

# Applications of HMMs in computational biology

---

**Jiří Kléma**

Department of Computer Science,  
Czech Technical University in Prague

Lecture based on Mark Craven's class at University of Wisconsin



<http://cw.felk.cvut.cz/wiki/courses/b4m36bin/start>

# Overview

---

- Hidden Markov models
  - the relationship between states and symbols remains hidden,
  - three important general tasks to be solved,
- two biological tasks solved with HMMs
  - characterization/classification of protein families,
  - gene finding.

# Hidden Markov models (HMMs)

---

- in the Markov models it is clear which state accounts for each symbol of the observed sequence,
- HMMs distinguish between the observed and hidden part of a problem
  - multiple states could account for each observed sequence symbol,
  - the link between states and symbols makes the hidden part of the problem,
- the parameters of an HMM
  - as in Markov chain models, we have transition probabilities

$$a_{kl} = P(\pi_i = l | \pi_{i-1} = k)$$

where  $\pi$  represents a path (sequence of states) through the model

- in addition, emission probabilities decouple states and symbols

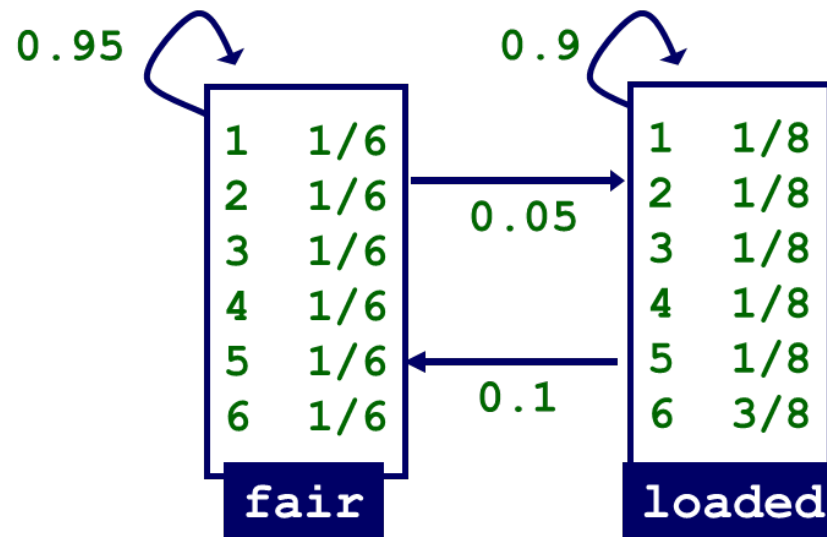
$$e_k(b) = P(x_i = b | \pi_i = k)$$

where  $e_k(b)$  is the probability of emitting character  $b$  in state  $k$ .

## Example HMM – dishonest casino

---

- Consider a dishonest casino
  - they use a fair die most of the time,
  - but occasionally they switch to a loaded die,
- HMM model
  - observable symbols = the outcomes of rolls =  $\{1,2,3,4,5,6\}$ ,
  - hidden states = the two dice =  $\{\text{fair}, \text{loaded}\}$ ,



# Three important HMM questions

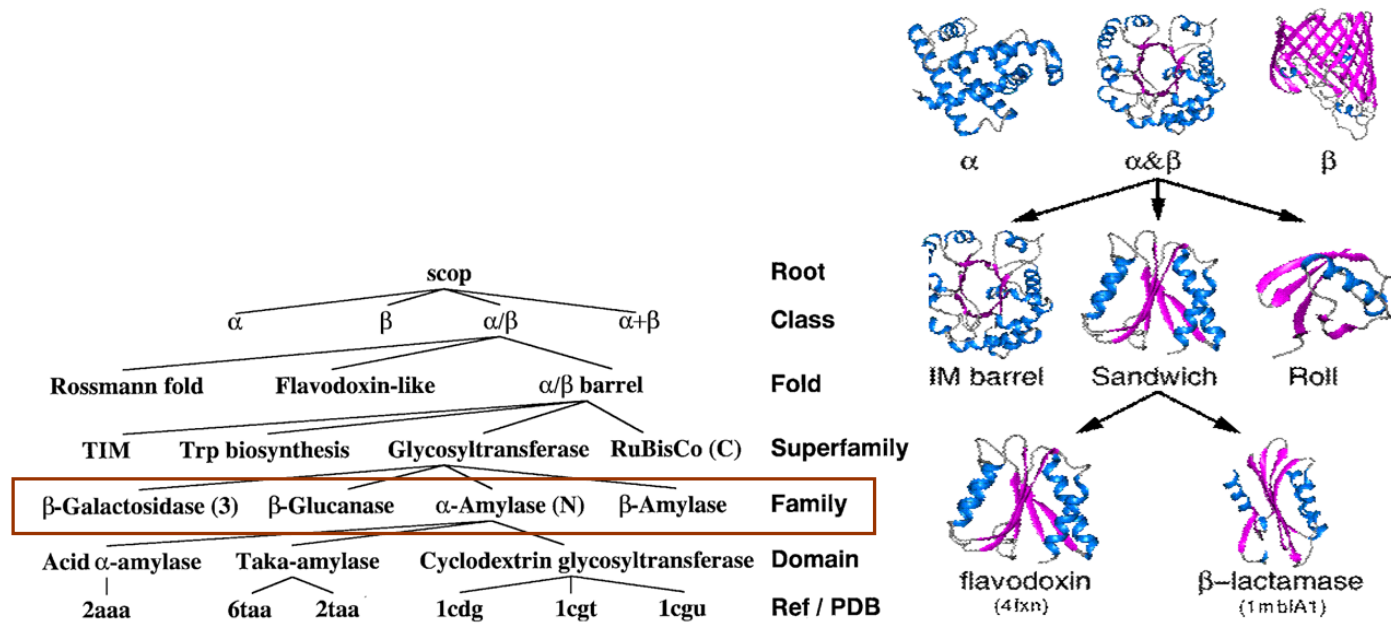
---

- How likely is a given sequence  $X$  given the model  $M$ ?
  - $P(X|M) = \sum_{\pi} P(X, \pi|M)$ , Forward algorithm,
  - assume a particular model of casino, calculate the probability of a certain sequence of rolls,
  - classification = when having more models, find the best match,
- What is the most probable “path” for generating a given sequence?
  - $\pi^* = \operatorname{argmax}_{\pi} P(X, \pi|M)$ , Viterbi algorithm,
  - having a sequence of rolls, decide which part is fair and which is not,
  - segmentation = “split” the sequence among states,
- How can we learn the HMM parameters given a set of sequences?
  - $\theta^* = \operatorname{argmax}_{\theta} P(X|\theta, M)$ , Forward-Backward (Baum-Welch) algorithm,
  - having a long sequence of rolls and a rough casino model, learn its probs,
  - learning = find a model that generalizes well to unseen sequences.

# The protein classification task

- Given: amino-acid sequence of a protein,
- Do: predict the family to which it belongs.

GDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL DNLKGT FATLSELHCDK LHVDPENFRLLGNVCVLAHHFGKEFTPPVQAAYAKVVAGVANALAHKYH



Marc Craven, BMI/CS 576, [www.biostat.wisc.edu/bmi576](http://www.biostat.wisc.edu/bmi576).

## Protein family – a simplified view

---

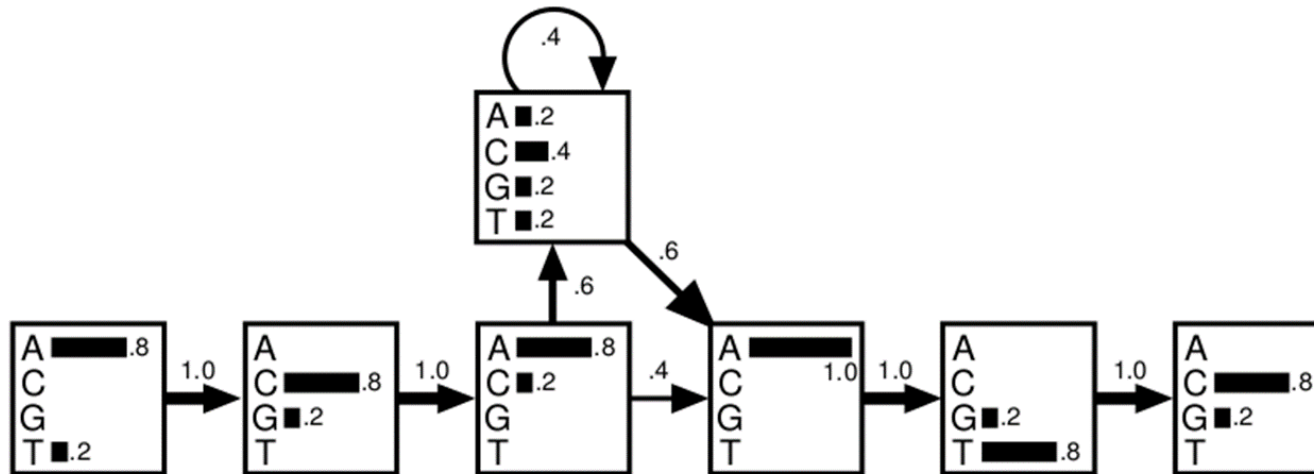
- Let us have a multiple sequence alignment for a protein family
  - how could we model the family?
  - do the aligned query sequences belong to the family?

|   |   |   |   |   |   |   |   |   |   |        |
|---|---|---|---|---|---|---|---|---|---|--------|
| A | C | A | - | - | - | A | T | G | } | family |
| T | C | A | A | C | T | A | T | C |   |        |
| A | C | A | C | - | - | A | G | C |   |        |
| A | G | A | - | - | - | A | T | C |   |        |
| A | C | C | G | - | - | A | T | C |   |        |

---

|   |   |   |   |   |   |   |   |   |         |
|---|---|---|---|---|---|---|---|---|---------|
| A | C | A | C | - | - | A | T | C | query 1 |
| A | A | A | C | - | - | A | T | C | query 2 |
| T | G | C | T | - | - | A | T | C | query 3 |

# Protein family – a reasonable HMM



|                    | Sequence          | Probability $\times 100$ | Log odds |
|--------------------|-------------------|--------------------------|----------|
| Consensus          | A C A C - - A T C | 4.7                      | 6.7      |
| Original sequences | A C A - - - A T G | 3.3                      | 4.9      |
|                    | T C A A C T A T C | 0.0075                   | 3.0      |
|                    | A C A C - - A G C | 1.2                      | 5.3      |
|                    | A G A - - - A T C | 3.3                      | 4.9      |
|                    | A C C G - - A T C | 0.59                     | 4.6      |
| Exceptional        | T G C T - - A G G | 0.0023                   | -0.97    |

Krogh: An Introduction to HMMs for Biological Sequences, CMMB 1998.



# Profile HMMs

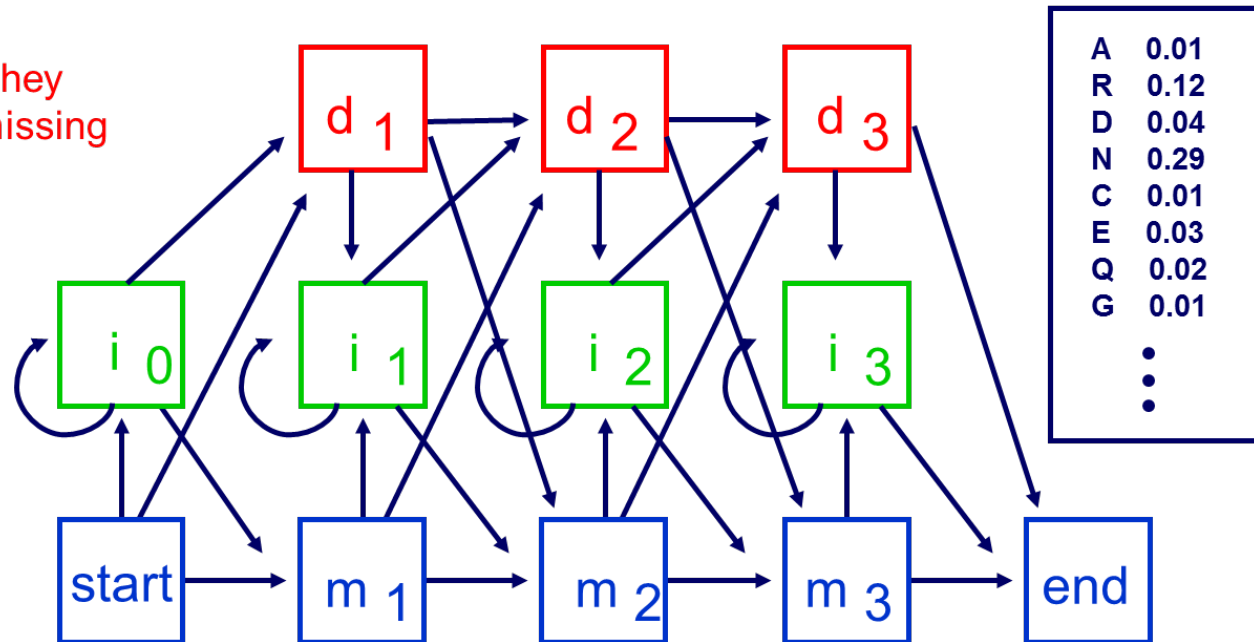
- profile HMMs are used to model families of sequences.

*Insert and match states have emission distributions over sequence characters*

*Delete states are silent; they account for characters missing in some sequences*

*Insert states account for extra characters in some sequences*

*Match states represent key conserved positions*



Marc Craven, BMI/CS 576, [www.biostat.wisc.edu/bmi576](http://www.biostat.wisc.edu/bmi576).

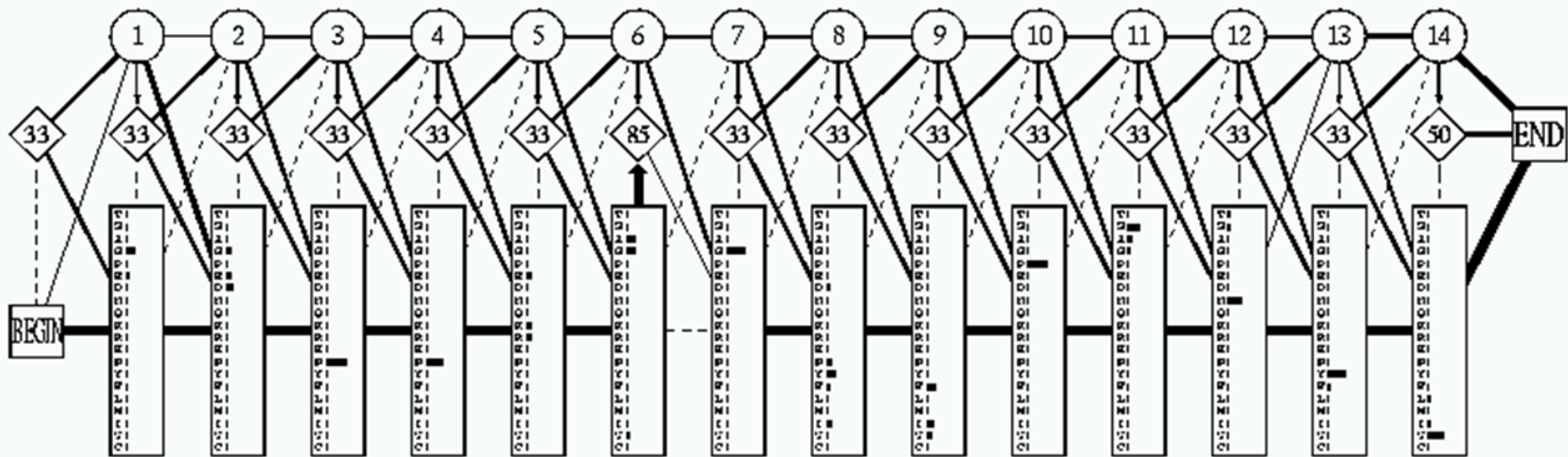
# Multiple alignment of SH3 domain

```
GGWWRGdy.ggkkkqLWFPSSNYV
IGWLNgynefttgerGGDFPSTYV
PNWWEgql..nnrrrGGIFPSTYV
DEWQAARR..deqqiGGIVPSK--
GEWKAARs..tqqeGGFIPFNFV
GDWLAARs..sqqqtGGYIPSTYV
GDWDAEL..kqgrrrGGKVPSTYL
-DWWEARsllssghrGGYVPSTYV
GDWYARsllitnseGGYIPSTYV
GEWKAARsllatrkGGYIPSTYV
GDWLAARsllvtgreGGYVPSTYV
GEWKAARsllskreGGFIPSTYV
GEWCEAARsllskreGGWVPSTYI
SDWWRVvnltrqqeGGLIPLNFFV
LPWWRARd.kngqqeGGYIPSTYI
RDWWEFRsktvypGGYYESGYV
EHWVKVkd.algnvGGYIPSTYV
IHWWRVqdr.ngheGGYVPSSYL
KDWVKVev..ndrqqGGFVPAAYV
VGWMPGlnerttrqrGGDFPSTYV
PDWWEGel..ngqrGGVFPASYV
ENWNGEei..gnrkGGIFPATYV
EEWLEGEec..kqkvGGIFPKVFFV
GGWKGdy.gttriqqQYFSTYV
DGWWRGSy..ngqvGGWFPSTYV
QGWWRGel..yqrvGGWFPANYSV
GRWKAARR..anggetGGIIPSTYV
GGWTQGel.ksgqkGGWAPTNYL
GDWWEARsn.tgenGGYIPSTYV
NDWWTGr t..ngkeGGIFPANYSV
```

Krogh: An Introduction to HMMs for Biological Sequences, CMMB 1998.

# A profile HMM trained for the SH3 domain

- delete states (silent) = upper line,
- insert states = middle line,
- match states = bottom line,
- line strength  $\approx$  transition probability.



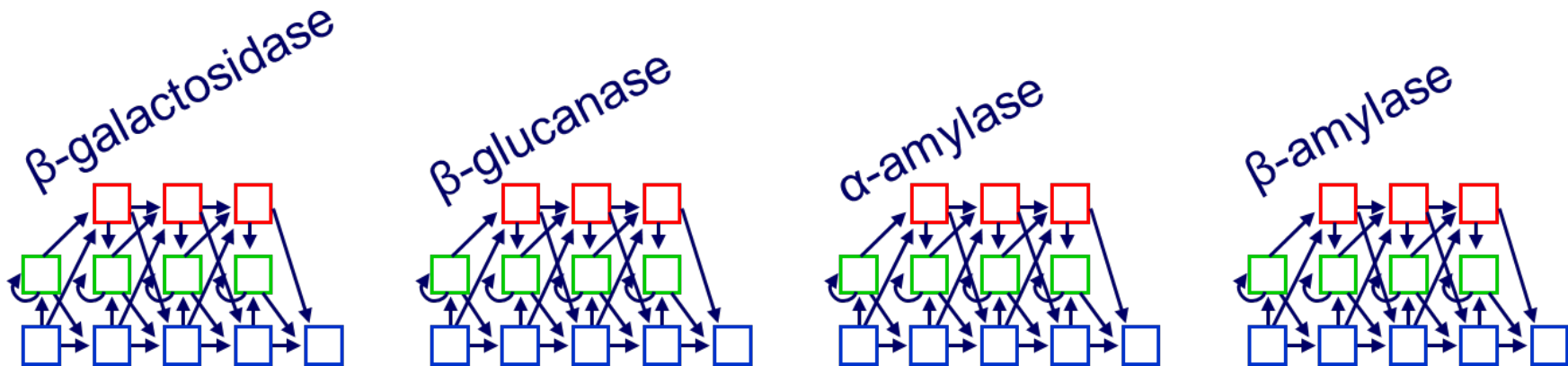
Krogh: An Introduction to HMMs for Biological Sequences, CMMB 1998.

# Profile HMMs

- to classify sequences according to family, we can train a profile HMM to model the proteins of each family of interest,
- given a sequence  $x$ , use Bayes' rule to make classification

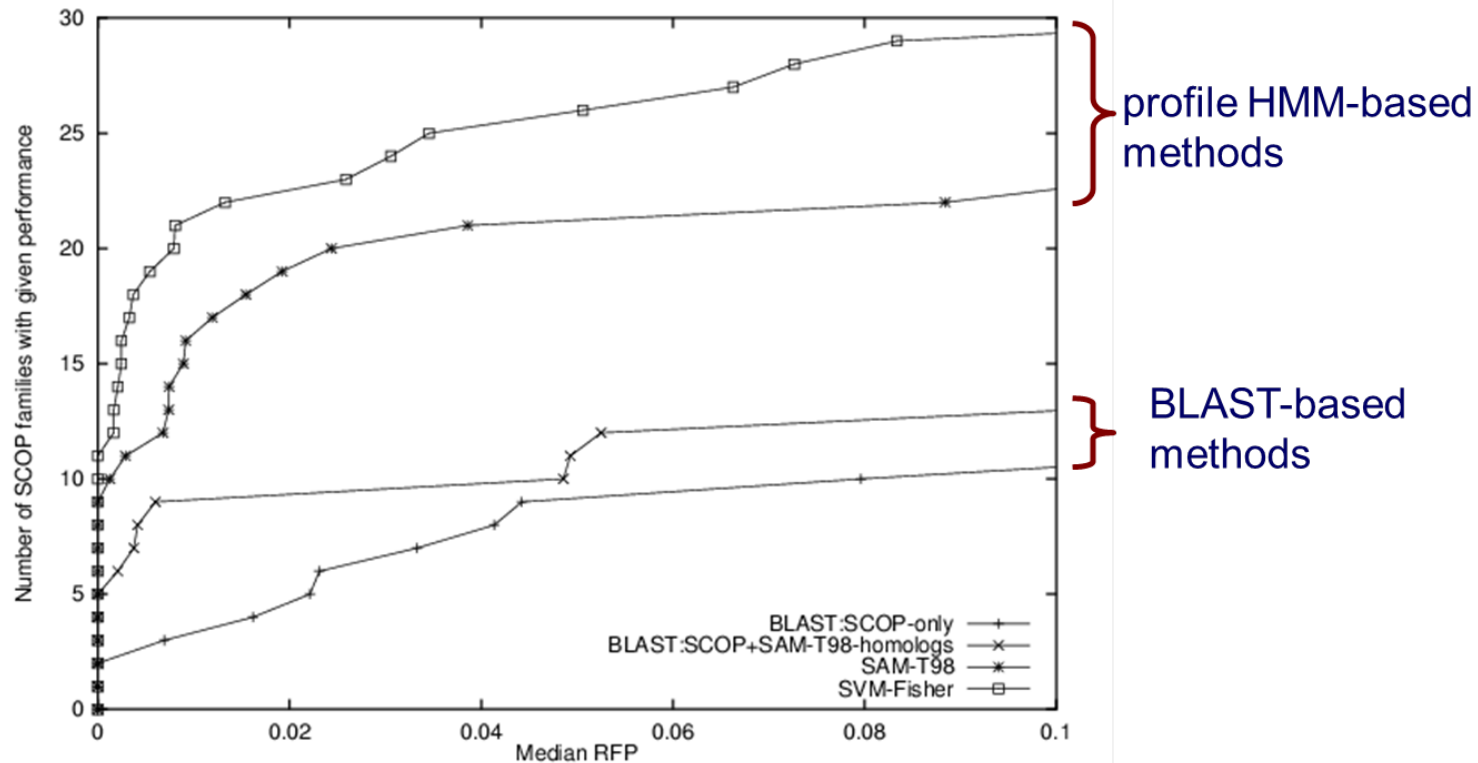
$$P(c_i|x) = \frac{P(x|c_i)P(c_i)}{\sum_j P(x|c_j)P(c_j)}$$

- use Forward algorithm to compute  $P(x|c_i)$  for each family  $c_i$ .



Marc Craven, BMI/CS 576, [www.biostat.wisc.edu/bmi576](http://www.biostat.wisc.edu/bmi576).

# Profile HMM accuracy



Jaakola et al., ISMB 1999.

- classifying 2447 proteins into 33 families,
- x-axis represents the median  $\#$  of negative sequences that score as high as a positive sequence for a given family's model.

# Pfam database – a large collection profile HMMs



HOME | SEARCH | BROWSE | FTP | HELP | ABOUT



## Pfam 34.0 (March 2021, 19179 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [Less...](#)

Proteins are generally composed of one or more functional regions, commonly termed **domains**. Different combinations of domains give rise to the diverse range of proteins found in nature. The identification of domains that occur within proteins can therefore provide insights into their function.

Pfam also generates higher-level groupings of related entries, known as **clans**. A clan is a collection of Pfam entries which are related by similarity of sequence, structure or profile-HMM.

The data presented for each entry is based on the [UniProt Reference Proteomes](#) but information on individual UniProtKB sequences can still be found by entering the protein accession. Pfam *full* alignments are available from searching a variety of databases, either to provide different accessions (e.g. all UniProt and NCBI GI) or different levels of redundancy.

### QUICK LINKS

[SEQUENCE SEARCH](#)

[VIEW A PFAM ENTRY](#)

[VIEW A CLAN](#)

[VIEW A SEQUENCE](#)

[VIEW A STRUCTURE](#)

[KEYWORD SEARCH](#)

[JUMP TO](#)

### YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...

Analyze your protein sequence for Pfam matches

View Pfam annotation and alignments

See groups of related entries

Look at the domain organisation of a protein sequence

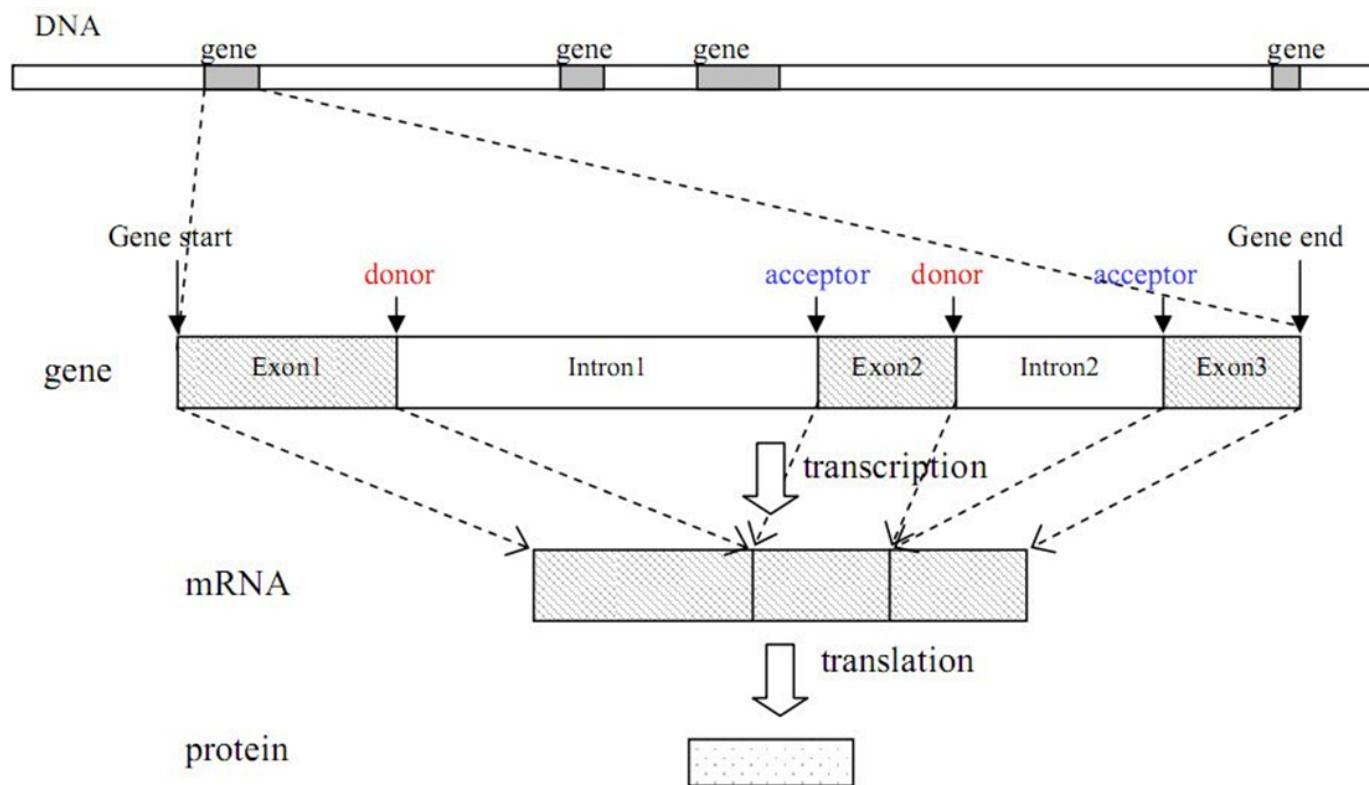
Find the domains on a PDB structure

Query Pfam by keywords

<http://pfam.xfam.org/>

# The gene finding task

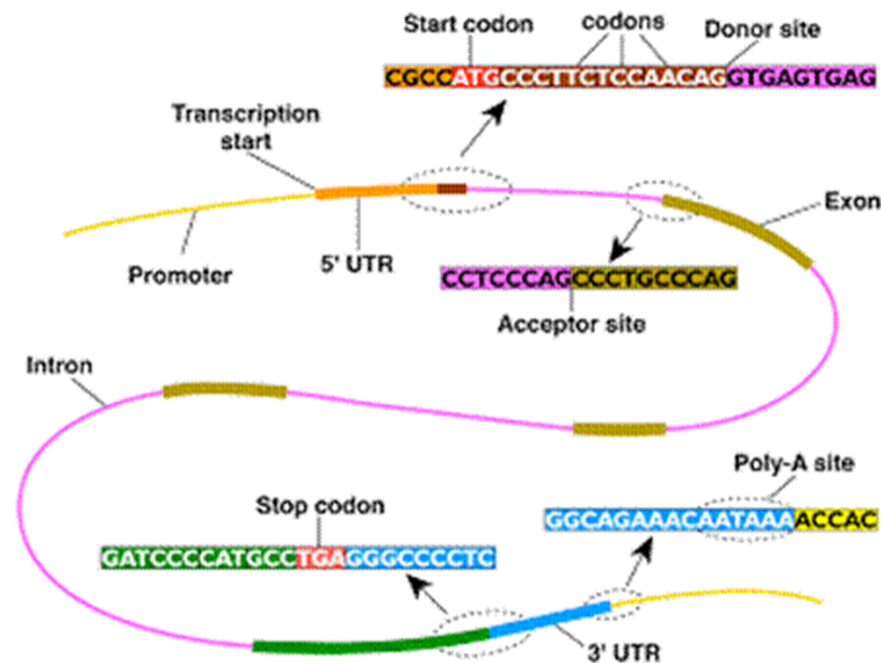
- Given: an uncharacterized DNA sequence,
- Do: locate the genes in the sequence, including the coordinates of individual exons and introns.



Marc Craven, BMI/CS 576, [www.biostat.wisc.edu/bmi576](http://www.biostat.wisc.edu/bmi576).

# Eukaryotic gene structure and evidence for gene finding

- **signals**: sequence signals (e.g. splice junctions) involved in gene expression,
- **content**: statistical properties that distinguish protein-coding DNA,
- **conservation**: signal and content properties that are conserved across related sequences (e.g. syntenic regions of the mouse and human genome).



Marc Craven, BMI/CS 576, [www.biostat.wisc.edu/bmi576](http://www.biostat.wisc.edu/bmi576).



# Gene finding: search by content

- Encoding a protein affects the statistical properties of a DNA sequence
    - some amino acids more frequent (Leu more prevalent than Trp),
    - different numbers of codons for different amino acids (Leu/6, Trp/1),
    - for a given amino acid, one codon often more frequent than others
- \* codon preference in E.coli and H. sapiens:

|            |            |            |            |
|------------|------------|------------|------------|
| UUU F 0.57 | UCU S 0.11 | UAU Y 0.53 | UGU C 0.42 |
| UUC F 0.43 | UCC S 0.11 | UAC Y 0.47 | UGC C 0.58 |
| UUA L 0.15 | UCA S 0.15 | UAA * 0.64 | UGA * 0.36 |
| UUG L 0.12 | UCG S 0.16 | UAG * 0.00 | UGG W 1.00 |
| CUU L 0.12 | CCU P 0.17 | CAU H 0.55 | CGU R 0.36 |
| CUC L 0.10 | CCC P 0.13 | CAC H 0.45 | CGC R 0.44 |
| CUA L 0.05 | CCA P 0.14 | CAA Q 0.30 | CGA R 0.07 |
| CUG L 0.46 | CCG P 0.55 | CAG Q 0.70 | CGG R 0.07 |
| AUU I 0.58 | ACU T 0.16 | AAU N 0.47 | AGU S 0.14 |
| AUC I 0.35 | ACC T 0.47 | AAC N 0.53 | AGC S 0.33 |
| AUA I 0.07 | ACA T 0.13 | AAA K 0.73 | AGA R 0.02 |
| AUG M 1.00 | ACG T 0.24 | AAG K 0.27 | AGG R 0.03 |
| GUU V 0.25 | GCU A 0.11 | GAU D 0.65 | GGU G 0.29 |
| GUC V 0.18 | GCC A 0.31 | GAC D 0.35 | GGC G 0.46 |
| GUA V 0.17 | GCA A 0.21 | GAA E 0.70 | GGA G 0.13 |
| GUG V 0.40 | GCG A 0.38 | GAG E 0.30 | GGG G 0.12 |

[Codon/a.a./fraction per codon per a.a.]  
E. coli K12 data from the Codon Usage Database

www.geneinfinity.org

|            |            |            |            |
|------------|------------|------------|------------|
| UUU F 0.46 | UCU S 0.19 | UAU Y 0.44 | UGU C 0.46 |
| UUC F 0.54 | UCC S 0.22 | UAC Y 0.56 | UGC C 0.54 |
| UUA L 0.08 | UCA S 0.15 | UAA * 0.30 | UGA * 0.47 |
| UUG L 0.13 | UCG S 0.05 | UAG * 0.24 | UGG W 1.00 |
| CUU L 0.13 | CCU P 0.29 | CAU H 0.42 | CGU R 0.08 |
| CUC L 0.20 | CCC P 0.32 | CAC H 0.58 | CGC R 0.18 |
| CUA L 0.07 | CCA P 0.28 | CAA Q 0.27 | CGA R 0.11 |
| CUG L 0.40 | CCG P 0.11 | CAG Q 0.73 | CGG R 0.20 |
| AUU I 0.36 | ACU T 0.25 | AAU N 0.47 | AGU S 0.15 |
| AUC I 0.47 | ACC T 0.36 | AAC N 0.53 | AGC S 0.24 |
| AUA I 0.17 | ACA T 0.28 | AAA K 0.43 | AGA R 0.21 |
| AUG M 1.00 | ACG T 0.11 | AAG K 0.57 | AGG R 0.21 |
| GUU V 0.18 | GCU A 0.27 | GAU D 0.46 | GGU G 0.16 |
| GUC V 0.24 | GCC A 0.40 | GAC D 0.54 | GGC G 0.34 |
| GUA V 0.12 | GCA A 0.23 | GAA E 0.42 | GGA G 0.25 |
| GUG V 0.46 | GCG A 0.11 | GAG E 0.58 | GGG G 0.25 |

[Codon/a.a./fraction per codon per a.a.]  
Homo sapiens data from the Codon Usage Database

www.geneinfinity.org

# The GENSCAN HMM for eukaryotic gene finding

Each shape denotes a functional unit of a gene or genomic region and is represented by a submodel in the HMM

Pairs of intron/exon units represent the different ways an intron can interrupt a coding sequence (after 1<sup>st</sup> base in codon, after 2<sup>nd</sup> base or after 3<sup>rd</sup> base)

Complementary submodel (not shown) detects genes on opposite DNA strand

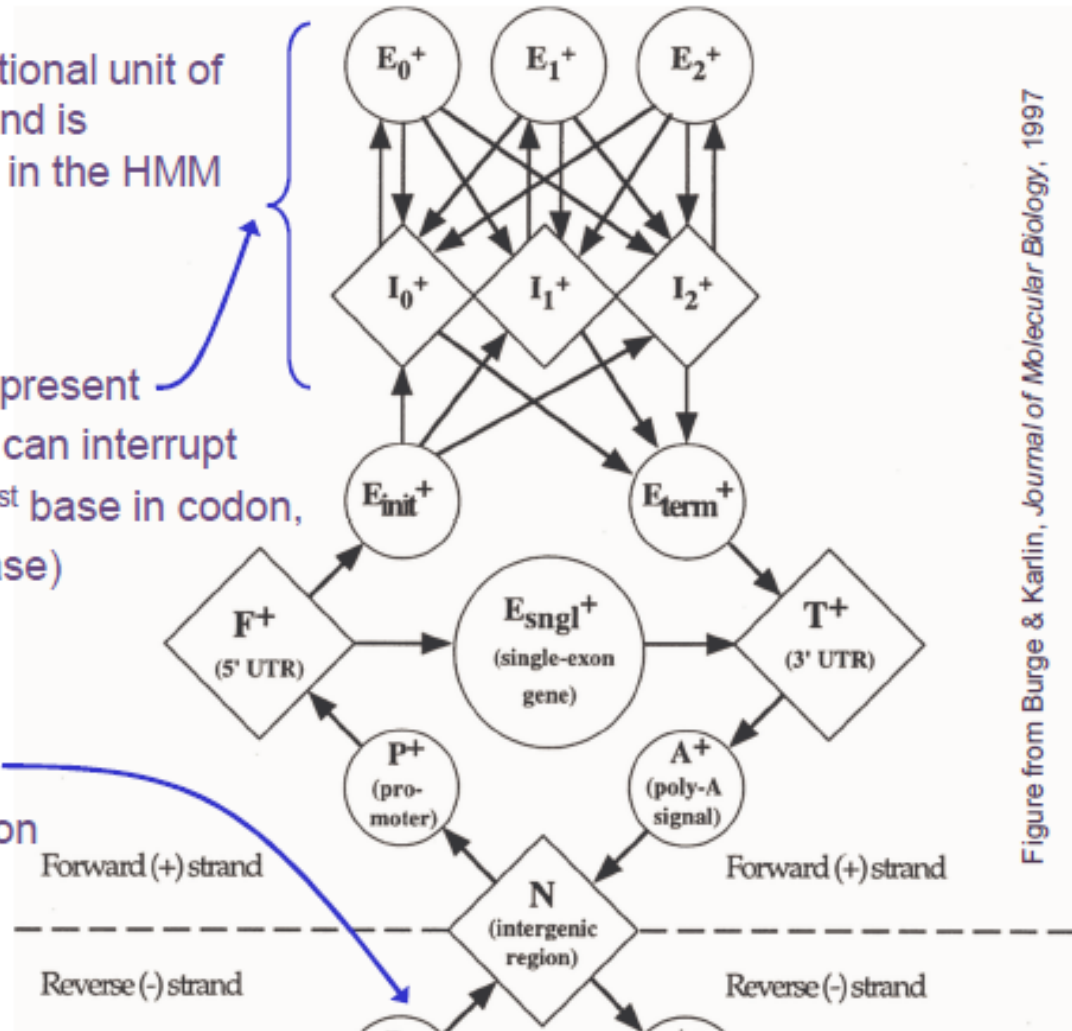


Figure from Burge & Karlin, *Journal of Molecular Biology*, 1997

# GENSCAN HMM and its submodels

## sequence

exons  
introns, intergenic regions  
poly-A, translation initiation, promoter  
splice junctions

## feature model

5th order inhomogenous  
5th order homogenous  
0th order, fixed-length  
tree-structured variable memory

### ■ Exon submodel

- for each “word”, consider its position with respect to the reading frame,
- use an inhomogeneous Markov chain.

reading frame

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

G C T A C G

G is in 3<sup>rd</sup> codon position

C T A C G G

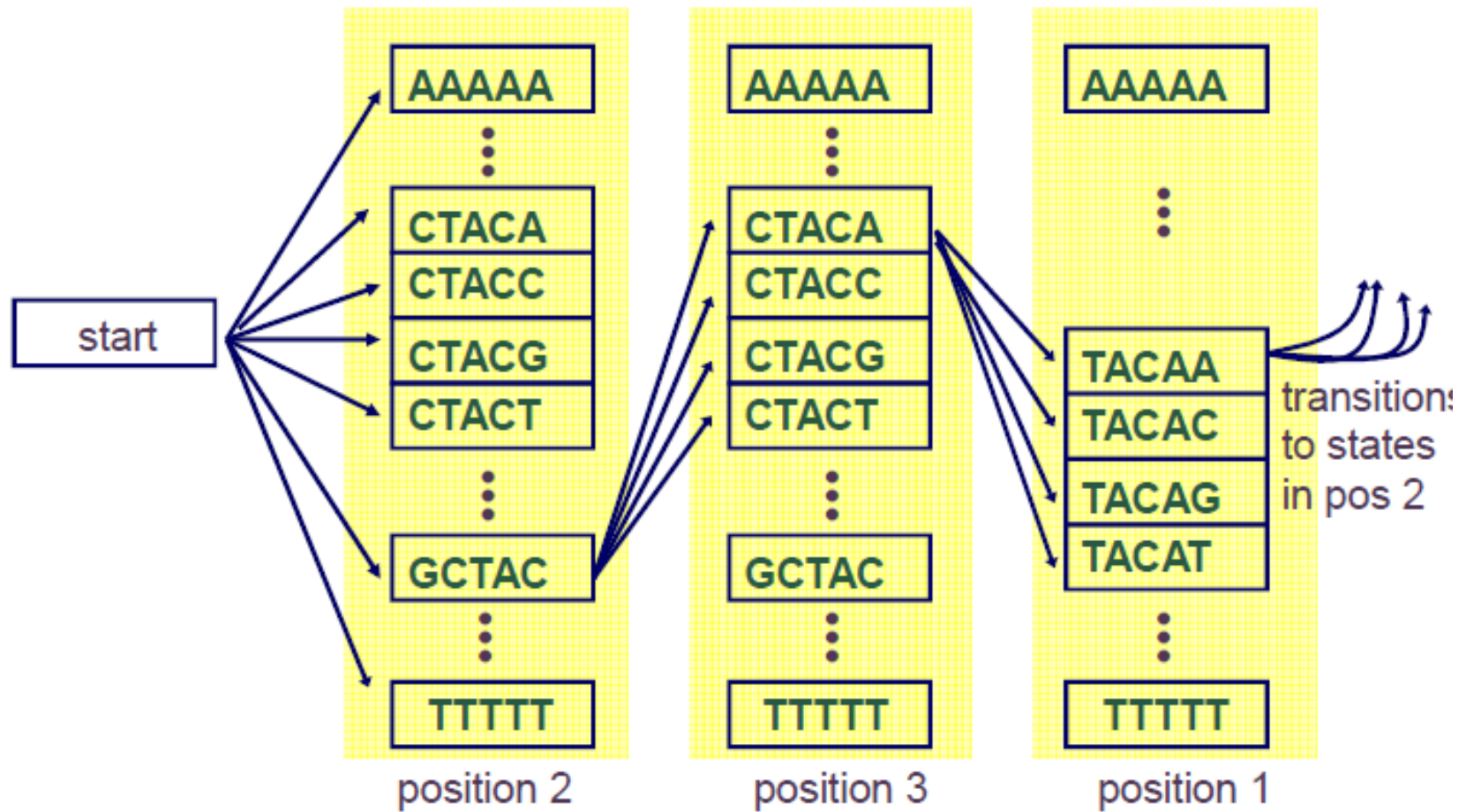
G is in 1<sup>st</sup> position

T A C G G A

A is in 2<sup>nd</sup> position



# A fifth-order inhomogeneous Markov chain



Marc Craven, BMI/CS 576, [www.biostat.wisc.edu/bmi576](http://www.biostat.wisc.edu/bmi576).



## Other issues in Markov models

---

- There are many interesting variants and extensions of the models/algorithms we considered here
  - separating length/composition distributions with semi-Markov models,
  - modeling multiple sequences with pair HMMs,
  - learning the structure of HMMs,
  - going up the Chomsky hierarchy: stochastic context free grammars,
  - discriminative learning algorithms (e.g. as in conditional random fields).