Introduction to Patterns, Profiles and Hidden Markov Models

Marco Pagni
Swiss Institute of Bioinformatics (SIB)

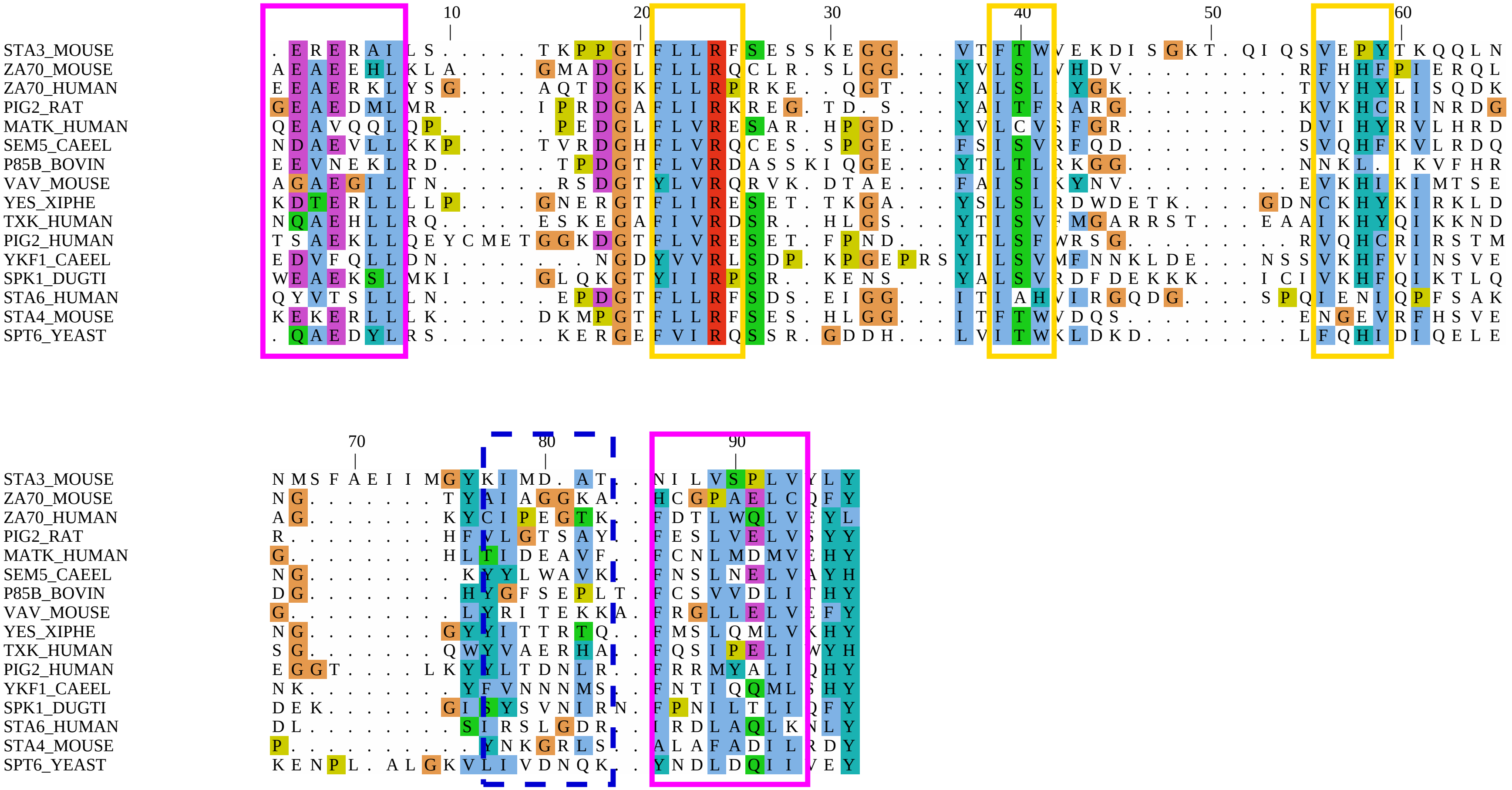
30th August 2002

Multiple alignments

Multiple sequence alignment (MSA)

- The alignment of multiple sequences is a method of choice to detect conserved regions in protein or DNA sequences. These particular regions are usually associated with:
    - Signals (promoters, signatures for phosphorylation, cellular location, ...);
    - Structure (correct folding, protein-protein interactions...);
    - Chemical reactivity (catalytic sites,...).
- The information represented by these regions can be used to align sequences, search similar sequences in the databases or annotate new sequences.
- Different methods exist to build *models* of these conserved regions:
    - Consensus sequences;
    - Patterns;
    - Position Specific Score Matrices (PSSMs);
    - Profiles;
    - Hidden Markov Models (HMMs),
    - ... and a few others.

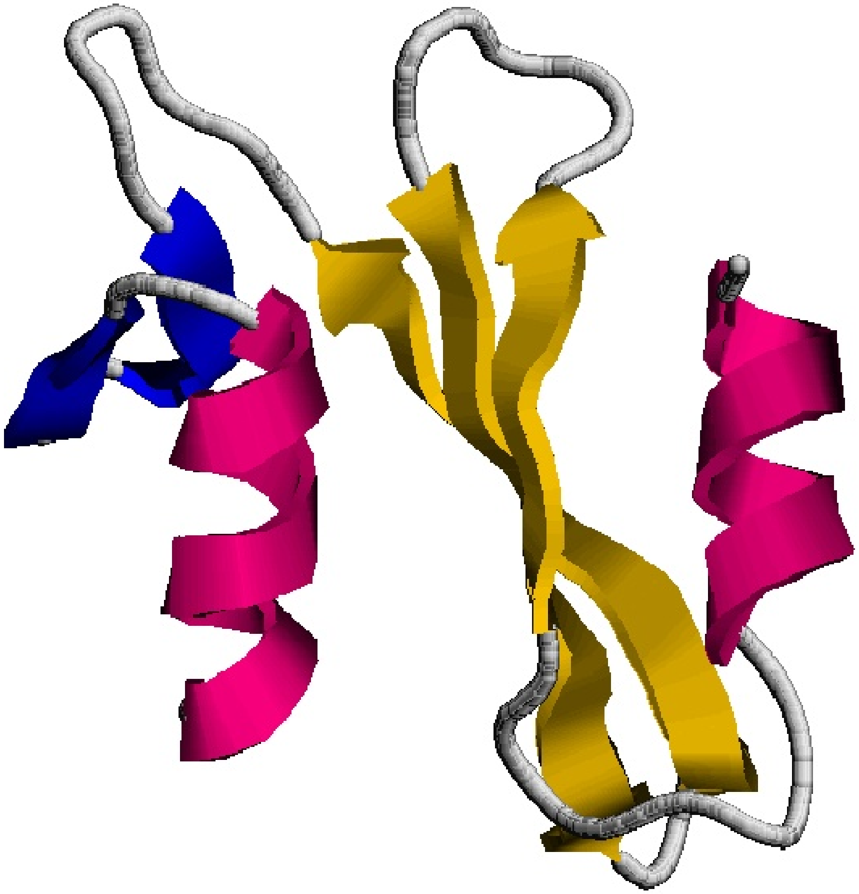
Multiple alignments reflect secondary structures



```
                              10                  20                  30                  40                  50                  60
STA3_MOUSE  . E R E R A I L S . . . . . T K P P G T F L L R F S E S S K E G G . . . V T F T W V E K D I S G K T . Q I Q S V E P Y T K Q Q L N
ZA70_MOUSE  A E A E E H L K L A . . . . G M A D G L F L L R Q C L R . S L G G . . . Y V L S L V H D V . . . . . . . . . R F H H F P I E R Q L
ZA70_HUMAN  E E A E R K L Y S G . . . . A Q T D G K F L L R P R K E . . Q G T . . . Y A L S L I Y G K . . . . . . . . . T V Y H Y L I S Q D K
PIG2_RAT    G E A E D M L M R . . . . . I P R D G A F L I R K R E G . T D . S . . . Y A I T F R A R G . . . . . . . . . K V K H C R I N R D G
MATK_HUMAN  Q E A V Q Q L Q P . . . . . . P E D G L F L V R E S A R . H P G D . . . Y V L C V S F G R . . . . . . . . . D V I H Y R V L H R D
SEM5_CAEEL  N D A E V L L K K P . . . . T V R D G H F L V R Q C E S . S P G E . . . F S I S V R F Q D . . . . . . . . . S V Q H F K V L R D Q
P85B_BOVIN  E E V N E K L R D . . . . . . T P D G T F L V R D A S S K I Q G E . . . Y T L T L R K G G . . . . . . . . . N N K L . I K V F H R
VAV_MOUSE   A G A E G I L T N . . . . . . R S D G T Y L V R Q R V K . D T A E . . . F A I S I K Y N V . . . . . . . . . E V K H I K I M T S E
YES_XIPHE   K D T E R L L L L P . . . . G N E R G T F L I R E S E T . T K G A . . . Y S L S L R D W D E T K . . . . G D N C K H Y K I R K L D
TXK_HUMAN   N Q A E H L L R Q . . . . . E S K E G A F I V R D S R . . H L G S . . . Y T I S V F M G A R R S T . . . E A A I K H Y Q I K K N D
PIG2_HUMAN  T S A E K L L Q E Y C M E T G G K D G T F L V R E S E T . F P N D . . . Y T L S F W R S G . . . . . . . . . R V Q H C R I R S T M
YKF1_CAEEL  E D V F Q L L D N . . . . . . . . N G D Y V V R L S D P . K P G E P R S Y I L S V M F N N K L D E . . . N S S V K H F V I N S V E
SPK1_DUGTI  W E A E K S L M K I . . . . G L Q K G T Y I I R P S R . . K E N S . . . Y A L S V R D F D E K K K . . . I C I V K H F Q I K T L Q
STA6_HUMAN  Q Y V T S L L L N . . . . . . E P D G T F L L R F S D S . E I G G . . . I T I A H V I R G Q D G . . . . S P Q I E N I Q P F S A K
STA4_MOUSE  K E K E R L L L K . . . . . D K M P G T F L L R F S E S . H L G G . . . I T F T W V D Q S . . . . . . . . . E N G E V R F H S V E
SPT6_YEAST  . Q A E D Y L R S . . . . . . K E R G E F V I R Q S S R . G D D H . . . L V I T W K L D K D . . . . . . . . L F Q H I D I Q E L E

                    70                  80                  90
STA3_MOUSE  N M S F A E I I M G Y K I M D . A T . . N I L V S P L V Y L Y
ZA70_MOUSE  N G . . . . . . . T Y A I A G G K A . . H C G P A E L C Q F Y
ZA70_HUMAN  A G . . . . . . . K Y C I P E G T K . . F D T L W Q L V E Y L
PIG2_RAT    R . . . . . . . . H F V L G T S A Y . . F E S L V E L V S Y Y
MATK_HUMAN  G . . . . . . . . H L T I D E A V F . . F C N L M D M V E H Y
SEM5_CAEEL  N G . . . . . . . . K Y Y L W A V K . . F N S L N E L V A Y H
P85B_BOVIN  D G . . . . . . . . H Y G F S E P L T . F C S V V D L I T H Y
VAV_MOUSE   G . . . . . . . . . L Y R I T E K K A . F R G L L E L V E F Y
YES_XIPHE   N G . . . . . . . G Y Y I T T R T Q . . F M S L Q M L V K H Y
TXK_HUMAN   S G . . . . . . . Q W Y V A E R H A . . F Q S I P E L I W Y H
PIG2_HUMAN  E G G T . . . . L K Y Y L T D N L R . . F R R M Y A L I Q H Y
YKF1_CAEEL  N K . . . . . . . . Y F V N N N M S . . F N T I Q Q M L S H Y
SPK1_DUGTI  D E K . . . . . . G I S Y S V N I R N . F P N I L T L I Q F Y
STA6_HUMAN  D L . . . . . . . . S I R S L G D R . . I R D L A Q L K N L Y
STA4_MOUSE  P . . . . . . . . . . Y N K G R L S . . A L A F A D I L R D Y
SPT6_YEAST  K E N P L . A L G K V L I V D N Q K . . Y N D L D Q I I V E Y
```

Multiple alignments reflect secondary structures

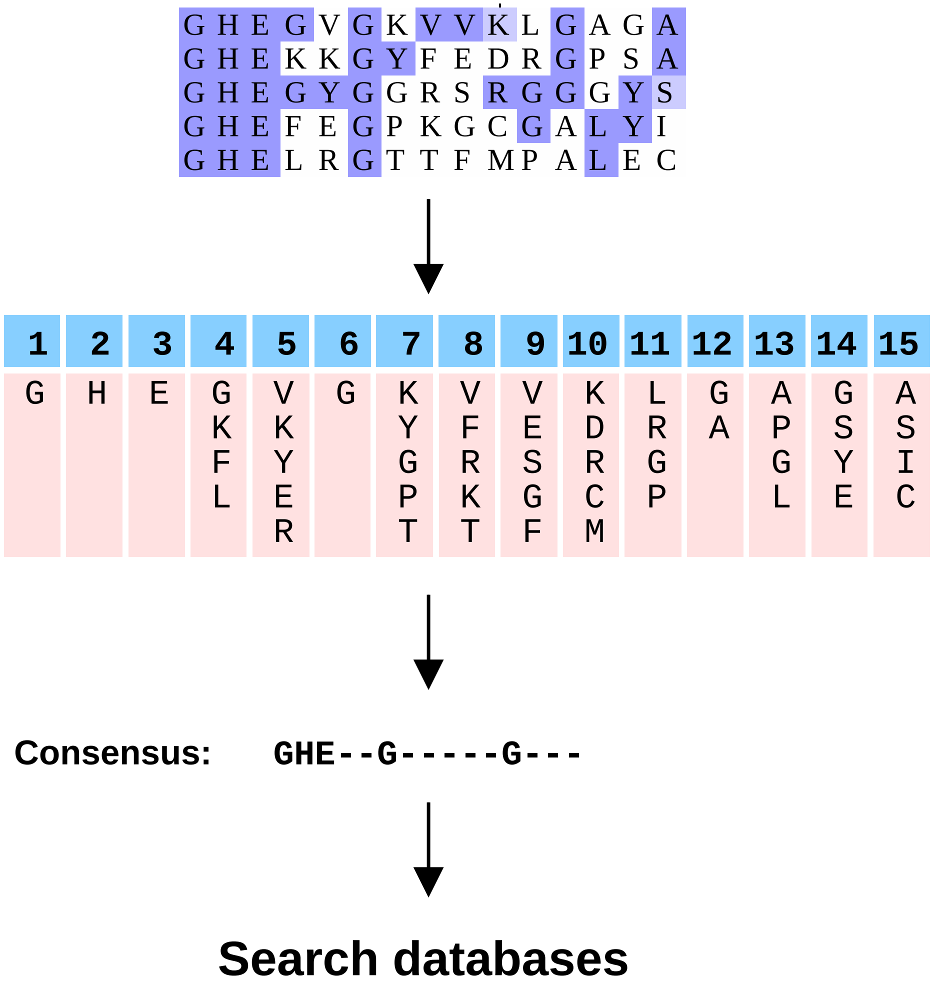


Consensus sequences

Consensus sequences

- The consensus sequence method is the simplest method to build a model from a multiple sequence alignment.
- The consensus sequence is built using the following rules:
    - Majority wins.
    - Skip too much variation.

How to build consensus sequences



```
GHEGVGKVVKLGAGA
GHEKKGYFEDRGPSA
GHEGYGGRSRGGGYS
GHEFEGPKGCGALYI
GHELRGTTFMPALEC
```

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
G	H	E	G	V	G	K	V	V	K	L	G	A	G	A
			K	K		Y	F	E	D	R	A	P	S	S
			F	Y		G	R	S	R	G		G	Y	I
			L	E		P	K	G	C	P		L	E	C
				R		T	T	F	M					

Consensus: GHE--G-----G---

Search databases

Consensus sequences

- Advantages:
  - This method is very fast and easy to implement.
- Limitations:
  - Models have no information about variations in the columns.
  - Very dependent on the training set.
  - No scoring, only binary result.
- When I use it?
  - May be of some use to find highly conserved signatures, as for example enzyme restriction sites for DNA.

Pattern matching

Pattern syntax

- A pattern describes a set of alternative sequences, using a single expression. In computer science, patterns are known as regular expressions.
- The Prosite syntax for patterns:
  - uses the standard IUPAC one-letter codes for amino acids (G=Gly, P=Pro, ...),
  - each element in a pattern is separated from its neighbor by a '-',
  - the symbol 'X' is used where any amino acid is accepted,
  - ambiguities are indicated by square parentheses '[ ]' ([AG] means Ala or Gly),
  - amino acids that are not accepted at a given position are listed between a pair of curly brackets '{ }' ({AG} means any amino acid except Ala and Gly),
  - repetitions are indicated between parentheses '( )' ([AG](2,4) means Ala or Gly between 2 and 4 times, X(2) means any amino acid twice),
  - a pattern is anchored to the N-term and/or C-term by the symbols '<' and '>' respectively.

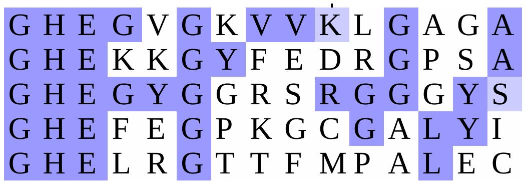
Pattern syntax: an example

- The following pattern

$$<\textbf{A-x-[ST](2)-x(0,1)-\{V\}}$$

- means:
  - an Ala in the N-term,
  - followed by any amino acid,
  - followed by a Ser or Thr twice,
  - followed or not by any residue,
  - followed by any amino acid except Val.

How to build a pattern



```
G H E G V G K V V K L G A G A
G H E K K G Y F E D R G P S A
G H E G Y G G R S R G G G Y S
G H E F E G P K G C G A L Y I
G H E L R G T T F M P A L E C
```

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
G	H	E	G	V	G	K	V	V	K	L	G	A	G	A
			K	K		Y	F	E	D	R	A	P	S	S
			F	Y		G	R	S	R	G		G	Y	I
			L	E		P	K	G	C	P		L	E	C
				R		T	T	F	M					

Profile: **`G-H-E-X(2)-G-X(5)-[GA]-X(3)`**

Search databases

Pattern examples

- Patterns and PSSMs are appropriate to build models of short sequence signatures.
- Example of short signatures:
    - Post-translational signatures:
        - Protein splicing signature: [DNEG]-x-[LIVFA]-[LIVMY]-[LVAST]-H-N-[STC]
        - Tyrosine kinase phosphorylation site: [RK]-x(2)-[DE]-x(3)-Y or [RK]-x(3)-[DE]-x(2)-Y
        - ...
    - DNA-RNA interaction signatures:
        - Histone H4 signature: G-A-K-R-H
        - p53 signature: M-C-N-S-S-C-[MV]-G-G-M-N-R-R
        - ...
    - Enzymes:
        - L-lactate dehydrogenase active site: [LIVMA]-G-[EQ]-H-G-[DN]-[ST]
        - Ubiquitin-activating enzyme signature: P-[LIVM]-C-T-[LIVM]-[KRH]-x-[FT]-P
        - ...

Patterns: Conclusion

- Advantages:
  - Pattern matching is fast and easy to implement.
  - Models are easy to design for anyone with some training in biochemistry.
  - Models are easy to understand for anyone with some training in biochemistry.
- Limitations:
  - Poor model for insertions/deletions (indels).
  - Small patterns find a lot of false positives. Long patterns are very difficult to design.
  - Poor predictors that tend to recognize only the sequence of the training set.
  - No scoring system, only binary response.
- When I use patterns?
  - To search for small signatures or active sites.
  - To communicate with other biologists.

Patterns: beyond the conclusion

- ▷ Patterns can be automatically extracted (discovered) from a set of unaligned sequences by specialized programs.
- ▷ Pratt, Splash and Teiresas are three of these specialized programs.
- ▷ Today *machine learning* is a very active research field
- ▷ Such automatic patterns are usually distinct from those designed by an expert with some knownledge of the biochemical litterature.

Position Specific Scoring Matrice (PSSM)

How to build a PSSM

- A PSSM is based on the frequencies of each residue in a specific position of a multiple alignment.

```
GHEGVGKVVKLGAGA
GHEKKGYFEDRGPSA
GHEGYGGRSRGGGYS
GHEFEGPKGCGALYI
GHELRGTTFMPALEC
```

→

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	0	0	0	0	0	0	0	0	0	0	0	2	1	0	2
C	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
D	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	5	0	1	0	0	0	1	0	0	0	0	1	0
F	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
G	5	0	0	2	0	5	1	0	1	0	2	3	1	1	0
H	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
K	0	0	0	1	1	0	1	1	0	1	0	0	0	0	0
L	0	0	0	1	0	0	0	0	0	0	1	0	2	0	0
M	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
Q	0	0	0	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	0	0	0
S	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
T	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
V	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0
W	0	0	0	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	0	1	0

-
  - Column 1: $f_{A,1} = \frac{0}{5} = 0,\ f_{G,1} = \frac{5}{5} = 1,\ \ldots$
  - Column 2: $f_{A,2} = \frac{0}{5} = 0,\ f_{H,2} = \frac{5}{5} = 1,\ \ldots$
  - ...
  - Column 15: $f_{A,15} = \frac{2}{5} = 0.4,\ f_{C,15} = \frac{1}{5} = 0.2,\ \ldots$

Pseudo-counts

- Some observed frequencies usually equal 0. This is a consequence of the limited number of sequences that is present in a MSA.
- Unfortunately, an observed frequency of 0 might imply the exclusion of the corresponding residue at this position position (this was the case with patterns).
- One possible trick is to add a small number to all observed frequencies. These small non-observed frequencies are refered to as a *pseudo-counts*.
- From the previous example with a pseudo-counts of 1:
  - Column 1: $f'_{A,1} = \frac{0+1}{5+20} = 0.04$, $f'_{G,1} = \frac{5+1}{5+20} = 0.24$, ...
  - Column 2: $f'_{A,2} = \frac{0+1}{5+20} = 0.04$, $f'_{H,2} = \frac{5+1}{5+20} = 0.24$, ...
  - ...
  - Column 15: $f'_{A,15} = \frac{2+1}{5+20} = 0.12$, $f'_{C,15} = \frac{1+1}{5+20} = 0.08$, ...
- There exist more sophisticated methods to produce more "realistic" pseudo-counts, and which are based on substitution matrix or Dirichlet mixtures.

Computing a PSSM

▷ The frequency of every residue determined at every position has to be compared with the frequency at which any residue can be expected in a *random sequence*.

▷ For example, let's postulate that each amino acid is observed with an identical frequency in a random sequence. This is a quite simplisitic null model.

▷ The score is derived from the ratio of the observed to the expected frequencies. More precisely, the logarithm of this ratio is taken and refered to as the log-likelihood ratio:

$$Score_{ij} = log(\frac{f'_{ij}}{q_i})$$

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

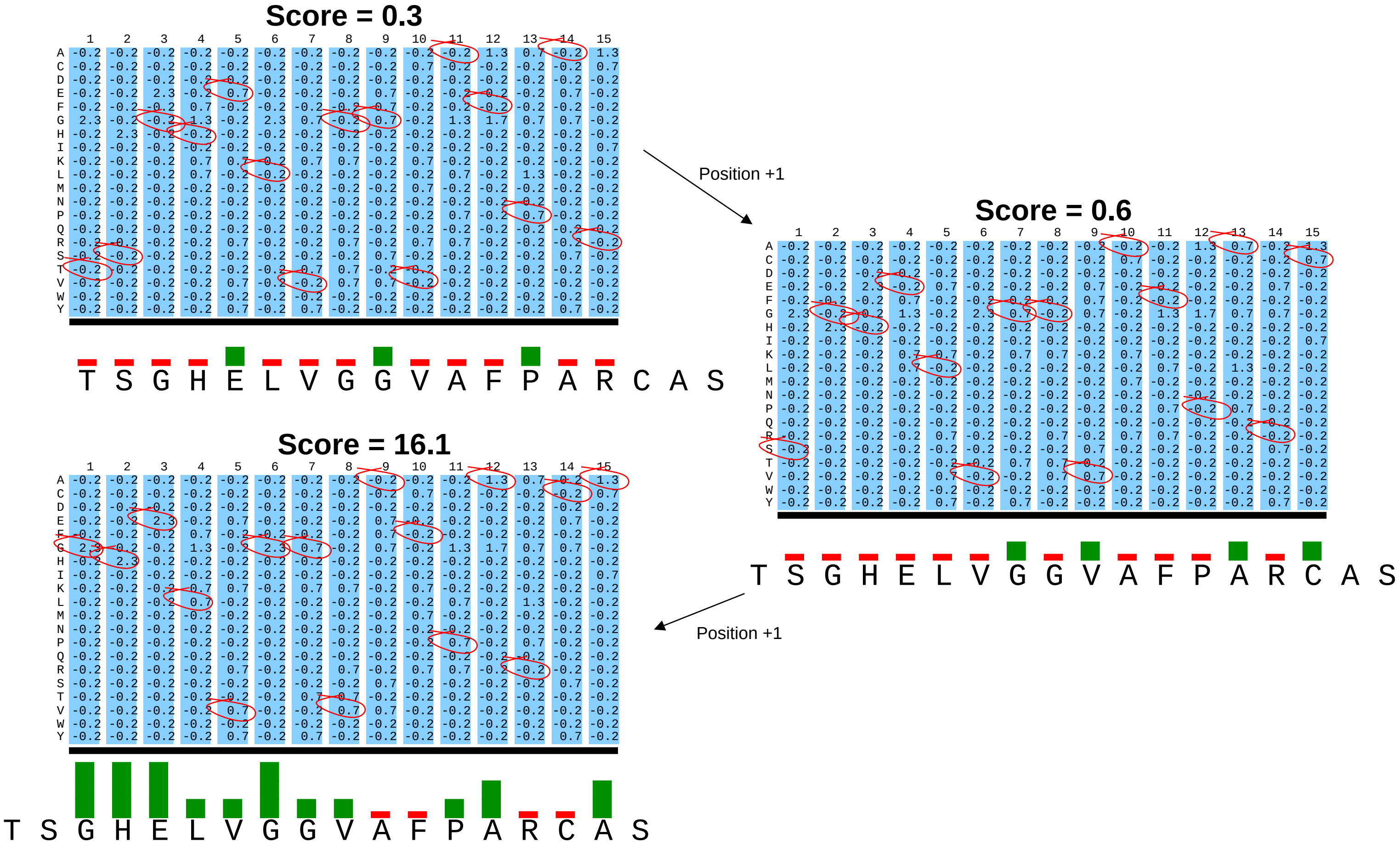
Example

- The complete position specific scoring matrix calculated from the previous example:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	1.3	0.7	-0.2	1.3
C	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	0.7
D	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
E	-0.2	-0.2	2.3	-0.2	0.7	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	0.7	-0.2
F	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
G	2.3	-0.2	-0.2	1.3	-0.2	2.3	0.7	-0.2	0.7	-0.2	1.3	1.7	0.7	0.7	-0.2
H	-0.2	2.3	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
I	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7
K	-0.2	-0.2	-0.2	0.7	0.7	-0.2	0.7	0.7	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2
L	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	1.3	-0.2	-0.2
M	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2
N	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
P	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	0.7	-0.2	-0.2
Q	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
R	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	0.7	-0.2	0.7	0.7	-0.2	-0.2	-0.2	-0.2
S	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	0.7	-0.2
T	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
V	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	0.7	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
W	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
Y	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2

How to use PSSMs

- The PSSM is applied as a sliding window along the subject sequence:
  - At every position, a PSSM score is calculated by summing the scores of all columns;
  - The highest scoring position is reported.



Score = 0.3

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	1.3	0.7	-0.2	1.3
C	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	0.7
D	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
E	-0.2	-0.2	2.3	-0.2	0.7	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	0.7	-0.2
F	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
G	2.3	-0.2	-0.2	1.3	-0.2	2.3	0.7	-0.2	0.7	-0.2	1.3	1.7	0.7	0.7	-0.2
H	-0.2	2.3	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
I	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7
K	-0.2	-0.2	-0.2	0.7	0.7	-0.2	0.7	0.7	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2
L	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	1.3	-0.2	-0.2
M	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2
N	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
P	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	0.7	-0.2	-0.2
Q	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
R	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	0.7	-0.2	0.7	0.7	-0.2	-0.2	-0.2	-0.2
S	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	0.7	-0.2
T	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
V	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	0.7	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
W	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
Y	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2

T S G H E L V G G V A F P A R C A S

Position +1

Score = 0.6

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	1.3	0.7	-0.2	1.3
C	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	0.7
D	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
E	-0.2	-0.2	2.3	-0.2	0.7	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	0.7	-0.2
F	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
G	2.3	-0.2	-0.2	1.3	-0.2	2.3	0.7	-0.2	0.7	-0.2	1.3	1.7	0.7	0.7	-0.2
H	-0.2	2.3	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
I	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7
K	-0.2	-0.2	-0.2	0.7	0.7	-0.2	0.7	0.7	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2
L	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	1.3	-0.2	-0.2
M	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2
N	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
P	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	0.7	-0.2	-0.2
Q	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
R	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	0.7	-0.2	0.7	0.7	-0.2	-0.2	-0.2	-0.2
S	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	0.7	-0.2
T	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
V	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	0.7	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
W	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
Y	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2

T S G H E L V G G V A F P A R C A S

Position +1

Score = 16.1

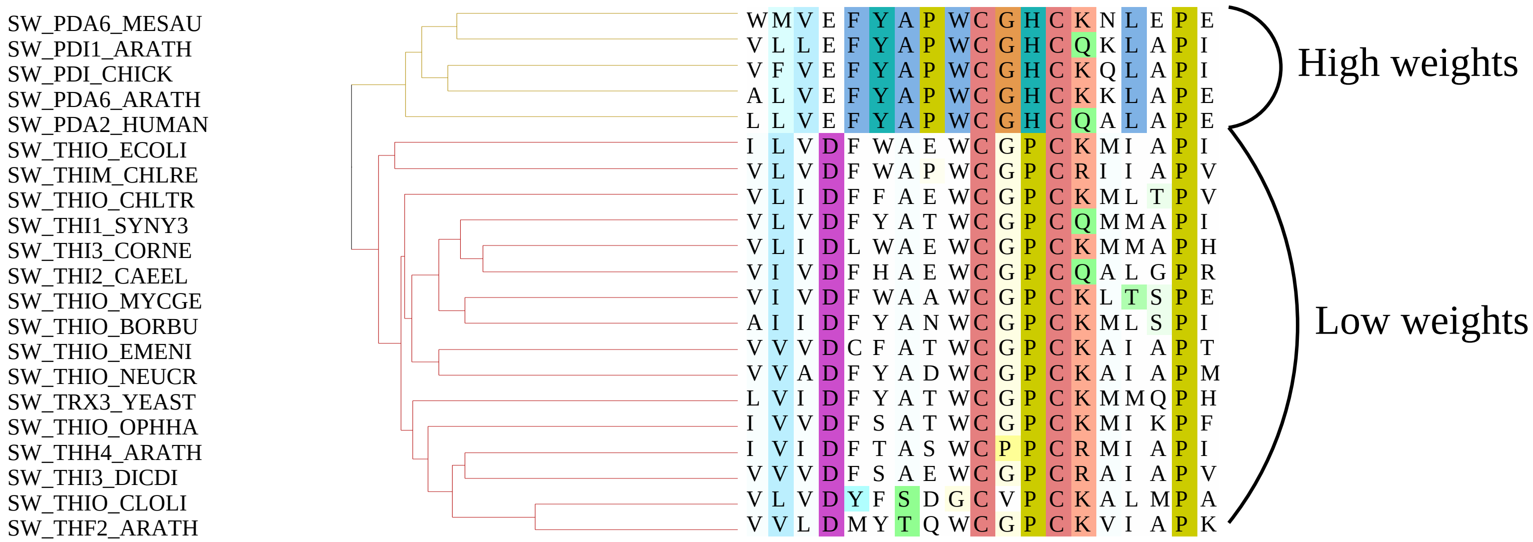
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	1.3	0.7	-0.2	1.3
C	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	0.7
D	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
E	-0.2	-0.2	2.3	-0.2	0.7	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	0.7	-0.2
F	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
G	2.3	-0.2	-0.2	1.3	-0.2	2.3	0.7	-0.2	0.7	-0.2	1.3	1.7	0.7	0.7	-0.2
H	-0.2	2.3	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
I	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7
K	-0.2	-0.2	-0.2	0.7	0.7	-0.2	0.7	0.7	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2
L	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	1.3	-0.2	-0.2
M	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2
N	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
P	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	0.7	-0.2	-0.2
Q	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
R	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	0.7	-0.2	0.7	0.7	-0.2	-0.2	-0.2	-0.2
S	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	0.7	-0.2
T	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
V	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	0.7	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
W	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
Y	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2

T S G H E L V G G V A F P A R C A S

Sequence weighting

- An MSA is often made of a few distinct sets of related sequences, or sub-families. It is not unusal that these sub-families are very differently populated, thus influencing observed residue frequencies.

- Sequences weighting algorithms attempt to compensate this sequence sampling bias.

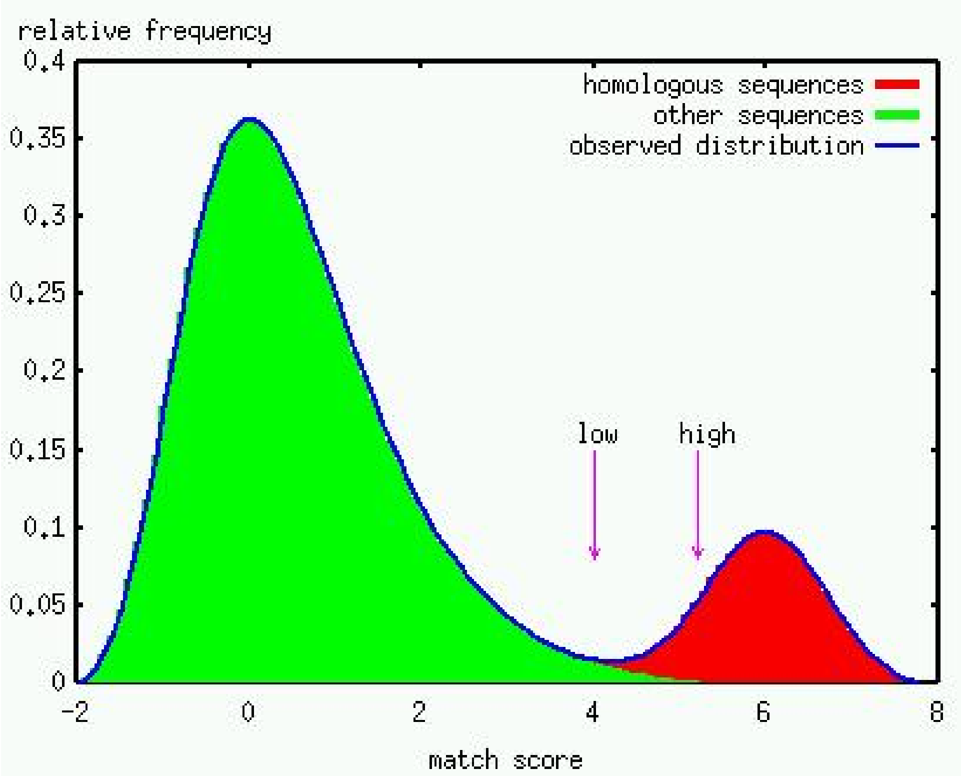


```
SW_PDA6_MESAU   W M V E F Y A P W C G H C K N L E P E
SW_PDI1_ARATH   V L L E F Y A P W C G H C Q K L A P I
SW_PDI_CHICK    V F V E F Y A P W C G H C K Q L A P I
SW_PDA6_ARATH   A L V E F Y A P W C G H C K K L A P E
SW_PDA2_HUMAN   L L V E F Y A P W C G H C Q A L A P E
SW_THIO_ECOLI   I L V D F W A E W C G P C K M I A P I
SW_THIM_CHLRE   V L V D F W A P W C G P C R I I A P V
SW_THIO_CHLTR   V L I D F F A E W C G P C K M L T P V
SW_THI1_SYNY3   V L V D F Y A T W C G P C Q M M A P I
SW_THI3_CORNE   V L I D L W A E W C G P C K M M A P H
SW_THI2_CAEEL   V I V D F H A E W C G P C Q A L G P R
SW_THIO_MYCGE   V I V D F W A A W C G P C K L T S P E
SW_THIO_BORBU   A I I D F Y A N W C G P C K M L S P I
SW_THIO_EMENI   V V V D C F A T W C G P C K A I A P T
SW_THIO_NEUCR   V V A D F Y A D W C G P C K A I A P M
SW_TRX3_YEAST   L V I D F Y A T W C G P C K M M Q P H
SW_THIO_OPHHA   I V V D F S A T W C G P C K M I K P F
SW_THH4_ARATH   I V I D F T A S W C P P C R M I A P I
SW_THI3_DICDI   V V V D F S A E W C G P C R A I A P V
SW_THIO_CLOLI   V L V D Y F S D G C V P C K A L M P A
SW_THF2_ARATH   V V L D M Y T Q W C G P C K V I A P K
```

PSSM Score Interpretation

▷ The E-value is the number of matches with a score equal to or greater than the observed score that are expected to occur *by chance*.

▷ The E-value depends on the size of the searched database, as the number of false positives expected above a given score threshold increases proportionately with the size of the database.



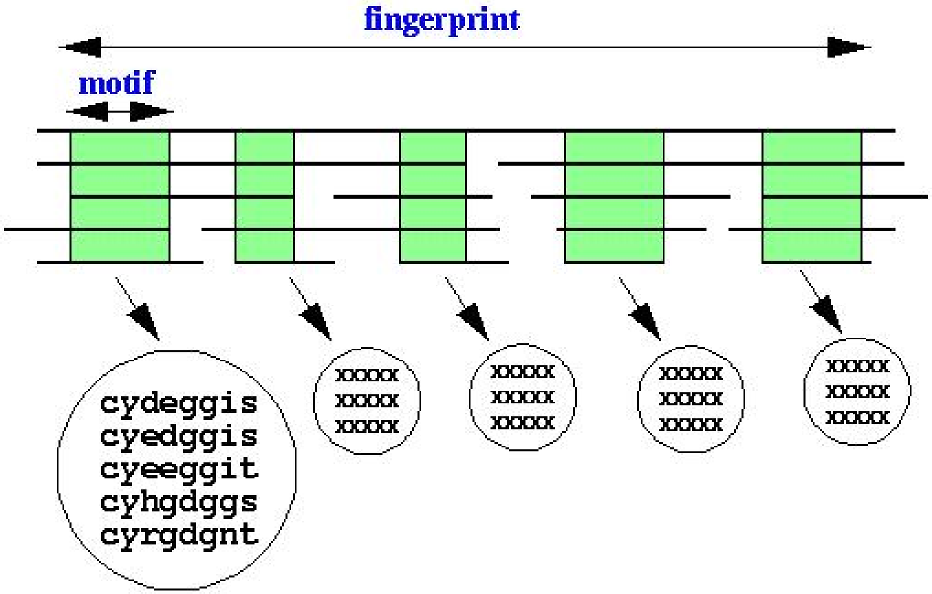
PSSM: Conclusion

- Advantages:
    - Good for short, conserved regions.
    - Relatively fast and simple to implement.
    - Produce match scores that can be interpreted based on statistical theory.
- Limitations:
    - Insertions and deletions are strictly forbidden.
    - Relatively long sequence regions can therefore not be described with this method.
- When I use it?
    - To model small regions with high variability but constant length.

PSSM: beyond the conclusion

- ▷ PSSMs can be automatically extracted (discovered) from a set of unaligned sequences by specialized programs. The program MEME is such a tool which is based on the *expectation-maximization algorithm*

- ▷ A couple of PSSMs can be used to describe the conserved regions of a large MSA. A datababase of such diagnostic PSSMs and search tools dedicated for that purpose are available.



Generalized profiles

The idea behind generalized profile

- One would like to generalize PSSMs to allow for insertions and deletions. However this raises the difficult problems of defining and computing an optimal alignment with gaps.

- Let us recycle the principle of dynamic programing, as it was introduced to define and compute the optimal alignments between a pair of sequences e.g. by the Smith-Waterman algorithm, and generalize it by the introduction of:

    - position-dependent match scores,
    - position-dependent gap penalties.

Generalized profiles are an extension of PSSMs

- The following information is stored in any generalized profile:
  - each position is called a match state. A score for every residue is defined at every match states, just as in the PSSM.
  - each match state can be omitted in the alignment, by what is called a deletion state and that receives a position-dependent penalty.
  - insertions of variable length are possible between any two adjacent match (or deletion) states. These insertion states are given a position-dependent penalty that might also depend upon the inserted residues.
  - every possible transition between any two states (match, delete or insert) receives a position-dependent penalty. This is primarily to model the cost of opening and closing a gap.
  - a couple of additional parameters permit to finely tune the behaviour of the extremities of the alignment, which can forced to be 'local' or 'global' at either ends of the profile and of the sequence.

Excerpt of an example of the generalized profile syntax

```
ID   THIOREDOXIN_2; MATRIX.
AC   PS50223;
DT          ? (CREATED); MAY-1999 (DATA UPDATE);        ? (INFO UPDATE).
DE   Thioredoxin-domain (does not find all).
MA   /GENERAL_SPEC: ALPHABET=’ABCDEFGHIKLMNPQRSTVWYZ’; LENGTH=103;
MA   /DISJOINT: DEFINITION=PROTECT; N1=6; N2=98;
MA   /NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=1.9370; R2=0.01816483; TEXT=’-LogE’;
MA   /CUT_OFF: LEVEL=0; SCORE=361; N_SCORE=8.5; MODE=1; TEXT=’!’;
MA   /DEFAULT: D=-20; I=-20; B1=-100; E1=-100; MM=1; MI=-105; MD=-105; IM=-105; DM=-105; M0=-6;
MA   /I: B1=0; BI=-105; BD=-105;
```

... many lines deleted ...

```
MA   /M: SY=’K’; M=-8,0,-25,1,8,-24,-14,-9,-22,19,-20,-11,0,-9,5,13,-3,-4,-16,-24,-13,6; D=-3;
MA   /I: I=-3; DM=-16;
MA   /M: SY=’P’; M=-6,-13,-26,-12,-9,-12,-19,-14,-5,-11,-5,-4,-12,8,-11,-13,-9,-6,-6,-25,-11,-12;
MA   /M: SY=’V’; M=-4,-22,-19,-24,-20,-2,-25,-21,11,-15,2,3,-20,-23,-17,-14,-9,-1,19,-11,-4,-19;
MA   /M: SY=’A’; M=28,-7,-15,-13,-6,-20,-2,-15,-15,-6,-14,-11,-5,-12,-6,-11,9,1,-6,-21,-17,-6;
MA   /M: SY=’P’; M=-6,-3,-27,2,2,-22,-14,-11,-20,-6,-24,-17,-5,25,-4,-11,3,1,-19,-29,-17,-3;
MA   /M: SY=’W’; M=-16,-27,-41,-28,-21,2,-13,-20,-20,-16,-19,-17,-26,-25,-15,-15,-26,-20,-26,93,19,-15;
MA   /M: SY=’C’; M=-9,-17,106,-26,-27,-20,-27,-28,-29,-28,-20,-20,-17,-37,-28,-28,-8,-9,-10,-48,-29,-27;
MA   /M: SY=’G’; M=-4,-12,-31,-9,-9,-27,24,-18,-27,-13,-25,-17,-7,14,-13,-17,-3,-13,-24,-24,-26,-13;
MA   /M: SY=’H’; M=-12,-10,-30,-8,-4,-14,-18,18,-17,-10,-18,-8,-7,16,-5,-11,-8,-10,-20,-22,-1,-8;
MA   /M: SY=’C’; M=-9,-19,111,-28,-28,-20,-29,-29,-28,-29,-20,-19,-18,-38,-28,-29,-8,-8,-9,-49,-29,-28;
MA   /M: SY=’R’; M=-12,-4,-27,-4,3,-22,-20,-2,-21,22,-19,-6,-2,-13,9,23,-9,-8,-16,-20,-6,4;
```

... many lines deleted ...

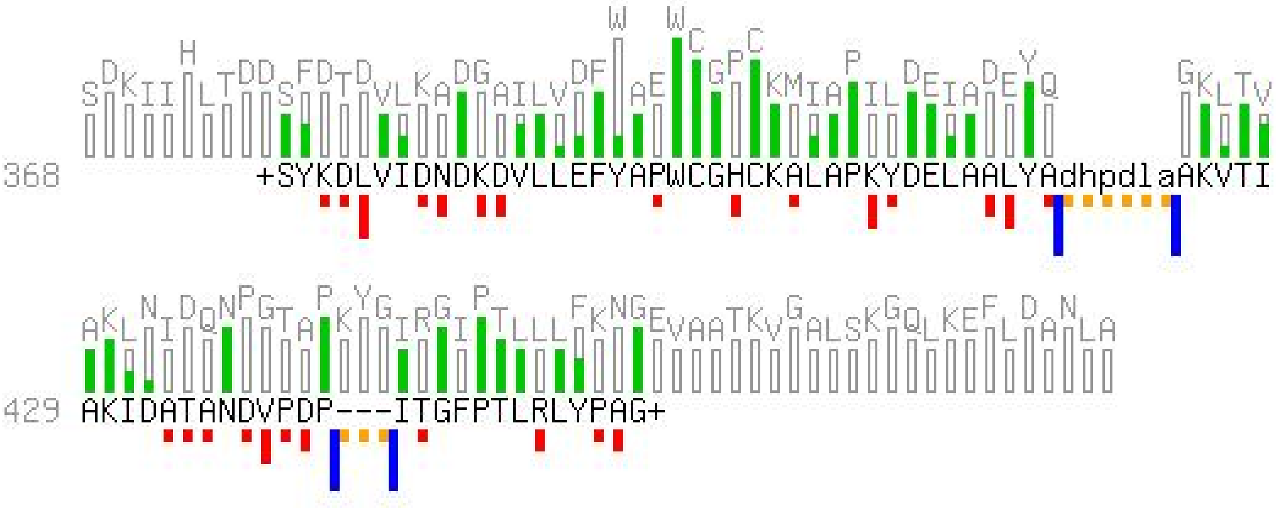
```
//
```

Details of the scores along an alignment I

- Smith-Waterman alignment of two thioredoxin domains:

```
THIO_ECOLI SFDTDVLKADGAILVDFWAEWCGPCKMIAPILDEIADEYQ------GKLTVAKLNIDQNP
           :.   :.  :  .:..:.: ::: :: .::  ::.:  :       .:.:.::..   :
PDI_ASPNG  SYKDLVIDNDKDVLLEFYAPWCGHCKALAPKYDELAALYADHPDLAAKVTIAKIDATAND

THIO_ECOLI GTAPKYGIRGIPTLLLFKNG
              :   : :.::: :.  :
PDI_ASPNG  VPDP---ITGFPTLRLYPAG
```

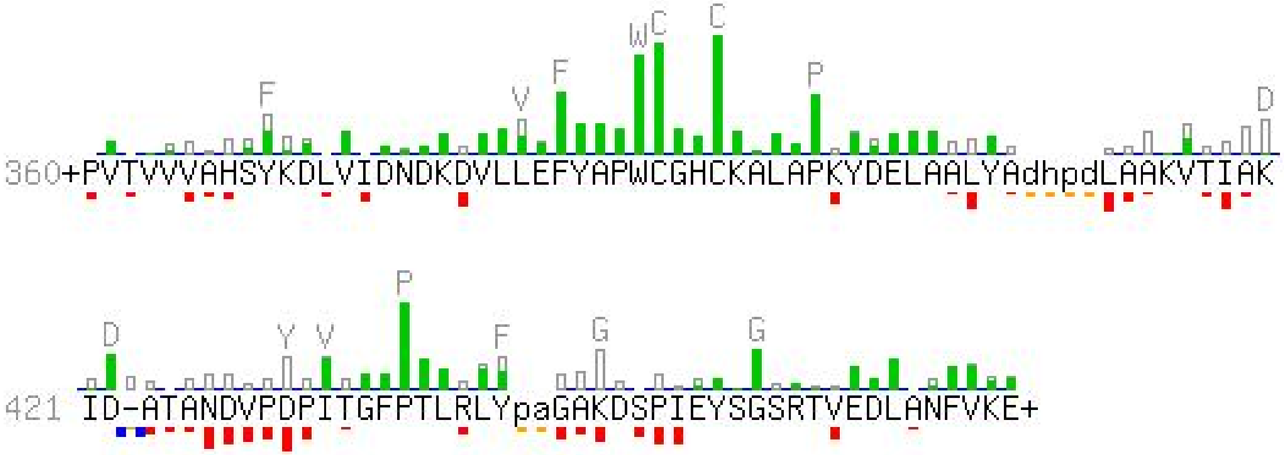


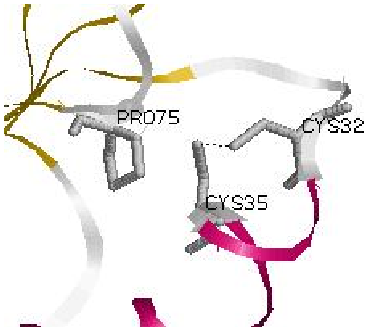
Details of the scores along an alignment II

- Alignment of a sequence of a thioredoxin domain on a profile built from a MSA of thioredoxins:

```
consensus   1 XVXVLSDENFDEXVXDSDKPVLVDFYAPWCGHCRALAPVFEELAEEYK----DBVKFVKV  -48
               : :         : : :  ::  : : ::::: : :    : :  :        :
PDI_ASPNG 360 PVTVVVAHSYKDLVIDNDKDVLLEFYAPWCGHCKALAPKYDELAALYAdhpdLAAKVTIA  -97

consensus  57 DVDENXELAEEYGVRGFPTIMFF--KBGEXVERYSGARBKEDLXEFIEK             -1
                :            : ::              : : :  : :  :
PDI_ASPNG 420 KID-ATANDVPDPITGFPTLRLYpaGAKDSPIEYSGSRTVEDLANFVKE             -49
```





Generalized profiles: Software

- ▷ Pftools is a package to build and use generalized profiles, which was developed by Philipp Bucher (http://www.isrec.isb-sib.ch/ftp-server/pftools/).
- ▷ The package contains (among other programs):
  - ▷ pfmake for building a profile starting from multiple alignments.
  - ▷ pfcalibrate to calibrate the profile model.
  - ▷ pfsearch to search a protein database with a profile.
  - ▷ pfscan to search a profile databse with a protein.

Generalized profiles: Conclusions

- Advantage:
  - Possible to specify where deletions and insertions occur.
  - Very sensitive to detect homology below the twilight zone.
  - Good scoring system.
  - Automatic building of the profiles.
  - Require more sophisticated software.
- Limitations:
  - Very CPU expensive.
  - Require some expertise to use proficiently.

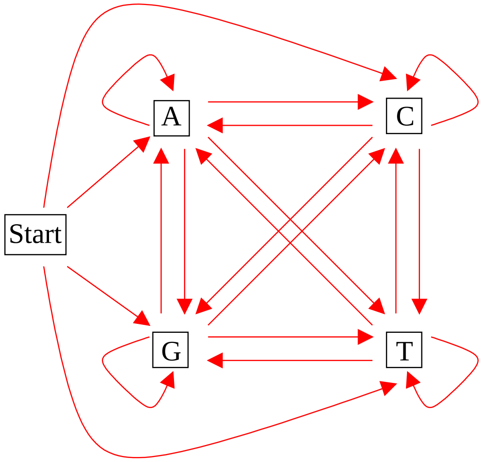
Hidden Markov Models: probabilistic models

Hidden Markov Models derive from Markov Chains

- Hidden Markov Models are an extension of the Markov Chains theory, which is part of the theory of probabilities.

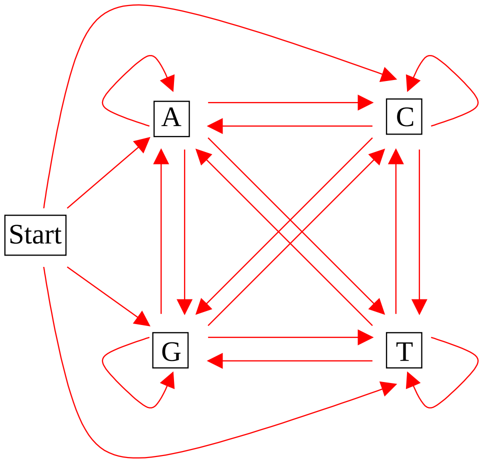
- A Markov Chain is a succession of **states** S_i $(i = 0, 1, \ldots)$ connected by **transitions**. Transitions from state S_i to state S_j has a probability of P_{ij}.

- An example of Markov Chain:
  - Transition probabilities:
    - $P(A|G) = 0.18$, $P(C|G) = 0.38$, $P(G|G) = 0.32$, $P(T|G) = 0.12$
    - $P(A|C) = 0.15$, $P(C|C) = 0.35$, $P(G|C) = 0.34$, $P(T|C) = 0.15$



How to calculate the probability of a Markov Chain

▷ Given a Markov Chain M where all transition probabilities are known:

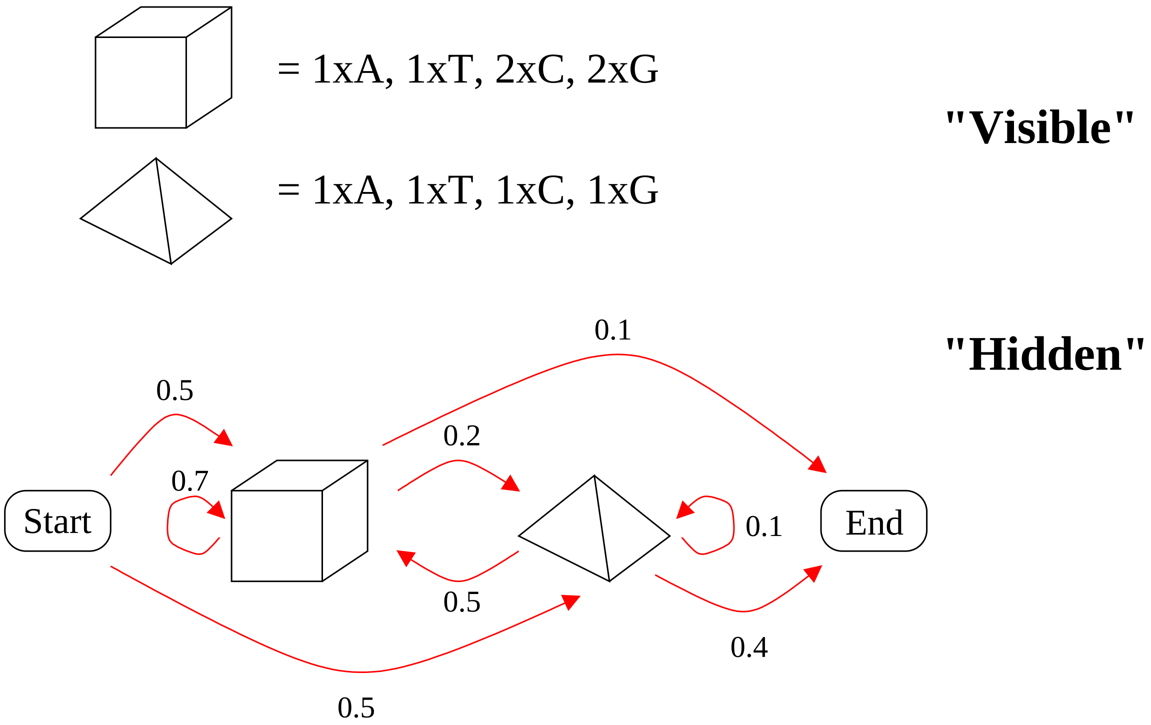


▷ The probability of sequence $x = GCCT$ is:

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

Hidden Markov Models are an extension of Markov Chains

- Hidden Markov Models (HMMs) are like Markov Chains: a finite number of **states** connected between them by **transitions**.
- But the major difference between the two is that the states of the Hidden Markov Models are not a symbol but a **distribution** of symbols. Each state can **emit** a symbol with a probability given by the distribution.



= 1xA, 1xT, 2xC, 2xG

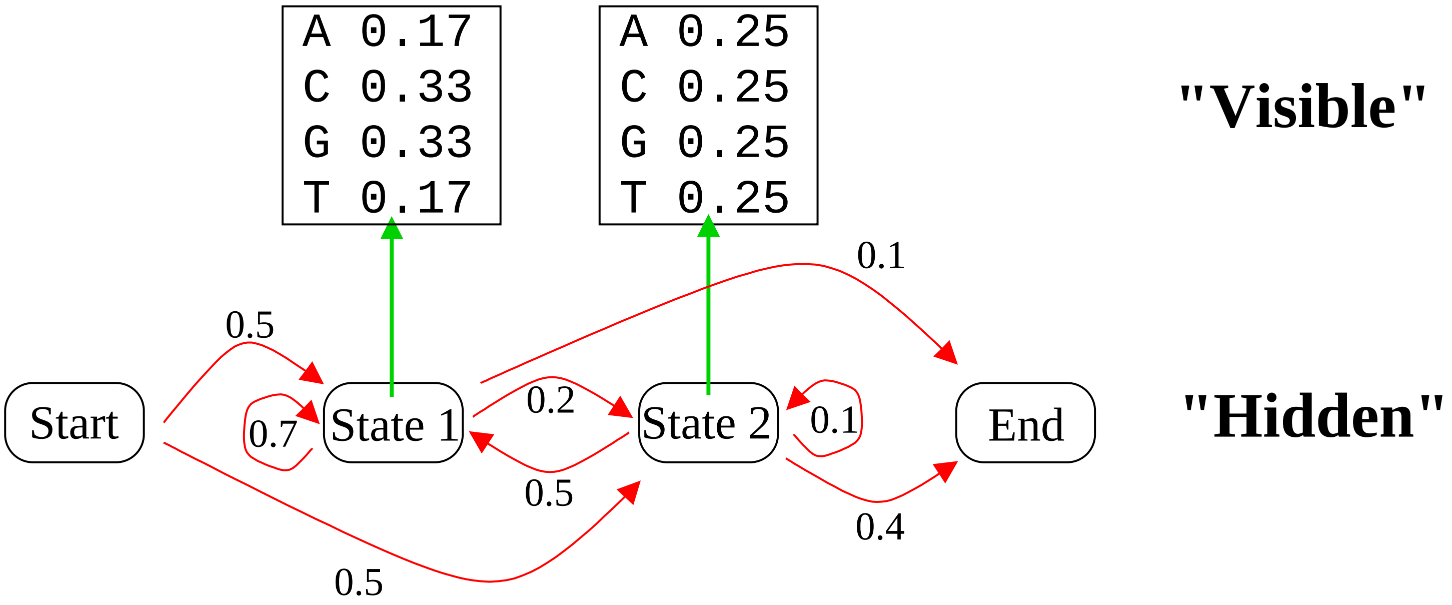
= 1xA, 1xT, 1xC, 1xG

"Visible"

"Hidden"

Example of a simple HMM

▷ Example of a simple Hidden Markov Model, generating GC rich DNA sequences:



```
A 0.17      A 0.25
C 0.33      C 0.25
G 0.33      G 0.25
T 0.17      T 0.25
```

```
START   1   1   1   1   2   2   1   1   1   2   END

         G   C   A   G   C   T   G   G   C   T
```

Hidden Markov Model parameters

- The parameters describing HMMs:
  - **Emission probabilities**. This is the probability of emitting a symbol x from an alphabet α being in state q.

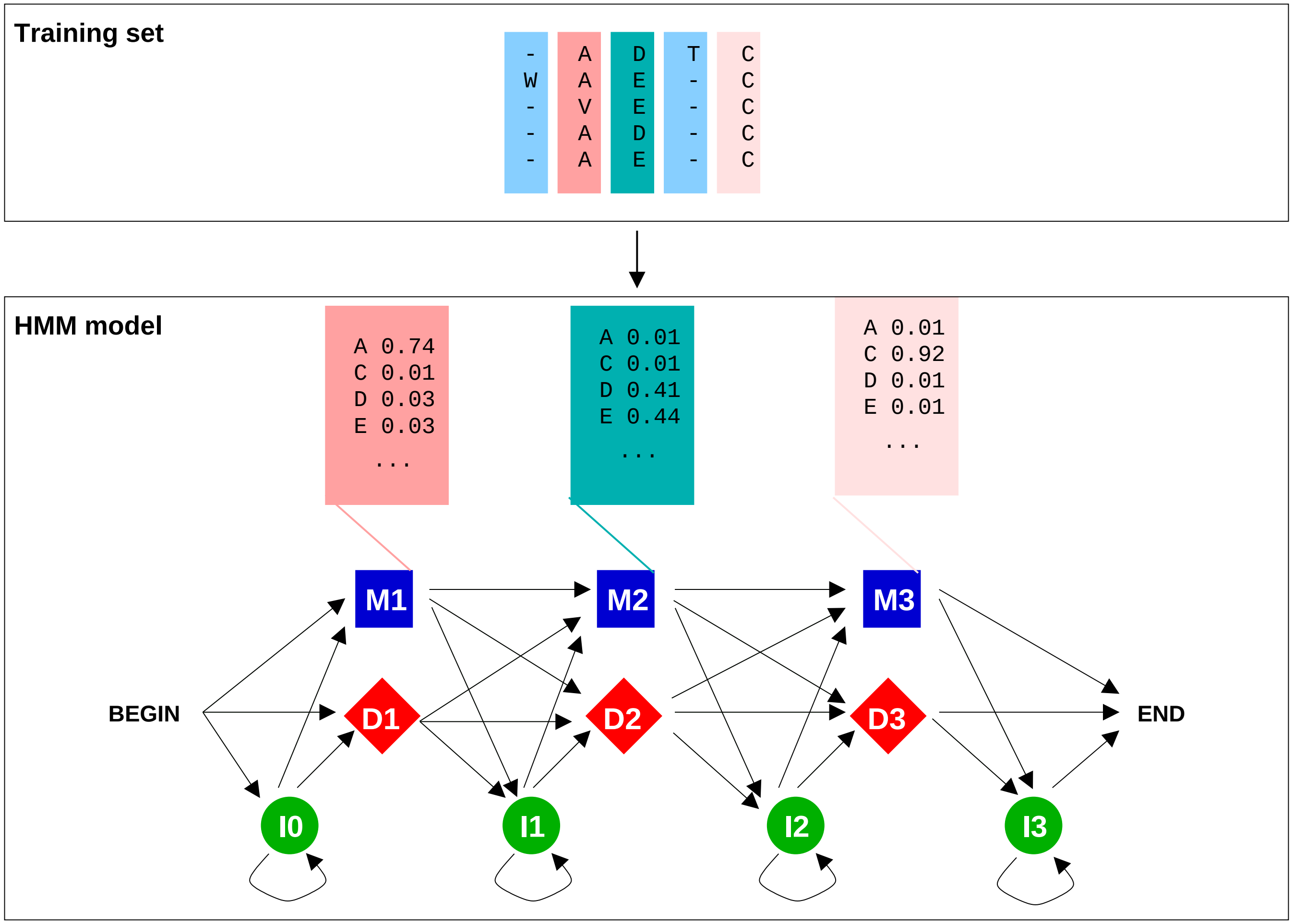
    $$E(x|q)$$

    - Residue emission probabilities are evaluated from the observed frequencies as for PSSMs.
    - Pseudo-counts are added to avoid emission probabilities equal to 0.
  - **Transition probabilities**. This is the probability of a transition to state r being in state q.

    $$T(r|q)$$

    - Transition probabilities are evaluated from observed transition frequencies.
- Emission and transition probabilities can also be evaluated using the *Baum-Welch training algorithm*.

HMMs are trained from a multiple alignment



Training set

```
- A D T C
W A E - C
- V E - C
- A D - C
- A E - C
```

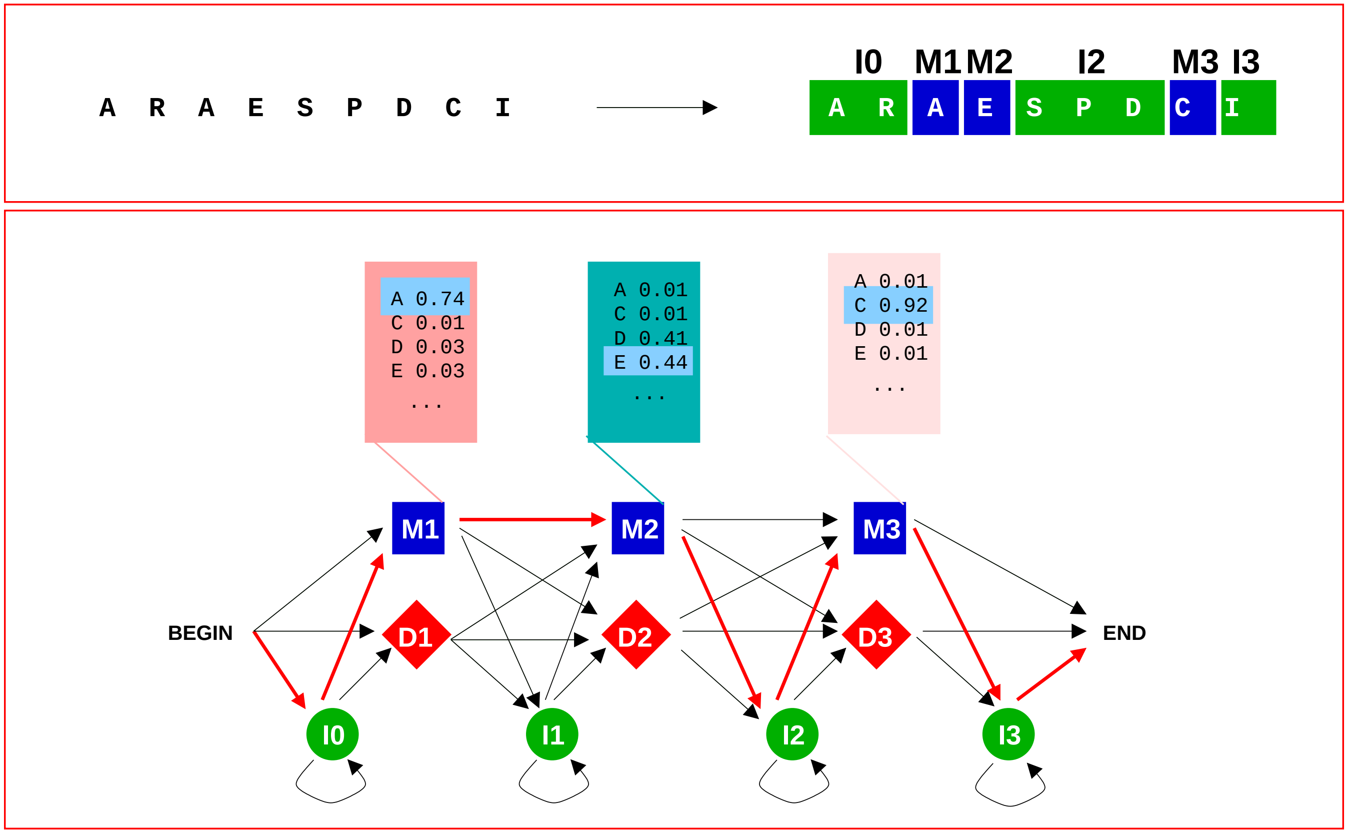
HMM model

```
A 0.74
C 0.01
D 0.03
E 0.03
...
```

```
A 0.01
C 0.01
D 0.41
E 0.44
...
```

```
A 0.01
C 0.92
D 0.01
E 0.01
...
```

Match a sequence to a model: find the best path



Algorithms associated with HMMs

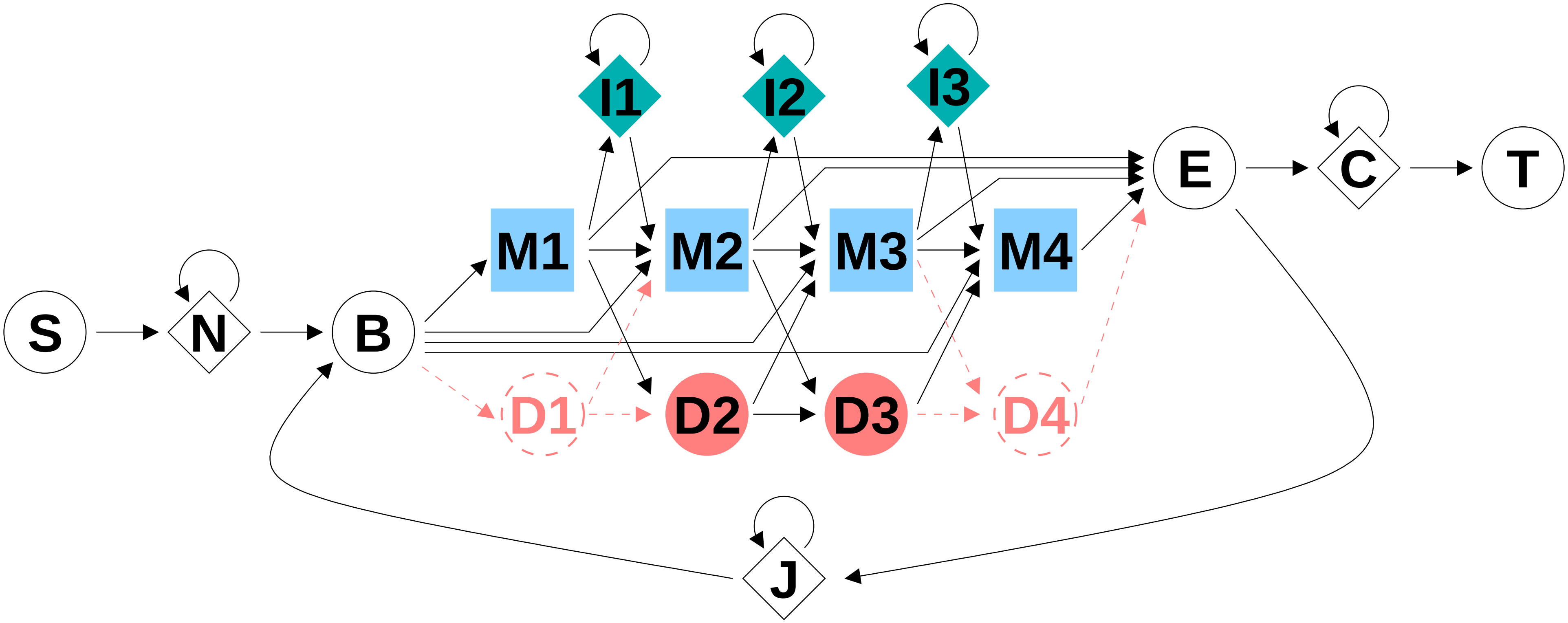
- ▷ Three important questions can be answered by three algorithms.
    - ▷ How likely is a given sequence under a given model?
        - ▷ This is the scoring problem and it can be solved using the **Forward algorithm**.
    - ▷ What is the most probable path between states of a model given a sequence?
        - ▷ This is the alignment problem and it is solved by the **Viterbi algorithm**.
    - ▷ How can we learn the HMM parameters given a set of sequences?
        - ▷ This is the training problem and is solved using the **Forward-backward algorithm** and the **Baum-Welch expectation maximization**.
- ▷ For details about these algorithms see:

  Durbin, Eddy, Mitchison, Krog.
  Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.
  Cambridge University Press, 1998.

Hidden Markov Models: Softwares

- HMMER2 is a package to build and use HMMs developed by Sean Eddy (http://hmmer.wustl.edu/).
- Software available in HMMER2:
  - hmmbuild to build an HMM model from a multiple alignment;
  - hmmalign to align sequences to an HMM model;
  - hmmcalibrate to calibrate an HMM model;
  - hmmemit to create sequences from an HMM model;
  - hmmsearch to search a sequence database with an HMM model;
  - hmmpfam to scan a sequence with a database of HMM models;
  - ...
- SAM is a similar package developed by Richard Hughey, Kevin Karplus and Anders Krogh (http://www.cse.ucsc.edu/research/compbio/sam.html).

The "Plan 7" architecture of HMMER2



Hidden Markov Models: Conclusions

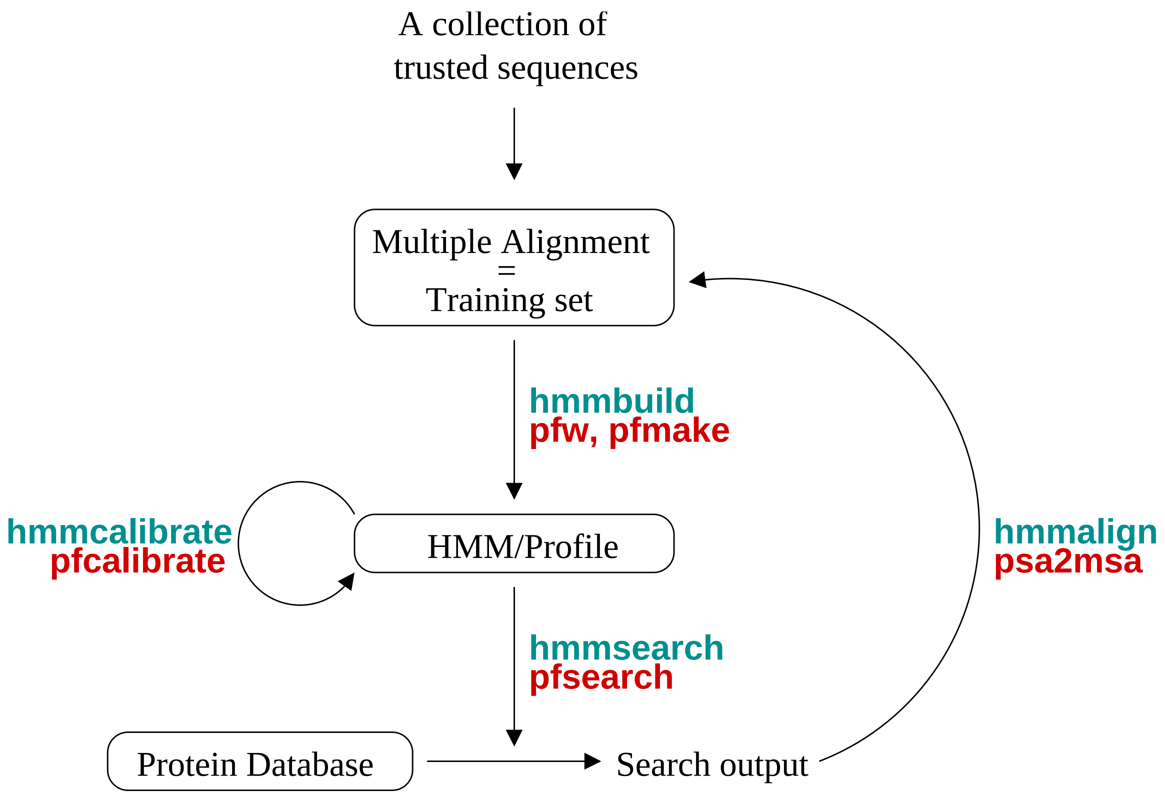
- Solid thoretical basis in the theory of probabilities.
- Other Advantages and limitations just like generalized profiles.

Generalized profiles and HMMs I

- Generalized profiles are *equivalent* to the 'linear' HMMs like those of SAM or HMMER2 (they are not equivalent to other HMMs of more complicated architecture).
- The optimal alignment produced by dynamical programming is *equivalent* to the Viterbi path on a HMM.
- There are programs to translate profiles from and into HMMs:
    - htop: HMM to profile.
    - ptoh: profile to HMM.

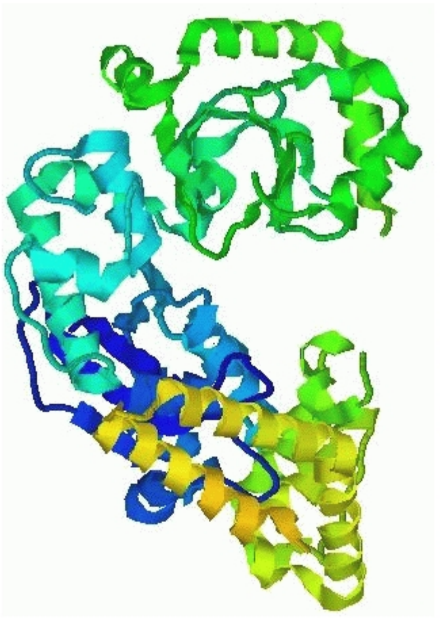
Generalized profiles and HMMs II

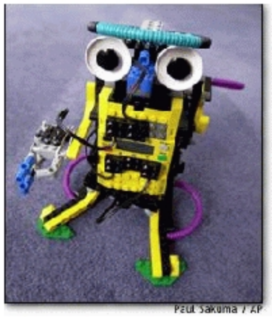
▷ Iterative model training with the PFTOOLS or HMMER2:



Generalized profiles and HMMs III

- HMMs and generalized profiles are very appropriate for the modelling of protein domains.
- What are protein domains:
    - Domains are discrete structural units (25-500 aa).
    - Short domains (25-50 aa) are present in multiple copies for structural stability.
    - Domains are functional units.



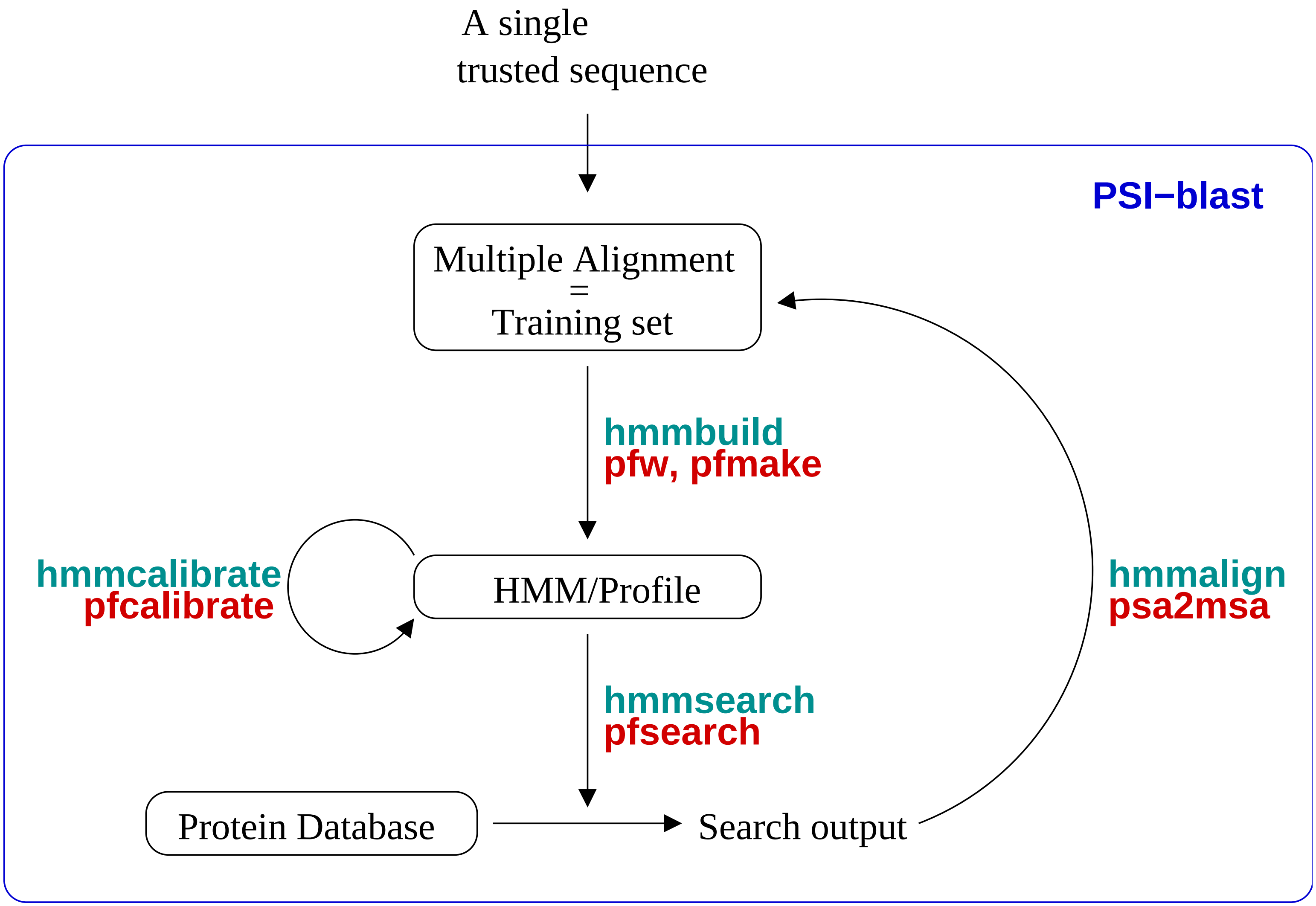


Paul Sakuma / AP

PSI-blast I

- PSSM could have simply been improved by the introduction of a position-independent affine gap cost model;
- This is less sophistication than the generalized profiles;
- But it is just this principle that is behind PSI-blast.
- The success and efficiency of PSI-blast has also much to do with:
  - the speed of the blast heuristic;
  - a particularily efficient algorithm for sequence weighting;
  - a very sophisticated statistical treatment of the match scores.

PSI-blast II


A single
trusted sequence
PSI−blast
Multiple Alignment
=
Training set
hmmbuild
pfw, pfmake
hmmcalibrate
pfcalibrate
HMM/Profile
hmmalign
psa2msa
hmmsearch
pfsearch
Protein Database
Search output

Databases

Patterns and PSSM databses

- Patterns database
  - Prosite
    - WEB access: http://www.expasy.ch/prosite/.
    - Contains also profiles.
    - Well documented.
    - Easy to test new patterns.
- PSSM databases:
  - BLOCKS
    PRINTS.
    - WEB access: http://www.blocks.fhcrc.org/
      http://bioinf.man.ac.uk/dbbrowser/PRINTS/.
    - Automatically produces PSSMs from families of sequences.
    - Easy to scan databases with the produced PSSMs.

Protein domain databases

- A non-exhaustive list of protein domain databases:
  - Pfam
    - http://www.sanger.ac.uk/Pfam.
    - Collection of protein domains and families (3071 entries in Pfam release 6.6).
    - Uses HMMs (HMMER2).
    - Good links to structure, taxonomy.
  - PROSITE
    - http://www.expasy.ch/prosite.
    - Collection of motifs, protein domains, and families (1494 entries in Prosite release 16.51).
    - Uses generalized profiles (Pftools) and patterns.
    - High quality documentation.

Protein domain databases

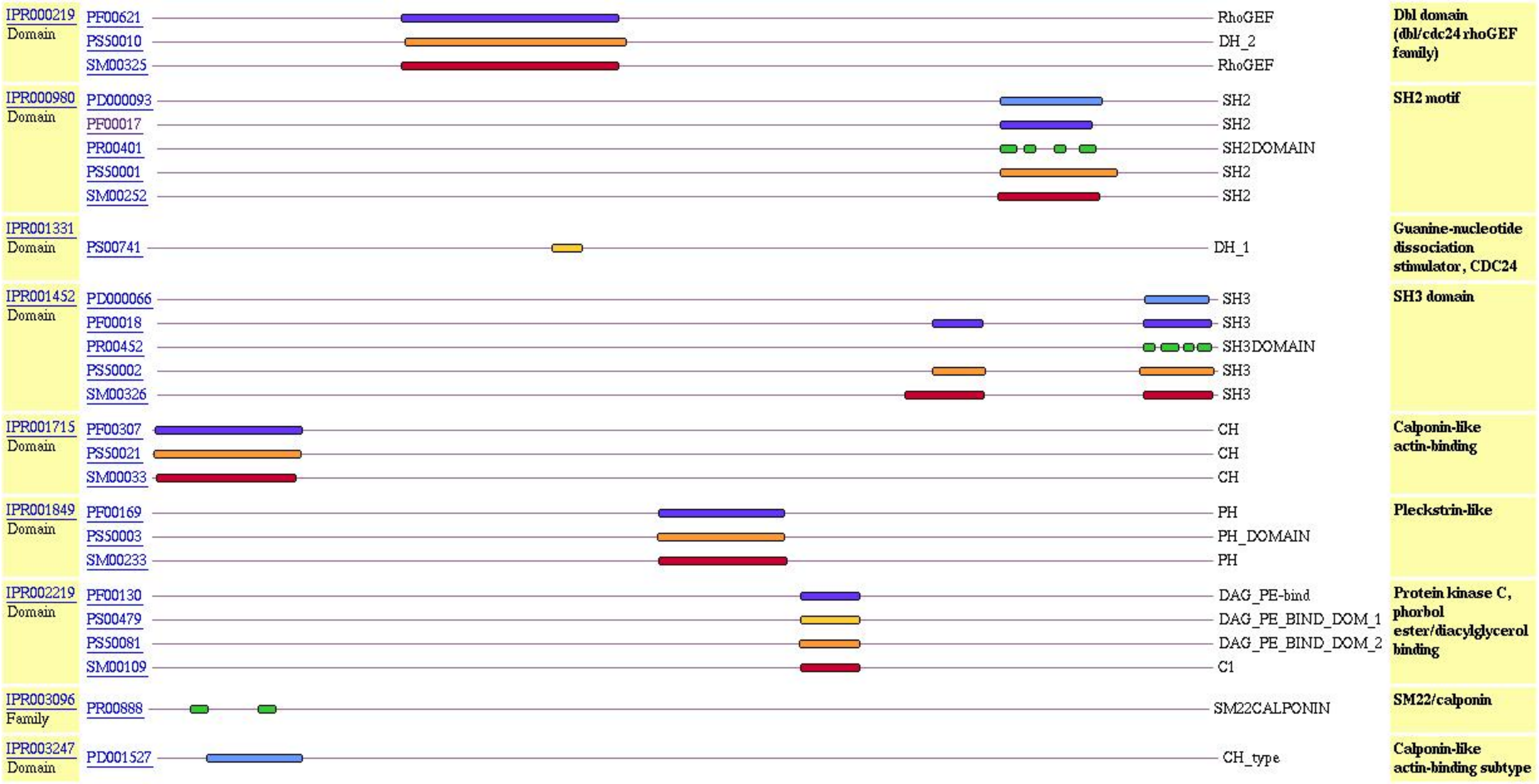
- A non-exhaustive list of protein domain databases (continued):
  - Prints
    - http://bioinf.man.ac.uk/dbbrowser/PRINTS.
    - Collection of conserved motifs used to characterize a protein.
    - Uses fingerprints (conserved motif groups).
    - Very good to describe sub-families.
    - Release 32.0 of PRINTS contains 1600 entries, encoding 9800 individual motifs.
  - ProDom
    - http://prodes.toulouse.inra.fr/prodom/doc/prodom.html.
    - Collection of protein motifs obtained automatically using PSI-BLAST.
    - Very high throughput ... but no annotation.
    - ProDom release 2001.2 contains 101957 families (at least 2 sequences per family).
  - ...

InterPro

- InterPro is an attempt to group a number of protein domain databases:
    - Pfam
    - PROSITE
    - PRINTS
    - ProDom
    - SMART
    - TIGRFAMs
- InterPro tries to have and maintain a high quality annotation.
- Very good accession to examples.
- InterPro web site: http://www.ebi.ac.uk/interpro.
- The database and a stand-alone package (iprscan) are available for UNIX platforms to locally run a complete Interpro analysis: ftp://ftp.ebi.ac.uk/pub/databases/interpro.

InterPro

- ▷ Example of a graphical output:



IPR000219 Domain	PF00621	RhoGEF	Dbl domain (dbl/cdc24 rhoGEF family)
	PS50010	DH_2	
	SM00325	RhoGEF	
IPR000980 Domain	PD000093	SH2	SH2 motif
	PF00017	SH2	
	PR00401	SH2DOMAIN	
	PS50001	SH2	
	SM00252	SH2	
IPR001331 Domain	PS00741	DH_1	Guanine-nucleotide dissociation stimulator, CDC24
IPR001452 Domain	PD000066	SH3	SH3 domain
	PF00018	SH3	
	PR00452	SH3DOMAIN	
	PS50002	SH3	
	SM00326	SH3	
IPR001715 Domain	PF00307	CH	Calponin-like actin-binding
	PS50021	CH	
	SM00033	CH	
IPR001849 Domain	PF00169	PH	Pleckstrin-like
	PS50003	PH_DOMAIN	
	SM00233	PH	
IPR002219 Domain	PF00130	DAG_PE-bind	Protein kinase C, phorbol ester/diacylglycerol binding
	PS00479	DAG_PE_BIND_DOM_1	
	PS50081	DAG_PE_BIND_DOM_2	
	SM00109	C1	
IPR003096 Family	PR00888	SM22CALPONIN	SM22/calponin
IPR003247 Domain	PD001527	CH_type	Calponin-like actin-binding subtype

The end