

An Introduction to Patterns, Profiles, HMMs, and PSI-BLAST

Course 2006

Marco Pagni and Lorenzo Cerutti

Swiss Institute of Bioinformatics, Lausanne

Outline

- Introduction
 - Reminder on pairwise alignments
 - Multiple alignments and their information content
- Models of multiple alignments and databases
 - Consensus sequences
 - Patterns and regular expressions
 - Position Specific Scoring Matrices (PSSMs)
 - Generalized Profiles
 - Hidden Markov Models (HMMs)
- PSI-BLAST and protein domain hunting

Outline

- Introduction
 - **Reminder on pairwise alignments**
 - Multiple alignments and their information content
- Models of multiple alignments and databases
 - Consensus sequences
 - Patterns and regular expressions
 - Position Specific Scoring Matrices (PSSMs)
 - Generalized Profiles
 - Hidden Markov Models (HMMs)
- PSI-BLAST and protein domain hunting

Pairwise alignments

- Pairwise alignments are used to compare pairs of sequences to find homologous regions.
- Various algorithms exist to build local or global pairwise alignments (Smith-Waterman, Needleman-Wunsch, BLAST, ...).
- However, they are limited to the primary sequence and do not inform about "hidden" features of the analyzed sequences.

```
seq1 WFHGSWTRQGAEHLL-LLKGEAGTFVLRECLSSPGQYVLSV--RYIGNHK--HCIISQHDRNGQFLIEDDRACDTFGMLLQHY
      :::   :.  ::  ::   .....:::  :  :  .  .:  .  . .  :  . . .  ::  :  :   :.  . . .
seq2 WYHGEIERSIAEGLLGQRRNNTGSFIVREALENIGAFSVTVYDKDISHPRVLHFRVNSNMNNG-FYIATKTCFRTIPYIIWFF
```

Outline

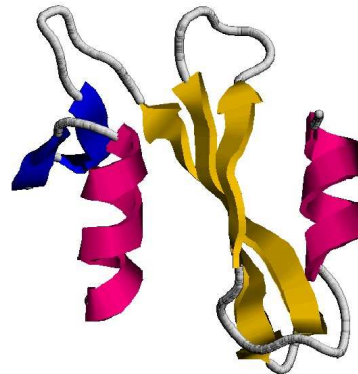
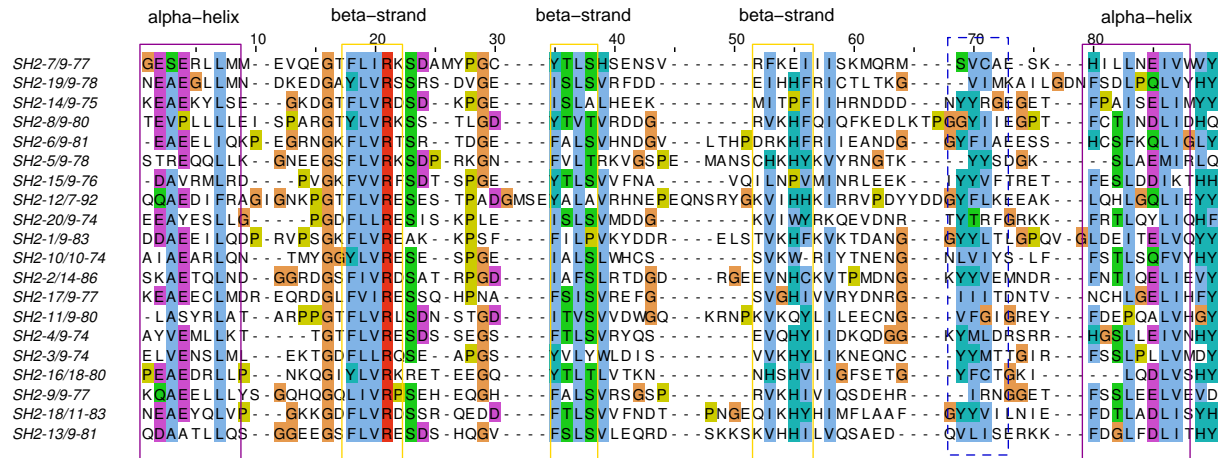
- Introduction
 - Reminder on pairwise alignments
 - Multiple alignments and their information content
- Models of multiple alignments and databases
 - Consensus sequences
 - Patterns and regular expressions
 - Position Specific Scoring Matrices (PSSMs)
 - Generalized Profiles
 - Hidden Markov Models (HMMs)
- PSI-BLAST and protein domain hunting

Multiple sequences alignment

- A **multiple sequence alignment** (MSA) has a higher information content than a pairwise alignment.
- MSA is a method of choice to detect conserved regions in DNA and proteins, usually associated with:
 - Signals (promoters, signatures for phosphorylation, cellular localization signals, ...)
 - Structure (folding, regions of interaction, ...)
 - Chemical reactivity (catalytic sites, ...)

MSA information content

- Example: MSA reflects secondary structure



Models of MSA

- We need a **model** to describe a MSA and its information content. The model will be used to re-align sequences, search databases, and transfer annotation.
- Various techniques exist to build a model of a MSA:
 - Consensus sequences
 - Patterns
 - Position Specific Scoring Matrices (PSSMs)
 - Profiles
 - Hidden Markov Models (HMMs)

Outline

- Introduction
 - Reminder on pairwise alignments
 - Multiple alignments and their information content
- Models of multiple alignments and databases
 - Consensus sequences
 - Patterns and regular expressions
 - Position Specific Scoring Matrices (PSSMs)
 - Generalized Profiles
 - Hidden Markov Models (HMMs)
- PSI-BLAST and protein domain hunting

Consensus sequences

- The **consensus sequence** method is the simplest way to build a model from a MSA.
- A consensus sequence is build using the following rules:
 - majority wins
 - skip to much variation

Consensus sequences

```
GHEGVGKVVKIG  
GHEKKGYFEDRG  
GHEGYGGRSRGG  
GHEFEGPKGCGA  
GHELRGTTFMPA
```



1	2	3	4	5	6	7	8	9	10	11	12
G	H	E	G	V	G	K	V	V	K	I	G
			K	K		Y	F	E	D	R	A
			F	Y		G	R	S	R	G	
			L	E		P	K	G	C	P	
				R		T	T	F	M		
<hr/>											
Consensus:	G	H	E	.	.	G

Consensus sequences: conclusion

- Advantages:
 - very fast and easy to implement (a simple word processor is enough).
- Limitations:
 - no information about variations in the columns of the MSA
 - highly dependent on the training set
 - no scores, only binary result (YES/NO)
- When to use consensus sequences?
 - to find highly conserved signatures, as for example restriction sites in DNA sequences

Outline

- Introduction
 - Reminder on pairwise alignments
 - Multiple alignments and their information content
- Models of multiple alignments and databases
 - Consensus sequences
 - **Patterns and regular expressions**
 - Position Specific Scoring Matrices (PSSMs)
 - Generalized Profiles
 - Hidden Markov Models (HMMs)
- PSI-BLAST and protein domain hunting

Patterns

- **Patterns** describe sets of alternative sequences using a single expression.
- In computer science patterns are known as **regular expressions** (regexp).
- To describe alternative sequences in a single expression we require a special syntax.

Pattern syntax

- aa are represented by single letter code
- each position is separated by a dash '-'
- 'x' represents any aa
- '['']' group of aa accepted for a position
- '{}' group of aa **not** accepted for a position
- '()' repetitions ([AG](2,4) means A or G between 2 and 4 times, x(2) means any aa twice)
- '<' anchor at the N-term
- '>' anchor at the C-term

Pattern vs. Regexp

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

How to build a pattern

```

GHEGVGKVVKIG
GHEKKGYFEDRG
GHEGYGGRSRGG
GHEFEGPKGCGA
GHELRGTTFMPA
  
```



	1	2	3	4	5	6	7	8	9	10	11	12
G	H	E	G	V	G	K	V	V	K	I	G	
			K	K		Y	F	E	D	R	A	
			F	Y		G	R	S	R	G		
			L	E		P	K	G	C	P		
				R		T	T	F	M			
Consensus:	G	H	E	.	.	G

Pattern: G-H-E-X(2)-G-X(5)-[GA]

Patterns: conclusion

- Advantages:
 - pattern matching is fast and easy to implement
 - models are easy to design and understand
- Limitations:
 - poor models for insertions/deletions (indels)
 - poor predictors: tend to recognize only sequences in the training set
 - no scores, only binary results (YES/NO)
- When to use patterns?
 - to search for relatively conserved and small signatures
 - to communicate to other biologists

Patterns: conclusion (2)

- Patterns can be automatically extracted (discovered) from a set of unaligned sequences by specialized software based on machine learning:
 - **Pratt** (<http://www.ebi.ac.uk/pratt/>)
 - **Splash** (<http://www.research.ibm.com/splash/>)
 - **Teiresias** (<http://cbcsrv.watson.ibm.com/Tspd.html>)
- Such automatic patterns are usually different from those designed by an expert with some knowledge of the biochemical literature.

Prosite: a patterns database

- Current version of **Prosite** contains 1329 patterns of protein motifs.
- Each pattern is associated with an exhaustive documentation.
- A **quality** value is associated to each pattern based on the true positive (TP), false negative (FN), and false positive (FP), found in SWISS-PROT.
- Frequently matching pattern are tagged with a special flag (SKIP_FLAG=TRUE).
- Web access: <http://www.expasy.org/prosite/>

Prosite: example

General information about the entry

Entry name	UCH_2_1
Accession number	PS00972
Entry type	PATTERN
Date	JUN-1994 (CREATED); DEC-2004 (DATA UPDATE); JAN-2006 (INFO UPDATE).
PROSITE documentation	PDOC00750

Name and characterization of the entry

Description	Ubiquitin carboxyl-terminal hydrolases family 2 signature 1.
Pattern	G-[LIVMFY]-x(1,3)-[AGCY]-[NASMQG]-x-C-[FYWC]-[LIVMFCA]-[NSTAD]-[SACV]-x-[LIVMSF]-[QF].

Numerical results

- ◆ UniProtKB/Swiss-Prot release number: **48.9**, total number of sequence entries in that release: **206586**.
- ◆ Total number of hits in UniProtKB/Swiss-Prot: **110 hits in 110 different sequences**
- ◆ Number of hits on proteins that are known to belong to the set under consideration: **110 hits in 110 different sequences**
- ◆ Number of hits on proteins that could potentially belong to the set under consideration: **0 hits in 0 different sequences**
- ◆ Number of false hits (on unrelated proteins): **0 hits in 0 different sequences**
- ◆ Number of known missed hits: **3**
- ◆ Number of partial sequences which belong to the set under consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences: **1**
- ◆ Precision (true hits / (true hits + false positives)): **100.00%**
- ◆ Recall (true hits / (true hits + false negatives)): **97.35%**

Comments

- ◆ Taxonomic range: **Eukaryotes, Eukaryotic viruses**
- ◆ Maximum known number of repetitions of the pattern in a single protein: **1**
- ◆ 'Interesting' site in the pattern: **7,active_site(?)**
- ◆ VERSION: **1**

Cross-references

True positive hits:

CYLD_HUMAN (Q9NQC7), CYLD_MOUSE (Q80TQ2), FAF_DROME (P55824),
UBP10_HUMAN (O14694), UBP10_MOUSE (P52479), UBP11_HUMAN (P51784).



Prosite: search and scan

Protein(s) to be scanned:

Enter one or more Swiss-Prot/TrEMBL accession number(s) [AC] (e.g. **P00747**) and/or sequence identifier(s) [ID] (e.g. **ENTK_HUMAN**), and/or PDB identifier, and/or paste **your own protein sequence(s)** in the box below:
(leave this box blank to scan PROSITE entries against selected protein databases)

Clear

General options:

Exclude motifs with a high probability of occurrence

Show low level score

Do not scan profiles [User Manual]

Show only sequences with at least hit(s)

Maximum of matched sequences

Output format

Retrieve complete sequences

Your e-mail (optional): (will send results by e-mail)

PROSITE pattern(s)/profile(s) to scan for:

Enter one or more PROSITE accession number(s) (e.g. **PS50240**), and/or identifier(s) (e.g. **CHEB**), and/or type **your pattern(s)** in PROSITE format in the box below:
(leave this box blank to scan sequence(s) against the entire PROSITE database)

and specify your search limits (only used if no protein data specified):

◆ **Protein database(s):** Swiss-Prot TrEMBL PDB databases
 including splice variants
randomize databases (to test a pattern, see help)

◆ **Taxonomic lineage (OC) / species (OS) filter:**

(see [NEWT Taxonomy](#); separate multiple taxa/species with a semicolon, e.g. *Eukaryota; Escherichia coli*; . Does not work on PDB sequences.)

◆ **Description (DE) filter:** e.g. *protease*

pattern options:

Allow at most X sequence characters to match a conserved position in the pattern

match mode (for patterns, see help)

MyHits: pattern search

user: anonymous
[log in](#)

Pattern Search

Pattern Input
Enter a pattern in PROSITE format
[Example](#)

```
[EQ]-x-L-Y-[DEQST]-x(3,12)-[LIV]-[ST]-Y-x-R-[ST]
```

[Clear input](#)
[Reset page](#)

This form lets you search the protein databases with a pattern written according to the [Prosite syntax](#).

Parameters

Database of sequences	<input type="checkbox"/> Swiss-Prot [sw] <input type="checkbox"/> TrEMBL [tr] <input type="checkbox"/> Swiss-Prot splice variants [sw_var] <input type="checkbox"/> trEST [te] <input type="checkbox"/> trGEN [tg] <input type="checkbox"/> trome [to] <input type="checkbox"/> Current ENSEMBL peptides for all species[ens] <input type="checkbox"/> Microbial complete proteomes[microb] <input type="checkbox"/> RefSeq Release[rs] <input type="checkbox"/> RefSeq weekly updates[rs_new]	search
Taxonomic restriction	Temporarily not available	

[Question or comment about this page.](#)

LC/MP-SIB-2006 – p.22/?

MyHits: pattern scan

user: anonymous
[log in](#)

Motif Scan

Protein Sequence Input
Enter a protein sequence in RAW or FASTA or Swiss-Prot format or a db:AC or db:ID identifier

FASTA format

```
#COMMENT: Enter a sequence in FASTA format
>Ubiquitin (FASTA)
MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQORLIFAGKQLED
GR.TLSDYNIQKESTLHLVLR.LRGG
```

[Clear input](#)
[Reset page](#)

Motif scanning means finding all known motifs that occur in a sequence. This form lets you paste a protein sequence, select the collections of motifs to scan for, and launch the search. Some general [documentation](#) is available about the Prosite and Pfam collections of motifs. Another [document](#) deals with the interpretation of the match scores. You should consult the home pages of [Prosite](#) on ExPASy, [Pfam](#) and [InterPro](#) for additional information.

Warning: The scan might take a few minutes, thus if your proteins of interest are already in the sequence databases (see [list](#)), the [Query by Protein](#) form is much faster, and the [Protein Hub](#) provides a collection of tools that you might find useful.

Parameters

Database of motifs (db description)	<input checked="" type="checkbox"/> PROSITE patterns <input checked="" type="checkbox"/> PROSITE patterns (frequent match producers) <input type="checkbox"/> PROSITE profiles <input type="checkbox"/> Pfile (more profiles) <input type="checkbox"/> Na-channel profiles <input type="checkbox"/> HAMAP profiles <input type="checkbox"/> Pfam HMMs (local models) <input type="checkbox"/> Pfam HMMs (global models)	search
--	--	------------------------

[Question or comment about this page.](#)

Outline

- Introduction
 - Reminder on pairwise alignments
 - Multiple alignments and their information content
- Models of multiple alignments and databases
 - Consensus sequences
 - Patterns and regular expressions
 - **Position Specific Scoring Matrices (PSSMs)**
 - Generalized Profiles
 - Hidden Markov Models (HMMs)
- PSI-BLAST and protein domain hunting

PSSM

- **Position Specific Scoring Matrices** (PSSMs) are based on the observed **frequencies** of each residue in each column of the MSA.
- Log-odds scores are derived from the observed frequencies:
 - log-odds are preferred for computational reasons.

PSSM: frequencies

```

GHEGVGKVVKIG
GHEKKGYFEDRG
GHEGYGGRSRGG
GHEFEGPKGCGA
GHELRGTTFMPA
  
```



	1	2	3	4	5	6	7	8	9	10	11	12
A	0	0	0	0	0	0	0	0	0	0	0	2
C	0	0	0	0	0	0	0	0	0	1	0	0
D	0	0	0	0	0	0	0	0	0	1	0	0
E	0	0	5	0	1	0	0	0	1	0	0	0
F	0	0	0	1	0	0	0	1	1	0	0	0
G	5	0	0	2	0	5	1	0	1	0	2	3
H	0	5	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	1	0
K	0	0	0	1	1	0	1	1	0	1	0	0
L	0	0	0	1	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	1	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	1	0	0	0	1	0
Q	0	0	0	0	0	0	0	0	0	0	0	0
R	0	0	0	0	1	0	0	1	0	1	1	0
S	0	0	0	0	0	0	0	0	1	0	0	0
T	0	0	0	0	0	0	1	1	0	0	0	0
V	0	0	0	0	1	0	0	1	1	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	1	0	1	0	0	0	0	0

$$f_{A,1} = \frac{0}{5} = 0, f_{G,1} = \frac{5}{5} = 1, \dots$$

$$f_{A,2} = \frac{0}{5} = 0, f_{H,2} = \frac{5}{5} = 1, \dots$$

...

$$f_{A,12} = \frac{2}{5} = 0.4, f_{G,12} = \frac{3}{5} = 0.6, \dots$$

Pseudo-counts

- Some frequencies equal 0. This reflects the limited number of sequences in the MSA.
- A frequency of 0 imply the exclusion of the corresponding residue at this position (this is the case with patterns).
- To avoid this we can add a small number to all observed counts. These small non-observed counts are referred to as **pseudo-counts**.
- **Substitution matrices** and **Dirichlet mixtures** can be used to produce more "realistic" pseudo-counts.

PSSM: pseudo-counts

```

GHEGVGKVVKIG
GHEKKGYFEDRG
GHEGYGGRSRGG
GHEFEGPKGCGA
GHELRTTFMPA
  
```



	1	2	3	4	5	6	7	8	9	10	11	12
A	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	2+1
C	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	1+1	0+1
D	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	1+1	0+1	0+1
E	0+1	0+1	5+1	0+1	1+1	0+1	0+1	0+1	1+1	0+1	0+1	0+1
F	0+1	0+1	0+1	1+1	0+1	0+1	0+1	1+1	1+1	0+1	0+1	0+1
G	5+1	0+1	0+1	2+1	0+1	5+1	1+1	0+1	1+1	0+1	2+1	3+1
H	0+1	5+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1
I	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	1+1	0+1
K	0+1	0+1	0+1	1+1	1+1	0+1	1+1	1+1	0+1	1+1	0+1	0+1
L	0+1	0+1	0+1	1+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1
M	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	1+1	0+1	0+1
N	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1
P	0+1	0+1	0+1	0+1	0+1	0+1	1+1	0+1	0+1	0+1	1+1	0+1
Q	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1
R	0+1	0+1	0+1	0+1	1+1	0+1	0+1	1+1	0+1	1+1	1+1	0+1
S	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	1+1	0+1	0+1	0+1
T	0+1	0+1	0+1	0+1	0+1	0+1	1+1	1+1	0+1	0+1	0+1	0+1
V	0+1	0+1	0+1	0+1	1+1	0+1	0+1	1+1	1+1	0+1	0+1	0+1
W	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1
Y	0+1	0+1	0+1	0+1	1+1	0+1	1+1	0+1	0+1	0+1	0+1	0+1

$$f_{A,1} = \frac{0+1}{5+20} = 0.04, f_{G,1} = \frac{5+1}{5+20} = 0.24, \dots$$

$$f_{A,2} = \frac{0+1}{5+20} = 0.04, f_{H,2} = \frac{5+1}{5+20} = 0.24, \dots$$

...

$$f_{A,12} = \frac{2+1}{5+20} = 0.12, f_{G,12} = \frac{3+1}{5+20} = 0.16$$

PSSM: score

- The frequency of each residue at each position of the MSA is compared to the frequency at which the residue is expected in a random sequence.
- The frequencies expected in random sequences are named a **null model**.
- A null model can be a simple uniform distribution, or a more complex distribution based on observations (ex. frequencies observed in SWISS-PROT).

PSSM: score (2)

- The **score** is derived from the ratio of the observed to the expected frequencies.
- The logarithm of this score is called **log-likelihood ratio**:

$$S_{ij} = \log\left(\frac{f'_{ij}}{q_i}\right) \quad (1)$$

where S_{ij} is the score for residue i at position j , f'_{ij} is the relative frequency for residue i at position j (corrected with pseudo-counts), and q_i is the expected relative frequency of residue i in the null model.

PSSM: score (3)

```

GHEGVGKVVKIG
GHEKKGYFEDRG
GHEGYGGRSRGG
GHEFEGPKGCGA
GHELRTTFMPA
  
```



	1	2	3	4	5	6	7	8	9	10	11	12
A	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	2+1
C	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	1+1	0+1
D	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	1+1	0+1	0+1
E	0+1	0+1	5+1	0+1	1+1	0+1	0+1	0+1	1+1	0+1	0+1	0+1
F	0+1	0+1	0+1	1+1	0+1	0+1	0+1	1+1	1+1	0+1	0+1	0+1
G	5+1	0+1	0+1	2+1	0+1	5+1	1+1	0+1	1+1	0+1	2+1	3+1
H	0+1	5+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1
I	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	1+1	0+1
K	0+1	0+1	0+1	1+1	1+1	0+1	1+1	1+1	0+1	1+1	0+1	0+1
L	0+1	0+1	0+1	1+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1
M	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	1+1	0+1	0+1
N	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1
P	0+1	0+1	0+1	0+1	0+1	0+1	1+1	0+1	0+1	0+1	1+1	0+1
Q	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1
R	0+1	0+1	0+1	0+1	1+1	0+1	0+1	1+1	0+1	1+1	1+1	0+1
S	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	1+1	0+1	0+1	0+1
T	0+1	0+1	0+1	0+1	0+1	0+1	1+1	1+1	0+1	0+1	0+1	0+1
V	0+1	0+1	0+1	0+1	1+1	0+1	0+1	1+1	1+1	0+1	0+1	0+1
W	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1	0+1
Y	0+1	0+1	0+1	0+1	1+1	0+1	1+1	0+1	0+1	0+1	0+1	0+1

Scores calculated in 1/3 bit:

...

$$S_{A,12} = \log \frac{\frac{2+1}{5+20}}{\frac{1}{20}} \times \frac{3}{\log 2} \approx 3.8,$$

$$S_{C,12} = \log \frac{\frac{0+1}{5+20}}{\frac{1}{20}} \times \frac{3}{\log 2} \approx -1,$$

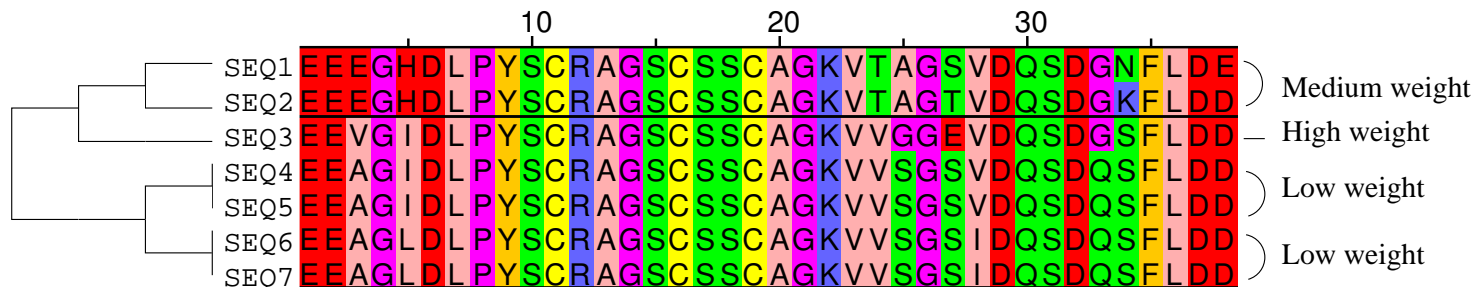
...

PSSM: example

	1	2	3	4	5	6	7	8	9	10	11	12
A	-1.0	-1.0	-1.0	-1.0	-1.0	-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	-1.0	-1.0	2.0	-1.0	-1.0
D	-1.0	-1.0	-1.0	-1.0	-1.0	-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	-1.0	2.0	-1.0	-1.0	-1.0
F	-1.0	-1.0	-1.0	2.0	-1.0	-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	2.0	-1.0	2.0	-1.0	3.8	5.0
H	-1.0	6.8	-1.0	-1.0	-1.0	-1.0	-1.0	-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	-1.0	-1.0	-1.0	-1.0	-1.0
K	-1.0	-1.0	-1.0	2.0	2.0	-1.0	2.0	2.0	-1.0	2.0	-1.0	-1.0
L	-1.0	-1.0	-1.0	2.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	2.0	-1.0
M	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	2.0	-1.0	-1.0
N	-1.0	-1.0	-1.0	-1.0	-1.0	-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	-1.0	-1.0	-1.0	2.0	-1.0
Q	-1.0	-1.0	-1.0	-1.0	-1.0	-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	2.0	-1.0	2.0	2.0	-1.0
S	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	2.0	-1.0	-1.0	-1.0
T	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	2.0	2.0	-1.0	-1.0	-1.0	-1.0
V	-1.0	-1.0	-1.0	-1.0	2.0	-1.0	-1.0	2.0	2.0	-1.0	-1.0	-1.0
W	-1.0	-1.0	-1.0	-1.0	-1.0	-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	2.0	-1.0	-1.0	-1.0	-1.0	-1.0

Sequence weighting

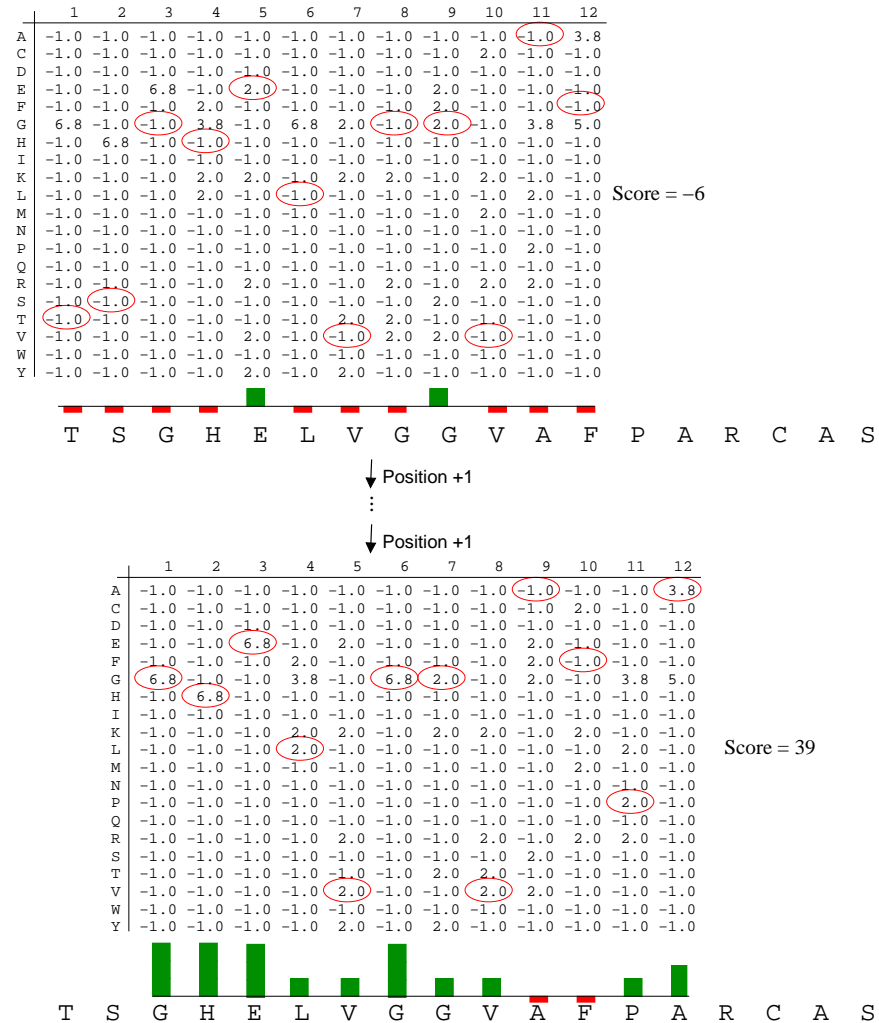
- Subfamilies in a MSA can be differently populated, thus influencing the observed residue frequencies.
- Sequences weighting algorithms attempt to compensate this sequence sampling bias.



PSSM: scoring a match

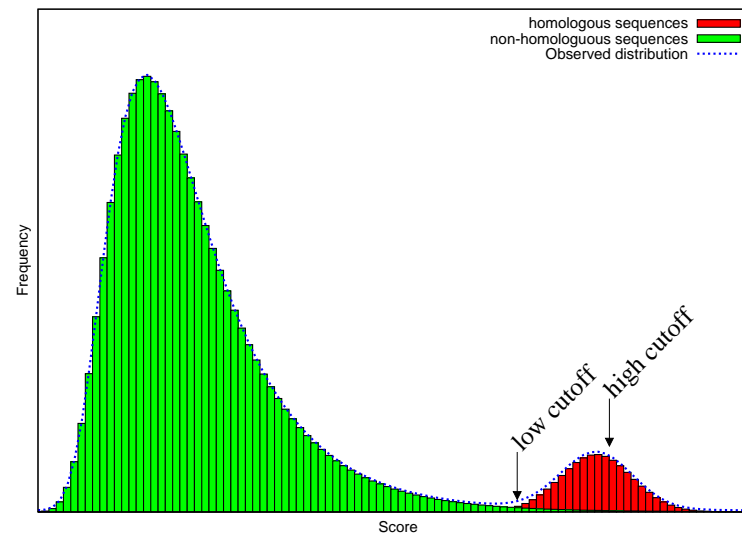
- The PSSM is applied as a **sliding window** along the subject sequence.
 - at each position, the score is obtained by summing the scores of all columns
 - the highest scoring position is reported

PSSM: scoring a match (2)



PSSM score interpretation

- We can estimate the score distribution of a PSSM on unrelated sequences.
- This allows to estimate the **E-value**: number of matches with equal or greater score than the observed that we expect to occur by chance.
- We must select a **cutoff** to ensure small E-values.

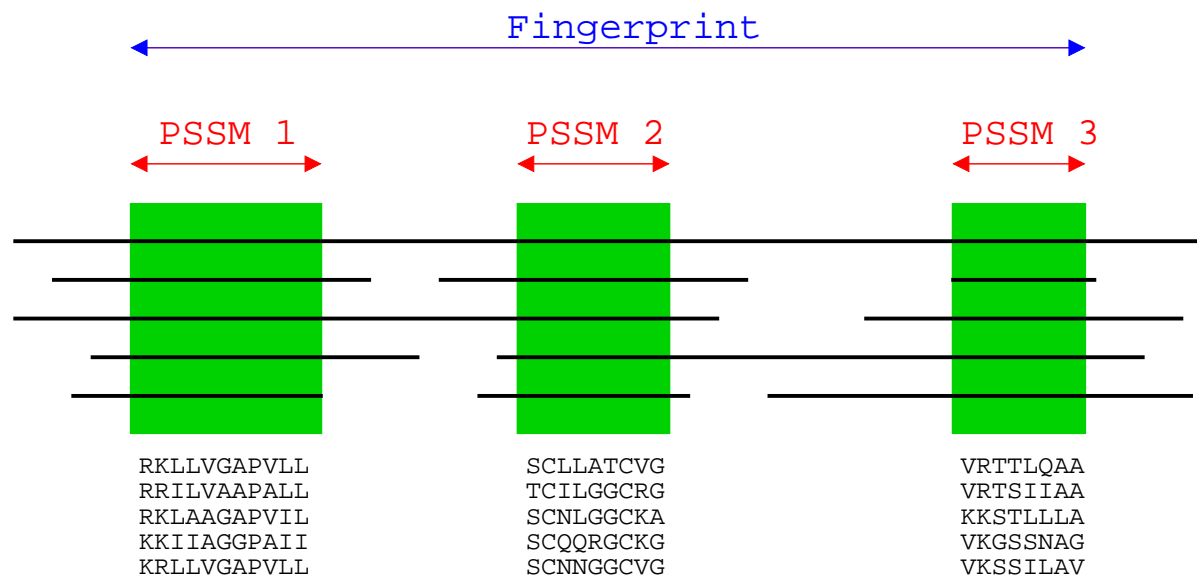


PSSM: conclusion

- Advantages:
 - good for short, relatively conserved regions
 - relatively fast and simple to implement
 - returns scores
- Limitations:
 - indels are forbidden: long regions cannot be described
- When to use PSSMs?
 - to model small regions with high variability but constant length

PSSM: conclusion (2)

- PSSMs can be automatically extracted from a set of unaligned sequences.
- **MEME** is an expectation-maximization algorithm that find PSSMs (<http://meme.sdsc.edu/meme/website/>).
- Two or more PSSMs can be used to describe long regions: **fingerprints**.



Fin erprints databases

- **PRINTS** is a collection of annotated fingerprints

(<http://umber.sbs.man.ac.uk/dbbrowser/PRINTS>).

Scan **PRINTS** with a PROTEIN query sequence; using an ID code from one of the following databases: {SWISSPROT SPTREMBL SWISSNEW TREMBLNEW} or by pasting it in as a raw sequence.
Please Note, DNA Sequences are NOT catered for in this software.

Important information concerning the E-value calculation [please read](#)

Please input a raw sequence:

The E-value threshold determines the level of significance of results in the 1st table

E-value threshold:

Select Database

Prints38_0 Prints36_0 Blocksplus11
 Prints37_0 Blocks11

Select Matrix

blos62
 blos45
 blos80

Distance variance:
 %

- **BLOCKS**: another fingerprints database (<http://blocks.fhcrc.org>).

Outline

- Introduction
 - Reminder on pairwise alignments
 - Multiple alignments and their information content
- Models of multiple alignments and databases
 - Consensus sequences
 - Patterns and regular expressions
 - Position Specific Scoring Matrices (PSSMs)
 - **Generalized Profiles**
 - Hidden Markov Models (HMMs)
- PSI-BLAST and protein domain hunting

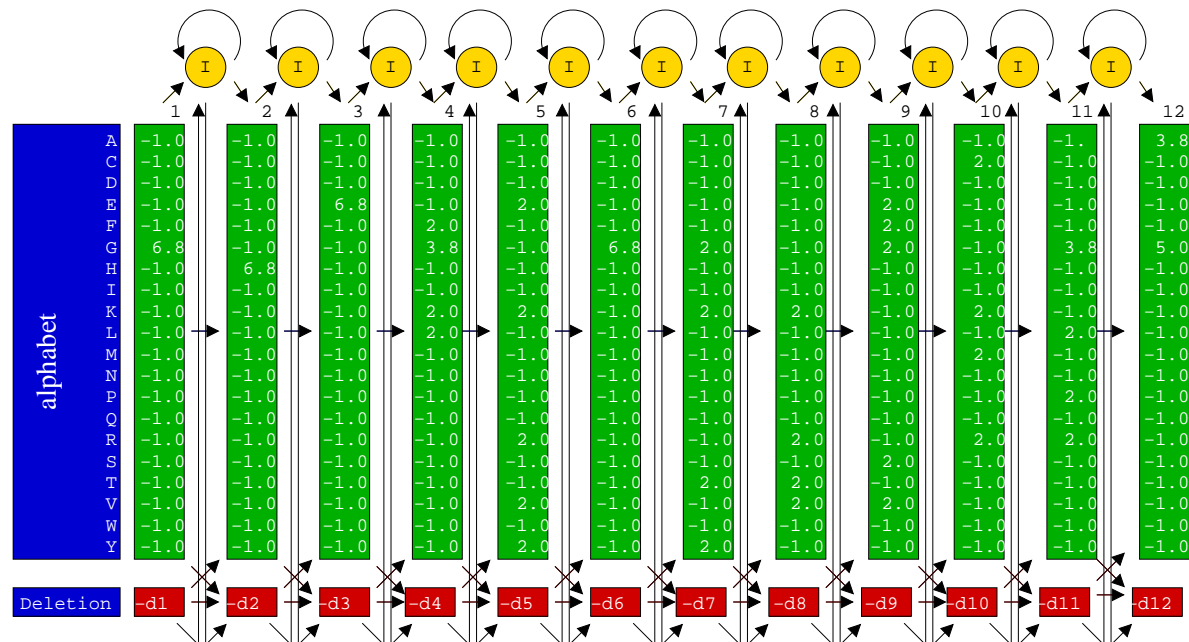
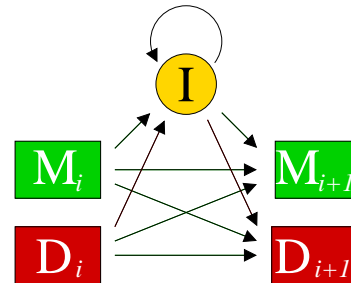
Generalized profiles

- **Generalized profiles** are an extension of the PSSMs, where position specific deletions and insertions penalties are considered.
- Generalized profiles are a generalization of the **dynamic programming** algorithm where:
 - the global substitution matrix is replaced by a PSSM;
 - gap penalties are replaced by position specific deletions and insertions penalties.

Generalized profiles: concepts

- **Match state**: a position dependent substitution score is associated with each residue as for PSSMs.
- **Deletion state**: at each position, a match state can be replaced by a deletion associated with a position specific penalty.
- **Insertion state**: variable length insertions between any two adjacent match/deletion states. They have a position specific penalty that might also depend upon the inserted residues.
- **Transitions**: transitions between states are associated with penalties, primarily to model the cost of opening and closing a gap.
- Some additional transitions permit to tune the model for local or global alignments.

Generalized profiles: concepts (2)



Prosite: a profiles database

- **Prosite** contains a collection of patterns and profiles describing protein domains and motifs.
- Entries in Prosite are associated to high quality annotation.
- Normally two cutoff scores are present: the first for trusted matches, the second for matches in the twilight zone interesting for discovery.

Prosite: example

Entry name	ABC_TM1
Accession number	PS50928
Entry type	MATRIX
Date	NOV-2003 (CREATED); NOV-2003 (DATA UPDATE); JAN-2006 (INFO UPDATE).
PROSITE documentation	PDOC00364
Name and characterization of the entry	
Description	ABC transporter integral membrane type-1 domain profile.
	<pre> /GENERAL_SPEC: ALPHABET='ABCDEFGHIJKLMNPQRSTVWYZ'; LENGTH=196; /DISJOINT: DEFINITION=PROTECT; N1=6; N2=191; /NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=2.4184906; R2=0.0217733; TEXT='-LogE'; /CUT_OFF: LEVEL=0; SCORE=280; N_SCORE=8.5; MODE=1; TEXT='!'; /CUT_OFF: LEVEL=-1; SCORE=188; N_SCORE=6.5; MODE=1; TEXT='?'; /DEFAULT: M0=-9; D=-50; I=-50; B1=-1000; E1=-1000; MI=-105; MD=-105; IM=-105; DM=-105; 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; BI=-105; BD=-105; /M: SY='L'; M= -4, -24, -19, -29, -22, 13, -25, -20, 15, -24, 21, 10, -21, -24, -21, -20, -16, -4, 11, -15, 2, -21; /M: M= -7, -19, -27, -21, -14, -5, -13, -16, -7, -9, -4, -3, -16, -12, -11, -6, -14, -10, -7, -2, -4, -13; /M: SY='N'; M= -5, 2, -22, -6, -8, -11, -11, -6, -8, -7, -11, -7, 9, -18, -6, -6, -2, -2, -9, -22, -9, -8; Match (M) SY='T'; M= 4, 0, -11, -6, -6, -14, -10, -13, -14, -10, -18, -13, 4, -11, -6, -10, 25, 32, -5, -33, -14, -6; /M: SY='L'; M= -6, -27, -19, -31, -23, 11, -28, -22, 20, -26, 28, 14, -25, -26, -22, -21, -20, -6, 16, -18, 0, -22; /M: SY='L'; M= -6, -19, -23, -21, -12, -1, -21, -16, 4, -13, 6, 3, -16, -20, -11, -11, -12, -5, 3, -12, 0, -12; /M: SY='I'; M= -8, -25, -21, -29, -22, 7, -29, -19, 19, -21, 18, 12, -22, -25, -18, -17, -17, -6, 16, -15, 4, -21; /M: SY='A'; M= 16, -10, -12, -16, -13, -13, -3, -19, -7, -14, -10, -7, -7, -15, -12, -16, 10, 9, 2, -26, -15, -13; /M: SY='L'; M= 0, -23, -17, -28, -22, 6, -19, -23, 12, -23, 13, 6, -20, -23, -22, -20, -11, -3, 13, -19, -4, -22; /M: SY='I'; M= 0, -21, -18, -26, -20, -1, -15, -22, 9, -21, 8, 4, -18, -20, -18, -20, -10, -4, 9, -19, -6, -20; /M: SY='A'; M= 15, -10, -13, -14, -12, -15, -1, -18, -10, -13, -13, -10, -6, -14, -11, -15, 11, 7, 0, -27, -16, -12; /M: SY='V'; M= 2, -18, -15, -24, -19, 1, -20, -21, 8, -18, 6, 4, -16, -20, -18, -17, -3, 7, 12, -21, -5, -18; /M: SY='I'; M= 0, -24, -18, -28, -21, 3, -22, -23, 15, -22, 14, 7, -21, -18, -21, -21, -12, -4, 15, -21, -6, -22; /M: SY='L'; M= -6, -26, -20, -31, -24, 8, -25, -23, 20, -26, 21, 11, -22, -23, -22, -22, -17, -7, 14, -18, -2, -23; /M: SY='A'; M= 13, -7, -12, -12, -9, -19, 4, -15, -15, -12, -17, -11, -3, -14, -7, -14, 11, 4, -8, -27, -18, -8; /M: SY='L'; M= -3, -22, -16, -27, -21, 4, -24, -21, 14, -22, 15, 9, -19, -23, -20, -19, -11, 0, 14, -19, -3, -21; /M: SY='I'; M= -1, -25, -19, -28, -21, 3, -24, -23, 15, -22, 13, 7, -22, -14, -21, -21, -13, -4, 15, -21, -6, -22; /M: SY='I'; M= -7, -28, -19, -33, -25, 10, -30, -24, 24, -26, 23, 14, -23, -25, -23, -23, -19, -7, 19, -18, 0, -25; /M: SY='G'; M= 12, -9, -20, -11, -15, -25, 38, -18, -28, -16, -24, -17, -2, -15, -15, -18, 7, -8, -18, -23, -25, -15; /M: SY='I'; M= -4, -23, -18, -28, -22, 6, -26, -21, 16, -22, 15, 9, -19, -23, -19, -20, -11, 0, 15, -19, 0, -21; /M: SY='L'; M= -3, -22, -22, -24, -17, -2, -18, -20, 3, -20, 5, 2, -19, 0, -18, -20, -12, -6, 1, -20, -9, -18; /M: SY='L'; M= 2, -22, -17, -27, -20, 3, -20, -21, 12, -22, 17, 8, -20, -23, -19, -20, -12, -4, 11, -19, -4, -20; /M: SY='G'; M= 24, -10, -18, -15, -14, -23, 27, -20, -21, -15, -18, -14, -5, -15, -14, -19, 6, -7, -12, -21, 23, -14; /M: SY='Y'; M= -8, -24, -22, -28, -22, 13, -27, -14, 14, -21, 13, 8, -21, -25, -19, -18, -17, -6, 10, -4, 16, -22; /M: SY='L'; M= 3, -22, -19, -27, -20, 3, -17, -18, 8, -20, 11, 6, -19, -22, -18, -19, -12, -6, 7, -13, 0, -19; Explicit I (I) I=-4; MD=-8; /M: SY='L'; M= 1, -19, -17, -23, -16, 1, -19, -17, 6, -16, 11, 7, -16, -21, -14, -12, -9, -3, 6, -19, -5, -15; /I: I=-4; MD=-8; </pre>

Explicit D

MyHits and profiles

- The **MyHits** service:
 - search and scan sequences with profiles and patterns;
 - build profiles starting from a MSA;
 - store personal sequences, patterns, and profiles;
 - other protein domain descriptors are available (HMMs);
 - very informative graphical representation of the alignments.

MyHits: motif scan

user: Icerutti
[log out](#)

Motif Scan

Protein Sequence Input
Enter a protein sequence in RAW or FASTA or Swiss-Prot format or a db:AC or db:ID identifier

Raw format

```
#COMMENT: RAW means plain-text alphabetic only (no r
MQIFVETLTGKTITLEVEPSDTIENVKAKIQDKEGIPPPDOORLIFAGKOLE
GRTLSDYNIQKESTLHLVLRLLGG
```

[Clear input](#)
[Reset page](#)

Motif scanning means finding all known motifs that occur in a sequence. This form lets you paste a protein sequence, select the collections of motifs to scan for, and launch the search. Some general [documentation](#) is available about the Prosite and Pfam collections of motifs. Another [document](#) deals with the interpretation of the match scores. You should consult the home pages of [Prosite](#) on ExPASy, [Pfam](#) and [InterPro](#) for additional information.

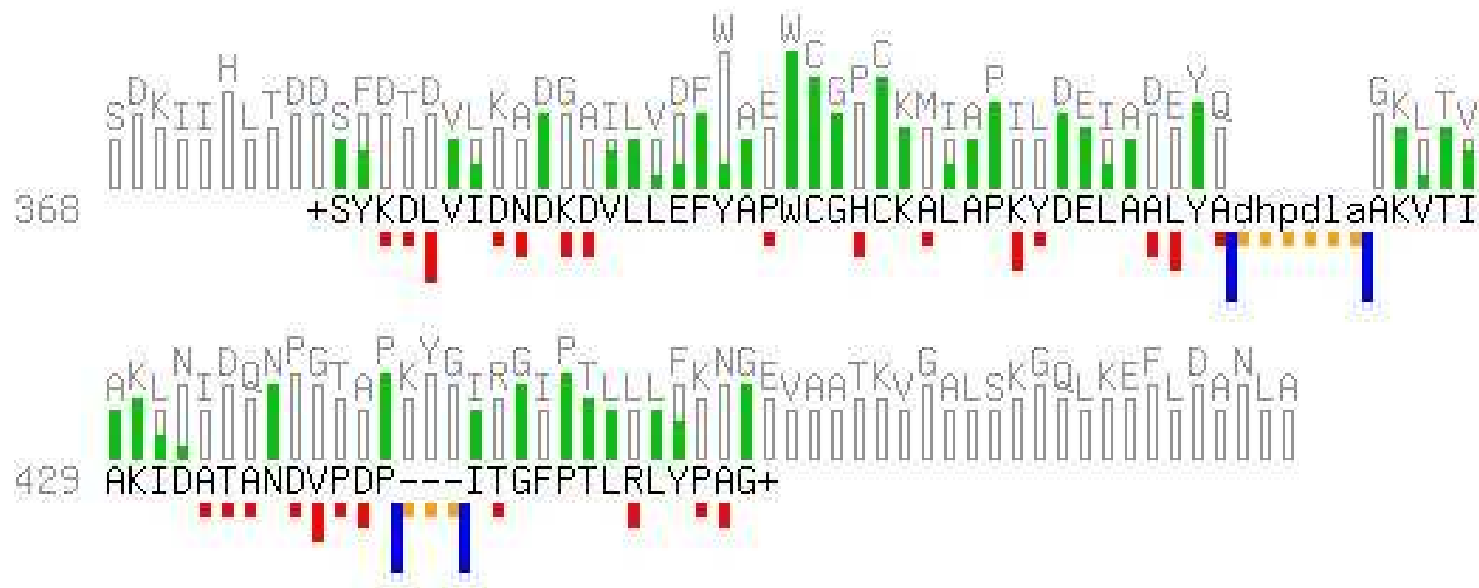
Warning: The scan might take a few minutes, thus if your proteins of interest are already in the sequence databases (see [list](#)), the [Query by Protein](#) form is much faster, and the [Protein Hub](#) provides a collection of tools that you might find useful.

Parameters

Database of motifs (db description)	<input type="checkbox"/> Still another private prf database of Icerutti <input type="checkbox"/> PROSITE patterns <input type="checkbox"/> PROSITE patterns (frequent match producers) <input checked="" type="checkbox"/> Another private prf database of Icerutti <input checked="" type="checkbox"/> PROSITE profiles <input checked="" type="checkbox"/> Prefile (more profiles) <input checked="" type="checkbox"/> HAMAP profiles <input type="checkbox"/> Pfam HMMs (local models) <input type="checkbox"/> Pfam HMMs (global models)	search
--	--	------------------------

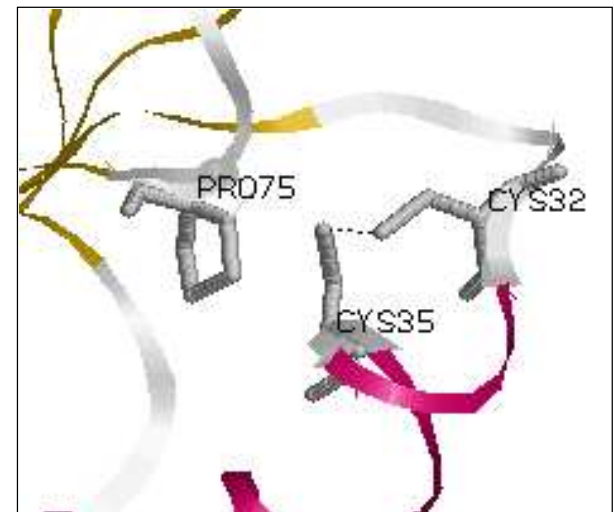
Pairwise alignment vs. Profile

- Smith-Waterman alignment of two thioredoxin domains:



Pairwise alignment vs. Profile (2)

- Thioredoxin domain aligned with a profile build from a MSA of thioredoxins:



Generalized profiles: software

- The package **Pftools** contains all the tools required to build and use generalized profiles (<http://www.isrec.isb-sib.ch/ftp-server/pftools/>)
- The package contains:
 - *pfmake* to build a profile from a MSA
 - *pfcalibrate* to calibrate the profile
 - *pfsearch* to search a protein database with a profile
 - *pfscan* to scan a protein against profiles
 - ...

Generalized profiles: conclusion

- Advantage:
 - deal with indels
 - very sensitive to detect homologies below the twilight zone
 - scoring system
 - tools for building and calibrating the profile
- Limitations:
 - sophisticated software
 - CPU expensive
 - user expertise is required

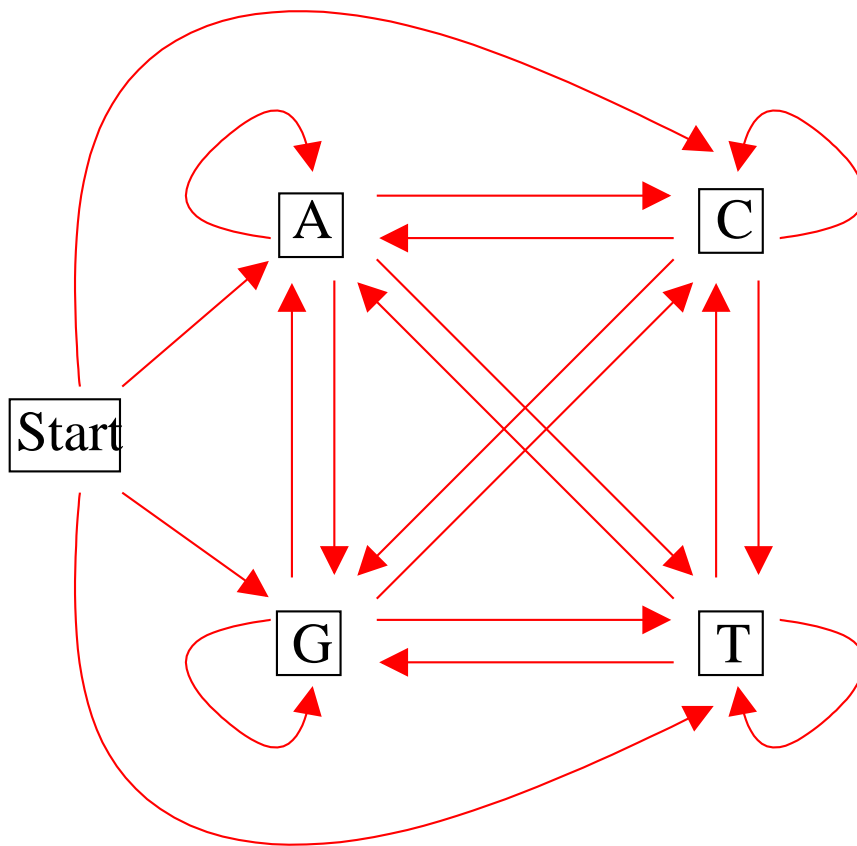
Outline

- Introduction
 - Reminder on pairwise alignments
 - Multiple alignments and their information content
- Models of multiple alignments and databases
 - Consensus sequences
 - Patterns and regular expressions
 - Position Specific Scoring Matrices (PSSMs)
 - Generalized Profiles
 - **Hidden Markov Models (HMMs)**
- PSI-BLAST and protein domain hunting

Hidden Markov Models (HMMs)

- Hidden Markov Models (HMMs) are an extension of the Markov Chain theory, which is part of the theory of probabilities.
- A Markov Chain is a succession of states s_i ($i = 0, 1, \dots$) connected by transitions.
- A probability P_{ij} is associated to each transitions from a state s_i to a state s_j .

Example of a Markov Chain



Transition probabilities

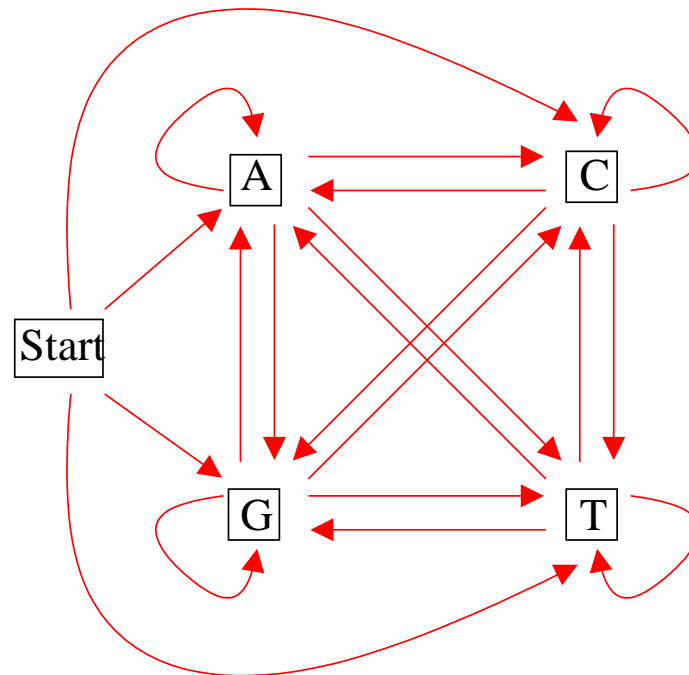
$$P(A|G) = 0.18, P(C|G) = 0.38,$$

$$P(G|G) = 0.32, P(T|G) = 0.12,$$

$$P(A|C) = 0.15, P(C|C) = 0.35,$$

...

Probability of a Markov Chain

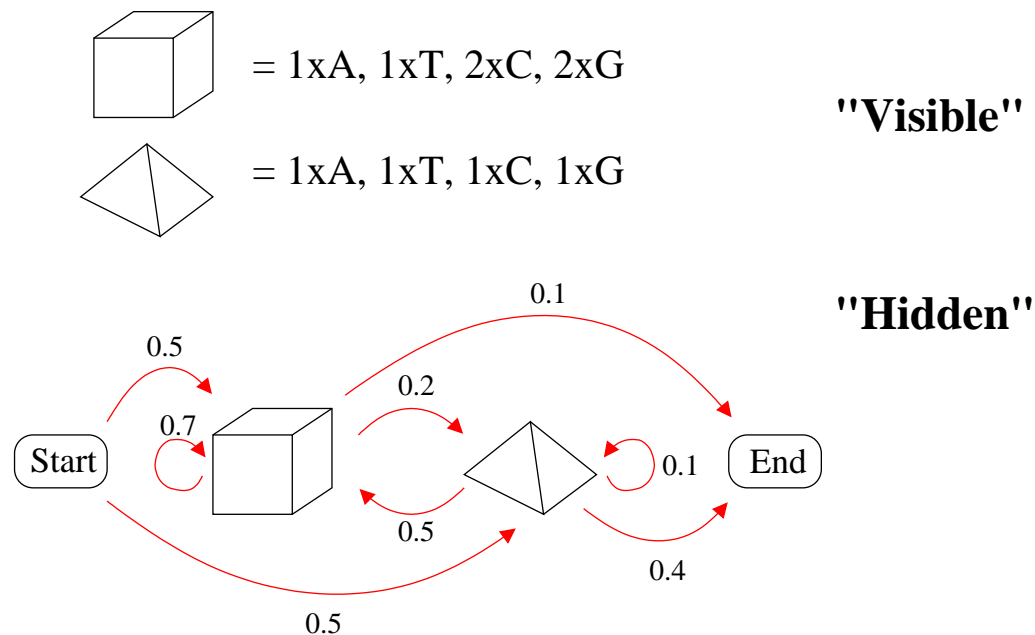


The probability of sequence $x = GCCT$ is:

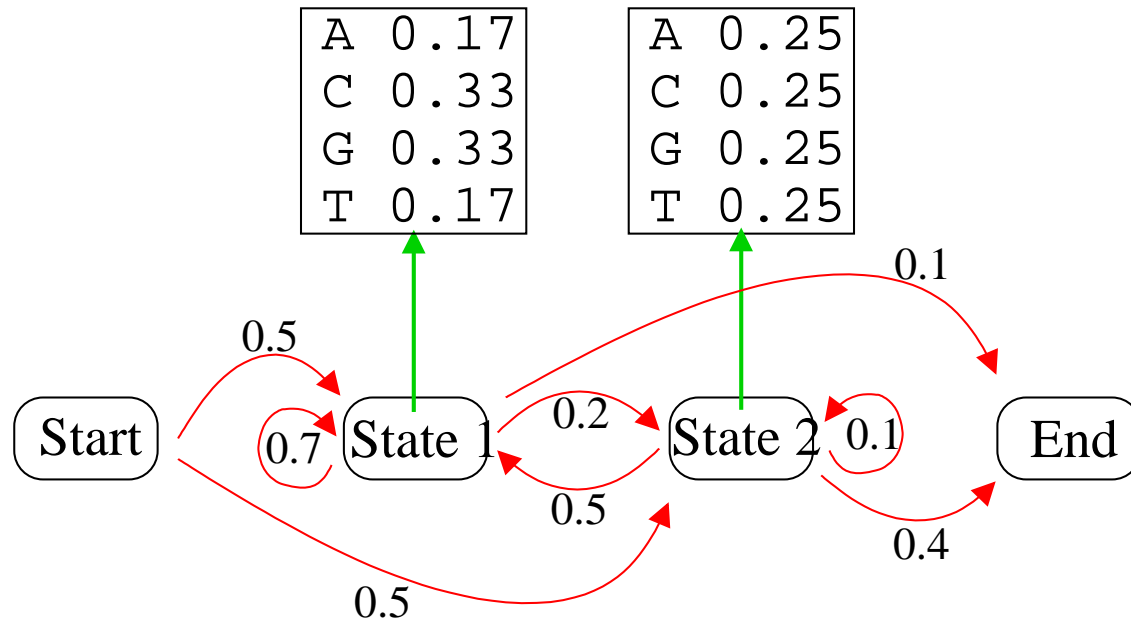
$$P(GCCT) = P(T|C)P(C|C)P(C|G)P(G)$$

Markov Chains to HMMs

- HMMs are like Markov Chains: a finite number of states connected by transitions, but ...
- States in a HMMs are not symbols but **distribution** of symbols.



HMM for GC rich DNA



"Visible"

"Hidden"

START	1	1	1	1	2	2	1	1	1	2	END
	G	C	A	G	C	T	G	G	C	T	

HMMs: parameters

- **Emission probabilities**: the probability of emitting a symbol x from an alphabet \mathcal{A} being in state q :

$$E(x|q)$$

- **Transition probabilities**: probability of a transition to state r being in state q :

$$T(r|q)$$

- **Initiation probabilities**: probability to start in state q :

$$I(q)$$

HMMs: algorithms

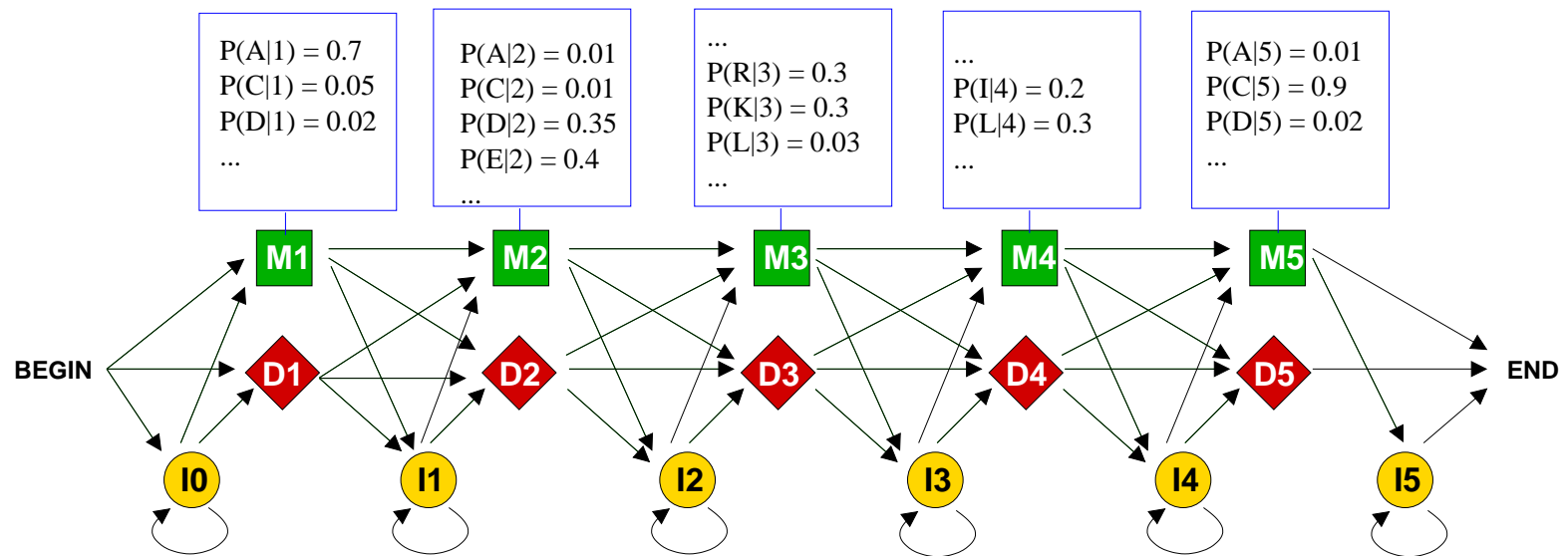
- How likely is a given sequence under a given model?
 - this is the scoring problem and can be solved using the **forward algorithm**
- which is the most probable path between states to model a sequence?
 - this is the alignment problem and can be solved using the **Viterbi algorithm**
- How can we learn the HMM parameters given a MSA?
 - this is the training problem and is solved using the **forward-backward algorithm** and the **Baum-Welch expectation maximization**.

HMMs: algorithms (2)

- For details about the algorithms see:
 - Durbin, Eddy, Mitchison, Krog
Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids.
Cambridge University Press, 1998
 - Baldi, Brunak
Bioinformatics: The Machine Learning Approach,
2nd edition.
The MIT Press, 2001

HMMs: example

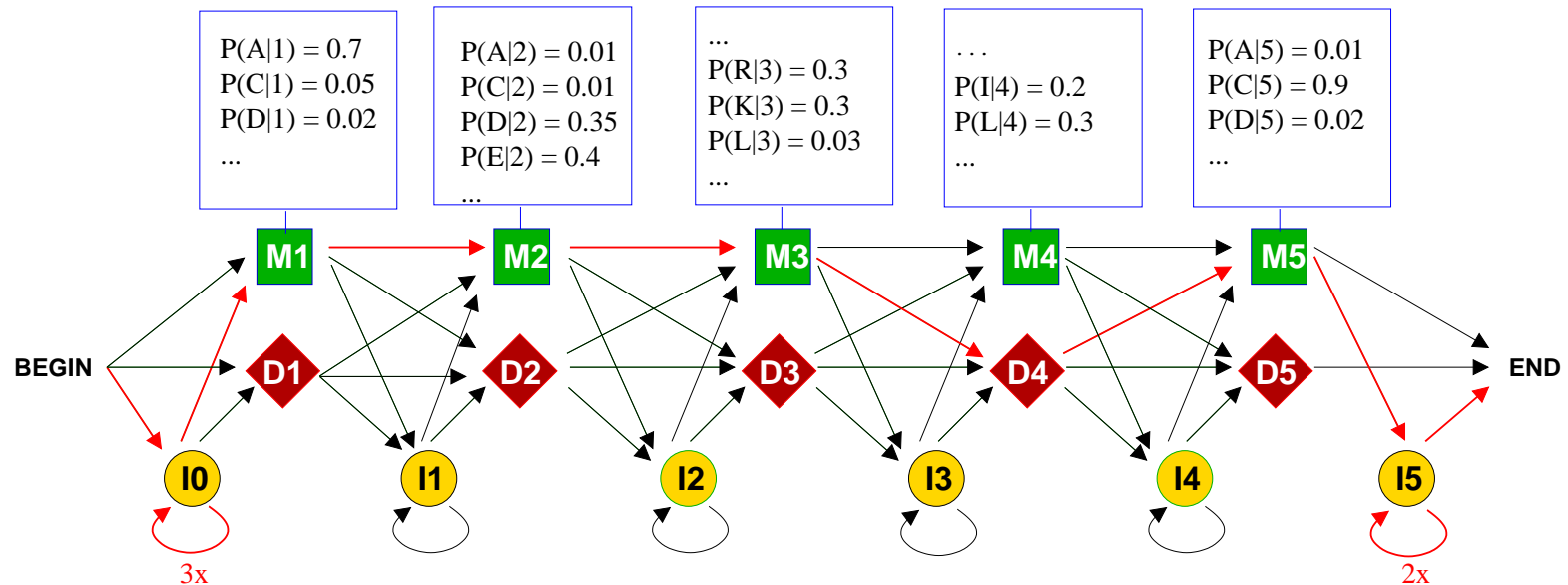
ADRL-C
AERIRC
VEKI-C
ADK--C
AEKL-C



HMMs: alignment example

VGGAERCSA

Align sequence to the model: Viterbi algorithm
(find the best path between states to model the sequence)



V G G A E R - C S A
 I0 (3x) M1 M2 M3 D4 M5 I5 (2x)

HMMs: software

- **HMMER2** is a package to build and use HMMs (<http://hmmer.wustl.edu/>).
 - *hmmbuild*: build HMM model from a MSA
 - *hmmcalibrate*: calibrate a HMM
 - *hmmsearch*: search a database with a HMM model
 - *hmmpfam*: scan a sequence with a HMM database
 - *hmmalign*: align sequences to a HMM
 - *hmmemit*: emit sequences from a HMM
- **SAM** is another package for HMMs (<http://www.cse.ucsc.edu/research/compbio/sam.html>)

Generalized profiles vs. HMMs

- Generalized profiles are equivalent to linear-HMMs like those of HMMER2 and SAM.
- The optimal alignment produced by dynamic programming with generalized profiles is equivalent to the Viterbi path on a HMM.
- The **Pftools** package contains translators:
 - *htop*: HMM to Generalized profile
 - *ptoh*: Generalized profile to HMM
- Generalized profiles allow manual tuning (by a well trained expert). This is very difficult with HMMs.

HMMs databases

- **Pfam** is a large collection of HMM models (8183 HMMs in Version 19.0), describing protein motifs, domains, and families (<http://www.sanger.ac.uk/Software/Pfam/>).
- **Smart** is another collection of HMM models for protein domains. Excellent taxonomic and localization information (<http://smart.embl-heidelberg.de/>).
- **tigrfam** is a database of HMMs for protein families (<http://www.tigr.org/TIGRFAMs/>).
- **SCOP Superfamily**: a collection of HMM models describing SCOP protein superfamilies (common structure) (<http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/>).

Pfam scan

The screenshot shows a web browser window with the address bar displaying `http://www.sanger.ac.uk/Software/Pfam/search.shtml`. The browser tabs include "UniFR course - schedule and teachers" and "Pfam: Search Pfam". The website header features the Wellcome Trust Sanger Institute logo on the left and the "Pfam" logo on the right, with a search input field and a "Search Pfam" button. A navigation menu below the header includes links for "Pfam Home", "Search by", "Browse by", "FTP", "iPfam", "Help", and "About".

The main content area is divided into two sections:

- By UniProt Identifier:** This section has a heading "By UniProt Identifier" and a sub-heading "Enter a UniProt name or accession number". It contains a text input field, a "Submit Query" button, a "Reset" button, and an "Example" button. To the right, a note states: "Pfam has pre-calculated the domain structure of the proteins in UniProt. If you know the name or accession number (e.g. [YAV_HUMAN](#) or [O91437](#)) then you can see the Pfam domains on the sequence instantaneously."
- By Protein sequence:** This section has a heading "By Protein sequence" and a sub-heading "Single sequence searches". It explains that if the UniProt identifier is unknown, a slower HMM search can be performed. It includes a large text area for "Cut and Paste your sequence here (This search will take 1-5 minutes)". To the right of this area are "Pfam Search Options" with a "Search type:" dropdown set to "Both Global & Fragment Pfam search" and an "Output format:" dropdown set to "Graphical output". A note below these options says: "* Searching against SMART and TIGR hmm's has been disabled. It should return shortly. *". There is also an "E-value cutoff level:" input field set to "1.0" and a link: "For help on the scores in Pfam, and the difference between standard and fragment searches, click [here](#)".

Below the "By Protein sequence" section, there is an alternative search method: "Or: Select the sequence file you wish to use". It includes a file input field, a "Browse..." button, and "Search Pfam", "Reset", and "Example" buttons. At the bottom, there is a section for "Other regions to search for:" with a checkbox for "low-complexity (seg)" which is currently unchecked.

At the very bottom, there is a section for "Large batch searches" with the text: "To do large scale searching against Pfam, you can upload a TEXT file (Not Word) in FASTA format. This resource is primarily for people who do not have access to large computing facilities or personnel to install HMMER locally."

MyHits Pfam scan

Protein Sequence Input
Enter a protein sequence in RAW or FASTA or Swiss-Prot format or a db:AC or db:ID identifier

Raw format

```
#COMMENT: RAW means plain-text alphabetic only (no numl
MQIFVKLTLTGKITLEVEPSDTIENVKAKIQDKEGIPPDQQRLLIFAGKQLED
GRTLSDYNIQKESLHLVLRLLRG
```

Clear input
Reset page

Motif scanning means finding all known motifs that occur in a sequence. This form lets you paste a protein sequence, select the collections of motifs to scan for, and launch the search. Some general [documentation](#) is available about the Prosite and Pfam collections of motifs. Another [document](#) deals with the interpretation of the match scores. You should consult the home pages of [Prosite](#) on ExPASy, [Pfam](#) and [InterPro](#) for additional information.

Warning: The scan might take a few minutes, thus if your proteins of interest are already in the sequence databases (see [list](#)), the [Query by Protein](#) form is much faster, and the [Protein Hub](#) provides a collection of tools that you might find useful.

Parameters

Database of motifs (db description)	<input type="checkbox"/> Still another private prf database of Icerutti [naprf] <input type="checkbox"/> PROSITE patterns [pat] <input type="checkbox"/> PROSITE patterns (frequent match producers) [freq_pat] <input type="checkbox"/> Another private prf database of Icerutti [lcprf] <input type="checkbox"/> PROSITE profiles [prf] <input type="checkbox"/> Profile (more profiles) [pre] <input type="checkbox"/> HAMAP profiles [hamap] <input checked="" type="checkbox"/> Pfam HMMs (local models) [pfam_fs]	search
--	--	--------

Smart scan

Sequence analysis

You may use either the Swissprot/Sptrembl/Ensembl sequence identifier (ID) / accession number (ACC) or the protein sequence itself to request the SMART service.

Sequence ID or ACC

Sequence

HMMER searches of the SMART database occur by default. You may also find:

- [Outlier homologues](#) and homologues of known structure
- [PFAM domains](#)
- [signal peptides](#)
- [internal repeats](#)
- [intrinsic protein disorder](#)

[Click here](#) to view your saved searches.

If you have multiple sequences to analyze, try [batch access](#) to SMART database.

Architecture analysis

You can search for proteins with combinations of [specific domains](#) in different species or taxonomic ranges. You can input the domains directly into "Domain selection" box, or use "GO terms query" to get a list of domains. See [What's New](#) for more info.

Domain selection

Example: TyrKc AND SH3 AND NOT SH2

GO terms query

Example: membrane AND signal transduction

Taxonomic selection

Select a taxonomic range via the selection box or type it into the text box below:

Examples: Dictyostelium discoideum, Porifera

You can try an [Advanced Query](#) if you're familiar with SQL.

Alert SMART

If you want to be automatically informed each time a new protein with a defined domain composition is deposited in the database, please use 'Alert SMART' (this facility is also available following an architecture analysis query).


InterPro

- The InterPro consortium attempts to group a number of protein domain databases:
 - PROSITE, Pfam, Prints, Smart, ProDom, TIGRFAMs, PANTHER, Gene3D, ...
 - PIR Superfamily (PIRSF): classification system based on evolutionary relationship
 - SCOP Superfamily: structure derived HMMs.
- High quality annotation.
- Good access to examples and taxonomy.
- <http://www.ebi.ac.uk/interpro/>

InterPro scan

InterProScan Sequence Search

This form allows you to query your sequence against InterPro. For more detailed information see the documentation for the perl stand-alone InterProScan package ([Readme file](#) or [FAQs](#)), or the InterPro [user manual](#) or [help pages](#).

 [Download Software](#)

YOUR EMAIL	RESULTS
<input type="text"/>	interactive ▾
APPLICATIONS TO RUN <input type="radio"/> Clear all <input checked="" type="radio"/> Check all	
<input checked="" type="checkbox"/> BlastProDom	<input checked="" type="checkbox"/> FPrintScan
<input checked="" type="checkbox"/> HMMTigr	<input checked="" type="checkbox"/> ProfileScan
<input checked="" type="checkbox"/> TMHMM	<input checked="" type="checkbox"/> HMMPanther
<input checked="" type="checkbox"/> HMMPiR	<input checked="" type="checkbox"/> ScanRegExp
<input checked="" type="checkbox"/> HMMPfam	<input checked="" type="checkbox"/> SuperFamily
<input checked="" type="checkbox"/> HMMSmart	<input checked="" type="checkbox"/> SignalPHMM
<input checked="" type="checkbox"/> Gene3D	
TRANSLATION TABLE (DNA/RNA only)	MIN. OPEN READING FRAME SIZE
None ▾	100 ▾

Enter or Paste a Sequence in any format:

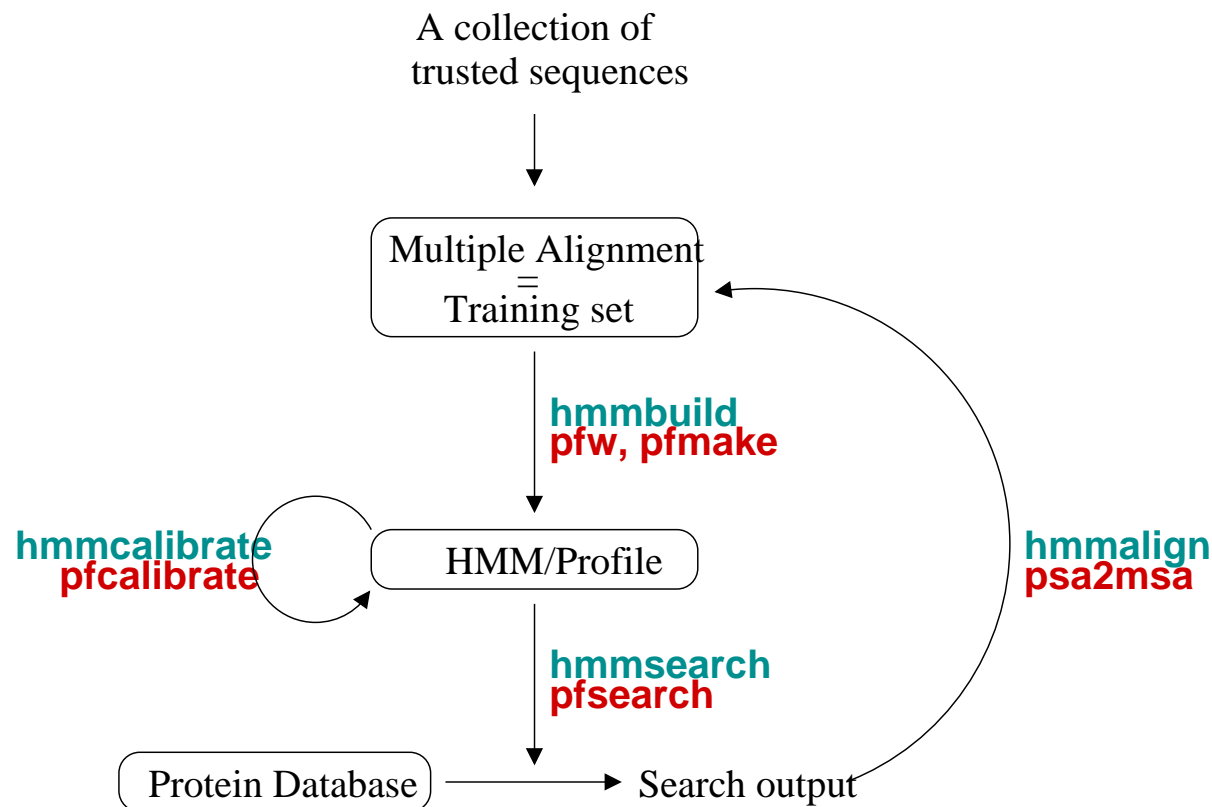
Upload a file:

Outline

- Introduction
 - Reminder on pairwise alignments
 - Multiple alignments and their information content
- Models of multiple alignments and databases
 - Consensus sequences
 - Patterns and regular expressions
 - Position Specific Scoring Matrices (PSSMs)
 - Generalized Profiles
 - Hidden Markov Models (HMMs)
- **PSI-BLAST and protein domain hunting**

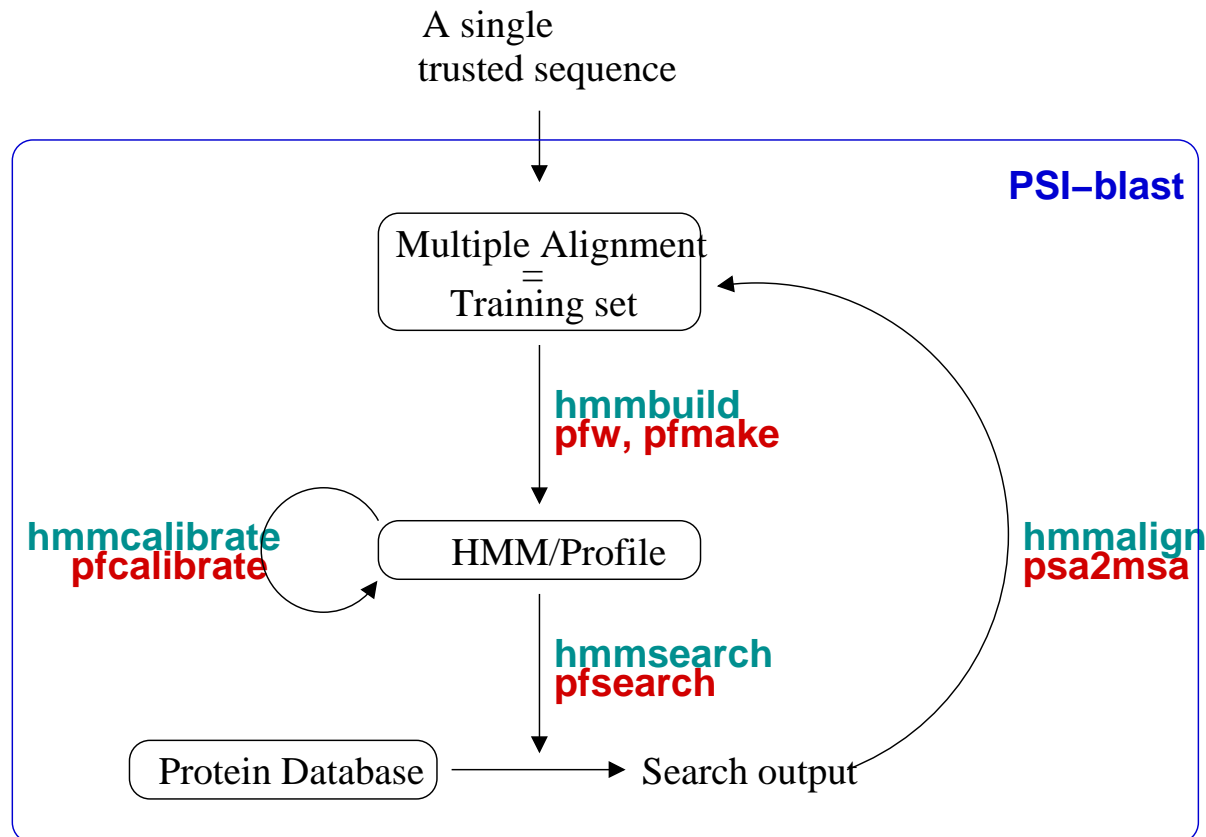
Protein domain hunting

- The **Pftools** and **HMMER2** packages can be used for protein domain hunting ... but CPU expensive.



Protein domain huntin (2)

- PSI-BLAST is faster and simpler to use ... but uses heuristics!



PSI-BLAST principle

1. A standard BLAST search is performed against a database using a substitution matrix (e.g. BLOSUM62).
2. A PSSM with position independent affine gap cost ([checkpoint](#)) is derived automatically from the alignments of the highest scoring hits.
3. The PSSM replaces the initial matrix to perform a new search in the database.
4. Step 2 and 3 can be repeated including the new detected sequences.
5. The PSI-BLAST has [converged](#) if no new sequences are included in the last cycle.

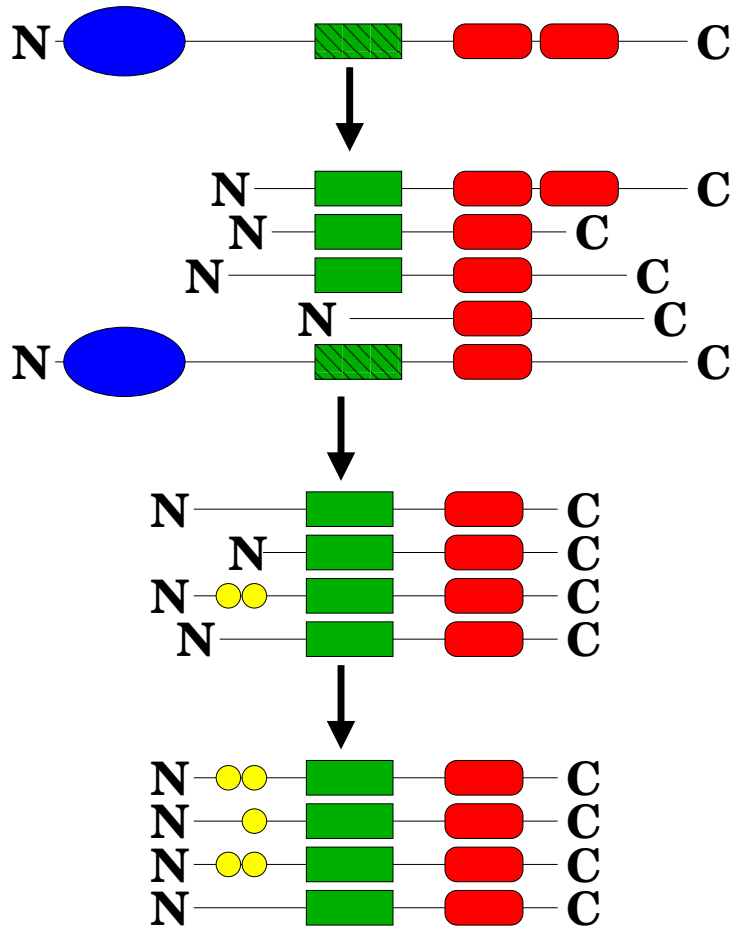
PSI-BLAST advantages

- Fast because of BLAST heuristics.
- Allows PSSMs searches on large databases.
- Efficient algorithm for sequence weighting.
- Sophisticated statistical treatment of the match scores.
- Single software.
- User friendly interface.

PSI-BLAST dan er

- Avoid too similar sequences: over fit!
- Can include false homologous. Check match sequences carefully and include/exclude sequences based on biological knowledge.
- The E-value reflects the significance of the match to the previous training set **not** the original sequence.
- Try reverse experiment to certify.
- No control on the multiple alignments produced at each cycle.

PSI-BLAST dan er (2)



**WRONG
ANNOTATION!**

MyHits for protein domain hunting

- MyHits service (<http://myhits.isb-sib.ch>) is an excellent environment for protein domain hunting.
- Full control of the PSI-BLAST:
 - user can check and re-align the selected sequences at each PSI-BLAST cycle
- User can build its own profiles and use them to search a database.
- Large number of sequences available.
- Alignments can be used to transfer annotation.

● ... lunch!