

Bioinformatics: Life Sciences on Your Computer

Motifs, Domains, Patterns

This is supplemental reading for BOTH the video lecture on Protein Motifs and for the video lecture on Prosite Patterns and Profiles.

Identifying proteins

Protein sequencing is about 60 years old. The method utilizes the Edman degradation, which is good for about 10 to 20 amino acids. It is still frequently used to measure short regions at the N-terminus of a protein. Gel electrophoresis is frequently used to estimate the molecular weight of a protein. 2-D gels can be used for isoelectric point determination.

Mass spectrometry (also called mass spectroscopy) is very sensitive and has become a hot research field for identifying new proteins and characterizing known proteins.

Protein domains and motifs

The video lectures have more on this subject but here are some definitions. Note that the definitions overlap a bit and are a bit subjective.

Profile: quantitative description of motif

Pattern: qualitative description of motif

Signature: protein category

Motif: short, conserved region, also called 'fingerprint' - often 10 to 20 amino acids (example: helix-turn-helix)

Domain: independent unit longer than a motif (>40 aa). example: ATPase domain

Pattern matching

Patterns can be used to categorize sequences. That allows for the searching of pattern databases. A protein sequence can be entered to see if there are any known pattern matches. Some patterns are short and easily matched.

The regular expression is used to designate the pattern. Some examples:

One protein phosphorylation motif (there are more than one) has the pattern, [ST]-X-[RK]. That means serine or threonine, followed by any amino acid, followed by arginine or lysine.

E-X(2)-[FHM]-X(4)-{P}-L is a pattern that translates as E, then any 2 amino acids, then F, H or M, then any 4 amino acids, then anything but P, then L.

x(2,4) means x-x or x-x-x or x-x-x-x

A < symbol before the pattern means "at the N-terminus" and a > symbol at the end of a pattern means "at the C-terminus."

Prosite

The Prosite database (<http://prosite.expasy.org/>) is a database of patterns and profiles that correspond to protein signatures. Prosite uses the exact match approach, but many patterns are too short to be specific enough to avoid false positives.

There is a green bar in the results section (middle of diagram below) that reads, "hits by patterns with a high probability of occurrence or by user-defined patterns." Matches under that bar may be false positives, because the pattern is short and occurs frequently by chance. As you would do with E values, do NOT conclude irrelevance if the matched pattern is below that bar. Just know that the pattern occurs frequently by chance and a match might not be biologically relevant. You will only see results under the bar if you unclick the box that reads, "exclude patterns with a high degree of occurrence. By default, it is checked, meaning short patterns will not display in the results.

hits by patterns: [2 hits (by 1 pattern) on 1 sequence]

Hits by [PS00141](#) ASP_PROTEASE Eukaryotic and viral aspartyl proteases active site :

GI-456252-EMB
-CAA82924-1- (418 aa)



105 - 116: [level tag: (0)] VIVDTGSSDLMI

Predicted feature:

ACT_SITE 108 By similarity [condition: none]

291 - 302: [level tag: (0)] VLLDSGTTIISYF

Predicted feature:

ACT_SITE 294 By similarity [condition: none]

hits by patterns with a high probability of occurrence or by user-defined patterns: [17 hits (by 7 distinct patterns) on 1 sequence]

GI-456252-EMB
-CAA82924-1- (418 aa)



[PS00005](#) PKC_PHOSPHO_SITE Protein kinase C phosphorylation site :

48 - 50: SxR

Predicted feature:

MOD_RES 48 Phosphoserine (By similarity) [condition: S]

67 - 69: TgR

Predicted feature:

MOD_RES 67 Phosphothreonine (By similarity) [condition: T]

148 - 150: Txx

Predicted feature:

MOD_RES 148 Phosphothreonine (By similarity) [condition: T]

There is a >20% error rate with Prosite. The results come with links to the Prosite documentation about the protein domain, motif or family represented by that pattern.

Discussion topic: Try Prosite with the sequence with gi|9506817. Unclick the box that excludes high-probability patterns. What long patterns (above the bar) match? What short patterns (below the bar)? What does it all mean?

E values

Patterns have no E values because patterns are qualitative, not quantitative. Profiles (PSSMs) show E values and HMMs also show E values. Both are quantitative methods and the likelihood of matching by chance can be calculated as it is in BLAST. Recall that HMMs allow for insertions and deletions while maintaining amino acid probabilities. The majority of protein signature databases use HMMs to quantify motifs, domains and families.

Prosite documentation

When a prosite pattern is matched, you see a link beginning with PS that is the accession number for that

pattern. Below is documentation for an aspartyl protease from *Candida albicans*. If you want to duplicate this, do a PROSITE scan with gi number 456252. You see a description, a list of proteins that fit the pattern and in the blue box at the bottom, you see the consensus pattern.

proSite PROSITE documentation PDOC00128 [for PROSITE entry PS00141]

[Home](#) [ScanProsite](#) [ProRule](#) [Documents](#) [Downloads](#) [Links](#) [Funding](#)

Eukaryotic and viral aspartyl proteases signature and profile

Description:

Aspartyl proteases, also known as acid proteases, (EC 3.4.23.-) are a widely distributed family of proteolytic enzymes [1,2,3] known to exist in vertebrates, fungi, plants, retroviruses and some plant viruses. Aspartyl proteases of eukaryotes are monomeric enzymes which consist of two domains. Each domain contains an active site centered on a catalytic aspartyl residue. The two domains most probably evolved from the duplication of an ancestral gene encoding a primordial domain. Currently known eukaryotic aspartyl proteases are:

- Vertebrate gastric pepsins A and C (also known as gastricsin)
- Vertebrate chymosin (rennin), involved in digestion and used for making cheese
- Vertebrate lysosomal cathepsins D (EC 3.4.23.5) and E (EC 3.4.23.36)
- Mammalian renin (EC 3.4.23.15) whose function is to generate angiotensin I from angiotensinogen in the plasma
- Fungal proteases such as asparticopepsin A (EC 3.4.23.18), candidapepsin (EC 3.4.23.24), mucoropepsin (EC 3.4.23.23) (mucor rennin), endothepepsin (EC 3.4.23.22), polyporopepsin (EC 3.4.23.29), and rhizopuspepsin (EC 3.4.23.21)
- Yeast saccharopepsin (EC 3.4.23.25) (protease A) (gene PEP4). PEP4 is implicated in posttranslational regulation of vacuolar hydrolases.
- Yeast barnespepsin (EC 3.4.23.35) (gene BART1), a protease that cleaves a factor and thus acts as an antagonist of the mating pheromone.
- Fission yeast sua 1 which is involved in degrading or processing the mating pheromones.

Most retroviruses and some plant viruses, such as bethnaviruses, encode for an aspartyl protease which is a homodimer of a chain of about 95 to 125 amino acids. In most retroviruses, the protease is encoded as a segment of a polyprotein which is cleaved during the maturation process of the virus. It is generally part of the pol polyproteins and, more rarely, of the gag polyprotein.

Conservation of the sequence around the two aspartyls of eukaryotic aspartyl proteases and around the single active site of the viral proteases allows us to develop a single signature pattern for both groups of protease. A profile was developed to specifically detect viral aspartyl proteases, which are missed by the pattern.

Key:

These proteins belong to families A1 and A2 in the classification of peptidases [4E1]

Last update:

December 2004 / Pattern and test revised.

Technical section:

PROSITE methods (with tools and information) covered by this documentation:

ASP_PROTEASE, PS00141: Eukaryotic and viral aspartyl proteases active site (PS12109)

Consensus pattern: [LIVMFGAC]-[LIVMTADN]-[LIVFSA]-D-[ST]-G-[STAV]-[STAPDENQ]-{GQ}-[LIVMFSTNC]-{EGK}-[LIVMFGTA]
D is the active site residue

Sequences known to belong to this class detected by the pattern: 411

Other sequence(s) detected in Swiss-Prot: 37

• Retrieve an alignment of Swiss-Prot true positive hits

[Clustal format, color, condensed view](#) / [Clustal format, color](#) / [Clustal format, plain text](#) / [Fasta format](#)

► [Produce a dot-plot comparison between this pattern and other patterns](#)

Note consensus pattern

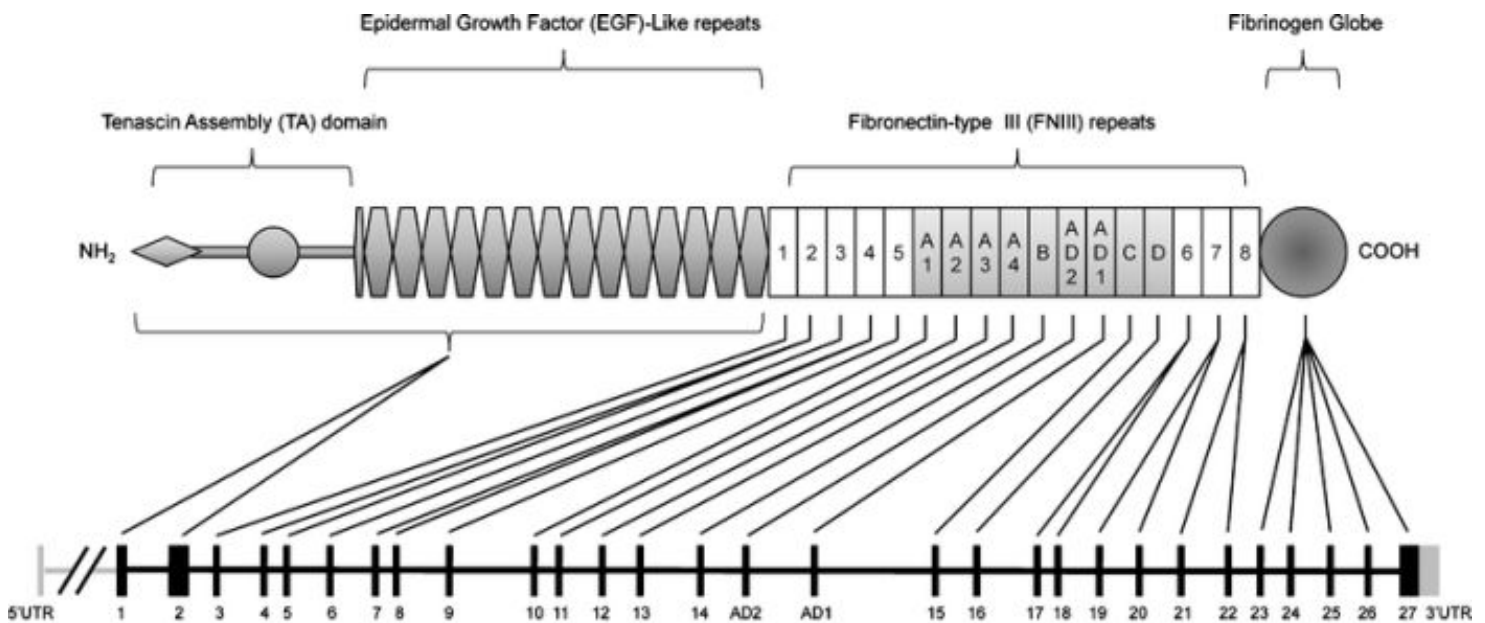
The pattern is [LIVMFGAC]-[LIVMTADN]-[LIVFSA]-D-[ST]-G-[STAV]-[STAPDENQ]-{GQ}-[LIVMFSTNC]-{EGK}-[LIVMFGTA]. The aspartic acid active site is the D at position 4.

You also see links to enzyme commission numbers. Aspartyl proteases begin with 3.4.23 because it's a class of proteases that may cleave proteins at different locations. The fourth number is assigned to specific aspartyl proteases.

Discussion topic: Look at the Prosite documentation for the long pattern matched by gi|9506817. What can you say about the pattern? Did you learn anything about this protein from the documentation.

Multidomain proteins

The diagram below shows a multidomain protein. It's quite common, especially in higher eukaryotes, for a protein to have many domains and even repeats of the same domain. The protein shown is Tenascin-C.



Guttery et al. (2010). *Cancer Metastasis Rev.* 29:595-606

Discussion topic: How do you think the existence of multidomain proteins relates to complexity in higher eukaryotes?

Profiles

HMMs and PSSMs generate quantitative results with E values. PSSM profiles are based on multiple alignment and a position matrix represents the motif. HMM profiles use a probabilistic model to represent a motif. Many databases use HMMs to classify motifs and other protein signatures. Here are some profile databases.

Summary: representing motifs

Pattern: Qualitative - regular expression

PSSM Profile: Quantitative - PSSM based on ungapped alignments

HMM Profile: Quantitative - HMM based on gapped or ungapped alignments

Profile Databases

PRINTS: <http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/>

ProDom: <http://prodom.prabi.fr/prodom/current/html/home.php>

Pfam: <http://pfam.janelia.org/>

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

PRINTS is based on protein fingerprints represented by PSSMs. It's a well-curated database at the University of Manchester in the UK. ProDom is based on PSI-BLAST and automated sequence clustering--essentially a PSSM method. Pfam is quite extensive and maintained by an international consortium of protein researchers. The database contains both the multiple sequence alignments and the HMMs that were generated from those MSAs. There is a curated and automated section. Protein sequences used in the automated section may eventually be promoted to the curated database. SMART (Simple Modular Architecture Research Tool) uses HMMER to build HMM models of protein signatures.

InterPro

InterPro (<http://www.ebi.ac.uk/interpro/>) is one-stop shopping at all of the above databases and PROSITE. [InterProScan](#) allows the user to enter a sequence and search 13 member databases.

Member database information

Signature database	Version	Signatures*	Integrated signatures**
CATH-Gene3D	3.5.0	<u>2626</u>	<u>1717</u>
HAMAP	201311.27	<u>1916</u>	<u>1912</u>
PANTHER	9.0	<u>59948</u>	<u>3399</u>
PIRSF	2.84	<u>3251</u>	<u>3225</u>
PRINTS	42.0	<u>2106</u>	<u>2028</u>
PROSITE patterns	20.97	<u>1308</u>	<u>1291</u>
PROSITE profiles	20.97	<u>1062</u>	<u>1038</u>
Pfam	27.0	<u>14831</u>	<u>14139</u>
PfamB	27.0	<u>20000</u>	<u>0</u>
ProDom	2006.1	<u>1894</u>	<u>1117</u>
SMART	6.2	<u>1008</u>	<u>998</u>
SUPERFAMILY	1.75	<u>2019</u>	<u>1369</u>
TIGRFAMs	13.0	<u>4284</u>	<u>4268</u>

* Some signatures may not have matches to UniProtKB proteins.

** Not all signatures of a member database may be integrated at the time of an InterPro release

Sequence database	Version	Count	Count of proteins matching	
			any signature	integrated signatures
UniProtKB	2014_05	56555610	49736553 (87.9%)	46556539 (82.3%)
UniProtKB/TrEMBL	2014_05	56010222	49205296 (87.9%)	46032040 (82.2%)
UniProtKB/Swiss-Prot	2014_05	545388	531257 (97.4%)	524499 (96.2%)

InterProScan allows you to enter a single sequence and search across 13 databases. The results are color coded based on the matched database. Below is the output from the aspartyl protease used in part 1 of the reading. Note how the colors match the database (click source database on the left menu to see database colors).

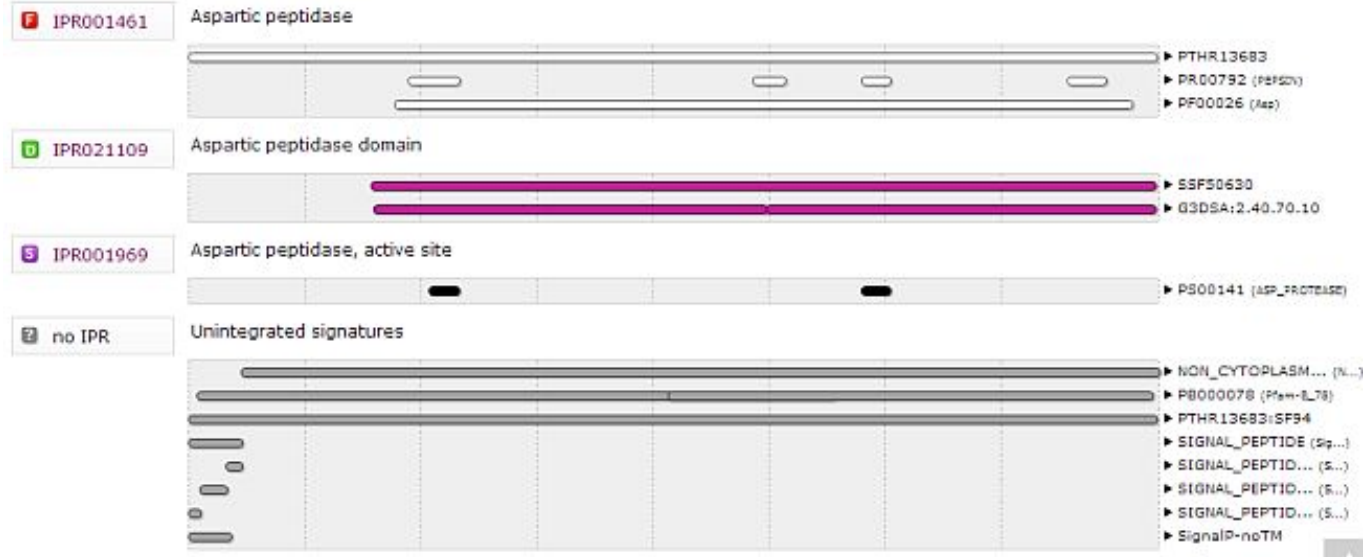
Protein family membership

Aspartic peptidase (IPR001461)

Domains and repeats



Detailed signature matches



The PR link stands for PRINTS; the PTHR for Panther, the PF for Pfam. The PS is for PROSITE, and the SSF for Superfamily. The IPR links at the left of each match link to the Interpro documentation for that motif.

External documentation

Below is the Pfam documentation for Peptidase A1, accessed by clicking on the link that reads "PF00026."

Family: Asp (PF00026)

42 architectures 4258 sequences 4 interactions 462 species 287 structures

Summary

Domain organisation
Clan
Alignments
HMM logo
Trees
Curation & model
Species
Interactions
Structures

Jump to...

Enter Class

Summary: Eukaryotic aspartyl protease

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

[Wikipedia: Aspartate protease](#) [Pfam](#) [InterPro](#)

This is the Wikipedia entry entitled "[Aspartate protease](#)". [More...](#)

Aspartate protease [Edit Wikipedia article](#)

Aspartic proteases are a family of protease enzymes that use an **aspartate** residue for catalysis of their peptide substrates. In general, they have two highly-conserved aspartates in the active site and are optimally active at acidic pH. Nearly all known aspartyl proteases are inhibited by pepstatin.

Aspartic endopeptidases EC 3.4.23. [4](#) of vertebrate, fungal and retroviral origin have been characterised^[2]. More recently, aspartic endopeptidases associated with the processing of bacterial type 4 prepilin^[3] and archaean preflagellin have been described^{[4][5]}. Eukaryotic aspartic proteases include **pepsins**, **cathepsins**, and **renins**. They have a two-domain structure, arising from ancestral duplication. Retroviral and retrotransposon proteases (Pfam PF00077 [4](#)) are much smaller and appear to be homologous to a single domain of the eukaryotic aspartyl proteases. Each domain contributes a catalytic Asp residue, with an extended active site cleft localized between the two lobes of the molecule. One lobe has probably evolved from the other through a gene duplication event in the distant past. In modern-day enzymes, although the three-dimensional structures are very similar, the amino acid sequences are more divergent, except for the catalytic site motif, which is very conserved. The presence and position of disulfide bridges are other conserved features of aspartic peptidases.

Contents [\[hide\]](#)

- 1 Examples
- 2 Catalytic Mechanism
- 3 Subfamilies
- 4 Propeptide
- 5 Human proteins containing this domain
- 6 Inhibition
- 7 External links
- 8 See also
- 9 References

Examples

- HIV-1 protease - a major drug-target for treatment of HIV
- Chymosin (or "rennin")
- Renin

Eukaryotic aspartyl protease



Structures of native and inhibited forms of human cathepsin D.[1]

Identifiers

Symbol	Asp
Pfam	PF00026 4
InterPro	IPR001461 4
PROSITE	PD000128 4
SCOP	1mpp 4
SUPERFAMILY	1mpp 4
OPM family	100 4
OPM protein	1lyb 4

Available protein structures: [show](#)

The documentation for Pfam provided by the Sanger institute has three tabs to toggle between three documentations from Wikipedia, from Pfam and from Interpro. The Pfam documentation is a simple description of the motif with external links. The wikipedia tab provides the wikipedia page with a button that allows the user to edit the wikipedia page.

Other external links from Interpro would take you directly to other databases, such as Prosite and Panther.