InterPro

Protein sequence analysis & classification

A CRITICAL GUIDE

Version: 1 August 2018

InterPro

Overview

This Critical Guide provides an introduction to the InterPro database, the largest, most comprehensive, integrated protein family database in the world. The rationale for creating the resource, the nature of its contributing databases and the kinds of information they provide are discussed, and the role of InterPro in protein classification and function-annotation projects is outlined.

Teaching Goals & Learning Outcomes

This Guide introduces the principal components of the InterPro database, the differences between them, and how their integration creates a resource whose diagnostic power is greater than the sum of its parts. On reading this Guide, you will be able to:

- explain how protein family databases are used to help annotate uncharacterised protein sequences;
- identify InterPro's constituent data resources and explain the main methods that underpin them;
- search InterPro using keywords and full sequences;
- analyse and interpret search results in terms of protein family hierarchies, their structural domains and functional features; and
- track the provenance of InterPro's annotations.

1 Introduction

Protein family databases emerged in the late 1980s/early 1990s, when researchers first became interested in cataloguing the functional features that characterise protein sequences. Then, inferring the functional specificity of newly determined sequences was a challenge: databases weren't sufficiently large or diverse, nor were their search tools sufficiently sensitive to be able to identify related sequences – **homologues** – reliably. In 1986, **Russell Doolittle** described a set of short amino acid **patterns** that characterised particular functional sites, reasoning that these could be collected and used to search for similar sites in query sequences[1]. **Amos Bairoch** was the first to do this. Scanning the literature for new patterns, however, he found their diagnostic power too poor to be useful[2], so he began building a collection of his own – this was to become the first publicly available database of protein family sites and patterns.

Bairoch's work led to the creation of several 'copy cat' databases, some of which went on to become pivotal tools for functional **annotation**. Elucidating protein function is key to understanding basic natural processes, the basis of disease, the interactions of species with their environments, *etc.* But characterising protein function experimentally is costly and slow, and can't keep pace with genomic and environmental sequencing projects, which can sequence complete genomes, or produce hundreds of millions of sequences from environmental samples, in a single experiment. For database curators, who must assign functions to the raw data, the annotation burdens are inconceivable. To give an idea, manually annotated **UniProtKB**/Swiss-Prot[3] is ~200 times smaller than its computer-annotated UniProtKB/TrEMBL counterpart; moreover, the peptide database behind the EBI's Metagenomics service[4] is predicted to contain millions of full-length sequences not yet represented in UniProtKB, and is likely to assimilate hundreds of millions or billions of sequences in future. Clearly, this mandates the use of computational approaches to facilitate the daunting work of curators.

Before the situation became so critical, an innovative function-annotation strategy was to exploit diagnostic signatures in protein family databases. Underpinned by conservation data from multiple sequence alignments, these offered a more sensitive and *scalable* alternative to pairwise sequence analysis methods like **BLAST**[5]. In 1999, to help annotate TrEMBL[6], several such databases were therefore integrated into a unified protein family resource: InterPro[7]. This Guide introduces InterPro and its component databases. It outlines their underlying analysis methods, and how, together, these create a uniquely powerful diagnostic tool for protein family annotation.

2 About this Guide

The following sections introduce InterPro, and present some of the main features of, and differences between, some of its constituent databases. The Guide gives a high-level content overview rather than a complete tour of the Web interface. Exercises are provided to help understand how to navigate, search and interpret the information stored in InterPro, and discover its provenance. Throughout the text, key terms – rendered in **bold** type – are defined in boxes. Additional information is provided in supplementary boxes.

KEY TERMS

Amos Bairoch: pioneering developer of Swiss-Prot, PROSITE, TrEMBL, InterPro, UniProt, *etc.*; he was also a co-founder of the SIB

Annotation: notes that make database entries informative & re-usable

BLAST: Basic Local Alignment Search Tool, a program for searching nucleotide or protein sequence databases with a query sequence

Homologue: a similar sequence that has shared evolutionary ancestry

Pattern: a consensus set of amino acid residues or nucleotide bases that characterises the region or family of sequences in which it is found

Russell Doolittle: a biochemist known for his work on the structure & evolution of proteins; he co-developed the hydropathy index

UniProtKB: UniProt Knowledgebase, a protein sequence database comprising UniProtKB/Swiss-Prot & UniProtKB/TrEMBL

3 What is InterPro?

Maintained at the **EBI**, InterPro (www.ebi.ac.uk/interpro) is the world's largest integrated protein family database. It was first released in October 1999, aiming to provide a basic level of annotation to the then new TrEMBL database, a computer-generated supplement to Swiss-Prot[6]. TrEMBL provided timely access to protein products emerging from genome projects, in a Swiss-Prot-like format, but offered little or no supporting information. In an attempt to add value to the deluge of uncharacterised data, the vision was therefore to annotate TrEMBL entries using signatures from several of the available protein family resources: specifically,

i) *PROSITE patterns*[8]: a collection of consensus regular expression-like patterns encoding functional sites and families;
ii) *PROSITE profiles*[9]: a compilation of **position-specific scoring matrices** describing protein families and superfamilies;
iii) *PRINTS*[10]: a compendium of protein **fingerprints** encoding protein subfamilies, families and superfamilies;
iv) *Pfam*[11]: a database of **Hidden Markov Models** (HMMs) describing protein domains and superfamilies; and
v) *ProDom*[12]: clusters of sequences encoding domain families.

The first integrated set of signatures[13] (1,370 patterns and 241 profiles from PROSITE 16.0, 1,157 fingerprints from PRINTS 23.1, and 1,465 entries from Pfam 4.0) was used to characterise 2,423 families and domains in Swiss-Prot 38.0 and TrEMBL 11.0. ProDom wasn't integrated until release 1.2 (to which it contributed 540 entries) because its clusters changed between different versions of Swiss-Prot, making it hard to assign stable accession numbers.

Over the next decade, many more databases joined the consortium[14], bringing major annotation and data-management issues. By 2018, with 14 contributing data sources (listed in Table 1), InterPro had acquired an almost inconceivable level a complexity.

Table 1 ***Composition of InterPro 69.0, 21 June 2018.*** *The version and number of source entries, and number/percentage of those that have been integrated, are shown.*

Partner database	Version	# Entries	# Integrated	% Integrated
Pfam	31.0	16,712	16,119	96
PANTHER	12.0	90,742	8,297	9
TIGRFAMs	15.0	4,488	4,445	99
PIRSF	3.02	3,285	3,223	98
CDD	3.16	12,805	2,868	22
HAMAP	2018_03	2,246	2,244	100
CATH-Gene3D	4.2.0	6,119	2,143	35
PRINTS	42.0	2,106	1,968	93
SUPERFAMILY	1.75	2,019	1,602	79
ProDom	2006.1	1,894	1,310	69
PROSITE patterns	2018_02	1,309	1,287	98
SMART	7.1	1,312	1,263	96
PROSITE profiles	2018_02	1,210	1,176	97
SFLD	3	303	147	48

Comparing columns 3 and 4 of Table 1 shows that, while most partner database contents are included, some aren't integrated: this can happen when i) there are problems reconciling sequence- and structure-based families (like those in **CATH-Gene3D**[15] and **CDD**[16]), whose conceptual models of protein families are very different; ii) partner resources contain vast amounts of (automatically generated) data relative to others (*e.g.*, **PANTHER**[17]), causing analysis and data-management problems, particularly if the signature collections change markedly between releases; or iii) signatures match unrelated sequences, outside their intended scope. Consequently, while the integration level of some partner databases is >90%, for others it is <50%. For each integrated entry, the curators must extract from the source database the associated family description, reconcile disparities between sources, or write new annotation if none exists – a formidable task. Despite the challenges, housing >34,000 families, domains and functional sites, InterPro 69.0 was able to provide some level of annotation for >80% of TrEMBL entries. Overall, the database has played, and continues to play, a valuable role in genome-annotation projects, including the human genome, and now offers the most comprehensive opportunity for protein family characterisation and functional annotation in the world[7].

To give an insight into some of the intricacy of InterPro, the nature and contributions of its founding partners are reviewed below.

4 InterPro's founding databases

As mentioned earlier, initially synthesised from different types of data from four sources, InterPro now amalgamates information from 14 databases. It is beyond the scope of this short Guide to discuss each of these in detail; instead, the founding partners are used as exemplars, to highlight the perspectives that the different resources bring, to give a uniquely powerful, integrated whole.

4.1 PROSITE

When Bairoch first created PROSITE, he began by devising short, consensus amino acid patterns that characterised various different binding sites, active sites, protein family groups, and so on; later, he augmented the resource with a set of more sophisticated profiles.

KEY TERMS

CATH-Gene3D: a resource that classifies protein folds according to their Class, Architecture, Topology & Homology; the Gene3D component uses CATH HMMs to predict domain boundaries & assign UniProtKB & Ensembl sequences into homologous superfamilies
EBI: European Bioinformatics Institute, the EMBL hub dedicated to the provision of bioinformatics services to the European community
Fingerprint: a set of ungapped sequence **motifs** whose number, order & separation form a characteristic signature for a given protein family, superfamily, subfamily or domain family
Hidden Markov Model (HMM): a probabilistic model comprising interconnecting states that denote the match, delete or insert status at each position within a sequence alignment
Motif: a contiguous, conserved set of amino acids within a sequence alignment, often denoting a key functional or structural feature
PANTHER: a database that uses HMMs to classify protein sequences into families, subfamilies & domains
Position-Specific Scoring Matrix (PSSM): or profile, a table of positional amino-acid weights & gap costs that encodes the conservation in (part of) a sequence alignment (*e.g.*, in a motif or domain), used to quantify the similarity between query sequences & the motif/domain
CDD: Conserved Domain Database, a resource that uses PSSMs to identify structural domains & classify protein sequences accordingly

PROSITE patterns

Developing patterns involved creating alignments of sequences that share particular functional or structural features, and identifying their conserved residues. The pattern of conservation was then encoded in a **regular expression**-like syntax that would, ideally, provide a consensus for all sequences sharing the same attributes.

The syntax of PROSITE patterns

In the dummy pattern shown here, amino acid residues are denoted using the **IUPAC single-letter notation**:

[DN]-[IVL]-x(2)-{PG}-[FY]-x(2,3)-E

[ ] : square brackets list amino acid residues allowed at a given position; a residue on its own is strictly conserved

- : dashes separate consecutive alignment positions

x : the wild-card x means that any amino acid can occur at that position

(n) : numbers in parentheses indicate how many times a residue, or residue group, can occur at a position; numbers separated by commas indicate a numerical range

{} : curly brackets list residues that are forbidden at a position

For each of these, Bairoch wrote detailed documentation, summarising the functional or structural site, domain or family grouping encapsulated by the expression. The first 58 patterns were released with a search tool called PROSITE as part of the commercial **PC/Gene** software package, in March 1988.

PROSITE was proficient at pinpointing potential protein functional sites and family relationships, and it was clear that it could be used to analyse both individual sequences and entire sequence collections, like Swiss-Prot[18]. Thus, in October 1989, it was split from PC/Gene and made publicly available for the first time; this version (4.0) held 202 patterns, some of the first of which are shown in Table 2.

Table 2 ***Some of PROSITE's first patterns.*** *The pattern length, the functional site it describes, and its associated identifiers are given. Asterisks denote patterns with a high probability of occurrence.*

Pattern	Length	Functional site	Pattern ID
N-{P}-[ST]-{P}	4	N-**glycosylation**	*PS00001
[RK](2)-x-[ST]	4	cAMP- & cGMP-dependent protein **kinase phosphorylation**	*PS00004
[ST]-x-[RK]	3	Protein kinase C phosphorylation	*PS00005
[ST]-x(2)-[DE]	4	Casein kinase II phosphorylation	*PS00006
[RK]-x(2,3)-[DE]-x(3)-Y	8-9	Tyrosine kinase phosphorylation	*PS00007
G-{EDRKHPFYW}-x(2)-[STAGCN]-{P}	6	N-**myristoylation**	*PS00008
x-G-[RK]-[RK]	4	**Amidation**	*PS00009
C-x-[DN]-x(4)-[FY]-x-C-x-C	12	Aspartic acid & asparagine **hydroxylation**	PS00010

As Table 2 shows, the earliest expressions were short – some only 3 or 4 residues long. While Swiss-Prot was fairly small, such patterns were quite potent; but the shorter a pattern and more variable it is (in terms of the range of allowed residues and wild-cards used) the weaker its diagnostic power. As Swiss-Prot grew, this began to limit the utility of many patterns. To investigate the problem, the specificity of PROSITE patterns current in 1994 was analysed with respect to Swiss-Prot 30.0 and to a randomised database derived from it[a]. Inevitably, the shortest, most variable patterns were found to be the most promiscuous. Of these, PS00001 to PS00009 were so noisy that they were flagged '*patterns with a high probability of occurrence*', a flag that allows search tools to ignore indiscriminate expressions.

Examining Table 2 shows that sites described by patterns PS00002 and PS00003 are absent. PS00002 encoded a glycosaminoglycan attachment site (S-G-x-G); this expression was qualified by the condition that at least two acidic amino acids should be found -2 to -4 residues relative to the serine attachment site. Similarly, PS00003 encoded a tyrosine sulphation site, whose consensus was described in terms of the acidic, hydrophobic and polar residues located 1 to 7 residues N- and C-terminally to the tyrosine. Originally, PS00002 and PS00003 were termed 'rules', as they used free-text assertions to complement or replace patterns that couldn't be efficiently encoded by the consensus-expression syntax. The only way to scan Swiss-Prot with such rules was thus to encode them directly in the search program, which many early PROSITE-search tools were unable to do[19].

With the expansion and growing diversity of Swiss-Prot, the diagnostic performance of several PROSITE rules and patterns started to fade. Some were revised to accommodate new family members in Swiss-Prot; others, like PS00002 and PS00003, were deleted. Moreover, some protein families, sites and domains (**globins**, **immunoglobulins**, **kringle domains**, *etc.*) were just too divergent to be captured effectively using consensus expressions. Part of the difficulty is that this is a brittle diagnostic approach: *i.e.*, matching is a binary event in which a sequence either matches exactly, or not at all. Trying to capture highly divergent sequence groups in this way tends to create patterns that either match significant numbers of **false-positives**, or lead to large numbers of **false-negatives**. Attempting to address this issue, two very different pattern-recognition techniques were developed: these were profiles[9] and protein fingerprints[10].

KEY TERMS

Amidation: attachment of an amide group to an organic compound

False-negative: a true member of a given data-set that fails to match a search query in that data-set

False-positive: a match to a search query in a given data-set that is not a true member of the data-set

Globin: the haem-containing protein component of haemoglobin and myoglobin, involved in binding & transporting oxygen

Glycosylation: attachment of a glycosyl moiety to a hydroxyl or other functional group in an organic compound

Hydroxylation: attachment of a hydroxyl group to an organic compound

Immunoglobulin: a major protein component of the immune systems of higher animals

IUPAC: International Union of Pure and Applied Chemistry, an international federation of organisations, whose work includes standardising nomenclature in chemistry & other scientific fields

Kinase: an enzyme that transfers phosphate groups from ATP to substrate molecules

Kringle domain: a small, autonomous protein domain found in a range of blood-clotting and fibrinolytic proteins; its name derives from the shape of a Scandinavian pastry, which it is said to resemble

Myristoylation: attachment of a myristoyl group to the α-amino group of an N-terminal glycine residue of a protein

PC/Gene: a commercial, MS-DOS based software package for protein & nucleotide sequence analysis

Phosphorylation: attachment of a phosphoryl group, generally to a serine, threonine or tyrosine side-chain in a protein

Regular expression: a sequence of characters defining a search pattern

Single-letter notation: individual letters used to denote the amino acids – D for aspartic acid, N for asparagine, S for serine, *etc.*

[a] ftp://ftp.vital-it.ch/pub/software/unix/prosite_scan/pst_stat.html

PROSITE profiles

Profiles, introduced into PROSITE in June 1994, are scoring tables that give numerical values (*e.g.,* expressed as residue frequencies, or mutation probabilities, *etc.*) to quantify the similarity of residues at equivalent positions in sequence alignments. They use a generalisation of the Gribskov *et al.*[20] and Lüthy *et al.*[21] data structure, whereby alignments are modelled as a series of components or 'states' (*i.e.*, beginning, B; match, M; insert, I; delete, D; end, E) and possible 'state transitions' between consecutive alignment positions: 16 state-transition scores, including those between identical states, are defined for all possible transitions (BI, BD, MI, MD, IM, *etc.*) – these state-transition scores are analogous to gap-opening penalties used in alignment algorithms. When aligning a sequence with a profile, the overall similarity score is then the sum of the component scores and penalties calculated relative to a given cut-off value.

By contrast with patterns, profiles are more tolerant of amino acid substitutions, insertions and deletions; from a practical perspective, this means they can model sequence relationships more realistically, and generally exhibit greater diagnostic power. Importantly, they can encompass full-length sequence alignments (whether these encode family-specific groups or family-agnostic domains), including conserved and divergent regions. As a result, unrelated sequences with only partially correct alignments can sometimes achieve significant scores. To minimise such false hits, sequences are required to align correctly to residues with experimentally verified structural or functional properties (*e.g.,* that mediate catalysis, ligand binding, protein-protein interaction, *etc.*) to be considered true matches[9].

Given their superior diagnostic performance, PROSITE began to accrete greater numbers of profiles, which now comprise ~50% of entries. However, although PROSITE is still popular, one of the most time-consuming of curators' tasks is manually documenting every database record; this has placed an inevitable brake on its growth.

4.2 PRINTS

The second technique devised to address the diagnostic limitations of some of PROSITE's patterns was fingerprinting[10]. While patterns use individual motifs to diagnose protein families or functional sites, fingerprints exploit the fact that, within sequence alignments, groups of motifs, and their inter-relationships, can create distinctive signatures for particular protein families or domains. This offers greater diagnostic flexibility, because it allows mismatches at the level of individual motifs, without compromising the discriminating power of the fingerprint as a whole.

Fingerprinting was devised as a tool to identify **rhodopsin**-like **GPCRs** in Swiss-Prot[22,23]. GPCRs' distinguishing characteristic is the presence of seven **transmembrane (TM) domains**, so the fingerprint encoded these motifs. The resulting discriminator was more robust than the existing PROSITE pattern, and led to its revision in 1991[22] (later, the pattern was complemented by a more potent profile).

The diagnostic power of the GPCR fingerprint spurred the development of many other protein fingerprints. These were collated into a new resource in October 1991, initially known as the Features Database[24], but re-branded as 'PRINTS' in 1994[10,25]. As in PROSITE, value was added to PRINTS' entries through manual annotation; the databases therefore shared similar annotation burdens. This led to a preliminary proposal to unify their formats[2], and sewed the seeds for creating the world's first integrated protein family resource.

4.2 ProDom

A very different analytical approach arose from an investigation of protein **modules**, aiming to identify these molecular 'building-blocks', and to cluster them into groups. The algorithm was applied to >21,000 sequences in Swiss-Prot 21.0, and the resulting clusters were automatically piped into a database – this was ProDom[12]. Alignments were also created for the groups, and consensus sequences derived from them in order to facilitate database searches.

A problem with this approach was that the clusters (and hence their alignments and consensus sequences) had to be re-calculated for every new release of Swiss-Prot. Computationally, this was onerous; more importantly, the resultant 'families' weren't consistent between successive Swiss-Prot releases. ProDom's contents were therefore difficult to index, preventing its integration into early releases of InterPro[13]. Adding to the difficulties, ProDom's target database soon also included TrEMBL[6] (which, at the time of its first release was twice the size of Swiss-Prot), bringing significant maintenance overheads. To give an idea of the scale of the challenge, the last version of ProDom, based on the December 2012 release of UniProtKB/Swiss-Prot and /TrEMBL[26], included more than 1.7 million domain families containing at least two sequences.

4.4 Pfam

In the decade after the release of PROSITE, database maintenance issues came to the fore, with increasing tension between users' desire for rapid access to sequence data and the need to ensure data quality and value. The manual approaches used to create high-quality alignments and detailed annotation in databases like PROSITE and PRINTS weren't scalable, so automated methods were used to create resources like ProDom. A hybrid approach was then introduced with Pfam[11], an HMM-based database comprising annotated (Pfam A) and automatically derived elements (Pfam B).

Like profiles, HMMs model sequence alignments as strings of match, insert and delete states. However, instead of using absolute scores, HMMs are probabilistic: match states contain amino acid probabilities in particular alignment columns; transition states give transition probabilities to and from insert and delete states, reflecting the propensity to insert or skip residues at given positions.

With its first release almost 10 years after PROSITE, Pfam grew rapidly. This was mostly owing to a change of emphasis towards automation, as Pfam adapted to cope with the engorgement of TrEMBL[27]. But annotating the growing numbers of entries became too arduous, so the Wikipedia model was adopted, effectively outsourcing annotation to the community. From release 25.0 onwards, Pfam B was retired, and the notes traditionally assigned to Pfam A entries were thence to be replaced by Wikipedia articles.

KEY TERMS

GPCR: G protein-coupled receptor, member of a ubiquitous family of membrane-bound, cell-surface proteins that transduce extracellular signals into intracellular responses by binding various ligands, & stimulating diverse intracellular pathways by coupling to different G proteins; they are targets of the majority of prescription drugs

Module: an autonomous folding domain believed to have arisen via genetic shuffling mechanisms (*e.g.*, kringle domain); modules are contiguous in sequence & often used as building blocks to confer a variety of complex functions on a parent protein, via multiple combinations of the same or different modules

Rhodopsin: a light-sensitive GPCR found in the rod-shaped photoreceptor cells of most vertebral retinas; it is an achromatic receptor, mediating vision in dim light

Transmembrane (TM) domain: a hydrophobic span of amino acids that crosses a membrane, usually ~20 residues in α-helical conformation

5 Sequence analysis with InterPro

The considerable versatility and diagnostic finesse of InterPro's multiple-database perspective is illustrated in Figure 1, which shows the result of querying the database for matches with the sequence of the human **melanocortin-5 receptor**.

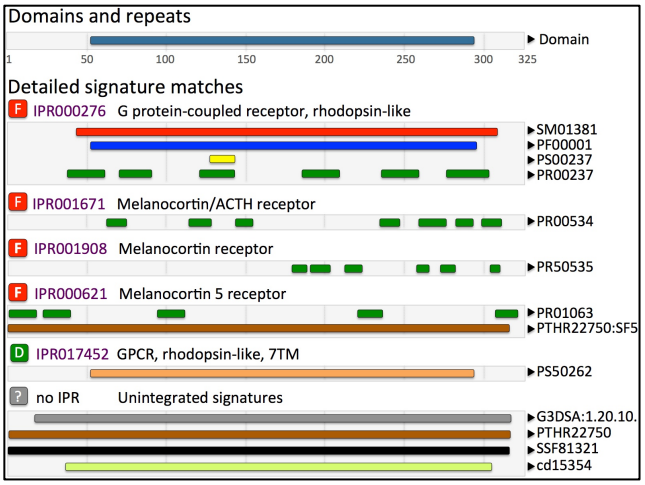

Figure 1 ***Excerpt from an InterPro look-up with UniProtKB entry MC5R_HUMAN (P33032).*** *Locations of matched domains, motifs & functional sites are shown. Colours denote the source database: red, SMART; blue, Pfam; yellow, PROSITE pattern; green, PRINTS fingerprint motifs; brown, PANTHER; orange, PROSITE profile. Matches to signatures that aren't yet integrated (e.g., CATH-Gene3D, grey; SUPERFAMILY, black) are also shown. InterPro accession numbers for each family (F) or domain (D) are on the left; source database accession numbers are on the right.*

In the complex result seen in Figure 1, each InterPro signature match is denoted as representing a family (red F) or a domain (green D), and has its own accession number (*e.g.*, *IPR000276*) and title (*e.g.*, *G protein-coupled receptor, rhodopsin-like*). Where source database signatures haven't yet been integrated into the resource, their family or domain status is denoted 'unknown' (grey ?), and no accession number is provided. Source database accession numbers are included to the right of each signature match (*e.g.*, SM01382).

This result shows a single domain (green D) and four levels of family hierarchy (four red Fs), from superfamily down to the constituent families and subfamilies. The difference in sequence coverage of the database matches is striking: some are hits with complete domains (*e.g.,* SMART (red), Pfam (blue), PANTHER (brown), and the PROSITE profile (orange)); others are hits with individual motifs (the PROSITE pattern (yellow)) or collections of motifs (the PRINTS fingerprints (dark green)). Thus, some of the superfamily- and subfamily-level matches (SMART, Pfam, PANTHER) are actually full domains

Here, the superfamily-level match diagnoses the sequence as a rhodopsin-like GPCR (*IPR000276*), and the matched domain as the characteristic 7TM region (*IPR017452*). The motif matches are more specific and informative, pinpointing unique functional and structural determinants at the superfamily, family and subfamily levels: the fingerprints hence reveal that not only is the sequence a rhodopsin-like GPCR, but it is a member of the ***melanocortin/ACTH receptor*** family, specifically a *melanocortin-5 receptor*.

Fingerprints offer information-rich layers to the family hierarchy and thus to the overall functional diagnosis, complementing the broad, domain-level views of most other member databases. Of course, the depth of its source database contents ultimately determines the specificity of InterPro's results; outputs may thus capture rather shallower hierarchies, as illustrated in Figure 2.

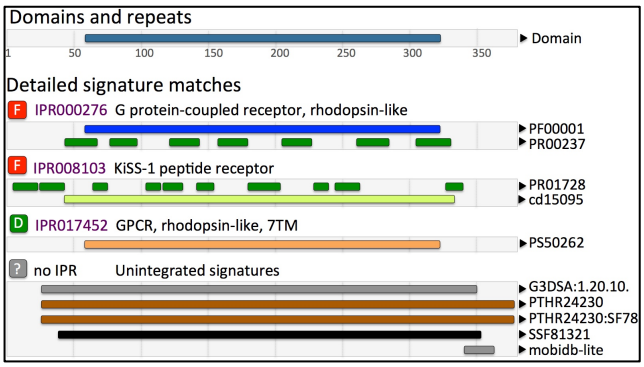

Figure 2 ***Excerpt from an InterPro look-up with UniProtKB entry KISSR_HUMAN (Q924U1).*** *For layout & colour-code details, see* **Figure 1**.

EXERCISES

1 On InterPro's home page, type 'MC5R_HUMAN' in the *'Search'* box (top right); press return. Note the number and type of matches (in the '*Filter view*' menu, tick the '*Colour by source database*' button).

2 Retrieve the sequence of MC5R_HUMAN from UniProtKB. Paste it into the *'Analyse your protein'* search box (middle of InterPro's home page); press '*Submit*'. Note the number & type of matches (again, tick the '*Colour by source database*' button). Compare the result with that obtained in (1). There appear to be more matches. What are they, & what further information do they provide?

3 The motifs of the rhodopsin-like GPCR superfamily fingerprint are derived from the 7 TM domains. From the result in (2), one of these motifs seems to be missing. Which is it? Why might this be so (hint: follow links to the source database & view the alignments)?

4 In both results, the PROSITE pattern aligns with one of the rhodopsin-like GPCR superfamily fingerprint motifs. In which TM domain does it lie? Explore the output: hover over the coloured bars & follow links to the source database – can you determine the functional significance of this motif from the source annotation?

5 In both results, the locations of the fingerprint motifs for the melanocortin/ACTH receptor family, melanocortin receptor subfamily & receptor 5 subtype appear to be different from those of the rhodopsin-like GPCR superfamily-level fingerprint. For each level of the family hierarchy, where are the motifs preferentially located? Follow links to the source database to see if you can determine the functional significance of some of these motifs.

6 How many levels of the family hierarchy are seen in Figure 2? Use KISSR_HUMAN to search InterPro. Follow the links to the source database to find a likely role for the KiSS receptor. Compare this with the annotation in InterPro. What does this suggest about the source of InterPro's annotation for this family?

KEY TERMS

ACTH: adrenocorticotropin hormone, a peptide hormone secreted by the anterior pituitary gland; it stimulates the secretion of various corticosteroids, & is used as a therapeutic & diagnostic agent

Melanocortin: generic name for peptide hormones melanotropin & corticotropin, derived from proopiomelanocortin in the pituitary gland; melanocortins are involved in regulation of food intake in mammals, exerting their effects by activating **melanocortin receptors**

Melanocortin receptor: a family of rhodopsin-like GPCRs with five known members, each having a different melanocortin specificity

Melanocortin-5 receptor: a member of the melanocortin receptor family, implicated in diverse physiological roles, including lipid metabolism & exocrine function

TAKE HOMES

1 InterPro is the world's largest integrated protein family database for sequence analysis & family classification;

2 InterPro's founding partners were PROSITE, PRINTS, ProDom & Pfam; today, the resource integrates 14 different data sources;

3 InterPro plays significant roles in annotating protein products of genome projects, & is a core component of the UniProtKB/TrEMBL-annotation pipeline; it is also available for sequence searches online;

4 InterPro's unique analytical power derives from the different diagnostic perspectives of its partner databases, combining single-motif, multiple-motif, domain-based and structure-based approaches into a result that is greater than the sum of its parts;

5 Integration of some of InterPro's partner database contents may be deferred for various reasons: *e.g.*, if the diagnostic power of a signature(s) erodes over time, if there are significant discrepancies between entries that purport to represent the 'same' family, *etc.*

6 A significant challenge for InterPro is in rationalising the different diagnostic outputs of its partner databases, and reconciling both their different nomenclatures & their different annotations;

7 The provenance of some of InterPro's annotation can be traced back to its partners (*e.g.*, to manual annotations in PROSITE & PRINTS).

6 References & further reading

1 Doolittle RF. (1986) ***Of Urfs and Orfs: a primer on how to analyze derived amino acid sequences.*** University Science Books, 20 Edgehill Road, Mill Valley, CA 94941, USA. ISBN 0-935702-54-7

2 Bairoch A. (2000) **Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times!** *Bioinformatics*, **16(1)**, 48-64.

3 The UniProt Consortium. (2017) **UniProt: the universal protein knowledgebase.** *Nucleic Acids Res.*, **45**(D1), D158-D169.

4 Mitchell AL *et al.* (2018) **EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies.** *Nucleic Acids Res.*, **46** (D1), D726–D735.

5 Altschul SF *et al.* (1990) **Basic local alignment search tool.** *J. Mol. Biol.*, **215**(3), 403-410.

6 Bairoch A & Apweiler R. (1996) **The SWISS-PROT protein sequence data bank and its new supplement TREMBL.** *Nucleic Acids Res.*, **24**(1), 21-25.

7 Finn RD *et al.* (2017) **InterPro in 2017 – beyond protein family and domain annotations.** *Nucleic Acids Res.*, **45**(D1), D190-9.

8 Bairoch A. (1991) **PROSITE: a dictionary of sites and patterns in proteins.** *Nucleic Acids Res.* **19**, 2241-2244.

9 Bairoch A & Bucher P. (1994) **PROSITE: recent developments.** *Nucleic Acids Res.* **22**(17), 3583-3589.

10 Attwood TK & Beck ME. (1994) **PRINTS - a protein motif fingerprint database.** *Protein Eng.* **7**(7), 841-848.

11 Sonnhammer EL *et al.* (1997) **Pfam: a comprehensive database of protein domain families based on seed alignments.** *Proteins* **28**(3), 405-420.

12 Sonnhammer EL & Kahn D. (1994) **Modular arrangement of proteins as inferred from analysis of homology.** *Protein Sci.* **3**(3), 482-492.

13 Apweiler R *et al.* (2001) **The InterPro database, an integrated documentation resource for protein families, domains and functional sites.** *Nucleic Acids Res.*, **29**(1), 37-40.

14 Hunter S *et al.* (2009) **InterPro: the integrative protein signature database.** *Nucleic Acids Res.* **37**(Database Issue), D211-15.

15 Dawson NL *et al.* (2017) **CATH: an expanded resource to predict protein function through structure and sequence.** *Nucleic Acids Res.* **45**(D1), D289–D295.

16 Marchler-Bauer A *et al.* (2017) **CDD/SPARCLE: functional classification of proteins via subfamily domain architectures.** *Nucleic Acids Res.* **45**(D1), D200-3.

17 Mi H *et al.* (2017) **PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements.** *Nucleic Acids Res.* **45**(D1), D183–9.

18 Bairoch A & Boeckmann B. (1991) **The SWISS-PROT protein sequence data bank.** *Nucleic Acids Res.*, **19** Suppl., 2247-2249.

19 Gattiker A *et al.* (2002) **ScanProsite: a reference implementation of a PROSITE scanning tool.** *Applied Bioinformatics* **1**(2), 107-108.

20 Gribskov M *et al.* (1987) **Profile analysis: detection of distantly related proteins.** *Proc. Natl. Acad. Sci. USA* **84**(13), 4355-4358.

21 Lüthy R *et al.* (1994) **Improving the sensitivity of the sequence profile method.** *Protein Sci.*, **3**(1), 139-146.

22 Attwood TK & Findlay JBC. (1993) **Design of a discriminating fingerprint for G-protein-coupled receptors.** *Protein Eng.*, **6**(2), 167-176.

23 Attwood TK & Findlay JBC. (1994) **Fingerprinting G-protein-coupled receptors.** *Protein Eng.*, **7**(2), 195-203.

24 Akrigg DA *et al.* (1992) **SERPENT - An information storage and analysis resource for protein sequences.** *CABIOS* **8**(3), 295-296.

25 Attwood TK *et al.* (2012) **The PRINTS database: a fine-grained protein sequence annotation and analysis resource - its status in 2012.** *Database*, 10.1093/database/base019.

26 The UniProt Consortium. (2013) **Update on activities at the Universal Protein Resource (UniProt).** *Nucleic Acids Res.* **41**(D1), D43-D47.

27 Finn RD *et al.* (2016) **The Pfam protein families database: towards a more sustainable future.** *Nucleic Acids Res.* **44**(D1), D279-D285.

7 Acknowledgements & funding

GOBLET Critical Guides marry ideas from the Higher Apprenticeship specification for college-level students in England (**www.contentextra.com/lifesciences/unit12/unit12home.aspx**) with the EMBnet Quick Guide concept.

This Guide was developed with the support of a donation from EMBnet to the GOBLET Foundation.

Design concepts and the Guide's front-cover image were contributed by CREACTIVE.

8 Licensing & availability

The Guide is freely available for download via the GOBLET portal (**www.mygoblet.org**) and EMBnet website (**www.embnet.org**).

9 Disclaimer

About the organisations

GOBLET

GOBLET (Global Organisation for Bioinformatics Learning, Education & Training) was established in 2012 to unite, inspire and equip bioinformatics trainers worldwide; its mission, to cultivate the global bioinformatics trainer community, set standards and provide high-quality resources to support learning, education and training.

GOBLET's ethos embraces:

- ***inclusivity****: welcoming all relevant organisations & people*
- ***sharing****: expertise, best practices, materials, resources*
- ***openness****: using Creative Commons Licences*
- ***innovation****: welcoming imaginative ideas & approaches*
- ***tolerance:*** *transcending national, political, cultural, social & disciplinary boundaries*

Further information about GOBLET and its Training Portal can be found at **www.mygoblet.org** and in the following references:

- Attwood *et al.* (2015) **GOBLET: the Global Organisation for Bioinformatics Learning, Education & Training.** *PLoS Comput. Biol.*, **11**(5), e1004281.
- Corpas *et al.* (2014) **The GOBLET training portal: a global repository of bioinformatics training materials, courses & trainers.** *Bioinformatics*, **31**(1), 140-142.

GOBLET is a not-for-profit foundation, legally registered in the Netherlands: CMBI Radboud University, Nijmegen Medical Centre, Geert Grooteplein 26-28, 6581 GB Nijmegen. For general enquiries, contact **info@mygoblet.org**.

EMBnet

EMBnet, the Global Bioinformatics Network, is a not-for-profit organisation, founded in 1988 as a network of institutions, to establish and maintain bioinformatics services across Europe. As the network grew, its reach expanded beyond European borders, creating an international membership to support and deliver bioinformatics services across the life sciences: **www.embnet.org**.

Since its establishment, a focus of EMBnet's work has been bioinformatics Education and Training (E&T), and the network therefore has a long track record in delivering tutorials and courses worldwide. Perceiving a need to unite and galvanise international E&T activities, EMBnet was one of the principal founders of GOBLET. For more information and general enquiries, contact **info@embnet.org**.

CREACTIVE

CREACTIVE, by Antonio Santovito, specialises in communication and Web marketing, helping its customers to create and manage their online presence: **www.gocreactive.com**.

About the author

Teresa K Attwood (orcid.org/0000-0003-2409-4235)

Teresa (Terri) Attwood is a Professor of Bioinformatics with more than 25 years' experience teaching introductory bioinformatics, in undergraduate and postgraduate degree programmes, and in *ad hoc* courses, workshops and summer schools, in the UK and abroad.

With primary expertise in protein sequence analysis, she created the PRINTS protein family database and co-founded InterPro (her particular interest is in the analysis of G protein-coupled receptors). She has also been involved in the development of software tools for protein sequence analysis, and for improving links between research data and the scientific literature (most notably, Utopia Documents).

She wrote the first introductory bioinformatics text-book; her third book was published in 2016:

- Attwood TK & Parry-Smith DJ. (1999) ***Introduction to Bioinformatics.*** Prentice Hall.
- Higgs P & Attwood TK. (2005) ***Bioinformatics & Molecular Evolution.*** Wiley-Blackwell.
- Attwood TK, Pettifer SR & Thorne D. (2016) ***Bioinformatics challenges at the interface of biology and computer science: Mind the Gap.*** Wiley-Blackwell.

Affiliation

School of Computer Science, The University of Manchester, Oxford Road, Manchester M13 9PL (UK).