

Hidden Markov Models (HMMs) and Profiles

Swiss Institute of Bioinformatics (SIB)

26-30 November 2001

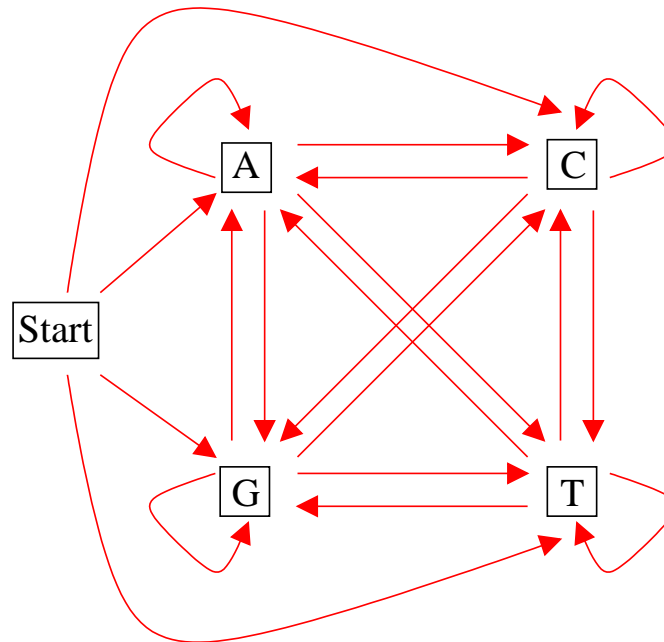
Markov Chain Models

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

An example of Markov Chain Model:

- Transition probabilities:

- ▷ $P(A|G) = 0.18$, $P(C|G) = 0.38$, $P(G|G) = 0.32$, $P(T|G) = 0.12$
- ▷ $P(A|C) = 0.15$, $P(C|C) = 0.35$, $P(G|C) = 0.34$, $P(T|C) = 0.15$



Markov Chain Models

Given a sequence x of length L , we can ask how probable the sequence is given a Markov Chain Model:

$$P(x) = (P(x_L), P(x_{L-1}), \dots, P(x_1)) = \\ P(x_L | x_{L-1}, \dots, x_1) P(x_{L-1} | x_{L-2}, \dots, x_1) \dots P(x_1)$$

Key property of a Markov Chain (of order 1):

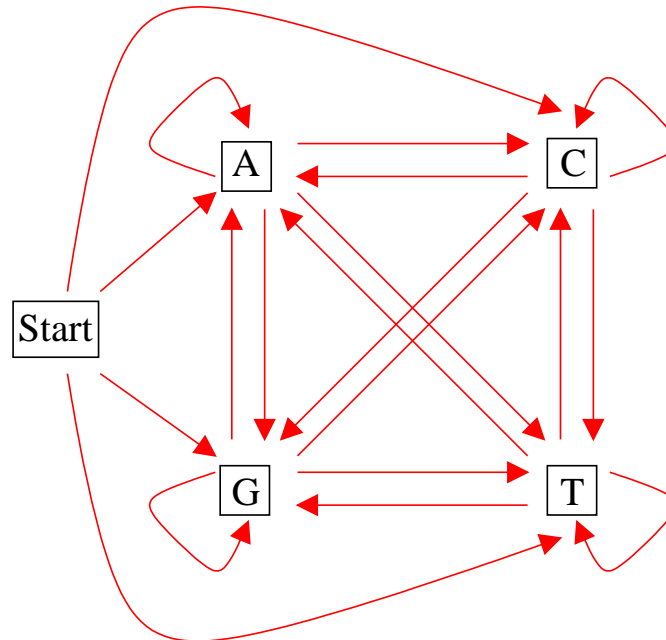
$$P(x_i | x_{i-1}, x_{i-2}, \dots, x_1) = P(x_i | x_{i-1}).$$

Therefore:

$$P(x) = P(x_L | x_{L-1}) P(x_{L-1} | x_{L-2}) \dots P(x_1) = \\ P(x_1) \prod_{i=2}^L P(x_i | x_{i-1})$$

What all this stuff means?

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



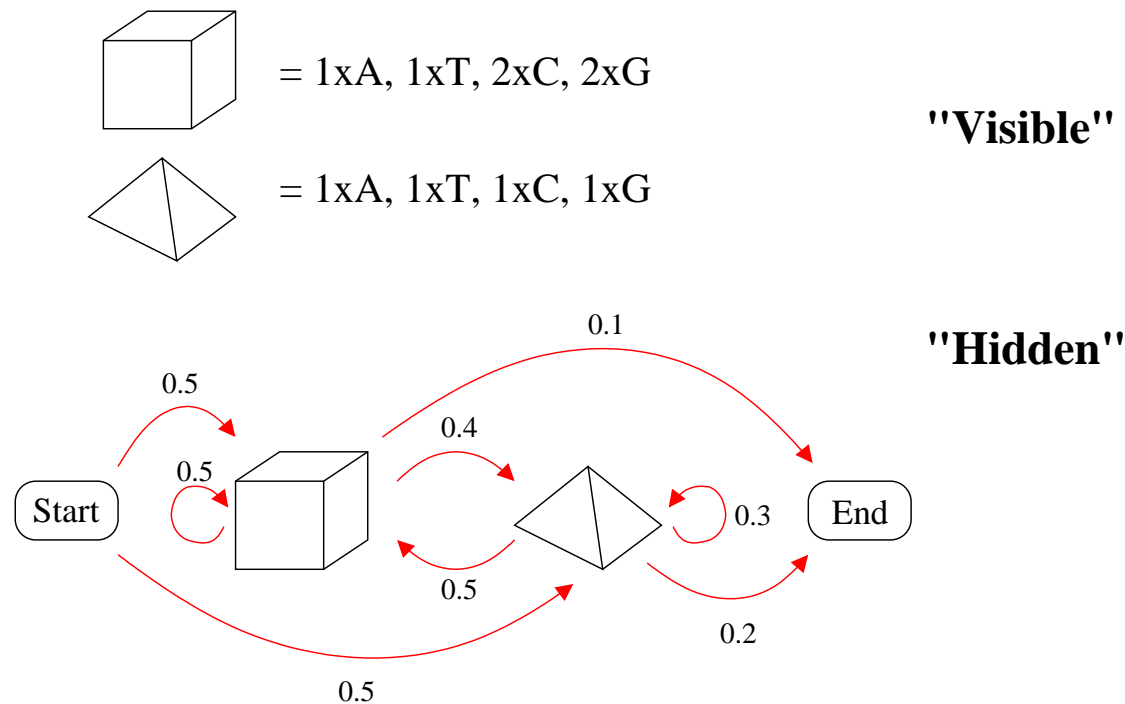
The probability of sequence $x = GCCT$ is:

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

Hidden Markov Models

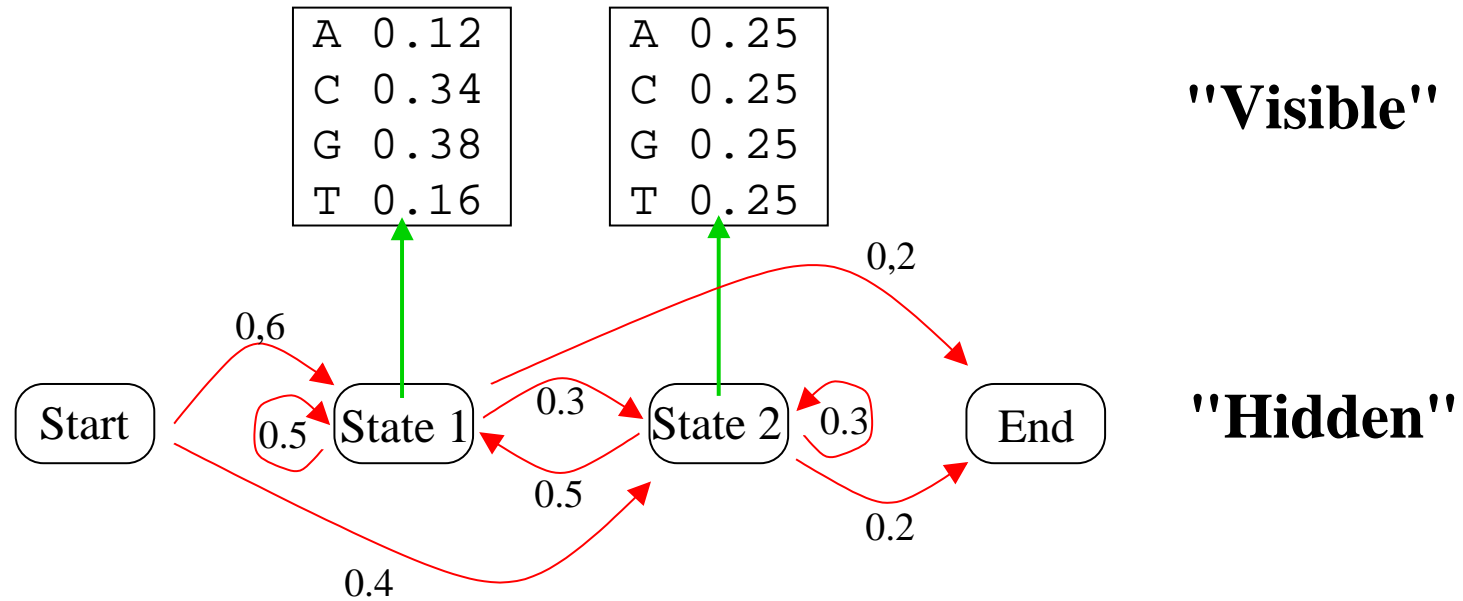
Hidden Markov Models (HMMs) are like Markov Chain Models: a finite number of **states** connected between them by **transitions**.

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.



Hidden Markov Model

Example of a simple Hidden Markov Model:



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

Hidden Markov Models

The parameters of the HMMs:

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

$$E(x|q)$$

- **Transition probabilities.** This is the probability of a transition from to a state r being in state q .

$$T(r|q)$$

Parameter estimation

How we can determine the parameters (emission and transition probabilities) of our model?

To estimate the parameters we use **Maximum Likelihood Estimation**.

Estimation of the probability of observing one symbol from an alphabet α (in our case $\alpha = \{A, C, G, T\}$):

- Given a set of sequences:

```
gccgcgcttg  
gcttggtggc  
tggccgttgc
```

- the maximum likelihood estimates are:

$$P(A) = \frac{0}{30} = 0$$

$$P(G) = \frac{13}{30} = 0.433$$

$$P(C) = \frac{9}{30} = 0.3$$

$$P(T) = \frac{8}{30} = 0.267$$

Parameters estimation

To avoid to set probabilities to 0 we can use different methods:

- For example add pseudo-counts:

$$P(A) = \frac{0+1}{34} = 0,029 \quad P(G) = \frac{13+1}{34} = 0.412$$

$$P(C) = \frac{9+1}{34} = 0.294 \quad P(T) = \frac{8+1}{34} = 0.265$$

A general form for priors estimation:

$$P(A) = \frac{n_a + p_A m}{(\sum_i n_i) + m}$$

where m is the number of virtual instances (pseudo-counts) and p_A is the prior probability of A .

Parameters estimation

Transition probabilities estimation between the symbols of the alphabet α (for example $\alpha = \{A, C, G, T\}$):

- Transition probabilities are estimated for each observed couple of symbols:

$$p_{ij} = \frac{n_{ij}}{\sum_{j'} n_{ij'}}$$

- where p_{ij} is the probability of transition from symbol i to symbol j , and n_{ij} is the number of transition from symbol i to symbol j observed.
- Is possible to add pseudo-counts as described previously.

For example for transition $A \rightarrow G$:

- Count all transitions $A \rightarrow A$, $A \rightarrow C$, $A \rightarrow G$, $A \rightarrow T$ in CpG sequences and non-CpG sequences.
- Transition probability is then calculated:

$$p_{AG} = \frac{n_{AG}}{n_{AA} + n_{AC} + n_{AG} + n_{AT}}$$

An example

Distinguish CpG islands from other sequences regions.

- Two models are required:
 - ▷ a model to represent CpG islands
 - ▷ a **null model** to represent the other regions

A sequences x is scored for being a CpG island:

$$score(x) = \log \frac{P(x|CpGmodel)}{P(x|nullmodel)}$$

Parameters for the CpG island and null models:

| CpG | A | C | G | T | Null | A | C | G | T |
|-----|------|------|------|------|------|------|------|------|------|
| A | 0.18 | 0.27 | 0.43 | 0.12 | A | 0.18 | 0.27 | 0.43 | 0.12 |
| C | 0.17 | 0.37 | 0.27 | 0.19 | C | 0.17 | 0.37 | 0.27 | 0.19 |
| G | 0.16 | 0.34 | 0.38 | 0.12 | G | 0.16 | 0.34 | 0.38 | 0.12 |
| T | 0.08 | 0.36 | 0.38 | 0.18 | T | 0.08 | 0.36 | 0.38 | 0.18 |

HMMs algorithms

Three important questions can be answered by three algorithms.

How likely is a given sequences 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.

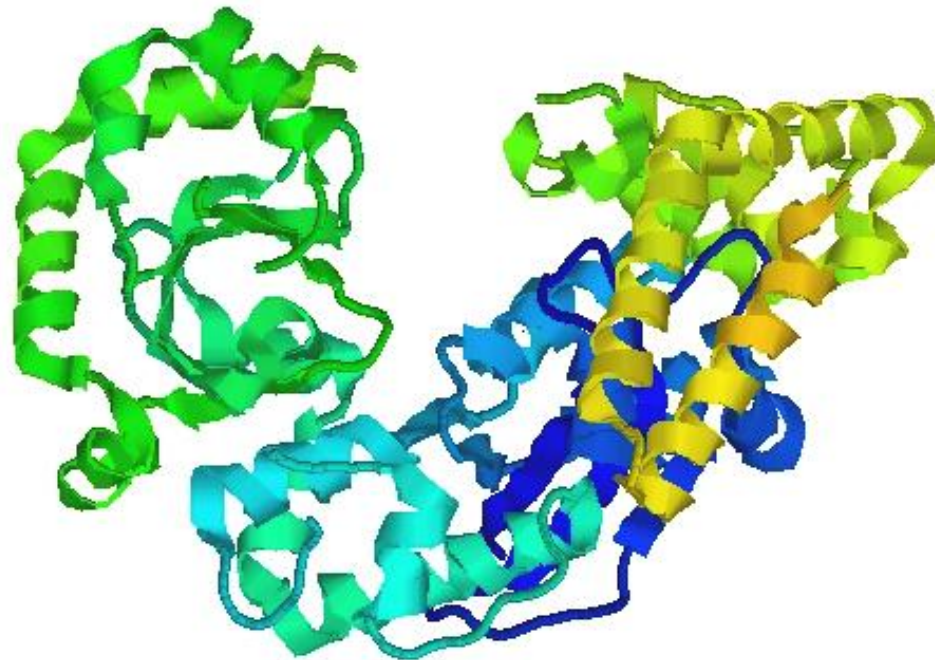
Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.
Cambridge University Press, 1998.

HMMs and protein domains

HMMs are very appropriate for modelization of protein domains.

What are protein domains:

- Domains are discrete structural units.
- Defined by structure.
- Domains are functional units.



HMMs and protein domains

Multiple alignments are used as training set for the building of HMMs.

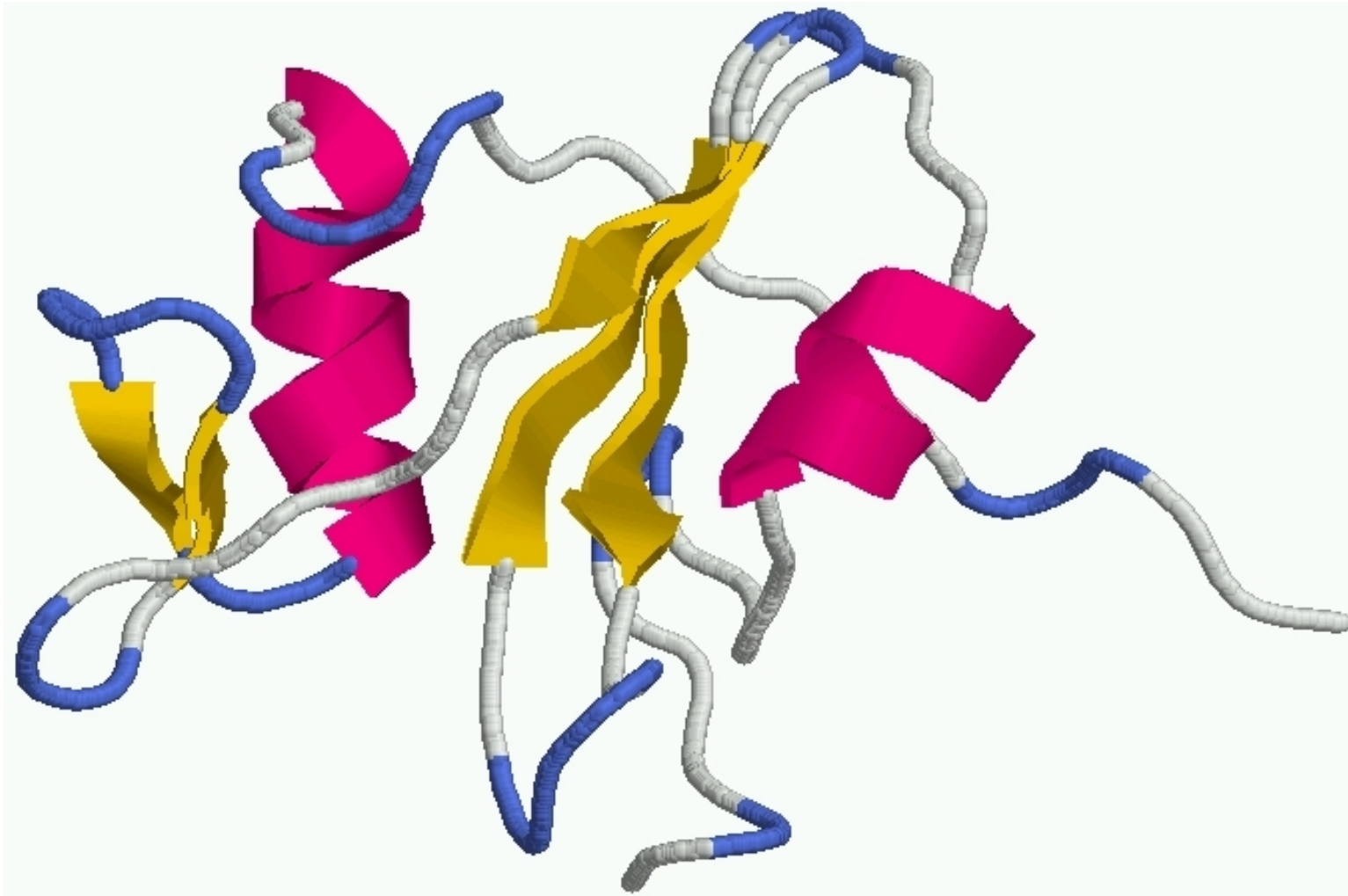
sh2.stk

```

ABL1_CAEL 194 WYHGKISRSDSEAILGS..GITGSFLVRESE.T.SIGQ...YTLVSRHDG.....RVFHYRINVDNTE.....KMFITQEVK..FRTLGEVLVHHH 269
P85A_HUMAN 624 WNVGSSNRNKAENLLRG..KRDGTFLVRESS..KQGC...YACSVVVDG.....EVKHCVINKTATG.....YGFAEPYNLYSSLKELVLHY 698
DRK_DROME 60 WYYGRITRADAIEKLLSN..KHEGAFLIRISES.SPGD...FSLSVKCPD.....GVQHFQVLRDAQS.....KFFLVVVK..FNSLNELVEYH 134
NCK_HUMAN 282 WYYGKVTRHQAEIMALNE.RGHEGDFLIRDSES.SPND...FSVSLKAQG.....KNKHFKVQLKET.....VYCIGQRK..FSTMEELVEHY 356
BLK_MOUSE 117 WFFRTISRKDAERQLLAPMKNKAGSFLIRESES.NKGA...FSLSVKDITTOG...EVVKHYKIRSLDNG.....GYYISPRIT..FPTLQALVQHY 198
YES_XIPHE 159 WYFGKLSRKDTERLLLPGNERGTFLIRESET.TKGA...YSLSLRDWDETK..GDNCKHYKIRKLDNG.....GYYITTRTQ..FMSLQMLVKHY 241
STK_HYDAT 126 WYFGDVKRAEAEKRLMVRGLPSGTFLIRKAET.AVGN...FSLSVRDGD.....SVKHYRVRKLDTG.....GYFITTRAP..FNSLYELVQHY 203
SRK1_SPOLA 122 WFLGKIKRVEAEKMLNQSFNQVGSFLIRDSET.TPGD...FSLSVKDQD.....RVRHYRVRRLDGD.....SLFVTRRST..FQILHELVDHY 199
SRC1_DROME 162 WFFENVLRKEADKLLLAENPRGTFLVRPSEH.NPNG...YSLSVKDWEDGR..GYHVKHYRIKPLDNG.....GYYIATNQT..FPSLQALVMAY 244
CRKL_HUMAN 14 WYMGPVSRQEAQTRLQG..QRHGMLVLRDSS.T.CPGD...YVLSVSENS.....RVSHYIINSLPNR.....RFKIGDQE..FDHLPALLEFY 88
CSK_CHICK 82 WFHGKITREQAERLLYP..PETGLFLVRETN.YPGD...YTLVCSCEG.....KVEHYRIIYSSSK.....LSIDEEVY..FENLMQLVEHY 156
MATK_HUMAN 122 WFHGKISGQEAQQOLQP..PEDGLFLVRESAR.HPGD...YVLCVSFGR.....DVIHYRVLHRDG.....HLTIDEAVF..FCNLMDMVEHY 196
CSW_DROME 6 WFHPTISGIEAEKLLQE.QGFDGSFLARLSS.NPGA...FTLSVRRGN.....EVTHIKIQNNGDF.....FDLYGGEK..FATLPELVQYY 81
CSW_DROME 111 WFHGNLSGKEAEKLLILE.RGKNGSFLVRESQS.KPGD...FVLSVRTDD.....KVTHVMIRWQDKK.....YDVGGGES..FGTSLSELIDHY 186
GTPA_HUMAN 181 WYHGKIDRTIAEERLRQ.AGKSGSYLIRESDR.RPGS...FVLSFSLQMN.....VNHFRIIAMCGD.....YYIGG.RR..FSSLSDLIGYY 256
PIP4_RAT 668 WYHASLTRAQAHEHMLMR.VPRDGAFLVRKRNE..PNS...YALSFRAGE...KIKHCRVQEGQT.....VMLGNSE..FDSLVDLISYY 741
KSYK_PIG 163 WFHGKISRDESEQIVLIGSKTNGKFLIRARDN...GS...YALGLLHEG.....KVLHYRIDKDKTG.....KLSIPGGKN..FDTLWQLVEHY 238
ZA70_HUMAN 163 WYHSSLTREEAERKLYSGAQTGKFLRPRKE..QGT...YALSIIYGK.....TVYHYLISQDKAG.....KYCIPEGTK..FDTLWQLVEYL 239
VAV_MOUSE 671 WYAGPMERAGAEGILTN..RSDGTYLVRQRVK.DTAE...FALSIVYNV.....EVKHIKIMTSEG.....LYRITEKKA.FRGLLELVEFY 745
KSYK_HUMAN 15 FFFGNITREEAEDYLQGGMSDGLYLLRQSRN.YLGG...FALSVAHGR.....KAHYTIERELNG.....TYAIAGGRT..HASPADLCHYH 92
SPK1_DUGTI 100 EAWREIQRWEAEKSLMKIGLQKGTYIIRPSR..KENS...YALSVRDFDEKKK.ICIVKHFQIKTLQDEK.....GISYSVNIRN.FPNILTLIQFY 184
BTK_HUMAN 281 WYSKHMTRSQAELQLLKQ.EGKEGGFIVRDS..KAGK...YTVSVFAKSTGDP.QGVIRHYVVCSTPQS.....QYYLAEKHL..FSTIPELINYH 362
TEC_MOUSE 246 WYCRNTNRSKAEQLLRT.EDKEGGFMVRDSS..QPGL...YTVSLYTKFGGEG.SSGFRHYHIKETATSP...KKYYLAEKHA..FGSIPEIIEYH 329
ITK_HUMAN 239 WYNKSLSRDKAEKLLLD.TGKEGAFMVRDSR..TAGT...YTVSVFTKAVVSENNPCIKHYHIKETNDNP...KRYVVAEKYV..FDSIPELLINYH 323
TXK_HUMAN 150 WYHRNITRNQAEHLLRQ.ESKEGAFIVRDSR..HLGS...YTLVFMGARRST.EAAIKHYQIKKNDG.....QWYVAERHA..FQSIPELIWYH 231
SRC2_DROME 214 WYVGYMSRQRAESLLKQ.GDKEGCFVVRKSS..TKGL...YTLVSLHTKVPQ...SHVKHYHIKQNARC.....EYVSEKHC..CETIPDLINYH 292
FER_HUMAN 460 WYHGAIPRIEAQELLK...KQGDFLVRESHG.KPGE...YVLSVYSDG.....QRRHFIIQYVNDM.....YRFEGTG..FSNIPQLIDHH 531
FPS_DROME 438 WFHGVLPREEVVRLLNN...DGDFLVRETIRNEESQ...IVLSVCWNGH.....KHFIVQTTGEG.....NFRFEGPP..FASIQELIMHQ 510
GTPA_HUMAN 351 WFHGKISKQEAYNLLMT.VGQVCSFLVRPSDN.TPGD...YSLYFRTNE.....NIQRFKICPTPN.....NQFMMGGRY..YNSIGDIIDHY 426
YKF1_CAEL 20 YFHGLIQREDVFQLLDN...NGDYVVRVLD.P.KPGEPRSYILSVMFNNKLDE.NSSVKHFVINSVENK.....YFVNNNMS..FNTIQQMLSHY 101
SPT6_YEAST 1258 YFFPFNGR.QAEDYLRS..KERGEFVIRQSSR.GDDH...LVITWKLDKD.....LFQHIDIQELEKENPLALGKVLIVDNQK..YNDLDQIIVEY 1339

```

HMMs and protein domains



HMMs and protein domains

A model based on three types of states is appropriate to modelize biological sequences:

- **match state:** Emits a symbol corresponding to a match.
- **insert state:** Emits a symbol between matches.
- **delete state:** Non-emitting silent state.

Match = M

Insertion = I

Deletion = D

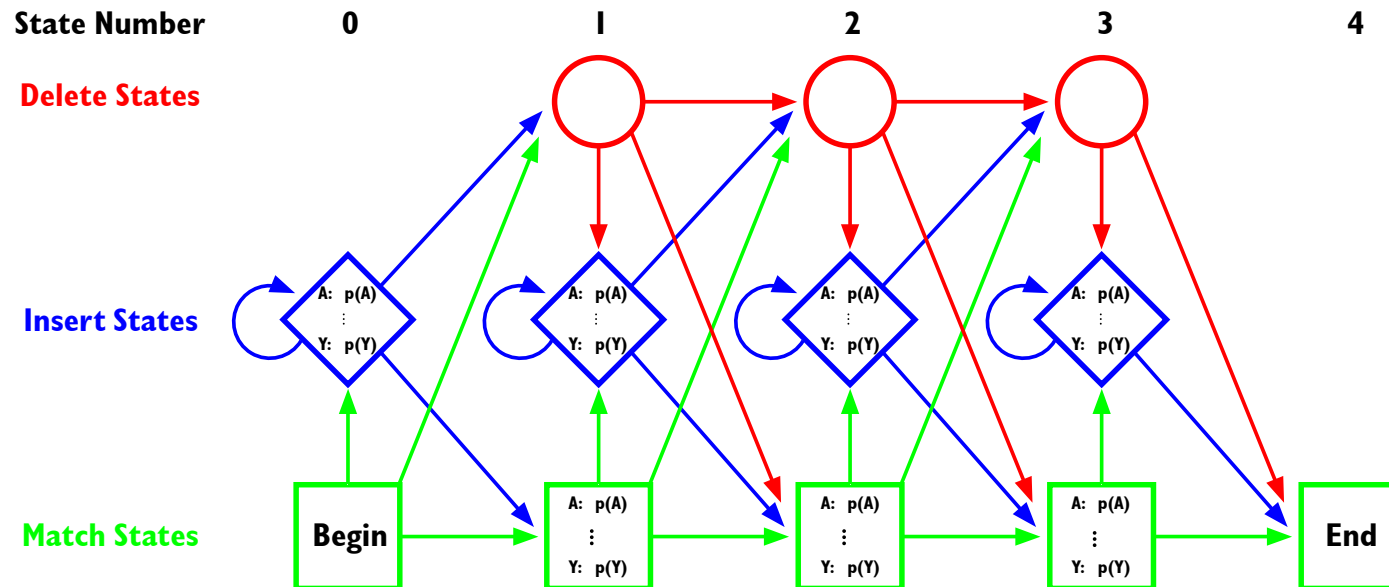
HMM

| | | | | | | | |
|-----|---|---------------------------------------------|---------------|------------|-------------|----------|----|
| HMM | 1 | WYHGKISR.SDSEAILGS.....GITGSFLVRESETSIGQYTI | SVRHDGRVFHYRI | INVDNTE... | KMFITQEVKFR | TLGELVHH | 76 |
| SEQ | 1 | WFHKKVEKRTSAEKLLQEYCMETGK.....VRESETFPNDYTL | SFWRSGRVQHCR | IRSTMEGGTL | KYYLTDNLRFR | MYALIQHY | 81 |

The diagram shows the alignment between the HMM sequence (top) and the SEQ sequence (bottom). The HMM sequence is: WYHGKISR.SDSEAILGS.....GITGSFLVRESETSIGQYTI SVRHDGRVFHYRI INVDNTE... KMFITQEVKFR TLGELVHH 76. The SEQ sequence is: WFHKKVEKRTSAEKLLQEYCMETGK.....VRESETFPNDYTL SFWRSGRVQHCR IRSTMEGGTL KYYLTDNLRFR MYALIQHY 81. The alignment is indicated by horizontal lines: green lines for matches (M), blue lines for insertions (I), and a red line for a deletion (D). The match states (M) are located under the following amino acid pairs: (W, W), (Y, Y), (H, H), (K, K), (I, I), (S, S), (R, R), (E, E), (A, A), (I, I), (L, L), (G, G), (S, S), (G, G), (I, I), (T, T), (G, G), (S, S), (F, F), (L, L), (V, V), (R, R), (E, E), (S, S), (I, I), (G, G), (Q, Q), (Y, Y), (T, T), (I, I), (S, S), (V, V), (R, R), (H, H), (D, D), (G, G), (R, R), (V, V), (F, F), (H, H), (Y, Y), (R, R), (I, I), (I, I), (N, N), (V, V), (D, D), (N, N), (T, T), (E, E), (K, K), (M, M), (F, F), (I, I), (T, T), (Q, Q), (E, E), (V, V), (K, K), (F, F), (R, R), (L, L), (G, G), (E, E), (L, L), (V, V), (H, H), (H, H). The insertion states (I) are located under the following amino acid pairs: (S, -), (D, -), (S, -), (D, -), (E, -), (A, -), (I, -), (L, -), (G, -), (S, -), (-, S), (-, D), (-, S), (-, F), (-, L), (-, V), (-, R), (-, E), (-, S), (-, I), (-, G), (-, Q), (-, Y), (-, T), (-, I), (-, S), (-, V), (-, R), (-, H), (-, D), (-, G), (-, R), (-, V), (-, F), (-, H), (-, Y), (-, R), (-, I), (-, I), (-, N), (-, V), (-, D), (-, N), (-, T), (-, E), (-, K), (-, M), (-, F), (-, I), (-, T), (-, Q), (-, E), (-, V), (-, K), (-, F), (-, R), (-, L), (-, G), (-, E), (-, L), (-, V), (-, H), (-, H). The deletion state (D) is located under the amino acid pair (G, -).

HMMs and protein domains

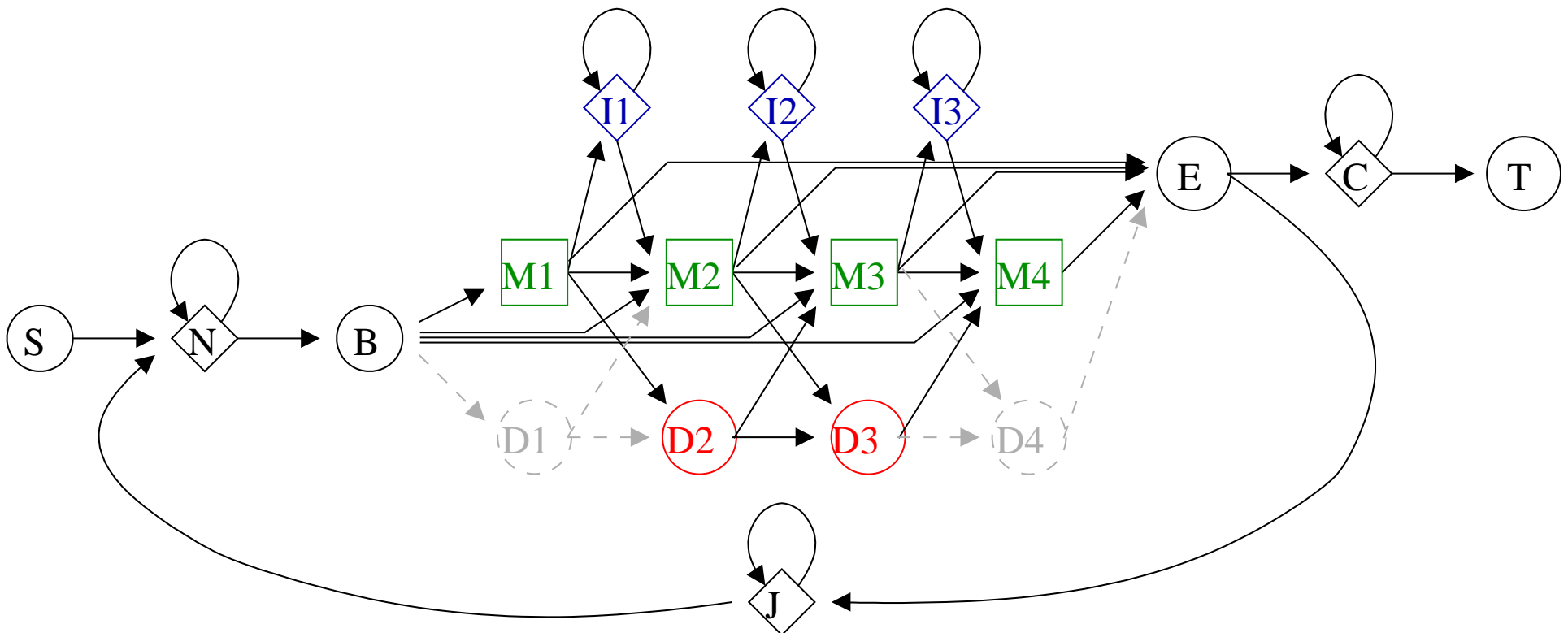
Simple HMM protein sequence model



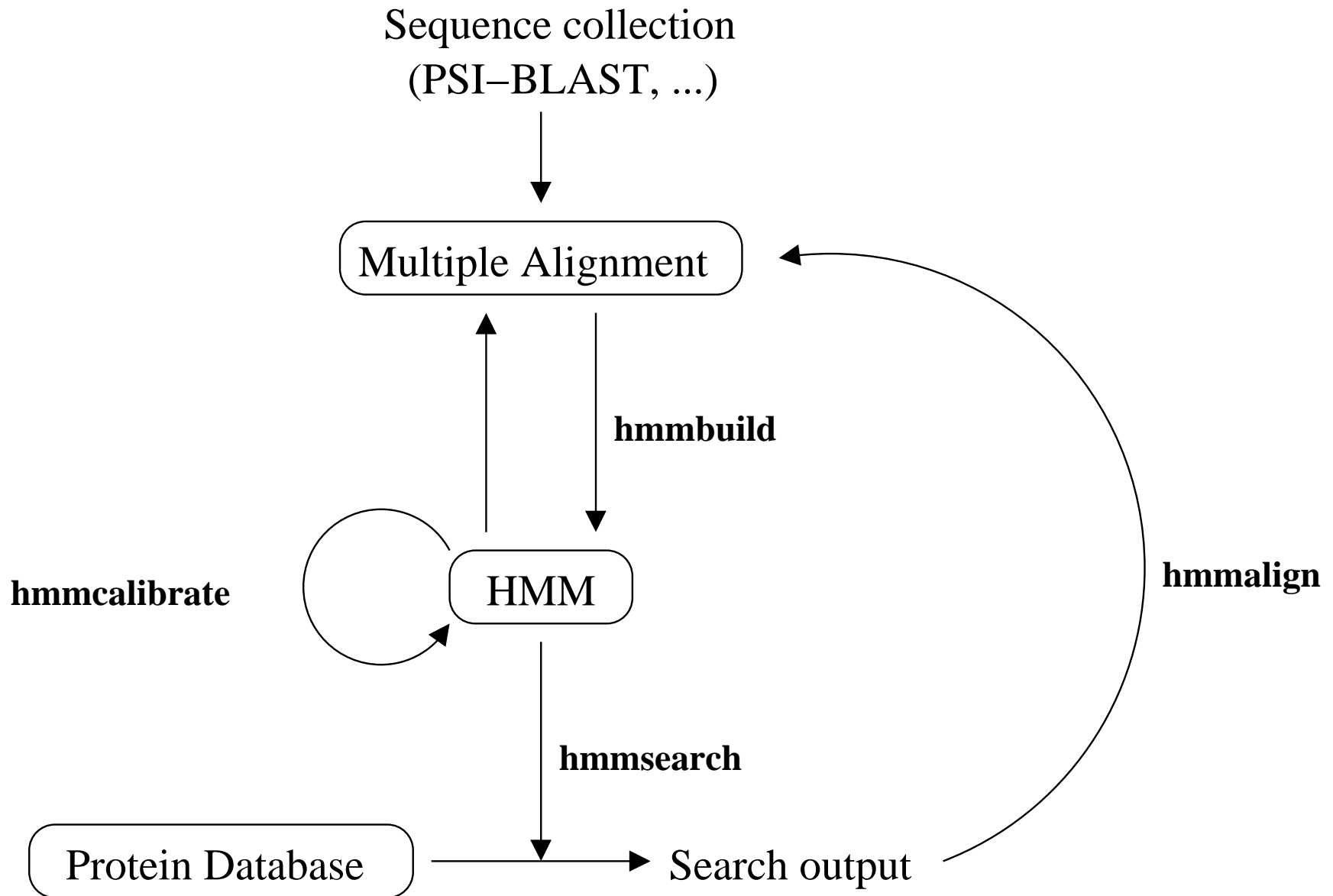
HMMER2

Package developed by Sean Eddy (<http://hmmer.wustl.edu/>).

HMMER uses the "Plan 7" architecture:



HMMER2



Generalized profiles

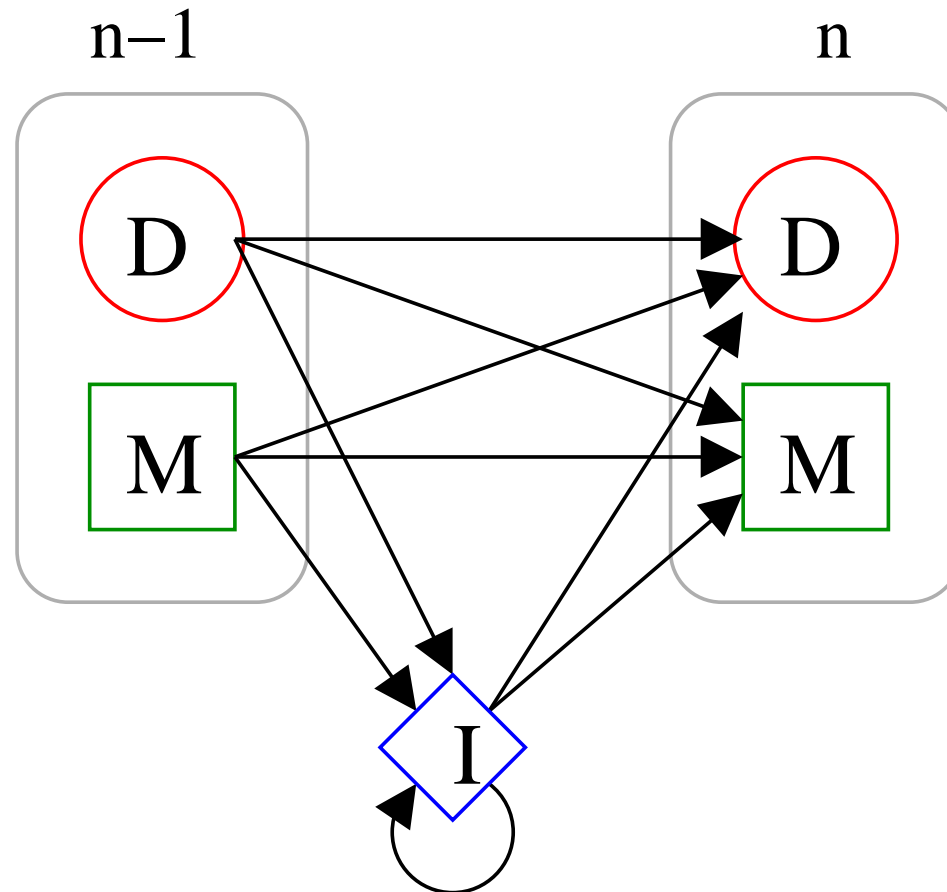
Generalized profiles combine aspects of Position Specific Score Matrices (PSSMs) and Hidden Markov Models (HMMs).

Generalized profiles are composed of alternating **match** and **insert** positions. Scores are associated with each position.

- The match position gives a residue specific match extension score and a deletion extension score.
- The insert position gives residue specific insertion scores.
- Scores are also associated to all possible transitions.

There is an equivalence between the structure of generalized profiles and a linear HMMs.

Generalized profiles



Pftools

The package has been developed by Philipp Bucher (<http://www.isrec.isb-sib.ch/ftp-server/pftools/>).

The package contains:

- `pfmake` for building a profile starting from multiple alignments.
- `pfsearch` to search a profile against a protein database.
- `pfscan` to search a protein against a profile database.

Two tools have been created to translate profiles from and into HMMs:

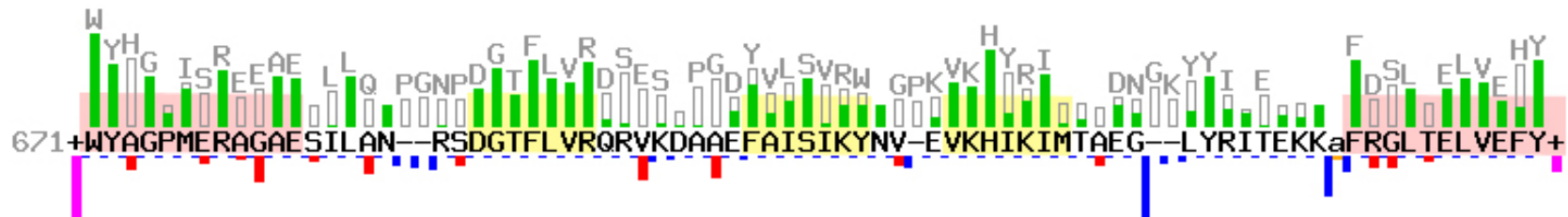
- `htop`: HMM to profile.
- `ptoh`: profile to HMM.

HMM-profiles and Generalized profiles

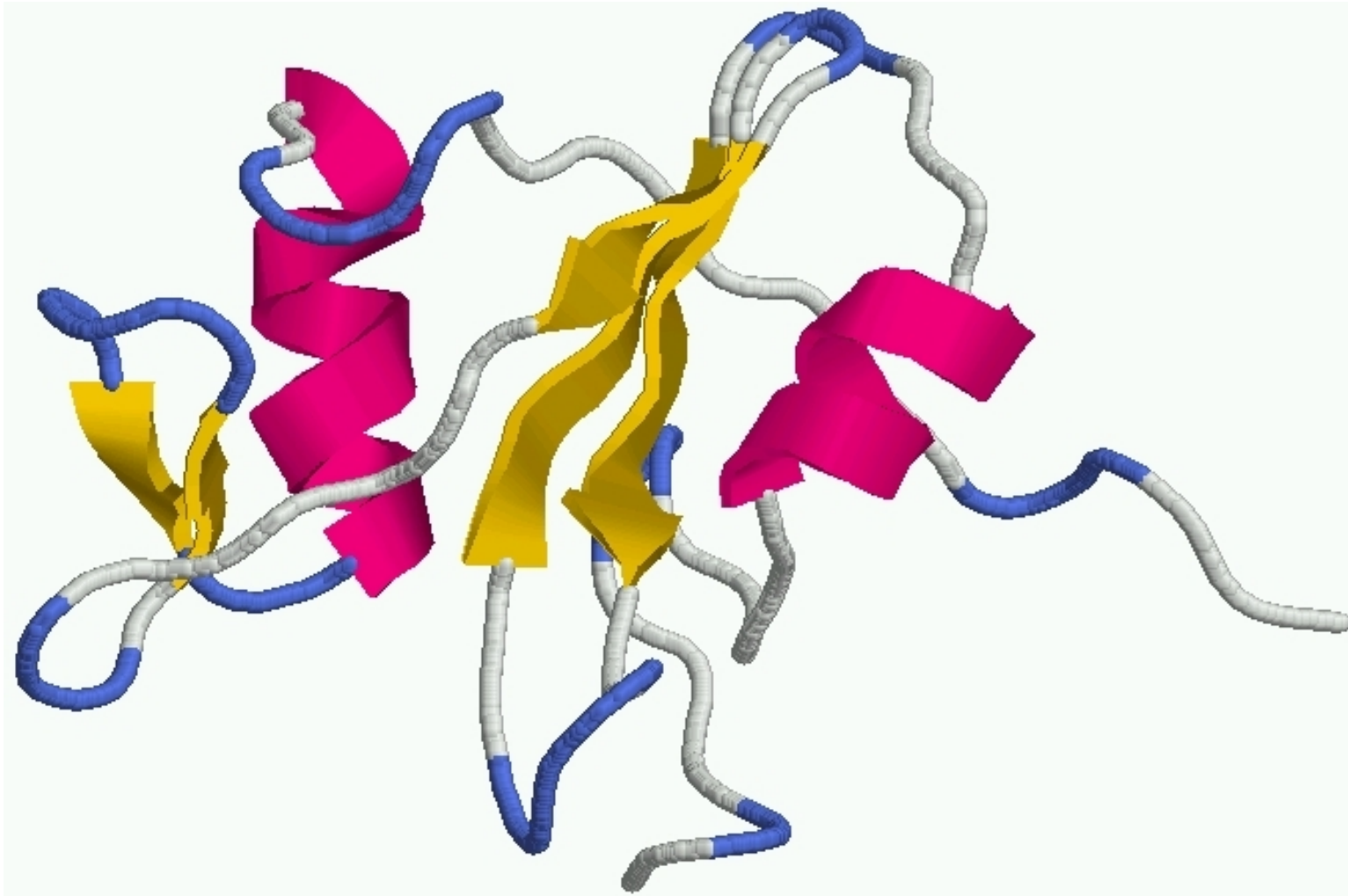
The meaning of the position specific scores in the generalized profiles:



The emission probabilities of HMM-profiles can be translated into a scoring system as for the generalized profiles:



HMMs and protein domains



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 (continue):

- Prints

- ▷ <http://bioinf.man.ac.uk/dbbrowser/PRINTS>.
- ▷ Collection of conserved motifs used to characterize a protein.
- ▷ Uses fingerprints (conserved motif group).
- ▷ 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 for family).

- ...

InterPro

InterPro is an attempt to group a number of protein domain databases:

- Pfam
- PROSITE
- PRINTS
- ProDom
- SMART
- TIGRFAMs

InterPro try to have and maintain a high quality annotation.

Very good accession to examples.

InterPro web site: <http://www.ebi.ac.uk/interpro>.

A stand alone package (iprscan) and the database are available for UNIX platforms to run a complete Interpro analysis: <ftp://ftp.ebi.ac.uk/pub/databases/interpro>.

InterPro

InterPro
Release 4.

Proteins belonging to InterPro entry IPR000980

To view the complete output click [here](#).

Related proteins [Q9VXH3](#)
"Phosphoinositide-specific phospholipase C (PLC)"

Related proteins [Q9YHU7](#)
"Phosphoinositide-specific phospholipase C (PLC)" "PI3"

Related proteins [Q9YHU6](#)
"Phosphoinositide-specific phospholipase C (PLC)"

Related proteins [PIG1_BOVIN](#)
[more proteins in list_1](#)
"Phosphoinositide-specific phospholipase C (PLC)"

Related proteins [PIG2_HUMAN](#)
[more proteins in list_2](#)
"Phosphoinositide-specific phospholipase C (PLC)"

InterPro Entry IPR000980

Src homology 2 (SH2) domain

| | |
|-----------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Database | InterPro |
| Accession | IPR000980; SH2 (matches 623 proteins) |
| Name | Src homology 2 (SH2) domain |
| Type | Domain i |
| Dates | 08-OCT-1999 (created) 16-FEB-2000 (last modified) |
| Signatures | PR00401 ; SH2DOMAIN (336 proteins) PS50001 ; SH2 (575 proteins) PF00017 ; SH2 (482 proteins) PD00093 ; SH2 (501 proteins) SM00252 ; SH2 (595 proteins) |
| Process i | intracellular signaling cascade (GO:0007242) |
| Abstract i | The Src homology 2 (SH2) domain is a protein domain of about 100 amino-acid residues first identified as a conserved sequence region between the oncoproteins Src and Fps [1]. Similar sequences were later found in many other intracellular signal-transducing proteins [2]. SH2 domains function as regulatory modules of intracellular signalling cascades by interacting with high affinity to phosphotyrosine-containing target peptides in a sequence-specific and strictly phosphorylation-dependent manner [3, 4, 5, 6]. They are found in a wide variety of protein contexts e.g., in association with catalytic domains of phospholipase Cy (PLCy) and the nonreceptor protein tyrosine kinases; within structural proteins such as fodrin and tensin; and in a group of small adaptor molecules, i.e Crk and Nck. In many cases, when an SH2 domain is present so too is an SH3 domain, suggesting that their functions are inter-related. The domains are frequently found as repeats in a single protein sequence. The structure of the SH2 domain belongs to the alpha-beta class, its overall shape forming a compact flattened hemisphere. The core structural elements comprise a central hydrophobic anti-parallel beta-sheet, flanked by 2 short alpha-helices. In the v-src oncogene product SH2 domain, the loop between strands 2 and 3 provides many of the binding interactions with the phosphate group of its phosphopeptide ligand, and is hence designated the phosphate binding loop. |
| Examples | <ul style="list-style-type: none"> P00543 FES_FSVST: Feline sarcoma virus Tyrosine-protein kinase transforming protein P42679 MATK_HUMAN: Human Megakaryocyte-associated tyrosine-protein kinase P03949 ABL1_CAEEL: C. elegans Tyrosine-protein kinase P29349 CSW_DROME: Drosophila Protein-tyrosine phosphatase P46109 CRKL_HUMAN: Human CRK-like protein P09851 RSG1_BOVIN: Bovine GTPase-activating protein |