

# PROSITE/HAMAP: Sequence Similarity Search With Methods Based on Multiple Sequence Alignments (Patterns and Profiles)

Christian Sigrist

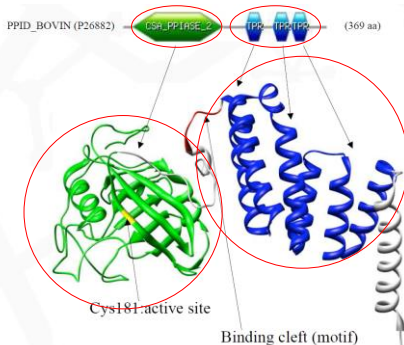


Swiss Institute of  
Bioinformatics

# General Definition on Conserved Regions

Conserved regions in proteins can be classified into 5 different groups:

- **Domains:** specific combination of secondary structures organized into a characteristic three dimensional structure or fold corresponding to a functional unit.

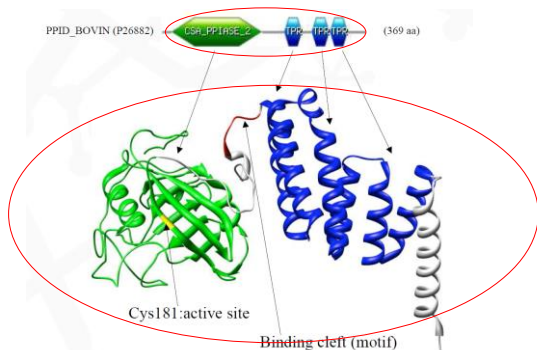


**PPID family:** 1 CSA\_PPIASE (cyclophilin-type peptidyl-prolyl cis-trans isomerase) domain + 3 TPR repeats (tetratricopeptide repeat).

## General Definition on Conserved Regions

Conserved regions in proteins can be classified into 5 different groups:

- **Families:** groups of proteins that have the same domain arrangement or that are conserved along the whole sequence.

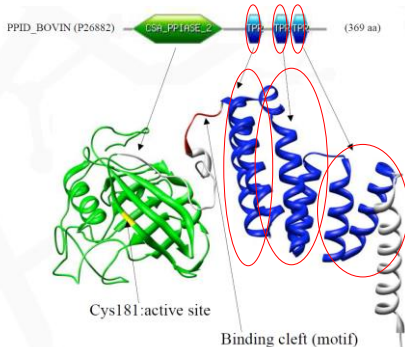


**PPID family:** 1 CSA\_PPIASE (cyclophilin-type peptidyl-prolyl cis-trans isomerase) domain + 3 TPR repeats (tetratricopeptide repeat).

# General Definition on Conserved Regions

Conserved regions in proteins can be classified into 5 different groups:

- **Repeats:** structural units always found in two or more copies that assemble in a specific fold. Assemblies of repeats might also be thought of as domains.

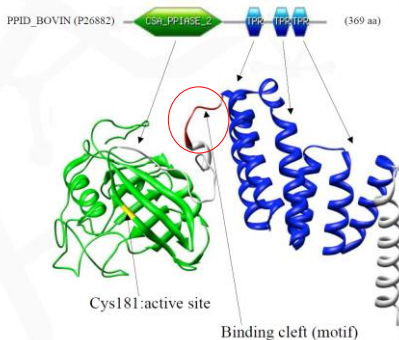


**PPID family:** 1 CSA\_PPIASE (cyclophilin-type peptidyl-prolyl cis-trans isomerase) domain + 3 TPR repeats (tetratricopeptide repeat).

# General Definition on Conserved Regions

Conserved regions in proteins can be classified into 5 different groups:

- **Motifs:** region containing conserved active- or binding-residues or short conserved regions present outside domains that may adopt folded conformation only in association with their binding ligands.

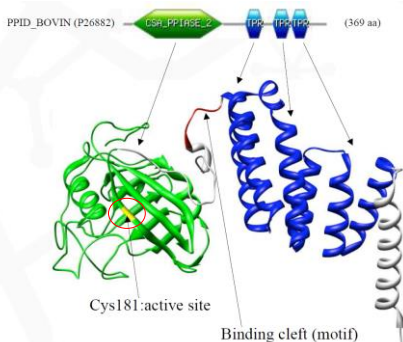


**PPID family:** 1 CSA\_PPIASE (cyclophilin-type peptidyl-prolyl cis-trans isomerase) domain + 3 TPR repeats (tetratricopeptide repeat).

# General Definition on Conserved Regions

Conserved regions in proteins can be classified into 5 different groups:

- **Sites:** functional residues (active sites, disulfide bridges, post-translationally modified residues)



**PPID family:** 1 CSA\_PPIASE (cyclophilin-type peptidyl-prolyl cis-trans isomerase) domain + 3 TPR repeats (tetratricopeptide repeat).

# Sequence identity and similarity

## ➤ Identity

Proportion of pairs of **identical** residues between two aligned sequences.

Generally expressed as a percentage.

This value strongly depends on how the two sequences are aligned.

## ➤ Similarity

Proportion of pairs of **similar** residues between two aligned sequences.

If two residues are similar is determined by a substitution matrix.

So this value depends strongly on the substitution matrix used.

**Sequence similarity searches can identify « homologous » proteins or genes by detecting excess similarity, i.e. statistically significant similarity that reflects common ancestry. Significant similarity is strong evidence that two sequences are related by evolutionary changes from a common ancestral sequence.**

# Sequence homology

**Sequence similarity is the observation, homology is the conclusion.**

## ➤ Homology

Two sequences are homologous if and only if they have a **common ancestor**. There is no percentage of homology! (It's either **yes** or **no**)

- Homologous sequences do not necessarily serve the same function...
- ... Nor are they always highly similar: structure may be conserved while sequence is not.
- **Orthologs** are homologous sequences that are the result of a **speciation event**.
- **Paralogs** are homologous sequences that are the result of a **duplication event**.
- **Xenologs** are homologous sequences that are the result of a **horizontal (or lateral) gene transfer event**.



## Similarity search: The quest of the Grail

- Sequence similarity searching is the most widely used, and most valuable strategy for characterizing newly determined sequences.

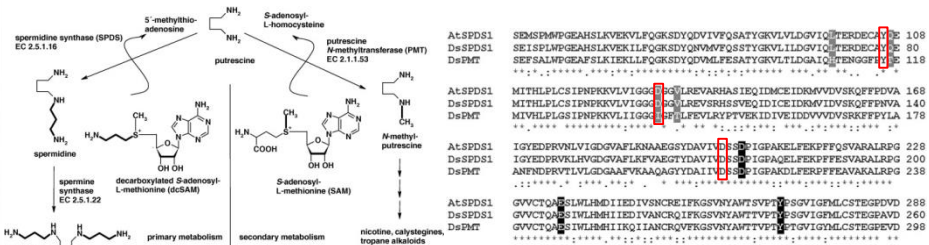
### How to find similarity between sequences?

- There are many traps:
  - Does the similarity reflect an homology or does it result from convergence?
  - Is the alignment the right one?
  - Is it an ortholog or only a paralog?
  - Is the function conserved?
  - ...
- Be careful!



# Sequence Homology vs Functional Convergence

- Homology often provides vital evidence in the prediction of molecular function, but does not necessarily mean that two homologous proteins possess common functions. It only means that they share a common ancestor.
- Ex: GATA zinc fingers, trypsin protease and haptoglobin, spermidine synthase (SPDS) and putrescine N-methyltransferase (PMT)
- PMT sequences are related more closely to those of plant SPDS than to any methyltransferases.



# BLAST

A popular way to identify similarities between proteins is to perform a pairwise alignment (Smith-Watermann, Needleman-Wunsch, BLAST, ...).

Check which part of the query sequence the BLAST retrieves!



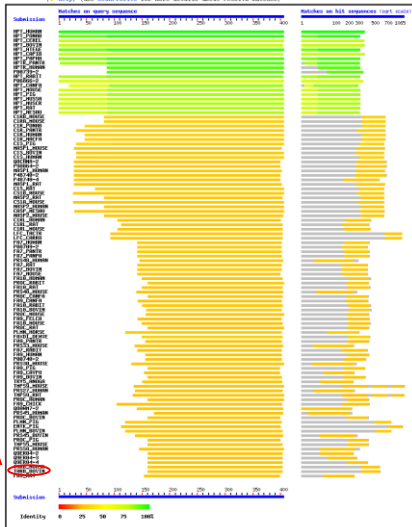
Prothrombin only matches the trypsin domain. The N-ter is completely different.



➔ Redo BLASTs with different parts (domains) of your query protein.

Graphical overview of the alignments

[Click here](#) to re-submit your query after making regions matching PROSITE profiles or Pfam HMMs  
[Help](#) (see [ScanProsite](#) for more details about PROSITE matches)



# Similarity Identification with Pairwise Alignments

- Normally, when the identity is higher than 40% this method gives good results.

```
>SCN2A_HUMAN_IQ repeat  
EEVSAIIIQRAYRRYLLKQKVKKVSSIIYKK
```

↓ Blast  
Fasta

```
sp Q9UQD0 Sodium channel protein type 8 subunit alpha (Sodium channel protein 1980 AA  
SCN8A_HUMAN type VIII subunit alpha) (Voltage-gated sodium channel align  
subunit alpha Nav1.6) [SCN8A] [Homo sapiens (Human)]
```

```
Score = 36.3 bits (78), Expect = 0.025  
Identities = 10/13 (76%), Positives = 12/13 (92%)
```

```
Query: 1 EEVSAIIIQRAYR 13  
EEVSA++ QRAYR  
Sbjct: 1895 EEVSAVVLQRAYR 1907
```

Only the N-ter of the query sequence matches and with a low score!

```
Score = 32.7 bits (67), Expect = 0.13, Method: Composition-based stats.  
Identities = 14/28 (50%), Positives = 21/28 (75%), Gaps = 2/28 (7%)
```

```
Query 1 EEVSAIIQRAYRRYLLKQKV--KKVSS 26  
EEVSA+++QRAYR +L ++ KK +S  
Sbjct 1895 EEVSAVVLQRAYRGLHARRGFICKKTT 1922
```

Even if you manually  
adjust the best  
substitution matrix!

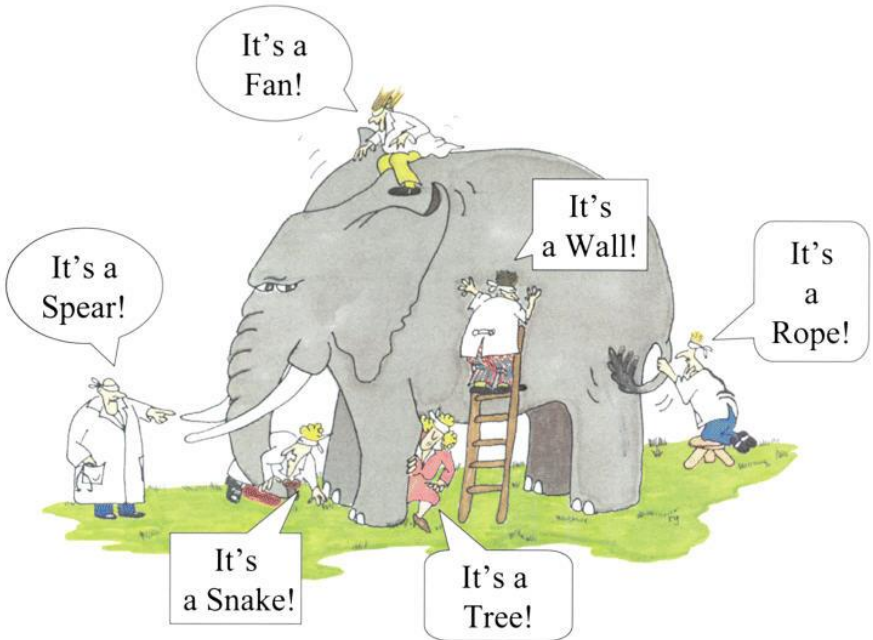
## Pairwise Sequence Alignments vs MSAs

```
          *           20          *
SCN2A_HS : EEVSATIIQRAYRRYLLKQKVKKVSLEYKK : 30
SCN8A_HS : EEVSAVVLQRAYRGHLARRGFICKKTTSNK : 30
          EEVSA666QRAYR L 4           3 K
```

- Another weakness of the pairwise alignment is that no distinction is made between an amino acid at a crucial position (like an active site) and an amino acid with no critical role.

```
          *           20          *
SCN2A_HS   : EEVSATIIQRAYRRYLLKQKVKKVSLEYKK : 30
SCN2A_RN   : EEVSAIVIQRAYRRYLLKQKVKKVSLEYKK : 30
SCN3A_HS   : EEVSAATIQRNRCYLLKQRLKNISSNYNK : 30
SCN8A_HS   : EEVSAVVLQRAYRGHLARRGFICKKTTSNK : 30
SCN8A_MM   : EEVSAVVLQRAYRGHLARRGFICKKTTSNK : 30
IQGA1_HS_1 : NEGLITRLQARCRGYLVRQEFRRSMNFKK : 30
IQGA1_HS_2 : QIPAITCIQSQRGYKQKAYQDRLAYLRS : 30
IQGA1_HS_3 : HKDEVVKIQSLARMHQARKRYRDRQLQYFRD : 30
IQGA1_HS_4 : HINDIIKIQAFIRANKARDDYKTLINAEDP : 30
IQGA1_MM_1 : NEGLITKLCACCRGYLVRQEFRRSMNFKK : 30
          6Q R 4
```

- A multiple sequence alignment (MSA) gives a more general view of a conserved region by providing a better picture of the most conserved residues, which are usually essential for the protein function. It can help to identify subfamilies. An MSA contains more information than a pairwise alignment and several tools have been developed to extract this information.



## Extracting Information from MSAs

Several models based on multiple sequence alignment have been developed in order to identify conserved regions (patterns, PSSMs, fingerprints, generalized profiles/HMMs). A search performed with such models is generally more sensitive than a pairwise alignment and can help identify very remote similarity (less than 20% of identity). They also offer a better alignment of important residues:

- **Consensus:** patterns / regular expressions
- **Profile:** weight matrices



# Patterns



# PROSITE patterns

- PROSITE patterns use a special syntax to describe the consensus of all the sequences present in the multiple alignment using a single expression.
- Used to describe small functional regions:
  - Enzyme catalytic sites;
  - Prosthetic group attachment sites (heme, PLP, biotin, *etc.*);
  - Amino acids involved in binding a metal ion;
  - Cysteines involved in disulfide bonds;
  - Regions involved in binding a molecule (ATP, DNA, *etc.*) or a protein.
- Excellent tool to annotate active sites in combination with profiles (ProRules).

# How to build a PROSITE pattern

```
sw:BMP1_HUMAN/547-572  EVDECS--RPNRGGCT--EQRCCLNTLGSYKC
sw:C1QR1_RAT/424-446   DIDECL-----GNPCDTLCINTDGSFRC
sw:CUBN_CANFA/167-196 DVNECQIYSGTPLG CQNGATCENTAGSYSC
sw:EGFL6_XENLA/180-206 DIDECA---VGKASCPINRRCVNTFGSYYC
sw:FBLN1_CAEEL/390-413 DVNECQ-----QGVCGSM ECLINLPGTYSKC
```

Pattern: [DE]-[VI]-[DN]-E-C-x(1,8)-[GS]-x(4,6)-C-x-N-[TL]-x-G-[ST]-[YF]-x-C

- Collect sequences known to contain the signature and produce a multiple sequence alignment of the region of interest.
- Build a pattern.
  - By hand
  - You can use automatic methods (e.g. <http://web.expasy.org/pratt/>) or a sequence logo to guide you

# How to build a PROSITE pattern

```

sw:BMP1_HUMAN/547-572  EVDECS--RPNRGGC--EQRCLENTLGSYKC
sw:C1QR1_RAT/424-446   DIDECL-----GNPCDTLCINTDGSFRC
sw:CUBN_CANFA/167-196 DVNECQIYSGTPLG CQNGATCENTAGSYSC
sw:EGFL6_XENLA/180-206 DIDECA---VGKASCPINRRCVNTFGSYYC
sw:FBLN1_CAEEL/390-413 DVNECQ-----QGVCGSMELINLPGTYKC
  
```

Pattern: [DE]-[VI]-[DN]-E-C-x(1,8)-[GS]-x(4,6)-C-x-N-[TL]-x-G-[ST]-[YF]-x-C

- Example using a sequence logo  
(<http://weblogo.berkeley.edu/logo.cgi>):



weblogo.berkeley.edu

# PROSITE patterns: the full syntax

- aa are represented by a single letter code (e.g. S)
- each position is separated by a dash '-' (e.g. S-P-R)
- 'X' represents any aa (e.g. S-X-R)
- '[]' group of aa accepted for a position (e.g. [ST]-X-[RK])
- '{}' group of aa not accepted for a position (e.g. [ST]-{PG}-[RK])
- '()' repetitions

Examples:

x(3) corresponds to x-x-x

x(2,4) corresponds to x-x or x-x-x or x-x-x-x

A(3) corresponds to A-A-A

Note: You can only use a range with 'x', i.e. A(2,4) is not a valid pattern element.

- '<' anchor at the N-term
- '>' anchor at the C-term

## PROSITE patterns syntax example

- Pattern: <M-X(0,1)-[ST](2)-X-{V}
- Regexp: `^M.?[ST]{2}.[~V]`
- Text:
  - The sequence must start with a methionine,
  - followed by any aa or nothing,
  - followed by a serine or threonine twice,
  - followed by any aa,
  - followed by any aa except a valine.

# Tricks to build a PROSITE pattern

```
sw:BMP1_HUMAN/547-572  EVDECS--RPNRGG C--EQRC LNTLG SYKC
sw:C1QR1_RAT/424-446   DIDECL-----GNPCDTLC INTDGSFRC
sw:CUBN_CANFA/167-196 DVNECQIYSGTPLG CQNGATCENTAGSYSC
sw:EGFL6_XENLA/180-206 DIDECA---VGKASCP INRRCVNTFGSYYC
sw:FBLN1_CAEEL/390-413 DVNECQ-----QGVCGSM E C I N L P G T Y K C
```

Pattern: [DE]-[VI]-[DN]-E-C-x(1,8)-[GS]-x(4,6)-C-x-N-[TL]-x-G-[ST]-[YF]-x-C

- For the construction of the pattern, it is useful to consider residues and regions proved/thought to be important to the biological function of that group of proteins (*e.g.* enzyme catalytic sites, *etc.*).
- A first pattern is built from the MSA of the most conserved residues. It is used to scan the database.
- If it picks up too many false positives, it is modified to make it more stringent.
- The difficulty resides in achieving a pattern which does not pick up too many false positives yet does not miss too many sequences (false negatives).
- In some cases this result can not be achieved and an optimal sequence pattern can not be built.

## How to Estimate the Quality of a Pattern

- We can not estimate the quality of a match with a pattern:  
PATTERNS don't produce a score, they match or not!
- But we can estimate the quality of the pattern.
- Two parameters can be computed to estimate the quality of a pattern: **precision** and **recall**.
  - False positives** = known false hits.
  - False negatives** = known missed hits.
  - Precision** = true hits/(true hits + false positives).
    - Precision = 1  $\Rightarrow$  no false positive.
    - Precision = 0.8  $\Rightarrow$  20% false positives.
  - Recall** = true hits/(true hits + false negatives).
    - Recall = 1  $\Rightarrow$  No missed hits.
    - Recall = 0.8  $\Rightarrow$  20% missed hits.
- To obtain these measures we require a well annotated protein databases (PROSITE uses UniProtKB/Swiss-Prot).

# PROSITE patterns: example of a pattern entry

[Home](#) | [ScanProSite](#) | [ProRule](#) | [Documents](#) | [Downloads](#) | [Links](#) | [Funding](#)



Entry: **PS00972**

## General information about the entry

Entry name <a href="#">[info]</a>	USP_1
Accession <a href="#">[info]</a>	PS00972
Entry type <a href="#">[info]</a>	PATTERN
Date <a href="#">[info]</a>	JUN-1994 (CREATED); DEC-2013 (DATA UPDATE); APR-2015 (INFO UPDATE).
PROSITE Doc. <a href="#">[info]</a>	PDOC00750
Associated ProRule <a href="#">[info]</a>	PRU10092

## Name and characterization of the entry

Description <a href="#">[info]</a>	Ubiquitin specific protease (USP) domain signature 1.
Pattern <a href="#">[info]</a>	$G-[LIVMFY]-x(1,3)-[AQC]-[NARWQ]-x-C-[FTWC]-[LIVMFCA]-[NSTAD]-[SACV]-x-[LIVRWF]-[QF]-.$

Active site

## Numerical results [\[info\]](#)

Numerical results for UniProtKB/Swiss-Prot release **2015\_06** which contains **548'586** sequence entries.

Total number of hits	282 in 282 different sequences
Number of true positive hits	282 in 282 different sequences
Number of 'unknown' hits	0
Number of false positive hits	0
Number of false negative sequences	26
Number of 'partial' sequences	1
Precision (true positives / (true positives + false positives))	100.00 %
Recall (true positives / (true positives + false negatives))	90.97 %

Number of true positives

Number of false positives

Number of false negatives

## Comments [\[info\]](#)

Taxonomic range <a href="#">[info]</a>	Eukaryotes, Eukaryotic viruses
Maximum number of repetitions <a href="#">[info]</a>	1
Site <a href="#">[info]</a>	active_site at position 7
Version <a href="#">[info]</a>	1



## Limitations of PROSITE pattern

```
sw:BMP1_HUMAN/547-572  EVDECS--RPNRGG C--EQRC LNTLG SYKC
sw:C1QR1_RAT/424-446   DIDECL-----GNPCDTLC INTDGSFRC
sw:CUBN_CANFA/167-196 DVNECQIYSGTPLG CQNGATCENTAGSYSC
sw:EGFL6_XENLA/180-206 DIDECA---VGKASCP INRRCVNTFGSYYC
sw:FBLN1_CAEEL/390-413 DVNECQ-----QGVCGSM E C I N L P G T Y K C
```

Pattern: [DE]-[VI]-[DN]-E-C-x(1,8)-[GS]-x(4,6)-C-x-N-[TL]-x-G-[ST]-[YF]-x-C

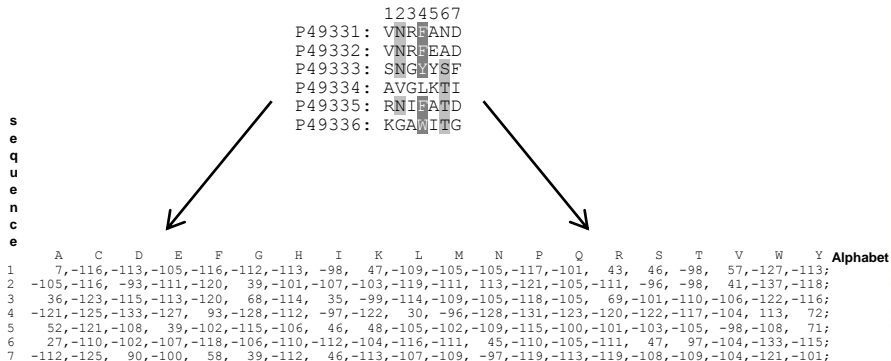
- OK to detect and annotate very conserved regions, but poor gap models
- residues at one position are considered equivalent in their frequencies
- Patterns are not predictive: if a symbol is not present at one position, this will exclude variants that have not yet been observed from being detected
- no score of the match is produced (you match or not)



# Position Specific Scoring Matrix (PSSM)

# Position Specific Scoring Matrix (PSSM)

A PSSM or a profile is based on the frequencies of each residue at a specific position in an MSA. The MSA is converted into a matrix where a **score** is given to each amino acid at each position of the MSA according to the observed frequency (positive scores for expected amino acids and negative scores for unexpected ones).



# Construction of a PSSM (1)

**First step:** weight sequences.

- When constructing a PSSM from an MSA it is a mistake to give all sequence of the alignment the same weight.
- A large set of closely related sequences carries little more information than a single member, but it will drastically influence the score of each amino acid at each position and decrease the influence of divergent sequences.
- To counteract this effect it is important to weight sequences, with those having many close relatives receiving smaller weight.

## Construction of a PSSM: Weight Sequences (1)

1	2	3	4	5	6	7	
A	S	T	A	M	P	V	W=0.25
A	T	S	L	M	V	T	W=0.25
S	S	S	L	M	L	T	W=0.25
A	T	P	A	M	S	S	W=0.25
A	T	A	L	L	S	A	W=0.125
A	T	A	L	L	S	A	W=0.125

- To compensate for this sampling bias, we can use **sequence weighting algorithms**, e.g.:
- based on phylogenetic tree: Gerstein, Sonnhammer and Chotia (GSC)
  - based on Voronoï algorithm: Sibbald and Argos

## Construction of a PSSM (2)

**2nd step:** count the number of occurrence of the different amino acids (or bases) at each position of the alignment

1	2	3	4	5	6	7
A	S	T	A	M	P	V
A	T	S	L	M	V	T
S	S	S	L	M	L	T
A	T	P	A	M	S	S
A	T	A	L	L	S	A

1	2	3	4	5	6	7
4a	3t	2s	3l	4m	2s	1v
1s	2s	1t	2a	1l	1l	2t
		1a			1v	1s
		1p			1p	1a

## Construction of a PSSM (3)

**3rd step:** derivation of the preliminary frequency matrix

	1	2	3	4	5	6	7
4a	3t	2s	3l	4m	2s	1v	
1s	2s	1t	2a	1l	1l	2t	
		1a			1v	1s	
		1p			1p	1a	

	1	2	3	4	5	6	7
A	0.8	0	0.2	0.4	0	0	0.2
L	0	0	0	0.6	0.2	0.2	0
M	0	0	0	0	0.8	0	0
V	0	0	0	0	0	0.2	0.2
P	0	0	0.2	0	0	0.2	0
S	0.2	0.4	0.4	0	0	0.4	0.2
T	0	0.6	0.2	0	0	0	0.4

## Construction of a PSSM (4)

### 4th step: correction of the sample bias.

- An MSA represent a sample of all proteins that contain such a conserved region, thus a **sample bias** can be observed: not all possibilities are represented in the MSA: some observed frequencies are equal 0 and thus will exclude the corresponding amino acid at this position (like in patterns).
- To circumvent this problem, one possibility is to add a small number to all observed frequencies, **pseudo-counts** to avoid null frequencies.
- A more elegant way is to modulate the pseudo-count for conservative substitutions using **substitution matrices** or **dirichlet mixtures**.
- The number of sequences in the MSA is also important. If there are a lot of sequences there is less sample bias and thus pseudo-count are less important.

(Usually logarithms of 'corrected' frequencies are used so as to speed up computation time).



## Pseudo-counts

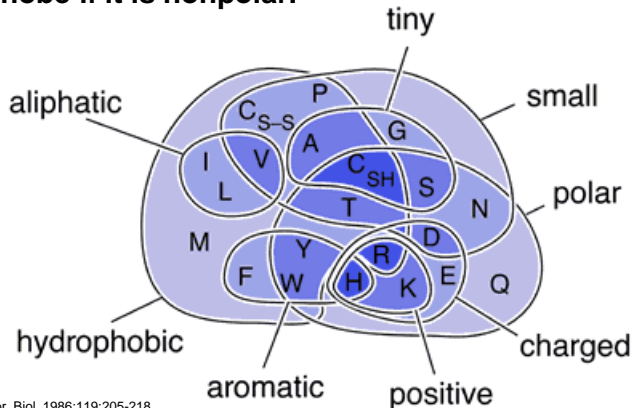
- For example we add 0.1 to all counts of the previous matrix and re-normalize to obtain frequencies:

$$f(A, 1) = \frac{0.8 + 0.1}{1 + 2} = 0.3, f(C, 1) = \frac{0.0 + 0.1}{1 + 2} = 0.0333, \dots, f(A, 7) = \frac{0.2 + 0.1}{1 + 2} = 0.1$$

	1	2	3	4	5	6	7
A	0.300	0.033	0.100	0.166	0.033	0.033	0.100
C	0.033	0.033	0.033	0.033	0.033	0.033	0.033
D	0.033	0.033	0.033	0.033	0.033	0.033	0.033
E	0.033	0.033	0.033	0.033	0.033	0.033	0.033
F	0.033	0.033	0.033	0.033	0.033	0.033	0.033
G	0.033	0.033	0.033	0.033	0.033	0.033	0.033
H	0.033	0.033	0.033	0.033	0.033	0.033	0.033
I	0.033	0.033	0.033	0.033	0.033	0.033	0.033
K	0.033	0.033	0.033	0.033	0.033	0.033	0.033
L	0.033	0.033	0.033	0.233	0.100	0.100	0.033
M	0.033	0.033	0.033	0.033	0.300	0.033	0.033
N	0.033	0.033	0.033	0.033	0.033	0.033	0.033
P	0.033	0.033	0.100	0.033	0.033	0.100	0.033
Q	0.033	0.033	0.033	0.033	0.033	0.033	0.033
R	0.033	0.033	0.033	0.033	0.033	0.033	0.033
S	0.100	0.166	0.166	0.033	0.033	0.166	0.100
T	0.033	0.233	0.100	0.033	0.033	0.033	0.166
V	0.033	0.033	0.033	0.033	0.033	0.100	0.100
W	0.033	0.033	0.033	0.033	0.033	0.033	0.033
Y	0.033	0.033	0.033	0.033	0.033	0.033	0.033

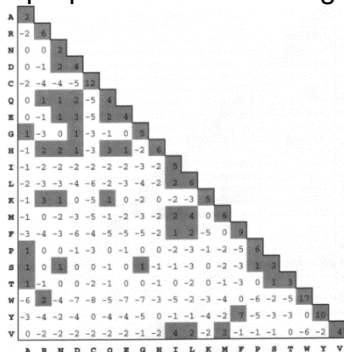
## Amino acid classification

- Amino acid side chains vary in size, shape, charge, hydrogen-binding capacity and chemical reactivity.
- The side-chain can make an amino acid a weak acid or a weak base, and a hydrophile if the side-chain is polar or a hydrophobe if it is nonpolar.



# Substitution Matrices

- “All amino acids are equal, but some amino acids are more equal than others.” Inspired from Georges Orwell
- In proteins some mismatches are more acceptable than others.
- Substitution matrices give a score for each substitution of one amino acid by another. These sets of numbers describe the propensities of exchanging one amino acid for another.



**Positive score:** the amino acid are similar.

(Mutations from one aa into the other occur more often than expected by chance during evolution).

**Negative score:** the amino acids are dissimilar.

(Mutations from one amino acids into the other occur less often than expected by chance during evolution).

- Examples: PAM, blosum, gonnet.

# Search a Database With a PSSM

- The sequence (MCFVNRFYSFCMP) is aligned to the PSSM:

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
M																					
C																					
F																					
V	1	12	-41	-20	5	-25	-42	-18	-18	33	-12	-12	-19	-41	42	9	2	9	16	-61	-11
N	2	-23	-54	-5	-24	-37	-19	-45	-3	7	-35	-38	59	-41	-12	-42	10	65	-17	-68	-15
R	3	-13	-62	-14	4	-53	78	-36	-65	-15	-64	-49	-14	-48	9	5	-10	-11	-63	-61	-42
F	4	-36	-68	-63	-36	60	-63	-38	-14	-47	3	-21	-52	-53	-34	-58	-39	-45	-26	138	36
Y	5	-22	-60	-54	-24	6	-43	0	30	13	0	-22	-27	-59	55	-9	-38	-11	37	-57	12
S	6	-35	-46	-18	14	-9	-51	-12	-19	34	-39	-28	36	-45	44	-9	-3	41	-27	-24	17
F	7	-33	-58	37	-6	-16	-39	-21	61	-23	-1	-28	-6	-58	-17	-54	-20	-9	14	-12	11
C																					
M																					
P																					

- Searching algorithm: sliding windows. At each position of the sliding window the score is obtained by summing the score of all columns
- Best score:  $16+59+5+60+12-3-16=133$

# Search a Database With a PSSM

- The sequence (MCFVNRFYSF) is aligned to the PSSM:

	A	C	D	E	F	G	H	I	K	L	M	N	Q	R	S	T	V	W	Y		
M																					
C																					
F																					
V	1	12	-41	-20	5	-25	-42	-18	-18	33	-12	-14	-19	-41	42	9	2	9	16	-61	-11
N	2	-23	-54	-5	-24	-37	-19	-45	-3	7	35	38	59	-41	-12	-42	10	65	-17	-68	-15
R	3	-13	-62	-14	4	-53	78	-36	-65	-15	34	-49	-14	-48	9	5	-10	-11	-63	-61	-42
F	4	-36	-68	-63	-36	60	-63	-38	-14	-47	3	-21	-52	-53	-34	-58	-39	-45	-26	138	36
Y	5	-22	-60	-54	-24	6	-43	0	10	13	0	-22	-27	-59	55	-9	-38	-11	37	-57	12
S	6	-35	-46	-18	14	-9	-51	-12	-19	34	-39	-28	36	-45	44	-9	-3	41	-27	-24	17
F	7	-33	-58	37	-6	-16	-36	-21	61	-23	-1	-28	-6	-58	-17	-54	-20	-9	14	-12	11
C																					
M																					
P																					

where do I put the cut off?

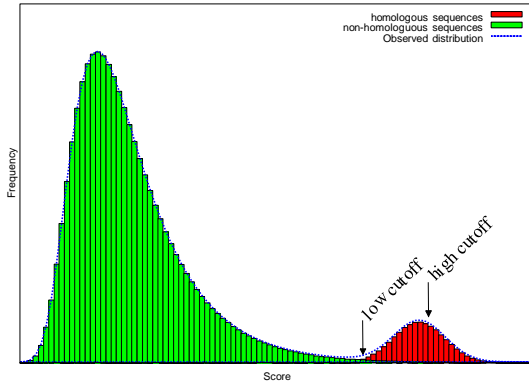
- Searching algorithm: sliding windows. At each position of the sliding window the score is obtained by summing the score of all columns
- Best score:  $16+59+5+60+12-3-16=133$

## Interpretation of the score

- How do I interpret the score produced by a profile? Which is the lowest score I consider to produce a true match?
- Only biological arguments tell you if a match is true or not.
- However, a statistical analysis can help us decide if a match is statistically significant (true positive) or not (false positive).

# Scores follow an EVD distribution

- The score distribution of a profile on **unrelated sequences** is approximated by an Extreme Value Distribution (EVD) (**green bars**).
- This property permits to calculate the **E-value**: the number of matches that we expect to occur by chance with a score  $\geq$  a given cut off.



# Advantages and limitations of PSSM

## **Advantages:**

- The score produced permits to estimate the quality of the match produced.
- Can be used to model short motifs.
- The method is relatively fast and simple to implement.

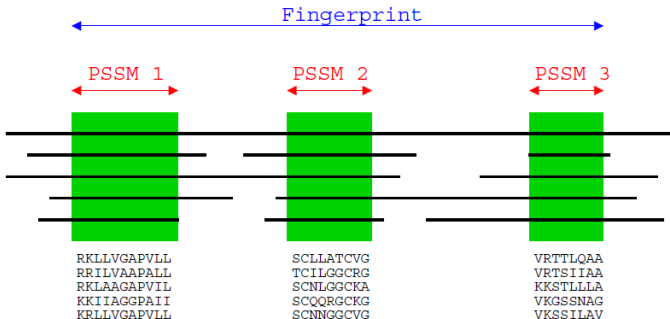
## **Limitations:**

- Insertions and deletions (Indels) are forbidden: long regions cannot be described.



# PSSM: Fingerprints

- To overcome the gap limitation of PSSMs (missing gap model), two or more PSSMs can be used to describe long regions. The combination of various PSSMs is called fingerprints.



- The PRINTS database is a collection of annotated fingerprints.



# Generalized profiles

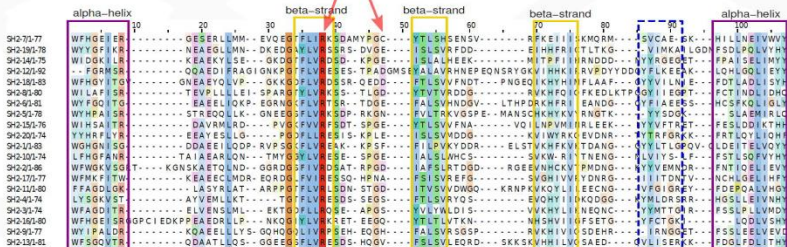
# Modeling of Profiles

	alpha-helix		beta-strand		beta-strand		beta-strand		alpha-helix															
SH2-71-77	WFHGEIER	10	.....	20	GESERLMM	30	EVOEG	40	FLIKSDAMYPGC	50	YTLSHSENSV	60	.....	70	RKFEIISKMRM	80	SVCAE-GK	90	.....	100	HILLNIEVWVY			
SH2-191-78	WYGFIKR	.....	NEAEG	LMN	DKEDG	YLVR	SSRS	DVGE	.....	ISLSVRFDD	.....	EIHFR	ICTLTKG	.....	VIMKAL	LGDM	FSDL	PQLV	YVHY	.....	FPAISELIMY			
SH2-141-75	WIDGKILR	.....	KEAEK	LSE	GKDD	FLVR	DSD	KPGE	.....	ISLALHEEK	.....	MITPF	IIRNDD	.....	YYRGE	ST	.....	.....	.....	.....	.....			
SH2-121-92	--FGMR	STR	.....	QQAEDI	FRAGIK	KNPQ	FLVR	ESES	TPADGMS	YALAVRHNEP	EQNSRY	GKVI	IHHK	IRV	PDYDDQ	YFLK	EEK	.....	.....	.....	LOHLG	LI EY		
SH2-181-83	WFHGYITG	.....	GNEAE	YQLVP	GKKK	FLVR	DSSR	QEDD	.....	FTLSVVFNDT	.....	PNGEQ	IKHYI	FLAAF	.....	YVYILN	E	.....	.....	.....	FDTLAD	LISYH		
SH2-81-80	WILAFISR	.....	TEVPL	LLLEI	SPARG	FLVR	KSS	TLGD	.....	YTVTVRDDG	.....	RVKHF	QIK	FKED	LKTP	GYII	IEGT	.....	.....	.....	FCTIN	DLIDHC		
SH2-81-81	WYFGITG	.....	EAEEL	IQKP	EGRNG	FLVR	TSR	TDGE	.....	FALSVHNDGV	.....	LTHPR	RKHFR	EANDG	.....	YFIAEE	SS	.....	.....	.....	HCSFK	QLIGL		
SH2-51-78	WYHPAISR	.....	STREQQ	LLK	GNEEG	FLVR	KSDP	RKGN	.....	FVLTAKVGSPE	MANSC	HKYKV	IRNGTK	.....	YYSDGK	.....	.....	.....	.....	.....	.....	SLAEM	IRLQ	
SH2-151-76	WIHSALTR	.....	DAVRM	LRD	PVG	FVVR	FSDT	SPGE	.....	YTLVSVFNA	.....	VQILNP	VMIRLEEK	.....	YVYVFTRE	.....	.....	.....	.....	.....	.....	FESLDD	KTHP	
SH2-201-74	YYHRFLYR	.....	EEAYE	ESLLO	P	DFLL	ESIS	KPLE	.....	ISLSVMDDG	.....	KVIW	FRKQ	EVDNR	.....	YTRFGR	K	.....	.....	.....	.....	FATLQ	YLQH	
SH2-171-83	WQWGNISG	.....	DDAEE	IQQD	RVPSS	FLVR	EAK	KPSF	.....	FILPWYDDR	.....	ELST	VKHF	TDANG	.....	YVLT	LGPOV	.....	.....	.....	.....	GDEIT	ELQYN	
SH2-101-74	LFHGFANR	.....	TAIAEAR	LQW	THYG	FLVRE	ESE	SPGE	.....	IALLSLWHCS	.....	SKKW	RITRENG	.....	FLVIYS	LF	.....	.....	.....	.....	.....	FSTLS	QPVYH	
SH2-211-86	WFWGKVSQ	T	.....	KGNSK	AETQ	LND	GGRDG	FVIVDSAT	RPQD	IAFSLRTDGD	.....	RGEEV	NHCKV	PHNDG	.....	YVYEMN	DR	.....	.....	.....	.....	FNTIQ	ELLEV	
SH2-171-77	WFMKFTW	.....	KEAEE	CLMDR	EQRDG	FVIRE	SSQ	HPNA	.....	FVSVRFEFG	.....	SVGH	VIVYDNRG	.....	IITDNY	V	.....	.....	.....	.....	.....	NCHL	GELHFY	
SH2-111-80	FFAGDLGK	.....	LASYR	LAT	ARPPG	FLVLR	LSDN	STGD	.....	ITVSVVDWQO	.....	KRNPK	VQYLI	EECNG	.....	VVFG	IGREY	.....	.....	.....	.....	FDEP	QALVHG	
SH2-41-74	LYSGKNST	.....	AYVEM	LKT	TG	FLVRES	SDS	SEGS	.....	FTLSVRYQS	.....	VOHY	IIDKQDGG	.....	YMLDR	SRR	.....	.....	.....	.....	.....	HGS	LLLVND	
SH2-31-74	WFGADISR	.....	ELVENS	LML	EKQ	DFLL	QSE	APGS	.....	YVLYALDIS	.....	VVKHY	LIN	NEQC	.....	YVMTG	R	.....	.....	.....	.....	FSSLP	LVNDY	
SH2-161-80	WFHGEISR	GPCI	EDKPP	EAEDR	LLP	NKGG	YLVR	KRET	EEQQ	YVLT	LVTKN	.....	NHSH	VIFSETG	.....	YFTG	GK	.....	.....	.....	.....	.....	LDQL	VSHY
SH2-91-77	WYIPALDR	.....	KQAE	ELL	LYS	GQHQQ	LIVP	SEH	EQGH	FALSVRSGSP	.....	RVKH	VIS	SDENR	.....	IRNG	ST	.....	.....	.....	.....	FSSLE	EELVEQ	
SH2-131-81	WFSQGVTR	.....	QDAAT	LQSS	GGEEG	FLVRES	SDS	HQGV	.....	FVLSVLEQRD	.....	SKKSK	VHHIL	VCSAED	.....	VLSER	K	.....	.....	.....	.....	.....	FDGL	FDLITH



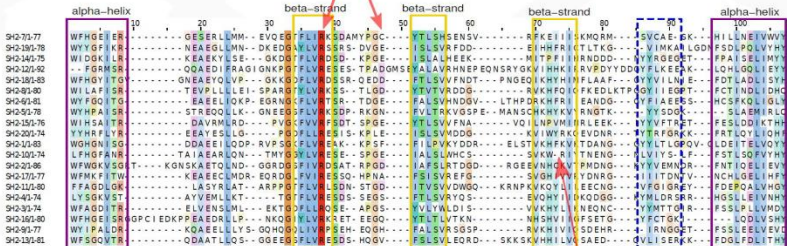
# Modeling of Profiles

model highly conserved columns vs columns of low conservation

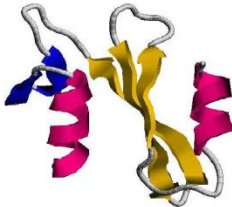


# Modeling of Profiles

model highly conserved columns vs columns of low conservation

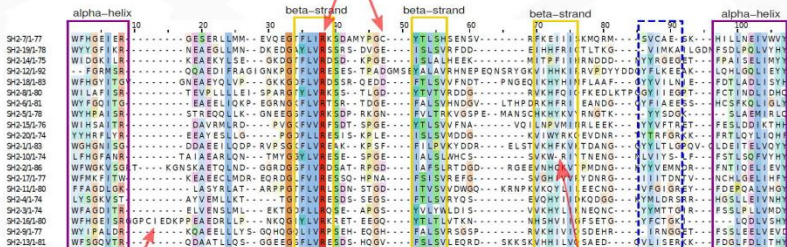


model deletions



# Modeling of Profiles

model highly conserved columns vs columns of low conservation

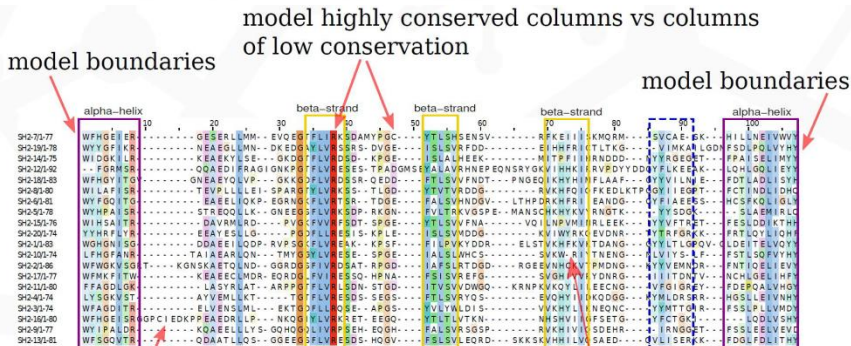


model insertions

model deletions



# Modeling of Profiles



## Build profiles: a pet example

A	S	T	A	M	P	V
A	T	S	L	M	V	T
S	S	S	L	M	L	T
A	T	P	A	M	S	S
A	T	A	L	L	S	A
A	T	A	L	L	S	A

- Sequence weighting: correct sampling bias.
- Residue counts: get the frequency of each residue at each position of the MSA.
- Pseudo-counts: avoid frequencies of 0  $\Rightarrow$  avoid exclusion of residues.
- Build the final scoring matrix: used to build and score alignments.



# Build profiles: model gaps

	1	2	3	4	5	6	7
A	-16	-4	6	10	-4	-4	6
C	-4	-4	-4	-4	-4	-4	-4
D	-4	-4	-4	-4	-4	-4	-4
E	-4	-4	-4	-4	-4	-4	-4
F	-4	-4	-4	-4	-4	-4	-4
G	-4	-4	-4	-4	-4	-4	-4
H	-4	-4	-4	-4	-4	-4	-4
I	-4	-4	-4	-4	-4	-4	-4
K	-4	-4	-4	-4	-4	-4	-4
L	-4	-4	-4	13	6	6	-4
M	-4	-4	-4	-4	16	-4	-4
N	-4	-4	-4	-4	-4	-4	-4
P	-4	-4	6	-4	-4	6	-4
Q	-4	-4	-4	-4	-4	-4	-4
R	-4	-4	-4	-4	-4	-4	-4
S	6	10	10	-4	-4	10	6
T	-4	13	6	-4	-4	-4	10
V	-4	-4	-4	-4	-4	6	6
W	-4	-4	-4	-4	-4	-4	-4
Y	-4	-4	-4	-4	-4	-4	-4

# Build profiles: model gaps

A	-16	-4	6	10	-4	-4	6
C	-4	-4	-4	-4	-4	-4	-4
D	-4	-4	-4	-4	-4	-4	-4
E	-4	-4	-4	-4	-4	-4	-4
F	-4	-4	-4	-4	-4	-4	-4
G	-4	-4	-4	-4	-4	-4	-4
H	-4	-4	-4	-4	-4	-4	-4
I	-4	-4	-4	-4	-4	-4	-4
K	-4	-4	-4	-4	-4	-4	-4
L	-4	-4	-4	13	6	6	-4
M	-4	-4	-4	-4	16	-4	-4
N	-4	-4	-4	-4	-4	-4	-4
P	-4	-4	6	-4	-4	6	-4
Q	-4	-4	-4	-4	-4	-4	-4
R	-4	-4	-4	-4	-4	-4	-4
S	6	10	10	-4	-4	10	6
T	-4	13	6	-4	-4	-4	10
V	-4	-4	-4	-4	-4	6	6
W	-4	-4	-4	-4	-4	-4	-4
Y	-4	-4	-4	-4	-4	-4	-4

# Build profiles: model gaps

M1 M2 M3 M4 M5 M6 M7

A	-16	-4	6	10	-4	-4	6
C	-4	-4	-4	-4	-4	-4	-4
D	-4	-4	-4	-4	-4	-4	-4
E	-4	-4	-4	-4	-4	-4	-4
F	-4	-4	-4	-4	-4	-4	-4
G	-4	-4	-4	-4	-4	-4	-4
H	-4	-4	-4	-4	-4	-4	-4
I	-4	-4	-4	-4	-4	-4	-4
K	-4	-4	-4	-4	-4	-4	-4
L	-4	-4	-4	13	6	6	-4
M	-4	-4	-4	-4	16	-4	-4
N	-4	-4	-4	-4	-4	-4	-4
P	-4	-4	6	-4	-4	6	-4
Q	-4	-4	-4	-4	-4	-4	-4
R	-4	-4	-4	-4	-4	-4	-4
S	6	10	10	-4	-4	10	6
T	-4	13	6	-4	-4	-4	10
V	-4	-4	-4	-4	-4	6	6
W	-4	-4	-4	-4	-4	-4	-4
Y	-4	-4	-4	-4	-4	-4	-4

# Build profiles: model gaps

M1	M2	M3	M4	M5	M6	M7
D1	D2	D3	D4	D5	D6	D7

A	-16	-4	6	10	-4	-4	6
C	-4	-4	-4	-4	-4	-4	-4
D	-4	-4	-4	-4	-4	-4	-4
E	-4	-4	-4	-4	-4	-4	-4
F	-4	-4	-4	-4	-4	-4	-4
G	-4	-4	-4	-4	-4	-4	-4
H	-4	-4	-4	-4	-4	-4	-4
I	-4	-4	-4	-4	-4	-4	-4
K	-4	-4	-4	-4	-4	-4	-4
L	-4	-4	-4	13	6	6	-4
M	-4	-4	-4	-4	16	-4	-4
N	-4	-4	-4	-4	-4	-4	-4
P	-4	-4	6	-4	-4	6	-4
Q	-4	-4	-4	-4	-4	-4	-4
R	-4	-4	-4	-4	-4	-4	-4
S	6	10	10	-4	-4	10	6
T	-4	13	6	-4	-4	-4	10
V	-4	-4	-4	-4	-4	6	6
W	-4	-4	-4	-4	-4	-4	-4
Y	-4	-4	-4	-4	-4	-4	-4
Del	-d1	-d2	-d3	-d4	-d5	-d6	-d7

# Build profiles: model gaps

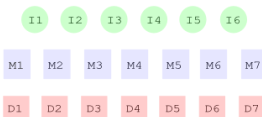


Diagram illustrating model gaps with negative values. The top row consists of six green circles labeled -I1, -I2, -I3, -I4, -I5, and -I6. Below this is a matrix of values. The matrix has 20 rows (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y) and 7 columns. The values are mostly -4, with some positive values (6, 10, 13, 16) and some negative values (-16, -13, -14, -15, -16). Below the matrix is a row of seven red boxes labeled -d1, -d2, -d3, -d4, -d5, -d6, and -d7.

A	-16	-4	6	10	-4	-4	6
C	-4	-4	-4	-4	-4	-4	-4
D	-4	-4	-4	-4	-4	-4	-4
E	-4	-4	-4	-4	-4	-4	-4
F	-4	-4	-4	-4	-4	-4	-4
G	-4	-4	-4	-4	-4	-4	-4
H	-4	-4	-4	-4	-4	-4	-4
I	-4	-4	-4	-4	-4	-4	-4
K	-4	-4	-4	-4	-4	-4	-4
L	-4	-4	-4	13	6	6	-4
M	-4	-4	-4	-4	16	-4	-4
N	-4	-4	-4	-4	-4	-4	-4
P	-4	-4	6	-4	-4	6	-4
Q	-4	-4	-4	-4	-4	-4	-4
R	-4	-4	-4	-4	-4	-4	-4
S	6	10	10	-4	-4	10	6
T	-4	13	6	-4	-4	-4	10
V	-4	-4	-4	-4	-4	6	6
W	-4	-4	-4	-4	-4	-4	-4
Y	-4	-4	-4	-4	-4	-4	-4
Del	-d1	-d2	-d3	-d4	-d5	-d6	-d7

# Build profiles: model gaps

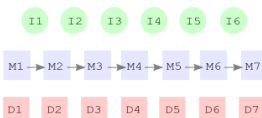
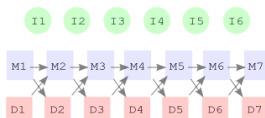


Diagram illustrating the mapping of model gaps (-I1 to -I6) to sequence positions (A to Y) and deletion positions (-d1 to -d7). The model gaps are represented by green circles, and the sequence positions are represented by letters. The deletion positions are represented by red boxes.

	-I1	-I2	-I3	-I4	-I5	-I6
A	-16	-4	6	10	-4	-4
C	-4	-4	-4	-4	-4	-4
D	-4	-4	-4	-4	-4	-4
E	-4	-4	-4	-4	-4	-4
F	-4	-4	-4	-4	-4	-4
G	-4	-4	-4	-4	-4	-4
H	-4	-4	-4	-4	-4	-4
I	-4	-4	-4	-4	-4	-4
K	-4	-4	-4	-4	-4	-4
L	-4	-4	-4	13	6	6
M	-4	-4	-4	-4	16	-4
N	-4	-4	-4	-4	-4	-4
P	-4	-4	6	-4	-4	6
Q	-4	-4	-4	-4	-4	-4
R	-4	-4	-4	-4	-4	-4
S	6	10	10	-4	-4	10
T	-4	13	6	-4	-4	10
V	-4	-4	-4	-4	-4	6
W	-4	-4	-4	-4	-4	-4
Y	-4	-4	-4	-4	-4	-4
Del	-d1	-d2	-d3	-d4	-d5	-d6

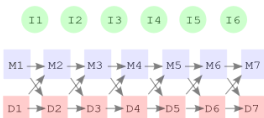
# Build profiles: model gaps



-11 -12 -13 -14 -15 -16

A	-16	-4	6	10	-4	-4	6
C	-4	-4	-4	-4	-4	-4	-4
D	-4	-4	-4	-4	-4	-4	-4
E	-4	-4	-4	-4	-4	-4	-4
F	-4	-4	-4	-4	-4	-4	-4
G	-4	-4	-4	-4	-4	-4	-4
H	-4	-4	-4	-4	-4	-4	-4
I	-4	-4	-4	-4	-4	-4	-4
K	-4	-4	-4	-4	-4	-4	-4
L	-4	-4	-4	13	6	6	-4
M	-4	-4	-4	-4	16	-4	-4
N	-4	-4	-4	-4	-4	-4	-4
P	-4	-4	6	-4	-4	6	-4
Q	-4	-4	-4	-4	-4	-4	-4
R	-4	-4	-4	-4	-4	-4	-4
S	6	10	10	-4	-4	10	6
T	-4	13	6	-4	-4	-4	10
V	-4	-4	-4	-4	-4	6	6
W	-4	-4	-4	-4	-4	-4	-4
Y	-4	-4	-4	-4	-4	-4	-4
Del	-d1	-d2	-d3	-d4	-d5	-d6	-d7

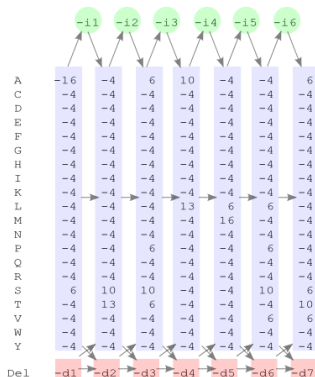
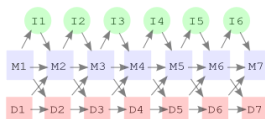
# Build profiles: model gaps



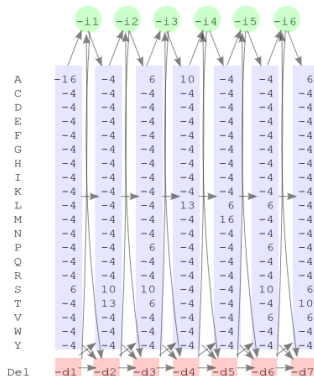
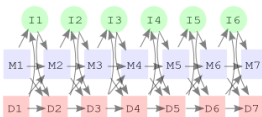
A	-16	-4	6	10	-4	-4	6
C	-4	-4	-4	-4	-4	-4	-4
D	-4	-4	-4	-4	-4	-4	-4
E	-4	-4	-4	-4	-4	-4	-4
F	-4	-4	-4	-4	-4	-4	-4
G	-4	-4	-4	-4	-4	-4	-4
H	-4	-4	-4	-4	-4	-4	-4
I	-4	-4	-4	-4	-4	-4	-4
K	-4	-4	-4	-4	-4	-4	-4
L	-4	-4	-4	13	6	6	-4
M	-4	-4	-4	-4	16	-4	-4
N	-4	-4	-4	-4	-4	-4	-4
P	-4	-4	6	-4	-4	6	-4
Q	-4	-4	-4	-4	-4	-4	-4
R	-4	-4	-4	-4	-4	-4	-4
S	6	10	10	-4	-4	10	6
T	-4	13	6	-4	-4	-4	10
V	-4	-4	-4	-4	-4	6	6
W	-4	-4	-4	-4	-4	-4	-4
Y	-4	-4	-4	-4	-4	-4	-4
Del	-d1	-d2	-d3	-d4	-d5	-d6	-d7



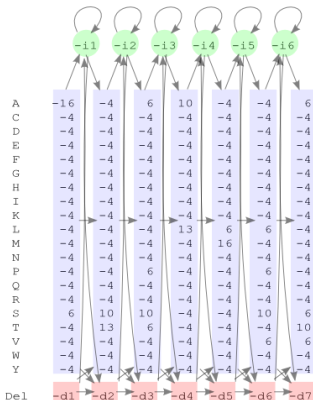
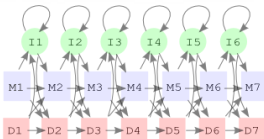
# Build profiles: model gaps



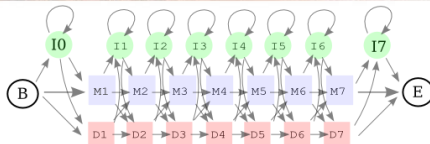
# Build profiles: model gaps



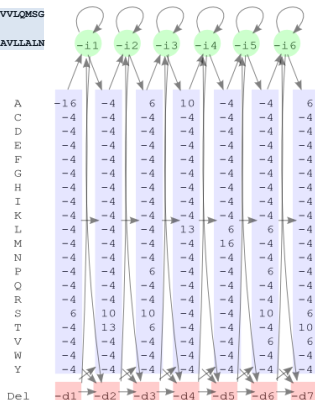
# Build profiles: model gaps



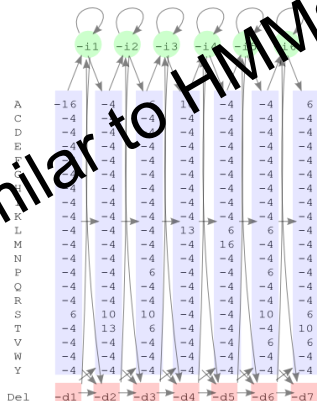
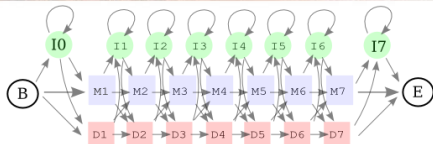
# Build profiles: model gaps




Profile  
 YLVDWDEFKSD--IYCSCRSFEYKGYLCRHAIVVLQMSG  
 Séquence  
 YTVQIDLDDDEKEXSCSCPXFE-HGXPKHILAVLLALN



# Build profiles: model gaps

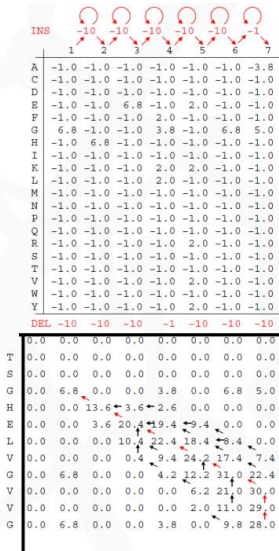


similar to HMMs!



# Search sequences with generalized profiles with dynamic programming

# Profiles: search and align with dynamic programming



Dynamic programming is a method for solving a complex problem by breaking it down into a collection of simpler subproblems, solving each of those subproblems just once, and storing their solutions - ideally, using a memory-based data structure.

ALIGNMENT: score 28

	1	2	3	4	5	6	-	-	7
T									
S	G	H	E	L	V	G	V	V	G

# Profiles: search and align with dynamic programming

INS

	1	2	3	4	5	6	7
A	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-3.8
C	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
D	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
E	-1.0	-1.0	6.8	-1.0	2.0	-1.0	-1.0
F	-1.0	-1.0	-1.0	2.0	-1.0	-1.0	-1.0
G	6.8	-1.0	-1.0	3.8	-1.0	6.8	5.0
H	-1.0	6.8	-1.0	-1.0	-1.0	-1.0	-1.0
I	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
K	-1.0	-1.0	-1.0	2.0	2.0	-1.0	-1.0
L	-1.0	-1.0	-1.0	2.0	-1.0	-1.0	-1.0
M	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
N	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
P	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
Q	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
R	-1.0	-1.0	-1.0	-1.0	2.0	-1.0	-1.0
S	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
T	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
V	-1.0	-1.0	-1.0	-1.0	2.0	-1.0	-1.0
W	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
Y	-1.0	-1.0	-1.0	-1.0	2.0	-1.0	-1.0
DEL	-1.0	-1.0	-1.0	-1	-1.0	-1.0	-1.0

	T	S	G	H	E	L	V	G	V	V	G
T	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
S	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
G	0.0	6.8	0.0	0.0	0.0	0.0	0.0	0.0	6.8	5.0	0.0
H	0.0	0.0	0.0	6.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
E	0.0	0.0	0.0	0.0	6.8	0.0	0.0	0.0	0.0	0.0	0.0
L	0.0	0.0	0.0	0.0	0.0	6.8	0.0	0.0	0.0	0.0	0.0
V	0.0	0.0	0.0	0.0	0.0	0.0	6.8	0.0	0.0	0.0	0.0
G	0.0	6.8	0.0	0.0	0.0	4.2	12.2	31.0	22.4	0.0	0.0
V	0.0	0.0	0.0	0.0	0.0	0.0	6.2	21.0	30.0	0.0	0.0
V	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	11.0	29.0	0.0
G	0.0	6.8	0.0	0.0	0.0	3.8	0.0	9.8	28.0	0.0	0.0

ALIGNMENT: score 28

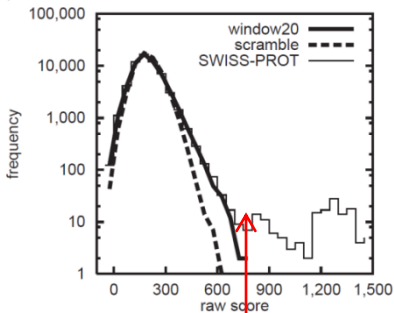
	1	2	3	4	5	6	-	-	7
T									
S									
G									
H									
E									
L									
V									
G									
V									
V									
G									

where do I put the cut off?



## Calibration and Determination of the Cut-off I

- The profile is scanned against a **randomized database** of proteins (Swiss-Prot 34 reversed or shuffled) to calibrate the profile and to deduce the cut-off value. (Look at the distribution of scores on a randomized database).
- The raw score is normalized to facilitate comparisons between results. (Normalized scores of different profiles can be compared)



Cut-off (usually E-value = 0.01)

E-value = relative frequency x size db

$$-\log \frac{N_{\text{matches}}(\text{score} > S)}{N_{DB}} = a + bS$$

## PROSITE profiles use normalized scores

- PROSITE profiles don't use directly E-values, but **normalized scores**, which are a linear transformation of the raw score

$$N_{score} = R_1 + R_2 \times Score$$

Where  $R_1$  and  $R_2$  are 2 parameters characterizing the right tail of the EVD.

- The  $N_{score}$  and the E-value are related by the following relationship

$$E(A) = A \times 10^{-N_{score}}$$

where  $A$  is the number of residues in the searched database.

- For example, for a database containing  $10^7$  residues, a normalized score of 9.0 corresponds to an E-value of 0.01.
- Pagni, M and Jongeneel, CV (2001) Briefings in Bioinformatics, 2, 51-67.

# PROSITE and HAMAP generalized profiles: Example



Entry: **PS50006**

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

## General information about the entry

Entry name	FHA_DOMAIN
Accession number	PS50006
Entry type	MATRIX
Date	NOV-1995 (CREATED); NOV-1995 (DATA UPDATE); JAN-2013 (INFO UPDATE).
PROSITE Documentation	<a href="#">PDOC50006</a>
Associated ProRule	<a href="#">PRU00086</a>

## Name and characterization of the entry

Description Forkhead-associated (FHA) domain profile.

Matrix / Profile

```
/GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=53; TOPOLOGY=LINEAR;  
/DISJOINT: DEFINITION=PROTECT; N1=6; N2=48;  
/NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=0.8473; R2=0.0187; TEXT='-LogE';  
/CUT_OFF: LEVEL=1; SCORE=43; N_SCORE=8.3; MODE=1; TEXT='!!!';  
/CUT_OFF: LEVEL=0; SCORE=329; N_SCORE=7.0; MODE=1; TEXT='!';  
/CUT_OFF: LEVEL=-1; SCORE=275; N_SCORE=6.0; MODE=1; TEXT='?';  
/DEFAULT: B1=-50; E1=-50; M1=-105; MD=-105; IM=-105; DM=-105; I=-20; D=-20;
```

Defaults values

```
          A B C D E F G H I K L M N P Q R S T V W Y Z  
/I:      B1=0; B1=-105; B0=-105;
```

Match state

```
/M: SY='V': M=-6,-26,-15,-30,-25, 5,-29,-24, 16,-19, 7, 5,-24,-26,-22,-19,-16,-6, 18,-6, 3,-24;  
/M: SY='T': M=-8,-12,-22,-16,-11,-6,-19,-10,-6,-2,-7,-3,-6,-18,-8, 3,-2, 5,-2,-21,-3,-11;  
/M: SY='I': M=-10,-29,-24,-35,-27, 16,-30,-25, 28,-29, 22, 13,-22,-25,-25,-24,-20,-10, 19,-15, 3,-27;  
/M: SY='G': M= 0,-10,-30,-10,-20,-30, 70,-20,-40,-20,-30,-20, 0,-20,-20,-20, 0,-20,-30,-20,-30,-20;  
/M: SY='R': M=-19,-8,-30,-8, 1,-21,-20, 5,-30, 29,-21,-9, 1,-19, 10, 62,-10,-11,-21,-21,-8, 1;  
/M: SY='N': M=-6, 4,-23,-2,-6,-16,-7, 1,-15,-3,-21,-12, 15,-10,-3, 2, 4,-3,-16,-27,-9,-7;  
/M: SY='S': M= 3,-1,-18,-4,-2,-18,-14,-14,-12,-3,-17,-12, 1,-8,-4,-5, 11, 11,-5,-30,-15,-4;  
/M: SY='D': M=-6, 16,-25, 24, 17,-29,-12,-7,-28, 3,-26,-21, 6, 1, 2,-6, 7, 1,-22,-34,-19, 9;
```

Implicit insertion  
(from defaults)

```
MT=-12; MD=-12; IM=-12; I=-4;
```

Explicit insertion  
Explicit deletion

```
/M: SY='R': M=-6, 1,-17, 0, 3,-16,-5,-2,-14, 4,-13,-8, 6,-6, 6, 7, 1,-4,-15,-19,-11, 4;
```

Implicit deletion  
(from defaults)

```
/I:      DM=-12;  
/M: SY='C': M=-4,-20, 45,-25,-23,-14,-22,-26,-9,-22,-10,-9,-18,-26,-23,-21,-6,-5, 5,-38,-21,-24;  
/M: SY='Q': M=-18, 42,-29, 56, 20,-38,-10, 2,-36, 1,-29,-26, 21,-11, 7,-7, 1,-9,-30,-38,-19, 13;
```

# Summary about patterns and profiles

## ➤ Patterns

- model a multiple sequence alignment using a compact string
- suited to model short and well conserved motifs
- good to describe functional residues
- easy to build, but not producing a score

## ➤ Profiles

- model a multiple sequence alignment using a numerical matrix representing the position specific distribution of the residues
- suited to model protein domains and gapped motifs
- excellent technology to detect distant homologies
- matches produce a score that can be interpreted using statistical methods

- Profiles and pattern can be used together (rules) to produce precise annotation



# Databases

# Family and domain identification tools and databases

- **General definition:** a given pattern or PSSM/profile specific for a domain is called a descriptor, descriptor motif, discriminator or predictor.
- **Domain databases:** PROSITE, Pfam, SMART, ProDom
- **Family databases:** HAMAP, PRINTS, PANTHER, PIRSF, TIGRFAM
- **Structural databases:** SCOP, CATH
- **Integrated databases:** CDD, InterPro
- **PSI-BLAST**



# Tips and tricks

# Validate your results

## Evaluate the score of the match

**As for wet lab results cross-validate your results with other methods:**

- Use an integrated database (**InterPro** or **CDD**): matches with different methods are likely to be true hits
- Perform a Blast with the matched region: matched sequences should contain the same motif
- Build a pattern (**ScanProsite**), generalized profile (**MyHits**) or HMM (**MyHits**) with the matched region: matched sequences should contain the same motif
- Look for additional information provided by the database: features like active or binding sites, PTMs or disulfides and read the documentation describing the motif



# MyHits: Build your own profile...



## PFSEARCH

search help

user: GUEST width: 600  
log in settings

This web interface to PFSEARCH is meant to be used in conjunction with the [MSA hub](#). To make a proficient use of this service, new users are warmly encouraged to read the [relative help page about PSI-BLAST](#).

### Tools

- Search ...
  - Pattern Search
  - BLASTP/PSI-BLAST
  - PFSEARCH (profile)
  - HMMER3 (profile-HMM)
- Motif Scan
- Query ...
  - by Protein
  - by Motif
- Align...
  - MAFFT
  - TCOFFEE
  - Profile Align
- Classify ...
  - JACOP
  - MkDom2
- Tools ...
  - Reformat MSA
  - Reformat SEQ
  - Dotlet

- Hub
- Results
- Misc
- Deprecated
- Privacy notice

multiple sequence alignment (MSA)

multiple FASTA format

clear input

seq\_source

- sw - Swiss-Prot
- bact - Some complete proteomes from Bacteria
- arch - Some complete proteomes from Archaea
- euk - Some complete proteomes from Eukaryota
- ur50 - UniRef50
- ur50\_wins20 - UniRef50 window shuffled (w=20)
- ecoli - Escherichia coli K12 proteome

Hide optional parameters

Taxonomic restriction  
([NCBI taxonomic browser](#))

- no restriction [all]-

Search method

global on query

Cutoff (normalized score) inclusion threshold

9.0 (e-value=0.021)

Cutoff (normalized score) report threshold

6.5 (e-value=6.7)

Cluster matches at approx. identity level

70 %

search

This service is very computer intensive, please be patient.

reset

# MyHits: ...or HMM



## HMMER3

search help

user: **GUEST** width: **600**

log in

settings

This form lets you build a HMMER3 profile-HMM from a multiple sequence alignment (MSA), and search the databases of protein sequences with it.

The [HMMER3](#) package was written by Sean Eddy. Only those options that are most useful to build a profile from an MSA are available below.

### Tools

- Search ...
  - Pattern Search
  - BLASTP/PSI-BLAST
  - PFSEARCH (profile)
  - HMMER3 (profile-HMM)
- Motif Scan
- Query ...
  - by Protein
  - by Motif
- Align...
  - MAFFT
  - TCOFFEE
  - Profile Align
- Classify ...
  - JACOP
  - MkDom2
- Tools ...
  - Reformat MSA
  - Reformat SEQ
  - Dotlet

### Hub

### Results

### Misc

### Deprecated

### Privacy notice

multiple sequence  
alignment (MSA)

examples

clear input

seq\_source

- sw - Swiss-Prot
- bact - Some complete proteomes from Bacteria
- arch - Some complete proteomes from Archaea
- euk - Some complete proteomes from Eukaryota
- ur50 - UniRef50
- ur50\_win20 - UniRef50 window shuffled (w=20)
- ecolli - Escherichia coli K12 proteome

▼ Hide optional parameters

Taxonomic restriction  
([NCBI taxonomic browser](#))

- no restriction [all]-

E-value inclusion threshold

1e-3

E-value report threshold

1

Cluster matches at approx. identity  
level

70 %

search

This service is  
very computer  
intensive,  
please be  
patient.

reset page

# ProRule

- **Additional information contained in a rule associated to each PROSITE descriptor increases its discriminatory power (combines advantages from both profiles and patterns).**

Acrosins are serine proteases of trypsin-like cleavage specificity.



Haptoglobins have lost active site residues and are therefore no longer catalytically active.



- **Requires a good profile/sequence alignment to be meaningful.**

## General rule information [?]

Accession	PRU00274
Dates	13-DEC-2003 (Created) 27-FEB-2009 (Last updated, Version 10)
Data class	Domain
Predictors	PROSITE, PS50240, TRYPsin_DOM
Name	Serine proteases, trypsin domain
Function	Cleaves preferentially Arg-I-Xaa, Lys-I-Xaa

## Propagated annotation [?]

### Description [?]

case <FTGroup> 1>  
+ Recline: EC<3.4.21->  
end case

### Comments [?]

Similarity	Belongs to the peptidase S1 family. Contains # peptidase S1 domain.
------------	--

### Cross-references [?]

case <FTGroup> 1>  
PROSITE [PS00134, TRYPsin\\_HIS, 1](#),  
[PS00135, TRYPsin\\_SER, 1](#);  
else  
[PS00134, TRYPsin\\_HIS, 0-1](#),  
[PS00135, TRYPsin\\_SER, 0-1](#);  
end case

### Gene Ontology [?]

case <FTGroup> 1>  
GO:0016787, Molecular function: hydrolase activity  
GO:0006233, Molecular function: peptidase activity  
GO:0004252, Molecular function: serine-type endopeptidase activity

### Keywords [?]

Hydrolase  
Protease  
Serine protease  
end case

case <FTag> disulf->  
Disulfide bond  
end case

### Features [?]

From: PS50240	From	To	Description	Tag	Condition	FTGroup
Key						
FORAID	11		Peptidase S1 #			
ACT_SITE	42	42	Change relay system (by similarity)		D	1
ACT_SITE	91	91	Change relay system (by similarity)		D	1
ACT_SITE	104	104	Change relay system (by similarity)		D	1
DISULFID	27	43	By similarity	disulf	C-x-x-C	
DISULFID	125	150	By similarity	disulf	C-x-x-C	
DISULFID	156	171	By similarity	disulf	C-x-x-C	
DISULFID	182	210	By similarity	disulf	C-x-x-C	