures*
- *Taxonomy*
- *Proteomes*
- *Structures*
- *AlphaFold*
- *Interactions*
- *Pathways*
- *Signature*
- *Subfamilies*
- *Alignment*
- *Curation*
- *Entries*
- *Sequence*
- *Similar proteins*

11.1 InterPro entry page

An InterPro entry represents a unique protein homologous superfamily, family, domain, repeat or important site based on one or more signatures provided by the *InterPro member databases*.

[Home](#) / [Browse](#) / [By Entry](#) / [InterPro](#) / [IPR000562](#) / [Overview](#)

D

IPR000562 **Fibronectin type II domain** ★

InterPro entry

Overview

- Proteins 16k
- Domain Architectures 792
- Taxonomy 3k
- Proteomes 735
- Structures 38
- AlphaFold 5k
- Pathways 156

Short name *FN_type2_dom*

Overlapping homologous superfamilies •

- Kringle-like fold (IPR013806)
- Fibronectin type II domain superfamily (IPR036943)

Description

Fibronectin is a multi-domain glycoprotein, found in a soluble form in plasma, and in an insoluble form in loose connective tissue and basement membranes, that binds cell surfaces and various compounds including collagen, fibrin, heparin, DNA, and actin. Fibronectins are involved in a number of important functions e.g., wound healing; cell adhesion; blood coagulation; cell differentiation and migration; maintenance of the cellular cytoskeleton; and tumour metastasis [3].

The major part of the sequence of fibronectin consists of the repetition of three types of domains, which are called type I, II, and III (also called FN1, FN2 and FN3 respectively) [1]. In fibronectin the type II domain is duplicated. Type II domain is approximately forty residues long, contains four conserved cysteines involved in disulphide bonds and is part of the collagen-binding region of fibronectin. Type II domains have also been found in a range of proteins including blood coagulation factor XII; bovine seminal plasma proteins PDC-109 (BSP-A1/A2) and BSP-A3 [4]; cation-independent mannose-6-phosphate receptor [6]; mannose receptor of macrophages [8]; 180 Kd secretory phospholipase A2 receptor [5], DEC-205 receptor [2]; 72 Kd and 92 Kd type IV collagenase (3.4.24.24) [9]; and hepatocyte growth factor activator [7].

[Add your annotation](#)

Contributing Member Database Entries

Conserved Domains

CDD: [cd00062](#)

proSite

PROSITE profiles: [PS51092](#)

proSite

PROSITE patterns: [PS00023](#)

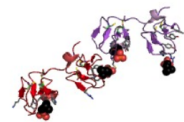
SMART

SMART: [SM00059](#)

Pfam

Pfam: [PF00040](#)

Representative structure



1h8p: Bull seminal plasma PDC-109 fibronectin type II module

Fig. 1: InterPro entry page for IPR000562.

InterPro entry pages give a brief description of the entry, name and unique InterPro identifier. The InterPro entry type (homologous superfamily, family, domain, repeat or site) is also indicated by an icon (e.g. a D with a green background for a domain).

Clicking on the star symbol next to the entry name will save the entry as a Favourite. The full list of saved entries is available in the *Favourites Entries component* in the homepage. More information about the data provided in an

On the right hand side, the **Add your annotation** button allows the user to suggest updates to the InterPro annotation and the page member databases contributing signatures to the entry are shown in a box. Below, the **Contributing Member Database Entry** integrated into the InterPro entry are listed with links to the corresponding *member database pages*. At the bottom of this column, if any experimentally solved structure is available, a **Representative structure** shows a small static 3D representation, the corresponding PDB ID and name, and a link to the *structure entry page*. The chosen representative structure is picked from structures that match the entry and have a resolution of less than 2

Angstroms. In this refined dataset, the representative structure is identified as the one exhibiting the highest coverage ratio for the entry, where a minimum of 50% of the residues in the structure are covered by the entry.

Overlapping homologous superfamilies and/or *Relationships to other entries* are indicated where available.

InterPro entry page can be found in the *InterPro Entries : essential information* section of the documentation.

Additional tabs in the left-hand side menu provide further information about the entry, and are displayed when the data is available. Types of data that may be available in the menu of an InterPro entry page include: *Proteins*, *Domain architectures*, *Taxonomy*, *Proteomes*, *Structures*, *AlphaFold*, *Pathways* and *Interactions*.

Although most InterPro entries remain carefully reviewed by our curators, some type Family entries containing signatures from PANTHER, NCBIfam or CATH-Gene3D which cover approximately the whole protein length are AI-generated. For these entries, the name, short-name and description have been generated automatically using a Large

Language Model. All AI-generated content is flagged as such with an **AI** tag. Please consider that this content has not been subjected to curator review when interpreting related results. More information on AI-generated content can be found in AI-generated content.

[Browse](#) / [By Entry](#) / [InterPro](#) / [IPR051632](#) / [Overview](#)



IPR051632
Rho guanine nucleotide exchange factor
AI

InterPro entry

Overview

- Proteins 8k
- Taxonomy 4k
- Proteomes 919
- Structures 65
- AlphaFold 3k
- Pathways 18

This entry contains information that has been generated using an AI language model. Please exercise discretion when interpreting the information provided.

[Read more on description generation](#) [Provide feedback](#)

Short name *Rho_GEF* AI

Description

AI-generated Unreviewed

This family of proteins includes Rho guanine nucleotide exchange factors (GEFs) that activate Rho-GTPases by facilitating the exchange of GDP for GTP. They play a pivotal role in various cellular processes such as cell motility, polarization, cytoskeletal dynamics, and signal transduction pathways. These proteins are involved in focal adhesion formation, axonal branching, synapse formation, dendritic morphogenesis, and B-lymphocyte activation. Some members are implicated in innate immune responses, acting as signaling intermediaries and participating in cytokine secretion. They also contribute to cell cycle regulation, apoptosis, and cancer progression. Additionally, they are involved in the development and differentiation of neuronal progenitor cells and the migration of precerebellar neurons. Certain members function as scaffold proteins, assembling signaling complexes downstream of G protein-coupled receptors, and are involved in cardiac hypertrophy and osteogenesis.

[Add your annotation](#)

Contributing Member Database Entry


PANTHER: [PTHR13944](#)

Representative structure



7g82: ARHGEF2 PanDDA analysis group deposition -- ARHGEF2 and RhoA in complex with Z1079512010

Fig. 2: InterPro AI-generated entry page for [IPR051632](#). Name, short-name and description have been generated using a Large Language Model and are flagged accordingly.

11.1.1 Proteins

List of proteins that are included in this entry displayed in a table. There is an option to display only proteins that have been manually curated in UniprotKB (**reviewed**), only proteins that have been automatically annotated (**unreviewed**), or all proteins (**both**, default).

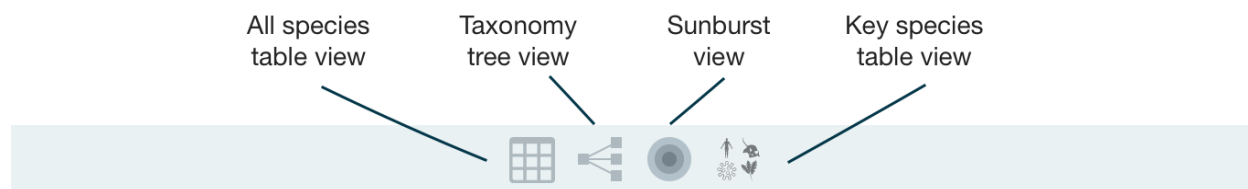
For each protein, the table displays the UniProt ID, name, corresponding gene, the organism where it is found, a link to the [AlphaFold structure prediction page](#) and a small protein viewer that highlights the region of the protein matched by the InterPro entry.

11.1.2 Domain architectures

Provides information about the different domains arrangements for the proteins matching this entry based on Pfam signatures. For InterPro entries, it provides information about where the domain is located in protein sequences and what, if any, combinations arise with other domains. Domain architectures can be downloaded in JSON and TSV formats through the **Export** button.

11.1.3 Taxonomy

List of species this entry is matching, based on data from [UniProt taxonomy](#). The information can be displayed in 4 different ways through the view options menu:



- Table with the list of all the species the proteins matching this entry are found in.
- Taxonomy tree of all the species the proteins matching this entry are found in.
- Sunburst view displays the taxonomy distribution of the proteins matching the entry, from the least specific at the centre to more specific going towards the outside.
- Table with the number of proteins found for key species, these are 12 model organisms commonly used in scientific research: *Oryza sativa subsp. japonica*, *Arabidopsis thaliana*, *Homo sapiens*, *Danio rerio*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Escherichia coli*, *Escherichia virus T4*, *Halobacterium salinarum*.

Sunburst is the default view of the subpage. A range of options can be selected to customise the view:

- The segment size can be adjusted based on the number of sequences matching a taxon (default) or by the number of species per taxon.
- The sunburst depth can be adjusted between 2 to 8 rings.

In the table views, for each organism, the taxonomy identifier and protein count information are provided. The ACTIONS column offers the possibility to:

- View all the protein matches in the [Proteins](#) tab
- Download a FASTA file of the protein matches
- View the taxonomy information in the [Taxonomy entry page](#)

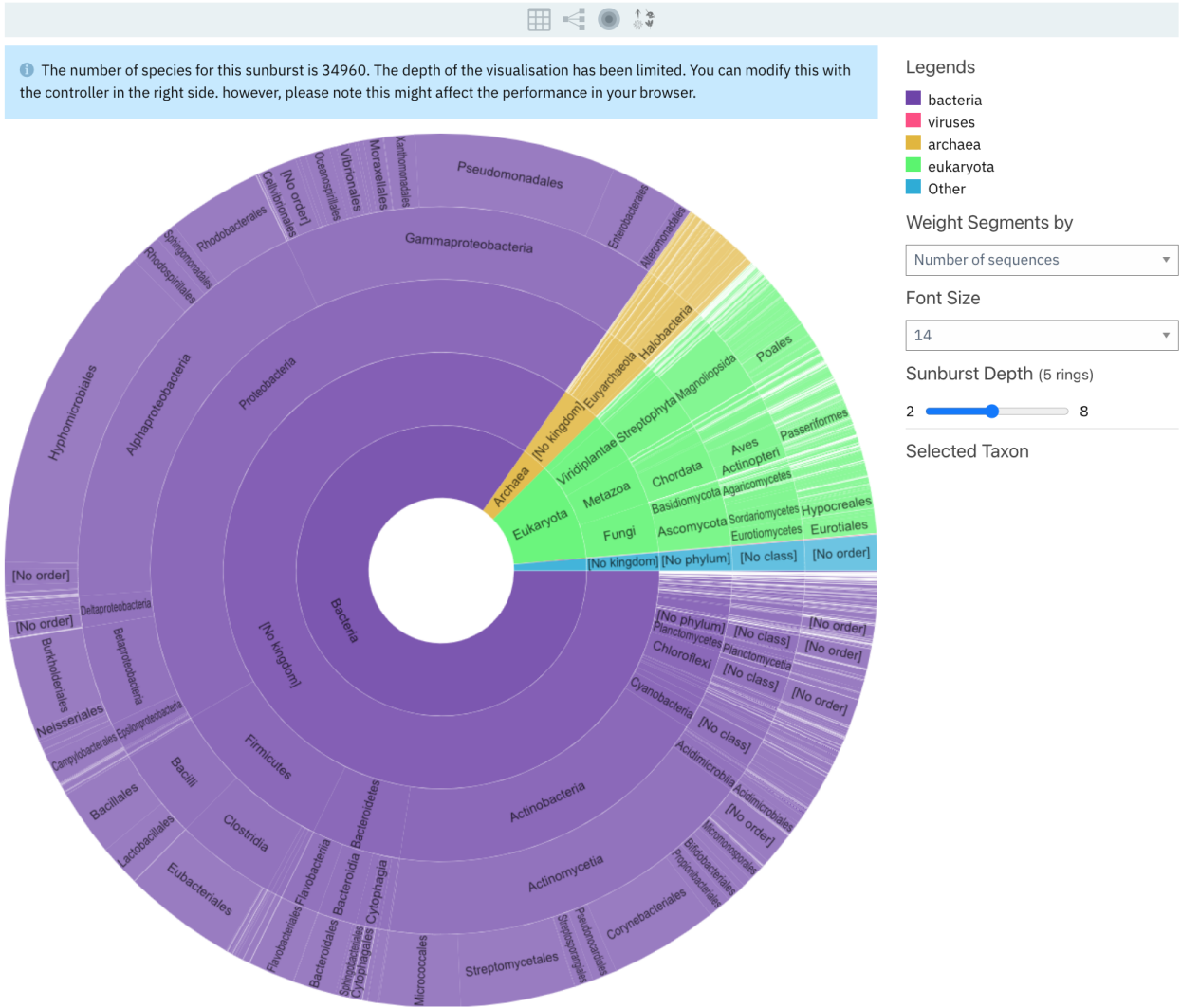


Fig. 3: Taxonomy sunburst view for PF00120

If the first option is selected, a table with all the corresponding proteins is displayed. For each protein, we can see the UniProt ID, name, corresponding gene, the organism where it is found, a link to the protein [AlphaFold structure prediction](#) and a small protein viewer that highlights the region of the protein matched by the InterPro entry.

11.1.4 Proteomes

List of proteomes whose members are represented by proteins matching this entry. A proteome represents a set of proteins whose genomes have been fully sequenced. A given taxonomy node may have one or more proteomes, for example, to reflect different assemblies of a genome. Proteome data is imported from [UniProt proteomes](#). For each proteome, the same set of actions are available than the ones in [Taxonomy](#), the taxonomy information being replaced by proteome information in the [Proteome entry page](#).

11.1.5 Structures

List of structures from the [PDB](#) database that match to protein sequences included in this entry.

11.1.6 AlphaFold

AlphaFold protein structure predictions are generated by [DeepMind](#) [4].

At the top of the page a 3D viewer (powered by [Mol*](#)) shows an interactive view of the predicted structure for one of the proteins matching the InterPro entry. The structure is coloured by per-residue pLDDT score, it can be zoomed in and out, and rotated. Clicking on a residue induces a zoom in effect and displays contacts with surrounding residues, clicking on the blank area around the structure zooms out.

The protein accession and organism are displayed on the left hand side, together with links to the corresponding [AlphaFold](#) and [UniProt](#) websites. The model confidence colour scale, determined using the pLDDT score, is also displayed, varying from dark blue (very high confidence) to orange (very low confidence).

The data can be downloaded in PDB or mmCIF format, by clicking on the corresponding buttons below the 3D viewer.

On an InterPro entry page, below the 3D viewer, a table containing the list of UniProt accessions matching the InterPro entry for which structure predictions have been generated is shown. For each protein it is possible to:

- Access the [Protein entry page](#) by clicking on the UniProt accession or name
- Access the [Taxonomy entry page](#) by clicking on the species
- Display the structure prediction on the current page by clicking on the **Show prediction** button

On a protein entry page, below the 3D viewer, the [protein sequence viewer](#) displays the member database signatures and InterPro entries matching the protein. Hovering over a match highlights the corresponding section in the predicted structure 3D view.

11.1.7 Pathways

List of pathways identified for protein sequences included in this entry. This information is provided by the [MetaCyc Metabolic Pathway Database](#) and the [Reactome database](#).

🏠 / Browse / By Entry / InterPro / IPR000562 / AlphaFold

D

IPR000562

Fibronectin type II domain ★

InterPro entry

Overview

Proteins 16k

Domain Architectures 792

Taxonomy 3k

Proteomes 735

Structures 38

AlphaFold 5k

Pathways 156

AlphaFold structure predictions

The protein structure below has been predicted by DeepMind with AlphaFold (Jumper, J et al. 2021). For more information and additional features, please visit this sequence's page at [AlphaFold DB](#).

This entry matches several proteins with structure predictions. Use the table below the structure viewer to select another protein.

Information

Protein
O60449
[View on AlphaFold DB](#) or [UniProtKB](#)

Organism
Homo sapiens

Model confidence

- Very High (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very Low (pLDDT < 50)

[PDB file](#)
[mmCIF file](#)
[Refresh](#)
[Fullscreen](#)
[Close](#)

Fig. 4: AlphaFold structure predictions tab for IPR000562, UniProt O60449.

11.1.8 Interactions

List of proteins characterised in experimentally proven data in which the proteins matching an entry are involved in protein:protein interactions.

11.2 Member database page

InterPro provides entry pages for each signature that a member database holds. This includes signatures that have not yet been, or can't be, integrated into InterPro (*unintegrated signatures*).

Member database signature entries provide information about which database the signature is from, the signature identifier, the type of entry as defined by the member database (e.g. family, domain or site), and the short name given to the entry by the member database.


Some member databases provide a description giving information about the family/domain or site function, when this is not the case and the signature is integrated in an InterPro entry, the InterPro description is displayed.

To address the absence of annotations for certain member database signatures that are not integrated into any InterPro entry, we've employed AI to automatically generate descriptions by extracting information from Swiss-Prot. It's important to note that these descriptions have not undergone curator review, and we advise regarding them as preliminary sources of information. Read more on AI-generated descriptions.

Some member databases create groups of families that are evolutionary related. Pfam calls them clans, CDD uses the term superfamily and, for PIRSF and Panther the concept is associated with the parent families of their hierarchy. We use the umbrella term Clan to refer to Pfam groups and Set to refer to the other groups. When available, the set/clan to which the signature belongs to is indicated.

The right hand side of the page provides links to the InterPro entry in which this signature has been integrated, and an external link to the signature on the member database's website when available. At the bottom of this column, if any experimentally solved structure is available, a **Representative structure** shows a small static 3D representation, the

[Home](#) / [Browse](#) / [By Entry](#) / [Ncbifam](#) / [NF012196](#) / [Overview](#)



NF012196 **Ig-like domain**

NCBIfam entry

Overview	Member database
Proteins 2k	NCBIfam (includes TIGRFAMs)
Taxonomy 1k	NCBIfam type domain
Proteomes 187	Short name <i>Ig_like_ice</i>
Structures 4	
Signature	
AlphaFold 720	

Description

This variant form of the Ig-like domain occurs as a repeat in a number of large adhesins, including a 1.5-MDa ice-binding adhesin, the *Marinomonas primoryensis* antifreeze protein.


Integrated to

[> IPR049826](#)

External Links

[View NF012196 in NCBIfam](#)

Representative structure



4p99: Ca²⁺-stabilized adhesin helps an Antarctic bacterium reach out and bind ice

Fig. 5: InterPro member database page for NCBIfam signature [NF012196](#).

corresponding PDB ID and name and a link to the [structure entry page](#). For Pfam signatures, the **Add your annotation** button allows the user to suggest updates to the Pfam annotation.

For signatures provided by the Pfam member database, a short extract of the wikipedia page is also displayed when available to complete the description.

In addition to the [Proteins](#), [Taxonomy](#), [Proteomes](#) and [Structures](#) tabs, member database pages may also display information in the following additional tabs: [Domain architectures](#), [AlphaFold](#), [Signature](#), [Alignment](#) and [Curation](#).

11.2.1 Signature

The signature representing the model that defines the entry is visualised in this page as a logo, using [Skyline](#). The logo data is displayed for the NCBIfam, Pfam, PANTHER, PIRSF, and SFLD member databases.

The visualisation displays the amino acid conservation for each residue in the model. To navigate large logos, the user can drag the rendered area to a desired position. Alternatively, the user can input a residue number to be viewed. When selecting a particular residue in the logo, the probabilities of each amino acid are displayed in the bottom part.

11.2.2 Alignment

This section allows users to view and download any available alignment file that is associated with the current member database signature. Currently, the alignment files are only available for the Pfam member database, but hopefully we will be able to include alignments for other member databases in the future.

First, one of the available alignments has to be selected. For example in the image below the user has selected the “seed” alignment. If the selected alignment has more than 1000 sequences, a warning message appears to inform users that big alignments can cause memory issues in the browser. A compressed file (gzip) of the current alignment is available by clicking on the **Download** button.

Interacting with the grey navigation bar over the sequences allows users to navigate the alignment; dragging the left and right limits of the navigation bar allows users to zoom to a particular position or adjust the zoom level. Alternatively,

🏠 / Browse / By Entry / Cathgene3d / G3DSA:1.10.10.10 / Overview

CATH

G3DSA:1.10.10.10

Winged helix-like DNA-binding domain superfamily/Winged helix DNA-binding domain

CATH-Gene3D entry

Overview	
Proteins	6M
Taxonomy	66k
Proteomes	14k
Structures	4k
Signature	
AlphaFold	5M
Subfamilies	2k

Member database [CATH-Gene3D](#)

CATH-Gene3D type homologous superfamily

Description Imported from IPR036388

Winged helix DNA-binding proteins share a related winged helix-turn-helix DNA-binding motif, where the "wings", or loops, are small β -sheets. The winged helix motif consists of two wings (W1, W2), three α -helices (H1, H2, H3) and three β -sheets (S1, S2, S3) arranged in the order H1-S1-H2-H3-S2-W1-S3-W2 [3]. The DNA-recognition helix makes sequence-specific DNA contacts with the major groove of DNA, while the wings make different DNA contacts, often with the minor groove or the backbone of DNA. Several winged-helix proteins display an exposed patch of hydrophobic residues thought to mediate protein-protein interactions.

Many different proteins with diverse biological functions contain a winged helix DNA-binding domain, including transcriptional repressors such as biotin repressor, LexA repressor and the arginine repressor [8]; transcription factors such as the hepatocyte nuclear factor-3 proteins involved in cell differentiation, heat-shock transcription factor, and the general transcription factors TFIIE and TFIIIF [2, 5]; helicases such as RuvB that promotes branch migration at the Holliday junction, and CDC6 in the pre-replication complex [4, 1]; endonucleases such as FokI and TnsA [6]; histones; and Mu transposase, where the flexible wing of the enhancer-binding domain is essential for efficient transposition [7].

References Imported from IPR036388

- Structure and function of Cdc6/Cdc18: implications for origin recognition and checkpoint control. Liu J, Smith CL, DeRyckere D, DeAngelis K, Martin GS, Berger JM. *Mol. Cell* 6, 637-48, (2000). [View article](#) PMID: [11030343](#)
- Hepatocyte nuclear factor 3/fork head or "winged helix" proteins: a family of transcription factors of diverse biologic function. Lai E, Clark KL, Burley SK, Darnell JE Jr. *Proc. Natl. Acad. Sci. U.S.A.* 90, 10421-3, (1993). [View article](#) PMID: [8248124](#)
- Winged helix proteins. Gajiwala KS, Burley SK. *Curr. Opin. Struct. Biol.* 10, 110-6, (2000). [View article](#) PMID: [10679470](#)
- Crystal structure of the RuvA-RuvB complex: a structural basis for the Holliday junction migrating motor machinery. Yamada K, Miyata T, Tsuchiya D, Oyama T, Fujiwara Y, Ohnishi T, Iwasaki H, Shinagawa H, Ariyoshi M, Mayanagi K, Morikawa K. *Mol. Cell* 10, 671-81, (2002). [View article](#) PMID: [12408833](#)
- The wing in yeast heat shock transcription factor (HSF) DNA-binding domain is required for full activity. Cicero MP, Hubl ST, Harrison CJ, Littlefield O, Hardy JA, Nelson HC. *Nucleic Acids Res.* 29, 1715-23, (2001). [View article](#) PMID: [11292844](#)
- Structure of the multimodular endonuclease FokI bound to DNA. Wah DA, Hirsch JA, Dorner LF, Schildkraut I, Aggarwal AK. *Nature* 388, 97-100, (1997). [View article](#) PMID: [9214510](#)
- The wing of the enhancer-binding domain of Mu phage transposase is flexible and is essential for efficient transposition. Clubb RT, Mizuuchi M, Huth JR, Omichinski JG, Savilahti H, Mizuuchi K, Clore GM, Gronenborn AM. *Proc. Natl. Acad. Sci. U.S.A.* 93, 1146-50, (1996). [View article](#) PMID: [8577730](#)
- Escherichia coli biotin holoenzyme synthetase/bio repressor crystal structure delineates the biotin- and DNA-binding domains. Wilson KP, Shewchuk LM, Brennan RG, Otsuka AJ, Matthews BW. *Proc. Natl. Acad. Sci. U.S.A.* 89, 9257-61, (1992). [View article](#) PMID: [1409631](#)

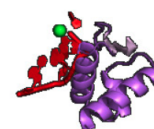
Integrated to

> [IPR036388](#)

External Links

[View G3DSA:1.10.10.10 in CATH-Gene3D](#)

Representative structure



4kmf: Crystal structure of Zalpha domain from *Carassius auratus* PKZ in complex with Z-DNA

Fig. 6: InterPro member database page for CATH-Gene3D signature G3DSA:1.10.10.10.

[Home](#) / [Browse](#) / [By Entry](#) / [Panther](#) / [PTHR13944](#) / [Overview](#)



PTHR13944

Rho guanine nucleotide exchange factor AI

PANTHER entry

Overview

Proteins	8k
Taxonomy	4k
Proteomes	919
Structures	65
Signature	
AlphaFold	3k
Subfamilies	5

This entry contains information that has been generated using an AI language model. Please exercise discretion when interpreting the information provided.

[Read more on description generation](#) [Provide feedback](#)

Member database [PANTHER](#)

PANTHER type family

Short name *Rho_GEF* AI

Description

AI-generated Unreviewed

This family of proteins includes Rho guanine nucleotide exchange factors (GEFs) that activate Rho-GTPases by facilitating the exchange of GDP for GTP. They play a pivotal role in various cellular processes such as cell motility, polarization, cytoskeletal dynamics, and signal transduction pathways. These proteins are involved in focal adhesion formation, axonal branching, synapse formation, dendritic morphogenesis, and B-lymphocyte activation. Some members are implicated in innate immune responses, acting as signaling intermediaries and participating in cytokine secretion. They also contribute to cell cycle regulation, apoptosis, and cancer progression. Additionally, they are involved in the development and differentiation of neuronal progenitor cells and the migration of precerebellar neurons. Certain members function as scaffold proteins, assembling signaling complexes downstream of G protein-coupled receptors, and are involved in cardiac hypertrophy and osteogenesis.

Integrated to

[> IPR051632](#)

External Links

[View PTHR13944 in PANTHER](#)

Representative structure



7g82: ARHGEF2 PanDDA analysis group deposition -- ARHGEF2 and RhoA in complex with Z1079512010

Fig. 7: InterPro member database page for PANTHER signature [PTHR13944](#). AI-generated content is accordingly flagged with an AI tag.

the zoom level can also be defined by scrolling up/down while holding the [ctrl] key. Scrolling up/down allows to move other sequences in the alignment into the visible area of the viewer.

11.2.3 Curation

This section provides information about the curation of the signature. Currently, it is only available for the Pfam member database. It is divided into 2 subsections:

- **Curation:** details about Pfam curators and Sequence ontology
- **HMM information:** displays the HMM building command used and offers the possibility to download the HMM profile defining the signature

Pfam PF00040 Fibronectin type II domain

Pfam entry [PF00001](#)

Overview	
Proteins	16k
Domain Architectures	792
Taxonomy	3k
Proteomes	721
Structures	38
Signature	
AlphaFold	5k
Alignment	
Curation	

Member database	Pfam ¹
Pfam type	domain
Short name	<i>fn2</i>
Clan	Kringle

Description [†]Imported from [IPR000562](#)

Fibronectin is a multi-domain glycoprotein, found in a soluble form in plasma, and in an insoluble form in loose connective tissue and basement membranes, that binds cell surfaces and various compounds including collagen, fibrin, heparin, DNA, and actin. Fibronectins are involved in a number of important functions e.g., wound healing; cell adhesion; blood coagulation; cell differentiation and migration; maintenance of the cellular cytoskeleton; and tumour metastasis [3].

The major part of the sequence of fibronectin consists of the repetition of three types of domains, which are called type I, II, and III (also called FN1, FN2 and FN3 respectively)^[1]. In fibronectin the type II domain is duplicated. Type II domain is approximately forty residues long, contains four conserved cysteines involved in disulphide bonds and is part of the collagen-binding region of fibronectin. Type II domains have also been found in a range of proteins including blood coagulation factor XII; bovine seminal plasma proteins PDC-109 (BSP-A1/A2) and BSP-A3^[4]; cation-independent mannose-6-phosphate receptor^[6]; mannose receptor of macrophages^[8]; 180 kD secretory phospholipase A2 receptor^[5]; DEC-205 receptor^[2]; 72 kD and 92 kD type IV collagenase (3.4.24.24)^[9]; and hepatocyte growth factor receptor^[6]; mannose receptor of macrophages^[8]; 180 kD secretory phospholipase A2 receptor^[5]. DEC-205 receptor^[2]; 72 kD and 92 kD type IV collagenase (3.4.24.24)^[9]; and hepatocyte growth factor activator^[7].

References

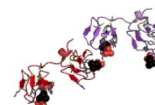
1. Complete primary structure of bovine plasma fibronectin. Skorstengaard K, Jensen MS, Sahl P, Petersen TE, Magnusson S. *Eur. J. Biochem.* 161, 441-53, (1986). [View article](#) PMID: 3780752
2. The receptor DEC-205 expressed by dendritic cells and thymic epithelial cells is involved in antigen processing. Jiang W, Swiggard WJ, Heuffer C, Peng M, Mirza A, Steinman RM, Nussenzweig MC. *Nature* 375, 151-5, (1995). [View article](#) PMID: 7753172
3. Cloning and analysis of the promotor region of the human fibronectin gene. Dean DC, Bowlin CL, Bourgeois S. *Proc. Natl. Acad. Sci. U.S.A.* 84, 1876-80, (1987). [View article](#) PMID: 3031656
4. Complete amino acid sequence of BSP-A3 from bovine seminal plasma. Homology to PDC-109 and to the collagen-binding domain of fibronectin. Seidah NG, Manjunath P, Rochement J, Sairam MR, Chretien M. *Biochem. J.* 243, 195-203, (1987). [View article](#) PMID: 3606570
5. Cloning and expression of a human precursor for secretory phospholipases A2. Lambeau G, Ancian P, Barhanin J, Lazdunski M. *J. Biol. Chem.* 269, 1575-8, (1994). [View article](#) PMID: 8294398
6. Structure and function of the mannose 6-phosphate/inositoline growth factor II receptors. Kornfeld S. *Annu. Rev. Biochem.* 61, 307-30, (1992). [View article](#) PMID: 1323236
7. Molecular cloning and sequence analysis of the cDNA for a human serine protease responsible for activation of hepatocyte growth factor: Structural similarity of the protease precursor to blood coagulation factor III. Miyazawa K, Shimomura T, Kitamura A, Kondo J, Morimoto Y, Kitamura N. *J. Biol. Chem.* 268, 10024-8, (1993). [View article](#) PMID: 7683665
8. Primary structure of the mannose receptor contains multiple motifs resembling carbohydrate-recognition domains. Taylor ME, Conary JT, Lennartz MR, Stahl PD, Drickamer K. *J. Biol. Chem.* 265, 12156-62, (1990). [View article](#) PMID: 2373685
9. H-ras oncogene-transformed human bronchial epithelial cells (TBE-1) secrete a single metalloprotease capable of degrading basement membrane collagen. Collier IE, Wilhelm SM, Eisen AZ, Marmor BL, Grant GA, Seltzer JL, Kronberger A, He CS, Bauer EA, Goldberg GL. *J. Biol. Chem.* 263, 6579-87, (1988). [View article](#) PMID: 2834383

 Add your annotation

Integrated to

> IPR000562

Representative structure

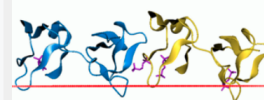


1h8p: Bull seminal plasma
PDC-109 fibronectin type II
module

Fibronectin type II domain [Wikipedia](#)

Fibronectin type II domain is a collagen-binding protein domain. Fibronectin is a multi-domain glycoprotein, found in a soluble form in plasma, and in an insoluble form in loose connective tissue and basement membranes, that binds cell surfaces and various compounds including collagen, fibrin, heparin, DNA, and actin. Fibronectins are involved in a number of important functions e.g., wound healing; cell adhesion; blood coagulation; cell differentiation and migration; maintenance of the cellular cytoskeleton; and tumour metastasis. The major part of the sequence of fibronectin consists of the repetition of three types of domains, which are called type I, II, and III.

Fibronectin type II domain



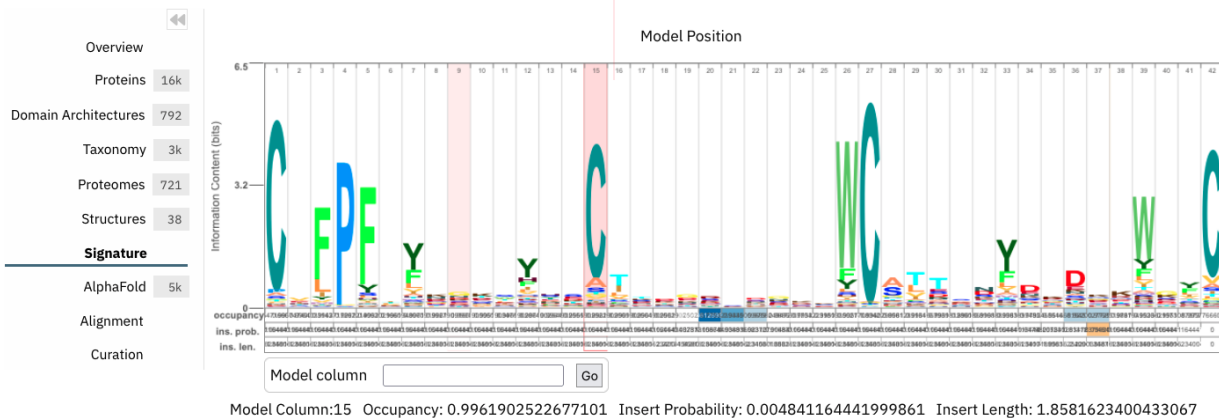
Collagen-binding type II domain of seminal plasma protein PDC-109.

Identifiers	
Symbol	fn2
Pfam	PF00040
InterPro	IPR000562
SMART	SM00059
PROSITE	PD0C00022
SCOP	1pdc
OPM family	
OPM protein	1h8p
CDD	cd00062

Fig. 8: InterPro member database page for Pfam signature PF00040.

[Home](#) / [Browse](#) / [By Entry](#) / [Pfam](#) / [PF00040](#) / [Logo](#)

Pfam PF00040 Fibronectin type II domain

[Pfam entry](#)[Home](#) / [Browse](#) / [By Entry](#) / [Pfam](#) / [PF00040](#) / [Entry Alignments](#)

Pfam PF00040 Fibronectin type II domain

[Pfam entry](#)

[Home](#) / [Browse](#) / [By Entry](#) / [Pfam](#) / [PF00040](#) / [Curation](#)

Pfam PF00040 Fibronectin type II domain

Overview

Proteins 16k

Domain Architectures 792

Taxonomy 3k

Proteomes 721

Structures 38

Signature

AlphaFold 5k

Alignment

Curation

Curation

Author [Sonnhammer ELL](#)

Sequence Ontology [SO:0000417](#)

HMM Information

HMM build commands
Build method: `hmmbuild -o /dev/null HMM SEED`
Search method: `hmmsearch -Z 75585367 -E 1000 --cpu 4 HMM pfamseq`

Gathering threshold
Sequence: 22.2
Domain: 22.2

Download
[Download](#) the raw HMM for this family

11.2.4 Subfamilies


This section provides a list of subfamilies derived from the signature and a link to get more information in the member database website. Currently, this list is available for the PANTHER and CATH-Gene3D member databases. For PANTHER subfamilies, the GO terms associated to them are also displayed.

11.3 Protein entry page

The Protein entry page contains information on a specific protein provided by [UniProt](#). Protein pages can be accessed either by entering a UniProt accession or identifier in a [Text search](#) or by clicking on a protein accession from the [Proteins](#) tab in an entry page.

The protein page provides the protein accession, the short name (identifier) given to the protein by Uniprot, the length of the protein sequence, species in which the protein is found, the proteome it belongs to, the gene encoding for the protein and a brief description of the protein's function where known. All the [InterPro family entries](#) this protein is matching are listed under “**Protein family membership**”. An external link to the protein entry in [Uniprot](#), as well as the export of the matches in TSV format and the possibility to perform a [HMMER search](#) or an [InterProScan search](#) are provided on the right hand side of the page.

The protein entry page also displays the [protein sequence viewer](#) to show the associated domains, sites etc.

When available, different isoforms of the protein can be selected to compare their InterPro matches with the consensus protein sequence. When an isoform is selected, a new [protein sequence viewer](#) corresponding to the selection is displayed and the url is update to reflect the change. The isoform matches can also be viewed side by side with the consensus protein sequence by clicking on the split icon  after selecting an isoform.

When available, GO terms associated to InterPro entries and PANTHER families are displayed at the bottom of the page. GO terms provide information about Biological processes, Molecular function and Cellular components.

The following tabs may be available: [Entries](#), [Structures](#), [Sequence](#), [Similar proteins](#) and [AlphaFold](#).

Home

Browse

By Protein

UniProt

O00167

Overview

O00167

Eyes absent homolog 2

UniProtKB/Swiss-Prot protein

Overview

Entries5

Structures7

AlphaFold1

Sequence

Similar Proteins214k

Short name

EYA2_HUMAN

Length

538 amino acids

Species

Homo sapiens (Human)

Proteome

UP000005640

Gene

EYA2

Function

Functions both as protein phosphatase and as transcriptional coactivator for SIX1, and probably also for SIX2, SIX4 and SIX5 (PubMed:12500905, PubMed:23435380). Tyrosine phosphatase that dephosphorylates 'Tyr-142' of histone H2AX (H2AXY142ph) and pro...

Show More

Family membership

Eyes absent family (IPR028472)

Isoforms

Select an isoform

External Links

UniProt

Search sequence with InterProScan

Download matches (TSV)

Download sequence (FASTA)

Fig. 9: Protein entry page for O00167.

11.3.1 Entries

List of InterPro entries that include this entity. The results can be filtered by member databases using the dropdown box located on the left side of the header of the result table. This functionality is available for all the tables presenting InterPro entries in the website.

Home

Browse

By Protein

UniProt

O00167

Entry

InterPro

O00167

Eyes absent homolog 2

UniProtKB/Swiss-Prot protein

Overview

Entries5

Structures7

AlphaFold1

Sequence

Similar Proteins214k

This protein matches these entries:

1 - 5 of 5 entries matching InterPro

Search

Export

Select your database:

InterPro5

NCBIfam1

PROSITE profiles0

AntiFam0

PANTHER1

PROSITE patterns0

CATH-Gene3D1

Pfam1

SFLD2

CDD1

PIRSF0

SMART0

HAMAP0

PRINTS0

SUPERFAMILY1

IPR038102

EYA domain superfamily

InterPro

200400

IPR042577

EYA domain, metazoan

InterPro

200400

Show20results

Previous1Next

56

Chapter 11. Browsing entries in the InterPro website

11.3.2 Sequence

This tab shows the protein FASTA sequence. The full sequence or part of the sequence (by selecting the region of interest) can be used to perform two types of search, available on the right side of the screen: [InterProScan search](#) or [HMMER search](#), which redirects to the corresponding pages.

11.3.3 Similar proteins

List of proteins that have the same domain architecture as this protein, including the Pfam/InterPro accession for each domain. The list can be filtered to either show all the protein matches or only the reviewed proteins from [UniProt](#). For each protein the UniProt ID, name, length, corresponding gene, the organism where it is found and a link to the protein [AlphaFold structure prediction page](#).

11.4 Structure entry page

InterPro provides entries for all the structures available in the [Protein Data Bank in Europe \(PDBe\)](#). A structure search can be performed by clicking on a structure provided in a results list or by entering the protein structure identifier in the [Quick search](#) box (magnifying glass symbol) or by performing a [Text search](#).

At the top of the structure page, general information about the structure is displayed: the structure's accession number (PDB ID), resolution, release date, the method used to determine the structure (e.g. "Xray") and the chains composing the structure. External links to [PDBe](#), [RCSB PDB](#), [PDBsum](#), [CATH](#), [SCOP](#), [ECOD](#) and [Proteopedia](#) are provided on the right hand side of the page.

Following, the general information section, a 3D viewer (powered by [Mol*](#)) shows an interactive view of the 3D structure. Hovering over a residue displays the name of the entry, the chain and residue information below the viewer. Clicking on a residue in the viewer induces a zoom in effect and displays contacts with surrounding residues, clicking on the blank area around the structure zooms out. Below it, the [protein sequence viewer](#) with the InterPro matches is displayed for each chain. It has an extra category representing the secondary structure information. Hovering over one of the tracks highlights the corresponding region of the protein structure in the 3D structure viewer.

More information is available on the corresponding [train online section](#).

The following tabs may be available: [Entries](#) and [Proteins](#).

11.5 Taxonomy entry page

Taxonomy pages display the name, taxonomy ID, lineage and children nodes for a particular taxon. Any reference to this taxon from another page throughout the website will link to this page.

The overview also includes a graphical representation of the lineage of the selected taxon. The nodes in the visualisation are also links, so you can jump to the page of a particular taxon of interest.

The following tabs may be available: [Entries](#), [Proteins](#), [Structures](#) and [Proteomes](#).

[Home](#) / [Browse](#) / [By Structure](#) / [PDB](#) / [1t2v](#) / [Overview](#)

1t2v Structural basis of phospho-peptide recognition by the BRCT domain of BRCA1, structure with phosphopeptide

pdb structure

OverviewEntries 4
Proteins 1

Accession	1t2v
Experiment type	X-Ray
Resolution	3.3 Å
Chains	A, B, C, D, E, F, G, H, I, J
Released	11 May 2004

External Links

[PDBe](#)
[RCSB PDB](#)
[PDBsum](#)
[CATH](#)
[SCOP](#)
[ECOD](#)
[Proteopedia](#)

Highlight Entry in the 3D structure

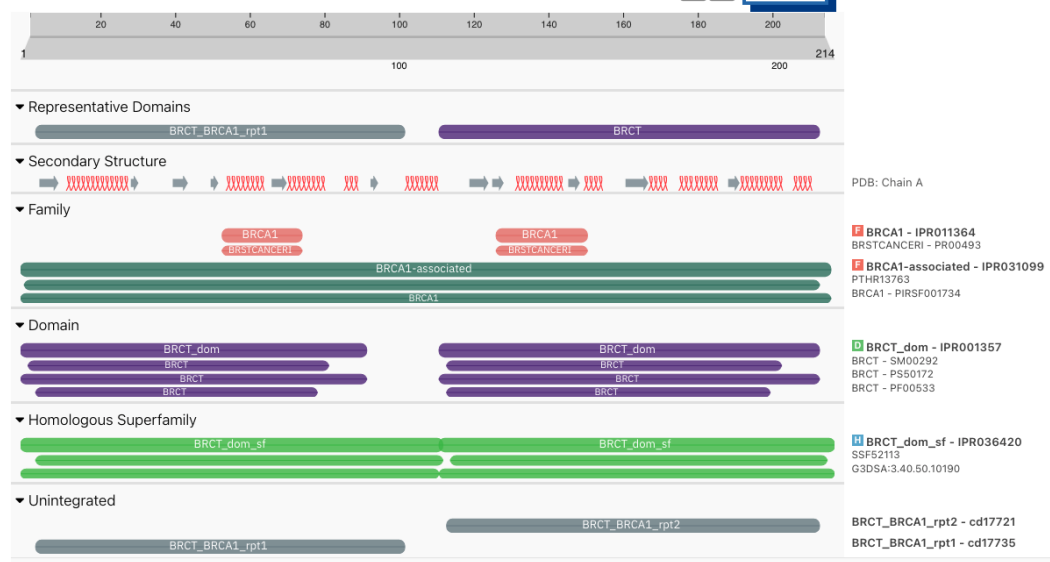
[PDB file](#) [mmCIF file](#) [Play](#) [Refresh](#) [Full Screen](#) [Print](#)Chain A ([P38398](#))Domains in the chain⁰

Fig. 10: Structure entry page for 1t2v.

Home / Browse / By Taxonomy / Uniprot / 6239 / Overview

Caenorhabditis elegans

UniProtKB taxonomy

Overview	
Entries	10k
Proteins	27k
Structures	469
Proteomes	1

Taxon ID	6239
Rank	Species
Lineage	root > cellular organisms > Eukaryota (eucaryotes) > Opisthokonta > Metazoa (metazoans) > Eumetazoa > Bilateria > Protostomia > Ecdysozoa > Nematoda (roundworms) > Chromadorea > Rhabditida > Rhabditina > Rhabditomorpha > Rhabditoidea > Rhabditidae > Peloderinae > Caenorhabditis > Caenorhabditis elegans
Children	None

External Links

UniProt

Entry Database: All

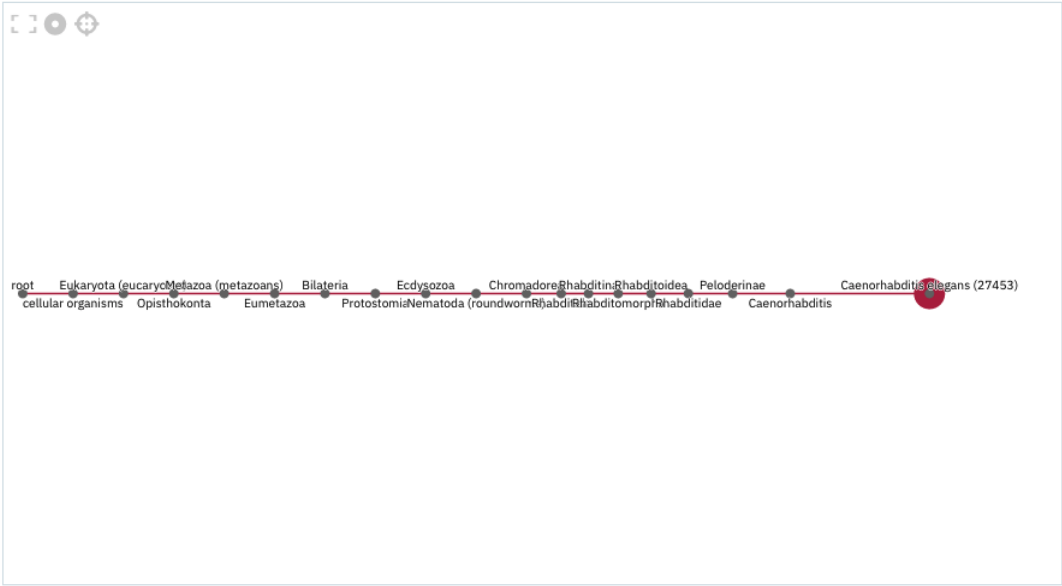


Fig. 11: Taxonomy entry page for *Caenorhabditis elegans*.

11.6 Proteome entry page

The proteome entry page displays general information provided by UniProt: its ID, strain, and a description of the organism. It also provides a link to the corresponding *taxonomy page*.

On the right-hand side, external links to the proteome page in UniProt and the genome page in Rfam are provided, when available.

The following tabs may be available: *Entries*, *Proteins* and *Structures*.

🏠 / Browse / By Proteome / Uniprot / UP000001940 / Overview

UP000001940
Caenorhabditis elegans
UniProtKB proteome

Overview

Entries 10k

Proteins 27k

Structures 469

Proteome ID	UP000001940
Strain	AB1
Taxonomy	Caenorhabditis elegans

External Links

[UniProt](#)
[Rfam](#)

Description

C. elegans is a nematode - a member of the phylum Nematoda which incorporates roundworms and threadworms, a phylum of smooth-skinned, unsegmented worms with a long cylindrical body shape tapered at the ends, including free-living and parasitic forms both aquatic and terrestrial. C. elegans is small, growing to about 1 mm in length, and lives in the soil - especially rotting vegetation - in many parts of the world, where it survives by feeding on microbes such as bacteria. It is of no economic importance to man. C. elegans is about as primitive an organism that exists however it shares many of the essential biological characteristics that are central problems of human biology. The worm is conceived as a single cell which undergoes a complex process of development, starting with embryonic cleavage, proceeding through morphogenesis and growth to the adult. It has a nervous system with a 'brain' (the circumpharyngeal nerve ring). It exhibits behavior and is even capable of rudimentary learning. It produces sperm and eggs, mates and reproduces. After reproduction it gradually ages, loses vigour and finally dies. Embryogenesis, morphogenesis, development, nerve function, behaviour and aging, and how they are determined by genes are some of the most fundamental mysteries of modern biology. C. elegans exhibits these phenomena, yet is only 1 mm long and may be handled as a microorganism - it is usually grown on petri plates seeded with bacteria. All 959 somatic cells of its transparent body are visible with a microscope, and its average life span is a mere 2-3 weeks. Thus C. elegans provides researchers with the ideal compromise between complexity and tractability. There are two sexes, a self-fertilizing hermaphrodite and a male. The adult essentially comprises a tube, the exterior cuticle, containing two smaller tubes, the pharynx and gut, and the reproductive system. Most of the volume of the animal is taken up by the reproductive system. Of the 959 somatic cells of the hermaphrodite some 300 are neurons. Neural structures include a battery of sense organs in the head which mediate responses to taste, smell, temperature and touch and although C. elegans has no eyes, it might respond slightly to light. Among other neural structures is an anterior nerve ring with a ventral nerve cord running back down the body. (There is also a smaller dorsal nerve cord.) There are 81 muscle cells. C. elegans moves by means of four longitudinal bands of muscle paired sub-dorsally and sub-ventrally. Alternative flexing and relaxation generates dorsal-ventral waves along the body, propelling the animal along. The development and function of this diploid organism is encoded by an estimated 17,800 distinct genes.

Fig. 12: Proteome entry page for UP000001940.

When clicking on the **Entries** tab, the list of InterPro entries matching any sequence in the proteome is displayed. By clicking on the dropdown menu in the table header, the list of entries from a member database can be displayed instead by selecting the database of interest.

11.7 Set/Clan entry page

Some *InterPro member databases* create groups of families that are evolutionary related, called sets/clans. This page offers an overview of a specific set/clan provided by a member database, it includes a short description and an interactive view of the signatures included in the set/clan. For the interactive view, different label types can be chosen through the **Label Content** menu: Accession, Name and Short name. For clans provided by the Pfam member database, an additional section provides literature references, when available.

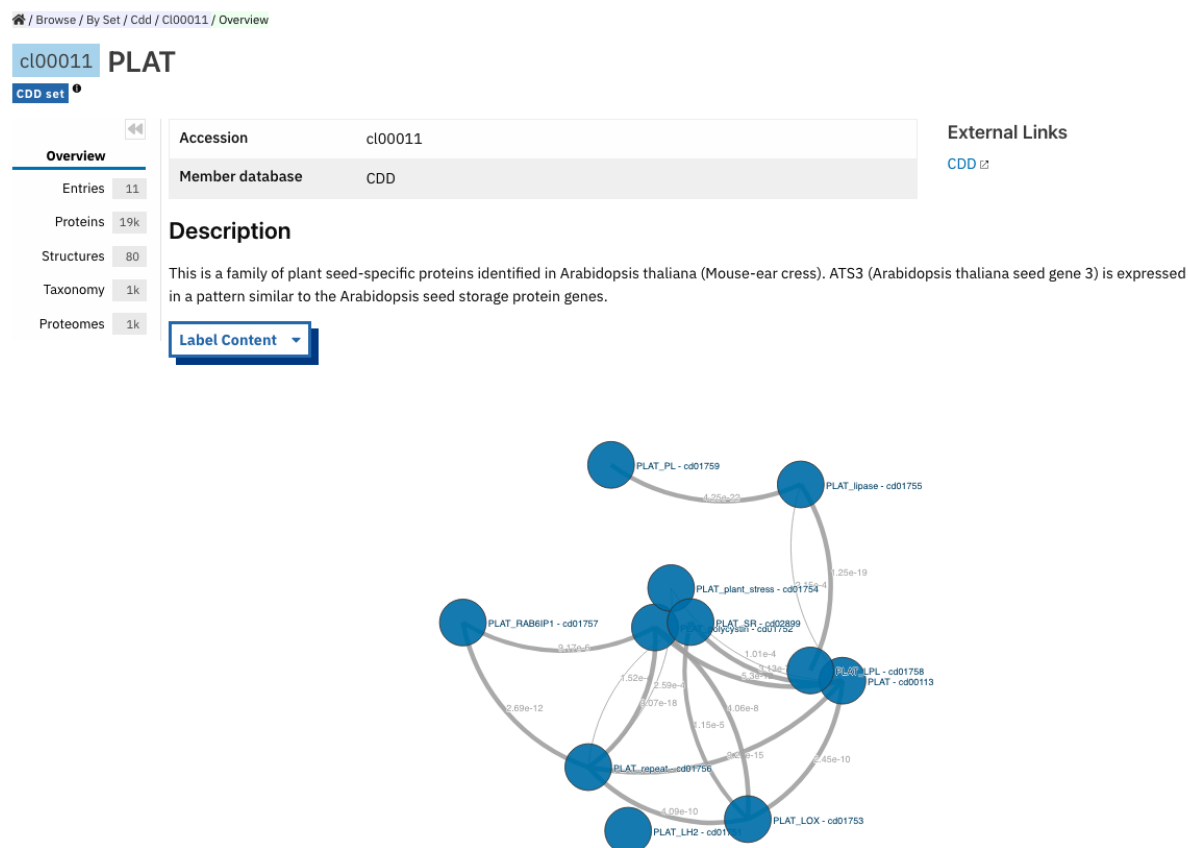


Fig. 13: Set entry page for cI00011 (CDD)

The following tabs may be available: *Entries*, *Proteins*, *Structures*, *Taxonomy*, *Proteomes* and *alignment_clan*.

11.7.1 Entries

Provides the list of signatures included in the set/clan (accession, name and short name).

For Pfam clans, the Entries tab contains the list of Pfam entries included in the clan and links to the entries SEED alignment and domain architectures pages.

HOW TO DOWNLOAD INTERPRO DATA?

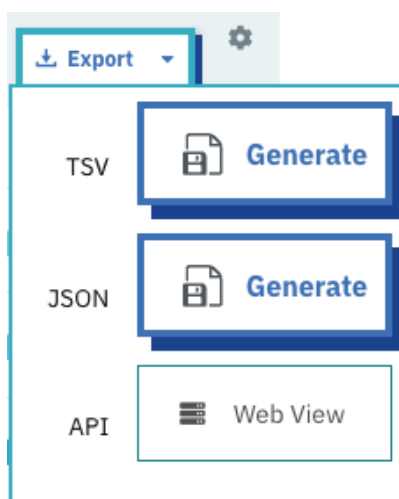
InterPro data and search tools are freely available for download. We provide bulk downloads, data exports on each relevant InterPro page and an API to allow easy access for user scripts.

12.1 Download page

This is available under the [Download](#) section in the [navigation menu](#). This page is divided into multiple tabs.

- The **InterPro** tab provides various files containing pre-calculated InterPro data for the current release that can be downloaded. Data from previous releases are available in the [InterPro ftp](#).
- The **InterProScan** tab provides a downloadable version of the [InterProScan software](#).
- The **Pfam** tab gives access various Pfam files. Data from previous Pfam releases are available in the [Pfam ftp](#).
- The **PRINTS** and **SFLD** tabs give access to the latest PRINTS and SFLD release files, available from the InterPro ftp.
- The **AntiFam** tab give access to the latest AntiFam release files.

12.2 Export button



The export button, found on various entry pages in InterPro, is located next to the *text filter* at the top of result tables. It allows data to be downloaded as JSON or Tab Separated Values (TSV). The data sent from the [InterPro Application Programming Interface \(API\)](#) to populate the table can also be viewed using this component. When the file to generate is too big (bigger than 10K entities) we recommend to use a script to get the information from the API. See [Your downloads](#) for more information on how to generate a script.

12.3 Your downloads

This page is accessible through the **Results** tab in the *navigation menu*, under “**Your downloads**” section.

The purpose of this page is to give the user a way to select and filter InterPro data. Filtered data can then be downloaded in different file formats (if the selection has less than 10K entities), using the provided API call or through a script generated automatically.

Select data

Main data type

Choose a main data type:

Protein

protein DB:

UniProtKB/Swiss-Prot

X

Add a modifier:

Please Choose...

Filters

Add a filter:

Structure

Taxonomy

Proteome

Set

entry

filter type:

entry

X

entry DB:

Interpro

X

entry accession:

IPR000001

X

For Example, the image above shows **Protein** as the main data type selected and it will only select proteins included in the database **UniProtKB/Swiss-Prot**; this selection is then filtered by the selection of the endpoint **entry** with **InterPro** as the database and accession **IPR000001**. In other words this will generate the list of SwissProt proteins that are matching IPR000001 (also available under the Proteins tab in the InterPro entry page for [IPR000001](#), with the reviewed option selected). The results are stored in the browser (IndexedDB), allowing to retrieve previous searches.

12.3.1 Output formats

The following output formats are currently supported, if the number of entities selected is lower than 10K:

- **Text**: a list of accessions, 1 per line
- **FASTA**: a single file with multiple sequences in Fasta format (only available for proteins)
- **JSON**: it reuses the format returned by the InterPro API.
- **TSV**: reformats the JSON from the API to create a TSV file.

After selecting the output format, clicking on the **Download** button at the bottom of the page will start the downloading.

12.3.2 Programming scripts

The script can be generated in 4 different languages: Python 2, Python 3, JavaScript and Perl, it allows the download of the filtered data directly from the *InterPro API* and can be integrated in the users own program.

12.4 InterPro Application Programming Interface (API)

The InterPro API provides programmatic access to all the InterPro entries and their related entities in Json format. The API has six main endpoints, which corresponds to the *InterPro data types*: entry, protein, structure, taxonomy, proteome and set.

An API call is formed of one or multiple endpoint blocks. An endpoint block consists of a data type, a source database and an accession (e.g. `api/datatype/sourcedb/accession`).

For example the URL `/entry/interpro` provides a pageable list of all the InterPro entries. And the URL `/protein/uniprot/p99999` returns all the details of the protein identified with the UniProt accession P99999.

The combined URL `/entry/interpro/protein/uniprot/p99999` returns the list of all the InterPro entries that match in the P99999 protein accession.

For more information on how to use the InterPro API, you can watch [this recorded webinar](#) or have a look at the API documentation on our [GitHub repository](#).

RELEASE NOTES

InterPro is updated approximately every 8 weeks. The [release notes page](#) provides information about the current InterPro release.

13.1 General information

The section at the top of the page gives details about the release version and date together with changes made in this release.

13.2 Other statistics

A range of statistics covering member databases, GO annotation, information about Proteins, Structures, Proteomes, Taxonomy and Sets are also available on this page.

Release notes

[Show Previous Releases](#)

InterPro 77.0 • 14th November 2019

InterPro 77.0 • 14th November 2019

Features include:

- The addition of 111 InterPro entries.
- Integration of 145 new methods from the CATH-Gene3D (134), SUPERFAMILY (11) databases.














Contents and coverage




InterPro protein matches are now calculated for all UniProtKB and UniParc proteins. InterPro release 77.0 contains **37,213** entries (last entry: [IPR043126](#)), representing:

H	Homologous Superfamily	3,263
F	Family	22,021
D	Domain	10,699
R	Repeat	317
S	Site	
	:.. Active Site	132
	:.. Binding Site	76
	:.. Conserved Site	688
	:.. PTM	17

Interpro cites 59894 publications in PubMed.

Member database information

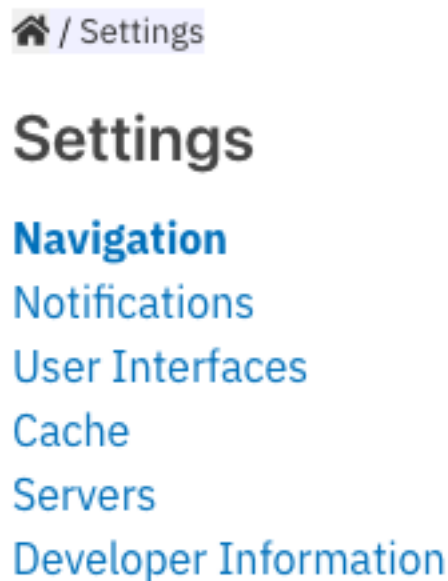
	SIGNATURE DATABASE	VERSION	SIGNATURES ⓘ	INTEGRATED SIGNATURES ⓘ
	CATH-Gene3D	4.2.0	6,119	2,581 (42.2%)
	HAMAP	2019_01	2,274	2,271 (99.9%)
	Pfam	32.0	17,929	17,407 (97.1%)
	PIRSF	3.02	3,285	3,216 (97.9%)
	PRINTS	42.0	2,106	1,950 (92.6%)
	PROSITE patterns	2019_01	1,310	1,286 (98.2%)
	PROSITE profiles	2019_01	1,232	1,173 (95.2%)
	PANTHER	14.1	123,151	9,403 (7.6%)
	SFLD	4	303	147 (48.5%)
	SMART	7.1	1,312	1,264 (96.3%)
	SUPERFAMILY	1.75	2,019	1,612 (79.8%)
	TIGRFAMs	15.0	4,488	4,431 (98.7%)
	CDD	3.17	14,908	3,180 (21.3%)

We use [MobiDB-lite](#) , a derivative of the [MobiDB](#)  database, to provide consensus annotation of long-range intrinsic disorder in protein sequences. Read more about MobiDB-lite in *Bioinformatics*, 33(9), 2017, 1402–1404, ([doi: 10.1093/bioinformatics/btx015](#) .

SETTINGS PAGE

On the [settings page](#), it is possible to select options that will persist beyond your current browsing session. Your choices are saved in the browser using a technology called IndexedDB. This allows you, for example, to choose 50 as the number of results to be included on the website data tables, and then this value will be remembered and the next time you visit our website, it will use 50 records for all the data tables.

All the settings are included on the same page and are organised in 7 sections. A menu on the left indicates which section you are currently displaying and allows you to jump directly to the one of your interest.



14.1 Navigation settings

- **Number of results per page:** Choose how many items should be listed in all the website data tables. The default is 20, we recommend using 20, 50, or 100 in order to get a more significant benefit from the server caching strategy.
- **Time (sec) to retry timed out queries:** Certain combinations of filters can create expensive, non-previously cached API queries. When the InterPro API times out, it keeps processing the query and once it is resolved, saves the result in the cache for future requests. This setting defines how long to wait before checking again if the result is ready, the default value is 10 seconds.

Number of results per page:



Time (sec) to retry timed out queries:



14.2 Notification settings

Allow us to send you browser notifications when one of your Jobs or Downloads finishes.

Enable notifications

There are few tips shown in the website on how to use features efficiently. It can be turned on/off.

Tip	Status
Taxonomy Tree navigation	<input checked="" type="checkbox"/>
Check connectivity	<input type="checkbox"/>
Use CTRL and scroll to zoom	<input checked="" type="checkbox"/>
Customise Settings	<input type="checkbox"/>
Help Links	<input type="checkbox"/>

- **Browser notifications:** this type of notifications are native to your browser, they allow the display of a notification outside the page. It is useful, for example, to let you know when an InterProScan search is completed. You can enable them by clicking the “Enable notifications” button. Unfortunately, we are not able to show a disable button because that change needs to be done from the setting of your browser directly.
- **Help tooltips:** in-page tooltip notifications try to make more visible parts or functions of the website that we think are not so obvious like this setting page, for example. These parameters allow you a granular selection of which tips will be enabled or disabled.

14.3 User interface settings

UI settings

Low graphics mode:

Recommended for low-end devices



Colour Domains:

Selection mode to colour by in the feature viewer

Label Content:

The content of the labels in the feature viewer

Label	Status
Accession	<input checked="" type="checkbox"/>
Name	<input type="checkbox"/>
Short Name	<input type="checkbox"/>

Display structure viewer all the time:

On some low-end devices, small screens, or under network or battery constraints, we might decide to not display the structure viewer by default. It will still be available on demand. Do you still want to display the viewer all the time?



- **Low graphics mode:** if you are visiting the InterPro website from a not too powerful device, you might benefit from selecting low graphics mode, which disables some animations and other visual effects that might cause poor performance on low-end devices.
- **Colour Domains:** defines the colouring strategy for the Protein sequence viewer. There are 3 options:
 - **Accession:** a unique colour for each accession in the graphic.
 - **Member database:** all entries of the same member database will have the same colour.
 - **Domain relationship:** InterPro entries will follow the accession strategy, but integrated signatures will be painted in the same colour as the linked InterPro entry.
- **Label Content:** applies to the Protein sequence viewer and the set's visualisation. You can choose the content of the labels of each entry by selecting at least 1 label from accession, name, or short name.
- **Display structure viewer all the time:** on some low-end devices, small screens, or under network or battery constraints, we might decide to not display the structure viewer by default. It will still be available on demand. With this option, you can set it to always display the viewer.

14.4 Cache settings

Caching:



In order to speed up the website we keep a local cache in your browser. It includes the API responses, since the last release of InterPro, and it gets dropped when a new version is released. You can disable the cache or clear it if, for instance, you think it is corrupted and is not displaying the latest data.

Clear Local Cache

14.5 Server settings

Servers settings

API Settings (modification temporarily disabled)

Protocol:	Hostname:	Port:	Root:	Status:
https://	www.ebi.ac.uk	443	/interpro/wwwapi/	Reachable

EBI Search Settings (modification temporarily disabled)

Protocol:	Hostname:	Port:	Root:	Status:
https://	www.ebi.ac.uk	443	/ebisearch/ws/rest/interpro7	Reachable

InterProScan Settings (modification temporarily disabled)

Protocol:	Hostname:	Port:	Root:	Status:
https://	www.ebi.ac.uk	443	/Tools/services/rest/iprscan5/	Reachable

Wikipedia Settings (modification temporarily disabled)

Protocol:	Hostname:	Port:	Root:	Status:
https://	en.wikipedia.org	443	/w/api.php	Reachable

AlphaFold API Settings (modification temporarily disabled)

Protocol:	Hostname:	Port:	Root:	Status:
https://	alphafold.ebi.ac.uk	443	/	Reachable

Reset settings to default values

To get all the data displayed, the InterPro website queries different API servers. Although the values in this section are read-only in the current version of the website, the information can be useful to identify any data errors on the website.

14.6 Developer Information

Information on the current build of the website. It is read-only but can help to investigate any errors in the website.

FREQUENTLY ASKED QUESTIONS (FAQS)

15.1 General Questions

15.1.1 Why is InterPro useful?

InterPro combines signatures from multiple, diverse databases into a single searchable resource, reducing redundancy and helping users interpret their sequence analysis results. By uniting the member databases, InterPro capitalises on their individual strengths, producing a powerful diagnostic tool and integrated resource.

15.1.2 What do people use InterPro for?

InterPro provides an easy route to many kinds of protein analysis, for example:

- Identify all the proteins that belong to a protein family or contain a particular domain
- Identify what domains and sites are found in a particular protein.
- Identify proteins that share a common domain, even when the names and activities of the proteins are highly variable.
- Examine the species in which a particular protein family or domain is found.
- Annotation of genomes with protein family information as well as GO terms.

15.1.3 Who uses InterPro?

InterPro is used by research scientists interested in the large-scale analysis of whole proteomes, genomes and metagenomes, as well as researchers seeking to characterise individual protein sequences. Within the EMBL-EBI, InterPro is used to help annotate protein sequences in UniProtKB. It is also used by the Gene Ontology Annotation group to automatically assign Gene Ontology terms to protein sequences.

15.1.4 What are entry types?

Each InterPro entry is assigned one of a number of types which tell you what you can infer when a protein matches the entry.

D Domain

Domains are distinct functional, structural or sequence units that may exist in a variety of biological contexts. A match to an InterPro entry of this type indicates the presence of a domain. Common examples of protein domains are the PH domain, Immunoglobulin domain or the classical C2H2 zinc finger.

F Family

A protein family is a group of proteins that share a common evolutionary origin reflected by their related functions, similarities in sequence, or similar primary, secondary or tertiary structure. A match to an InterPro entry of this type indicates membership of a protein family.

H Homologous Superfamily

A homologous superfamily is a group of proteins that share a common evolutionary origin, reflected by similarity in their structure. Since superfamily members often display very low similarity at the sequence level, this type of InterPro entry is usually based on a collection of underlying hidden Markov models, rather than a single signature. Homologous superfamilies usually comprise signatures from the SUPERFAMILY and CATH-Gene3D databases.

R Repeat

A short sequence that is typically repeated within a protein. Repeats are often relatively short <50 amino acids in length. Common repeats examples are Leucine Rich Repeats or WD40 repeats.

S Site

InterPro contains data for the following types of sites:

- **Active site** - A short sequence that contains one or more conserved residues, which allow the protein to bind to a ligand and carry out a catalytic activity.
- **Binding site** - A short sequence that contains one or more conserved residues, which form a protein interaction site.
- **Conserved site** - A short sequence that contains one or more conserved residues.
- **PTM site** - A short sequence that contains one or more conserved residues some of which are the site of a Post-translational modification.

U Unintegrated

In addition to signatures that have been grouped into InterPro entries, you can also find signatures from member databases that are “unintegrated” in InterPro. These signatures might not yet be curated or might not reach InterPro’s standards for integration. However, they can still provide important information about a protein of interest.

15.1.5 What are entry relationships?

InterPro organises its content into hierarchies, where possible. Entries at the top of these hierarchies describe broad families or domains that share higher level structure and/or function, while those entries at the bottom describe more specific functional subfamilies or structural/functional subclasses of domains.

For example, steroid hormone receptors constitute a family of nuclear receptors responsible for signal transduction mediated by steroid hormones, and can be sub-classified into different groups, including the liver X receptor subfamily. This subfamily consists of nuclear receptors that regulate the metabolism of several important lipids, including oxysterols.

15.1.6 What are overlapping entries?

On the entry page, the relationship between homologous superfamilies and other InterPro entries is calculated by analysing the overlap between matched sequence sets. An InterPro entry is considered related to a homologous superfamily if its sequence matches overlap (i.e., the match positions fall within the homologous superfamily boundaries) and either the Jaccard index (equivalent) or containment index (parent/child) of the matching sequence sets is greater than 0.75.

15.1.7 What do the colours mean in the graphical view of matches to my protein?

The graphical view of InterPro matches show where the signatures that match your protein appear on the sequence. There are two ways that these graphical “blobs” can be coloured. If you select “Colour by: domain relationship”, in the left hand menu, the domains that are from the same or related InterPro entries will be coloured the same, allowing easy visualisation of domains we know to be related. Unintegrated signatures will always be grey blobs, family signatures will always be shown as white, and sites will always be black when this option is selected.

If you select “Colour by: member database”, each blob in the sequence features section will be coloured according to the member database that provides the signature, as shown in this diagram. However, the sequence summary view will retain the domain relationship colour scheme.

15.1.8 Why are there no e-values associated with InterPro entries?

The signatures contained within InterPro are produced in different ways by different member databases, so their e-values and/or scoring systems cannot be meaningfully compared or combined. For this reason, we do not show e-values on the InterPro web site. However, e-values can be obtained via the downloadable InterProScan software package, which outputs detailed individual results for each member database sequence analysis algorithm.

15.1.9 How are InterPro entries mapped to GO terms?

The assignment of GO terms to InterPro entries is performed manually, and is an ongoing process (view related *publication*).

15.1.10 How do I contribute to InterPro?

We welcome your contributions. To report errors or problems with the database, please [get in touch via EBI support](#).

15.2 Sequence searches (InterProScan)

15.2.1 How can I ensure privacy for my sequence searches?

We adhere to EMBL standards on data privacy which can be found [here](#). However, if you have privacy concerns about submitting sequences for analysis via the web, the InterProScan software package can be downloaded for local installation from the [downloads page](#).

15.2.2 Can I access InterProScan programmatically?

InterProScan can be accessed programmatically via Web services that allow up to one sequence per request, and up to 25 requests in parallel (both [SOAP](#) and [REST](#) -based services are available).

15.2.3 How do I interpret my InterProScan results?

Please see the [Sequence search](#) section.

15.2.4 Can I trust my sequence search results?

We make every effort to ensure that signatures integrated into InterPro are accurate. Before being integrated, signatures are manually checked by curators to ensure that they are of a high quality (i.e., they match the proteins they are supposed to and hit as few incorrect proteins as possible).

While matches to InterPro should therefore be trustworthy, there are some caveats. Most proteins are currently uncharacterised, so quality checks can only ever be based on the subset of characterised proteins that match the signature. It is therefore possible that signatures can match false positives that have not been detected.

A useful rule of thumb is that the more signatures within an InterPro entry that match a protein, the more likely it is that the match is correct. Matches within the same hierarchy would also tend to increase confidence, as they all imply membership of a particular group.

Nevertheless, please bear in mind that the member database signatures are computational predictions. If you think one of our signatures matches false positives, please [contact us](#).

15.3 Web Interface

15.3.1 Which browsers are supported by the InterPro website?

For the best user experience, we recommend the use of the browsers and versions listed in the table below:

Browser	Version
Chrome	61 - 117
Edge	79 - 114
Mozilla Firefox	60 - 117
Safari	10.1 - 17
Opera	48 - 100
Android	99, 4.4.3 - 4.4.4
Chrome For Android	114
Firefox For Android	115
QQ Browser	13.1
Opera Mobile	73
iOS Safari	10.3 - 15.4
Samsung Internet	8.2 - 21

15.3.2 How do I view entry names instead of accessions in the graphical protein viewer?

The **Options** dropdown at the top right corner of the protein viewer above the protein scroll bar has labelling options grouped under “**Label by**”. Please select the **Name** option to see Entry names.

15.3.3 How do I explore the Taxonomy Tree viewer?

The taxonomy tree viewer can be navigated by clicking on nodes or using keyboard arrow keys.

15.3.4 I have selected a node in the Taxonomy tree viewer, how do I see data matching my selected taxonomy?

The information bar above the taxonomy viewer contains links on the right which lead to data filtered to match the selected taxonomy node.

15.4 Application Programming Interface (API)

15.4.1 How do I get started using the REST API?

Documentation for the API is available at our [GitHub repository](#).

If you’d like to see some example scripts in Perl, Python 3 or Javascript we have a script generator. Please follow the steps below:

1. Click on the [Results](#) tab in the *navigation menu*.
2. Click the [Your downloads](#) section.
3. Select the filters you’d like to apply.
4. Click on the **Copy code to clipboard** or **Download script file** buttons.

You can select the data type you’re interested in and apply filters to your query on this page. The corresponding API call is given under the **Results** section. The Code snippet section shows an example of code which you can run on your computer to fetch the data from the InterPro API.

15.4.2 Why do I get HTTP timeouts (code 408) when running queries?

Certain queries of the InterPro API may take a long time to run. Any request that takes longer than a few minutes is moved to run in the background and the API will return the HTTP status code 408 corresponding to a timeout. The query will continue to run in the background and the data will eventually become available.

The [Select and Download InterPro data](#) page shows examples of code which handles these timeout codes to allow fetching of data from the API.

15.5 Troubleshooting

15.5.1 Why doesn't the website work properly in Web Browser private/incognito mode?

Some functionality of the InterPro website, particularly InterProScan searches and downloading data make use of Browser storage. These functions require the user to agree to EMBL-EBI cookies and are incompatible with browser Incognito/Privacy modes.

Please grant permission for cookies and browse the site in a standard user session to fully enable functionality of the InterPro website.

Click on the “hamburger” icon above the magnifying glass icon to open the InterPro Menu sidebar. The **Connection status**, provides information on the status of the different resources used by InterPro. If all the lights are green it means all the resources are working as expected, otherwise you can see which resource has an issue.

15.6 Additional help

[Submit a ticket](#) to our helpdesk if you cannot find the answer to your questions here.

INTERPROSCAN

InterProScan is the software package that allows sequences to be scanned against InterPro's member database signatures.

Users who have novel nucleotide or protein sequences that they wish to functionally characterise can use InterProScan to run the scanning algorithms against the InterPro database in an integrated way.

16.1 Documentation

For more information on downloading, installing and running InterProScan please see the [InterProScan documentation](#).

16.2 Web services

Programmatic access to InterProScan is possible via a number of different web service protocols, that allow up to 100 sequences to be analysed per request.

16.2.1 REST

We provide access to InterProScan via [RESTful](#) services.

16.2.2 SOAP

We also provide access to InterProScan via [SOAP-based](#) web services.

16.3 Web based tools

Web access using the [Sequence search](#) box on the InterPro website, for the analysis of single protein sequences in FASTA format with a maximum length of 40,000 amino acids.

16.4 Source code

You can find, clone, and download the full InterProScan source code on the [Github repository](#).

16.5 Previous releases

To ensure you have the latest data and software enhancements we always recommend you download the latest version of InterProScan. However all previous releases are archived on the [FTP site](#).

16.6 License

The InterProScan software is distributed under the open source [Apache License](#), as are the included scanning tools (except SignalP and TMHMM). Therefore, you do not need a special license for commercial use but please cite the resource and keep the Copyright statement with your installation.

16.7 Follow us & reporting bugs

If you want to get updates on InterProScan follow InterPro on X [@InterProDB](#) or [LinkedIn](#).

If you want to submit a question or report a bug, please [contact us](#), providing as much information as possible so that we can recreate the problem.

INTERPRO CONSORTIUM MEMBER DATABASES

InterPro is the world's most comprehensive resource for protein family and domain information, but InterPro is only possible due to the amazing classification work of our collaborators. InterPro integrates protein signatures from 13 member databases, which use a variety of different methods to classify proteins. Each of the databases has a particular focus (e.g. protein domains defined from structure, or full length protein families with shared function). We strive to integrate the signatures from the member databases into InterPro entries to identify where different member database entries are the same entity.

17.1 CATH-Gene3D



The CATH-Gene3D database describes protein families and domain architectures in complete genomes. Protein families are formed using a Markov clustering algorithm, followed by multi-linkage clustering according to sequence identity. Mapping of predicted structure and sequence domains is undertaken using hidden Markov models libraries representing CATH and Pfam domains. CATH-Gene3D is based at University College, London, UK.

17.2 CDD



CDD is a protein annotation resource that consists of a collection of annotated multiple sequence alignment models for ancient domains and full-length proteins. These are available as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences via RPS-BLAST. CDD content includes NCBI-curated domain models, which use 3D-structure information to explicitly define domain boundaries and provide insights into sequence/structure/function relationships, as well as domain models imported from a number of external source databases.

17.3 HAMAP



<https://hamap.expasy.org/>

HAMAP stands for High-quality Automated and Manual Annotation of Proteins. HAMAP profiles are manually created by expert curators. They identify proteins that are part of well-conserved protein families or subfamilies. HAMAP is based at the SIB Swiss Institute of Bioinformatics, Geneva, Switzerland.

17.4 MobiDB Lite



<http://old.protein.bio.unipd.it/mobidblite/>

MobiDB offers a centralized resource for annotations of intrinsic protein disorder. The database features three levels of annotation: manually curated, indirect and predicted. The different sources present a clear tradeoff between quality and coverage. By combining them all into a consensus annotation, MobiDB aims at giving the best possible picture of the “disorder landscape” of a given protein of interest.

17.5 NCBIfam



https://www.ncbi.nlm.nih.gov/genome/annotation_prok/evidence/

NCBIfam is a collection of protein families, featuring curated multiple sequence alignments, hidden Markov models (HMMs) and annotation, which provides a tool for identifying functionally related proteins based on sequence homology. NCBIfam is maintained at the National Center for Biotechnology Information (Bethesda, MD). NCBIfam includes models from TIGRFAM, another database of protein families developed at The Institute for Genomic Research, then at the J. Craig Venter Institute (Rockville, MD, US).

17.6 PANTHER



<http://www.pantherdb.org/>

PANTHER is a large collection of protein families that have been subdivided into functionally related subfamilies, using human expertise. These subfamilies model the divergence of specific functions within protein families, allowing more accurate association with function, as well as inference of amino acids important for functional specificity. Hidden Markov models (HMMs) are built for each family and subfamily for classifying additional protein sequences. PANTHER is based at University of Southern California, CA, US.

17.7 Pfam

<https://pfam.xfam.org/>

Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains. Pfam is based at EMBL-EBI, Hinxton, UK. Since 2022, Pfam annotations are hosted by the InterPro website.

17.8 PIRSF



<https://proteininformationresource.org/pirsf/>

PIRSF protein classification system is a network with multiple levels of sequence diversity from superfamilies to subfamilies that reflects the evolutionary relationship of full-length proteins and domains. PIRSF is based at the Protein Information Resource, Georgetown University Medical Centre, Washington DC, US.

17.9 PRINTS



<https://interpro-documentation.readthedocs.io/en/latest/prints.html>

PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family or domain. PRINTS is based at the University of Manchester, UK.

17.10 PROSITE profiles

<https://prosite.expasy.org/>

PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family a new sequence belongs. PROSITE is based at the Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland.

17.11 SFLD



<http://sfld.rbvi.ucsf.edu/archive/django/index.html>

SFLD (Structure-Function Linkage Database) is a hierarchical classification of enzymes that relates specific sequence-structure features to specific chemical capabilities.

17.12 SMART



<http://smart.embl-heidelberg.de/>

SMART (a Simple Modular Architecture Research Tool) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. SMART is based at EMBL, Heidelberg, Germany.

17.13 SUPERFAMILY



Superfamily

<https://supfam.mrc-lmb.cam.ac.uk/>

SUPERFAMILY is a library of profile hidden Markov models that represent all proteins of known structure. The library is based on the SCOP classification of proteins: each model corresponds to a SCOP domain and aims to represent the entire SCOP superfamily that the domain belongs to. SUPERFAMILY is based at the MRC Laboratory of Molecular Biology, Cambridge, UK.

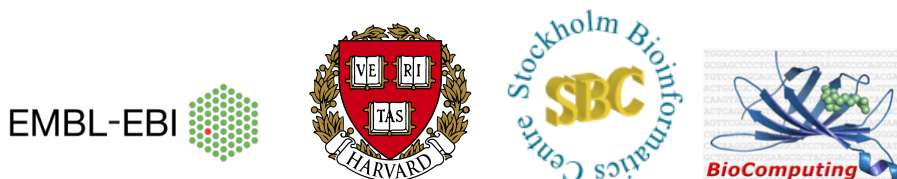
ABOUT PFAM

Pfam version 36.0 was produced at the [European Bioinformatics Institute](#) using a sequence database called *Pfamseq*, which is based on [UniProt](#) release 2022_05.

If you find Pfam useful for your research, please cite the latest Pfam publication found in [this list](#). Links to the Pfam documentation and publications are available in the [Help/Documentation section](#) in the InterPro website.

Pfam is freely available under the [Creative Commons Zero](#) (“CC0”) licence.

Pfam is powered by the [HMMER3](#) package written by Sean Eddy and his group at [HHMI/ Harvard University](#).



Pfam is supported by the following organisations:

EMBL is EMBL-EBI's parent organisation. It provides core funding (staff, space, equipment) for Pfam.



The Wellcome Trust has supported Pfam since the database inception, via core funding when based at the Wellcome Trust Sanger Institute. As well as providing and maintaining the campus on which the EMBL-EBI is located, the Wellcome Trust also now provides significant funding for Pfam (grant 221320/Z/20/Z). The current grant runs from October 2020 to September 2025.

Supported by
wellcometrust



BBSRC is supporting Pfam activities (BB/X012492/1) from January 2024 to December 2027, and has previously supported Pfam activities via grants BB/L024136/1, BB/N00521X/1, and BB/S020381/1.

The Howard Hughes Medical Institute supports the Eddy group.



Many organisations have supported Pfam activities in the past.

For more information, please contact the [Pfam helpdesk](#).

ABOUT PRINTS

PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family; its diagnostic power is refined by iterative scanning of a SwissProt/TrEMBL composite. Usually the motifs do not overlap, but are separated along a sequence, though they may be contiguous in 3D-space. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs, full diagnostic potency deriving from the mutual context provided by motif neighbours. PRINTS was previously hosted at the University of Manchester Bioinformatics Education and Research, but has been retired and InterPro serves as an archive for this resource.

For more information about PRINTS, please refer to its latest publication:

Attwood TK, Coletta A, Muirhead G, Pavlopoulou A, Philippou PB, Popov I, Romá-Mateo C, Theodosiou A, Mitchell AL. The PRINTS database: a fine-grained protein sequence annotation and analysis resource—its status in 2012. *Database (Oxford)*. 2012;2012 bas019. doi:10.1093/database/bas019. PMID: 22508994.

ABOUT SFLD

The Structure-Function Linkage Database ([SFLD](#)) is a hierarchical classification of enzymes that relates specific sequence-structure features to specific chemical capabilities. It was developed by the Babbitt Laboratory in collaboration with the UCSF Resource for Biocomputing, Visualization, and Informatics. As of April 2019, the database is in static format, and will not be updated.

ABOUT ANTIFAM

AntiFam is a resource of profile-HMMs designed to identify spurious protein predictions. AntiFam profile-HMMs come from two sources:

1. A number of spurious Pfam families have been built in the past. These were based on erroneous gene predictions. These protein families have been deleted from Pfam, but new proteins may be predicted. More recently proteins identified as Shadow ORFs and their homologues have been used to create new AntiFam families.
2. Profile-HMMs have been created from translations of commonly occurring non-coding RNAs such as tRNAs.

This collection of profile-HMM models is designed to be used as a quality control step for the UniProt sequence database as well as metagenomic projects.

Note that AntiFam models may hit proteins which are extended at the N-terminus due to the wrong initiator methionine being selected. Proteins which have known Pfam domains are unlikely to be spurious proteins.

Release	# Entries
1.0	8
1.1	23
2.0	49
3.0	54
4.0	67
5.0	72
6.0	250
7.0	263

AntiFam is freely available under the Creative commons Zero (CC0) licence. <http://creativecommons.org/publicdomain/zero/1.0/>

21.1 How to use AntiFam

AntiFam is composed of a collection of alignments found in the file AntiFam.seed. Using the HMMER3 software a library of profile-HMMs was built. This library is found in the file AntiFam.hmm.

To use the hmm library you must first make index files with the following command

```
hmmcompress AntiFam.hmm
```

To search AntiFam against a set of sequences you run the following command

```
hmmsearch --cut_ga AntiFam.hmm yourseq.fasta
```

Any reported matches are very likely to be spurious gene predictions.

21.2 Superkingdom-specific sets

AntiFam includes superkingdom-specific sets of HMMs:

- AntiFam_Eukaryota.hmm
- AntiFam_Bacteria.hmm
- AntiFam_Archaea.hmm
- AntiFam_Virus.hmm

These contain HMMs that we have found to identify spurious proteins in each of the superkingdoms, unidentified includes unclassified organisms. One HMM may identify spurious proteins from multiple superkingdoms, and therefore may be present in more than one of these superkingdom-specific sets.

21.3 Acknowledgements

We would like to thank Wolfram Hoeps who made AntiFam release 5.0 and Syed Muktadir Al Sium who generated the large number of new families added to release 6.0.

21.4 How to cite AntiFam

If you use AntiFam in your work please cite the following paper:

Ruth Y Eberhardt, Dan Haft, Marco Punta, Maria Martin, Claire O'Donovan, Alex Bateman. (2012) AntiFam: A tool to help identify spurious ORFs in protein annotation. Database:bas003. PMID:[22434837](#).

INTERPRO TEAM

The [InterPro](#) resource is curated and maintained at the [European Bioinformatics Institute](#) in Cambridge, UK.

22.1 Team members

- [Alex Bateman](#) - Team Leader
- [Antonina Entcheva Andreeva](#) - Biocurator
- [Matthias Blum](#) - Development Project Leader
- [Laise Cavalcanti Florentino](#) - Software Developer
- [Sara Chuguransky](#) - Biocurator
- [Tiago Grego](#) - Software Developer
- [Emma Hobbs](#) - Bioinformatics Developer
- [Beatriz Lazaro Pinto](#) - Biocurator
- [Typhaine Paysan-Lafosse](#) - Curation Project Leader
- [Gustavo Salazar-Orejuela](#) - Senior Software Developer

22.2 Previous contributors

- [Rob Finn](#) - Team Leader
- [Sarah Hunter](#) - Team Leader
- [Nicky Mulder](#) - Team Leader
- [Rolf Apweiler](#) - Team Leader
- [Luis Sanchez Pulido](#) - Biocurator
- [Swaathi Kandasaamy](#) - Web Developer
- [Matloob Qureshi](#) - Lead Web Developer
- [Hsin-Yu Chang](#) - Biocurator
- [Gift Nuka](#) - Senior Software Developer
- [Lowri Williams](#) - Biocurator
- [Alex Mitchell](#) - Curation Coordinator

- Lorna Richardson - Curation Coordinator
- Simon Potter - Development Coordinator
- Matthew Fraser - Software Developer
- Sebastien Pesseat - Web Developer
- Aurelien Luciani - Web Developer
- Amaia Sangrador Vegas - Biocurator
- Siew-Yit Yong - Bioinformatician/Production Manager
- Neil Rawlings - Biocurator
- Louise Daugherty - Biocurator
- Phil Jones - Senior Software Developer
- Craig McAnulla - Senior Bioinformatician
- Antony Quinn - Senior Software Developer
- Sandra Orchard - Biocurator
- Alex Kanapin - Senior Software Developer
- Wolfgang Fleischmann - Group Coordinator
- Evgeny Zdobnov - Software Developer
- Margaret Biswas
- Tom Oinn
- Florence Servant
- David Binns - Software Developer
- David Lonsdale - Curation Coordinator
- Rupinder Singh Mazara - Software Developer
- Jennifer McDowell
- Ujjwal Das - Database Production Manager
- John Maslen - Senior Software Developer
- Paul Bradley

FUNDING



InterPro is supported by EMBL, with additional funding from the Biotechnology and Biological Sciences Research Council (BBSRC grant BB/X012492/1) and the Wellcome Trust (grant 221320/Z/20/Z).

CHAPTER TWENTYFOUR

PRIVACY

Our privacy policy complies with the changes brought by the European Union data protection law (GDPR). You can find more information on the [Privacy Notice for EMBL-EBI Public Website](#). If you have any questions about this privacy policy, please [contact us via EBI support](#).

LICENSE

All of the InterPro, Pfam, PRINTS and SFLD downloadable data provided on the InterPro website is freely available under CC0 1.0 Universal ([CC0 1.0](#)) Public Domain Dedication.

The InterProScan software is distributed under the open source [Apache License](#). The included scanning tools and signature collections may be under different license terms. You do not need a special license for commercial use but please cite the resource and relevant individual member databases and keep the Copyright statement with your installation.

How to cite us

LITERATURE REFERENCES

1. Baek M, DiMaio F, Anishchenko I, et al. [Accurate prediction of protein structures and interactions using a three-track neural network](#). *Science* 373, 871-876 (2021).
2. Hiranuma, N. et al. [Improved protein structure refinement guided by deep learning based accuracy estimation](#). *Nature Communications* 12, 1340 (2021).
3. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. [IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests](#). *Bioinformatics* 29, 2722–2728 (2013).
4. Jumper, J., Evans, R., Pritzel, A. et al. [Highly accurate protein structure prediction with AlphaFold](#). *Nature* (2021)

PROTEIN FAMILIES CARD GAME

27.1 Protein families game

The Protein families game contains 42 cards divided in 7 families (6 protein cards each), the goal is to collect the maximum number of families by asking the other players for the protein cards you are missing in your hand to complete your families. The game logic is similar to the Happy families and Go Fish games.

The game is available to play online by clicking on the image below, or you can put a [request to organise the Protein families game activity in person](#).

27.1.1 Game rules

The game rules in English can be [downloaded](#).

The video below explains how to interact with the different objects in the online platform.

27.1.2 Translation

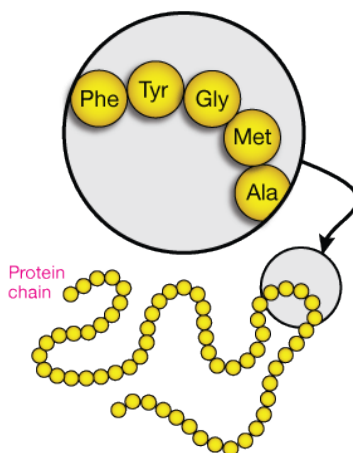
We are looking for volunteers to help us translate the game in different languages to increase its accessibility. Please [contact us](#) if this is something you'd like to do.

27.2 Understanding the biology

27.2.1 What is a protein?

A protein is a long molecule made up of small units known as amino acids. You can visualise a protein as a pearl necklace where each pearl is an amino acid. These amino acids are found mainly in food. The amino acids required to make proteins can be obtained from the proteins we eat, or produced by the body.

The image above is an illustration of the protein amino acid sequence chain. Source: <http://xaktly.com/Proteins.html>



27.2.2 What are proteins made of?

There are 20 amino acids. However, 9 of them are called “essentials” as they can’t be produced by the human body and we obtain them by eating certain protein-rich foods (meat, poultry, fish, dairy products, eggs, and soy), hence it is important to have a good diet with enough protein intake.

27.2.3 How are proteins formed?

The amino acids in a protein are ordered in a specific way. This sequence of amino acids determines the shape and function of the protein and its called its primary structure. Proteins can vary in size ranging from 15 to 30,000 amino acids.

One of the smallest proteins is called Aspartame (it is an artificial sweetener used as a sugar substitute in foods and beverages) and is made of only 2 amino acids.

On the contrary, the Titin protein is a giant protein made of 30,000 amino acids, that plays an important role in muscle elasticity.

In addition to the primary structure, proteins have higher order structural levels such as the secondary, tertiary and quaternary structure which define their three dimensional structure and provide them with different functions.

Illustration of the protein folding process from the amino acid sequence to the quaternary structure. Source: <https://cdn.kastatic.org/>

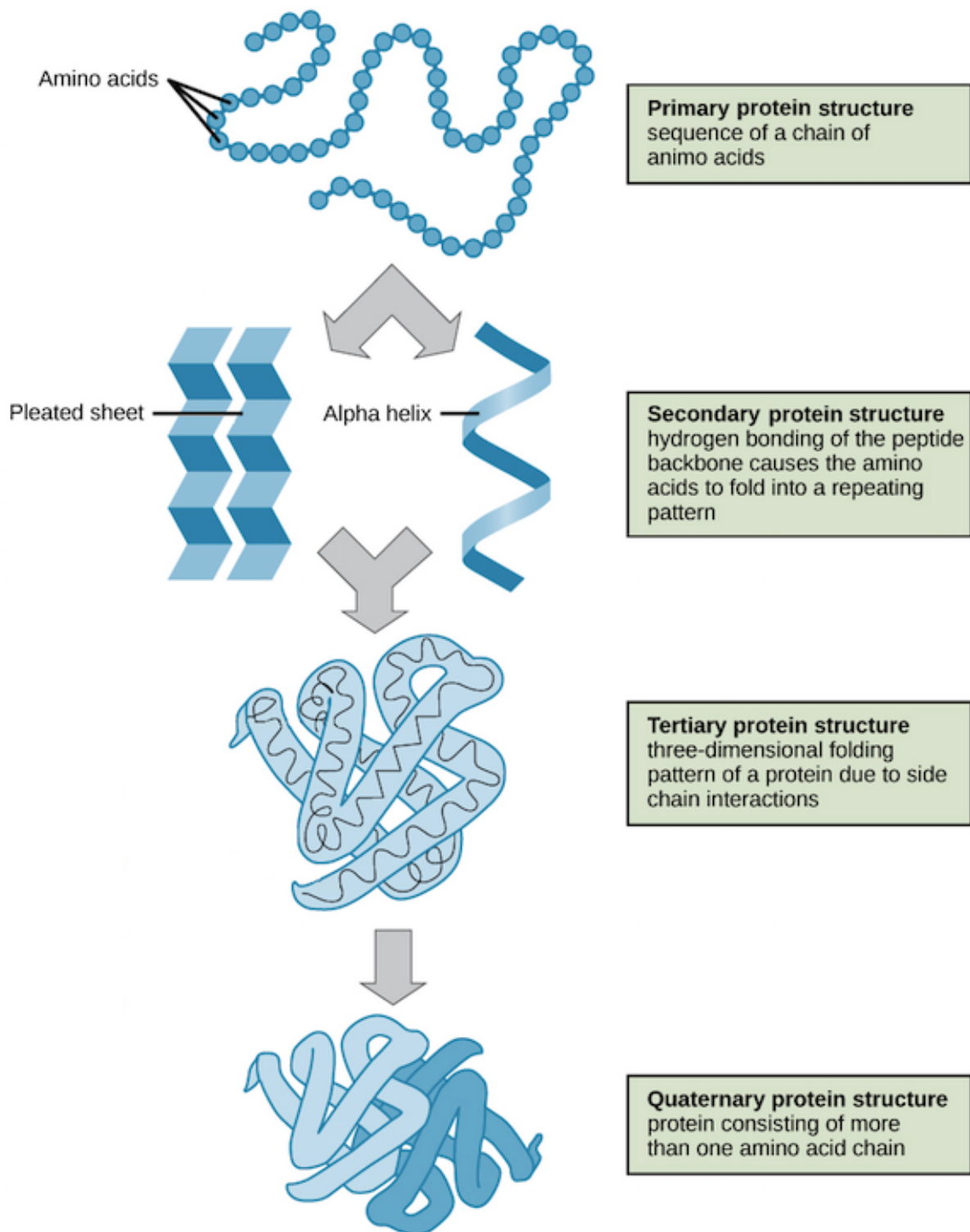
27.2.4 Where do proteins come from?

The way amino acids are organised in the protein isn’t random. Indeed, each sequence is very important, and if an amino acid is replaced by another one (by mistake) the protein might not work properly. The chains of amino acids forming proteins are determined by DNA.

The video below explains how proteins are produced from the DNA sequence.

Source: www.yourgenome.org

As you might have noticed, proteins are necessary for the body to work properly and represent about 60% of the components of a cell. They are always renewed and found in all living cells. They are essential for the cell function and responsible for diverse functions, like cellular structure (collagen), molecule transport (hemoglobin), cell activity regulators (insulin), helping molecules transformation.



27.2.5 What are proteins used for?

A human body needs proteins to perform many different functions. Some proteins help control processes in the body. Others transport, or carry, substances from one place in the body to another. Some proteins make up collagen, which helps give structure to cells. Antibodies, which fight infections and diseases, are proteins. Enzymes are also proteins, they help the body digest food and build new cells.

27.2.6 Why are proteins classified?

Proteins can be classified into groups when they have a similar chain of amino acids or a similar tertiary structure. These groups often contain well characterised proteins whose function is known. Thus, when a novel protein is identified, its functional properties can be proposed based on the group to which it is predicted to belong.

27.2.7 How are protein classified?

Proteins can be classified into different groups based on the families to which they belong, the domains they contain, or the sequence features they possess.

Protein family

A protein family is a group of proteins that share a common evolutionary origin (they have a common ancestor), we can identify them as they have related functions and similarities in their amino acid sequence or structure.

Example of a protein family: Nuclear hormone receptors

Nuclear hormone receptors constitute an important family of transcription regulators that are involved in diverse physiological functions. Members of the family include the steroid hormone receptors and receptors for thyroid hormone, retinoids, vitamin D3 and many other ligands. Nuclear hormone receptors are extremely important in medical research, a large number of them is being implicated in diseases such as cancer, diabetes, and hormone resistance syndromes.

List of a few members of the Nuclear hormone receptors family obtained from InterPro [IPR001723](#).

3D Structures of 4 Nuclear hormone receptors: Thyroid hormone (PDB [4lnw](#)), Vitamin D (PDB [3a40](#)), Retinoic acid (PDB [5k13](#)) and Estrogen (PDB [6vjd](#)) receptors.

Protein domains

Domains are distinct functional and/or structural units in a protein. Usually, they are responsible for a particular function or interaction, contributing to the overall role of a protein. Domains may exist in a variety of biological contexts, where similar domains can be found in proteins with different functions.

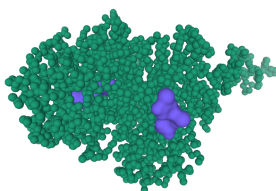
Example of a protein domain: Globins

Globins are involved in binding and/or transporting oxygen. They have evolved from a common ancestor and can be divided into three groups: single-domain globins, and two types of chimeric globins, flavohaemoglobins and globin-coupled sensors.

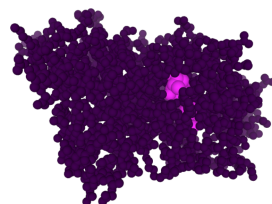
The major types of globins include:

- Neuroglobin is found in vertebrate brain and retina
- Hemoglobin transports oxygen from lungs to other tissues in vertebrates
- Protoglobin is found in archaea
- Cytochrome b5 is an oxygen sensor

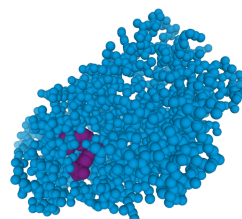
- ▼ **F** **Nuclear hormone receptor (IPR001723)**
 - F** Retinoid X receptor/HNF4 (IPR000003)
 - ... **F** Progesterone receptor (IPR000128)
 - F** Vitamin D receptor (IPR000324)
 - F** Glucocorticoid receptor (IPR001409)
 - F** Thyroid hormone receptor (IPR001728)
 - F** Ecdysteroid receptor (IPR003069)
 - F** Orphan nuclear receptor (IPR003070)
 - F** Peroxisome proliferator-activated receptor (IPR003074)
 - F** Retinoic acid receptor (IPR003078)
 - F** Nuclear receptor ROR (IPR003079)
 - F** Nuclear hormone receptor family 5 (IPR016355)
 - F** Liver X receptor (IPR023257)
 - F** Estrogen receptor/oestrogen-related receptor (IPR024178)
 - F** Nuclear receptor subfamily 0 group B member 1/2 (IPR033544)



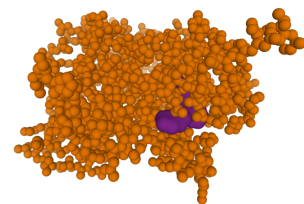
Thyroid hormone receptor



Vitamin D receptor

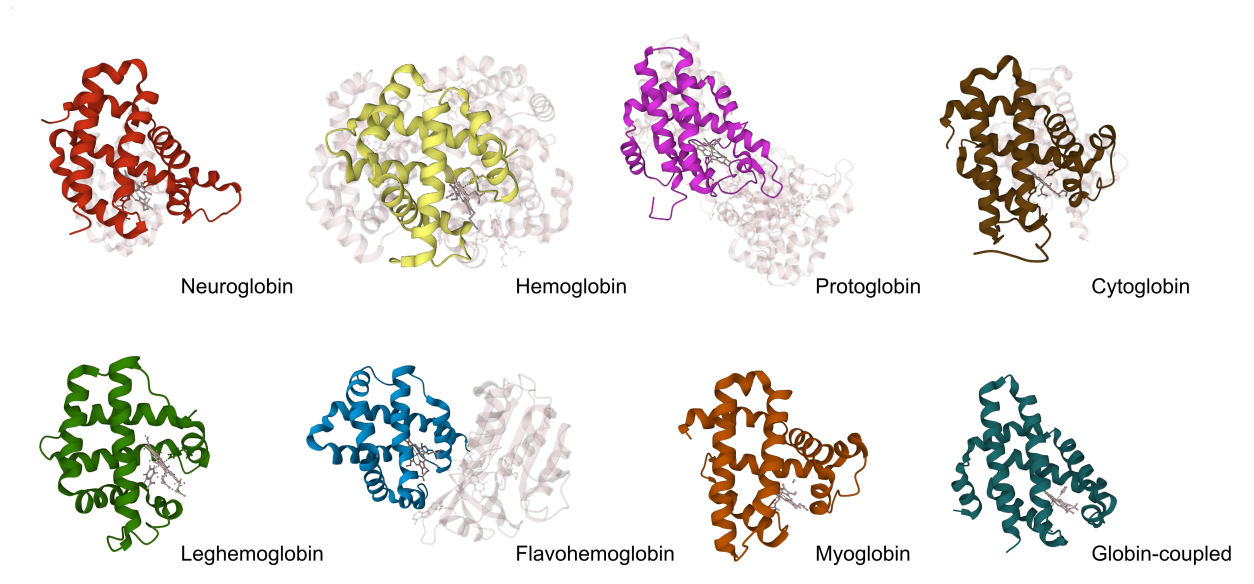


Retinoic acid receptor



Estrogen receptor

- Leghemoglobin is found in leguminous plants
- Flavohemoglobin provides protection against nitric oxide
- Myoglobin is responsible for oxygen storage in vertebrate muscle
- Globin-coupled sensors



Cartoon representation of the globins domains structures generated using [mol*](#). They are all made of eight alpha helices.

Family- and domain-based classifications are not always straightforward and can overlap, since proteins are sometimes assigned to families by virtue of the domain(s) they contain.

Sequence features

Sequence features are groups of amino acids that confer certain characteristics upon a protein, and may be important for its overall function. Sequence features differ from domains in that they are usually quite small (often only a few amino acids long), whereas domains represent entire structural or functional units of the protein. Sequence features are often nested within domains.

27.2.8 Protein classification in InterPro

Multiple groups of scientists work on protein classification and are using different methods and criteria to generate their categorisation. InterPro is the main resource for protein classification at the European Bioinformatic Institute. It regroups the protein classification from multiple databases into a single searchable resource. Having all this information available in a single location is very convenient and time saving for the scientific community, as the researchers don't have to look for information in different places. InterPro also provides a tool, called InterProScan, to help the function prediction of newly discovered proteins.

27.3 Ask questions or give feedback

Do you have questions about protein or protein classification?

Suggestions to improve the protein families game?

Would like us to run the Protein families game activity in your school or get a printed copy?

Send us your question(s) or requests.