

- The HMMs shown before model an emission sequence based on the **region** they reflect (e.g. a genomic region like GpC islands)
- Often we want to model classes of proteins or domains more specifically.
- What information do we have to model **position-specific emission probabilities**?
- **Profiles again!** ... extracted from an alignment of relevant sequences.
- The resulting linear HMMs are called **profile HMMs**.

Profile HMM

The profile P of length n based on alphabet Y is the matrix

$$[e_i(y) : i = 1, \dots, n \text{ and } y \in Y]$$

of probabilities.

$e_i(y)$ is the probability that y occurs on position i in the sequence.

SSAPLRTVKEVQF	S	$e_1(S)$	$e_2(S)$	$e_3(S)$	$e_4(S)$	$e_5(S)$	$e_6(S)$...
SACPLRTIKRVQF	A	$e_1(A)$	$e_2(A)$	$e_3(A)$	$e_4(A)$	$e_5(A)$...	
EAKVKKQIKSIQF	K	$e_1(K)$	$e_2(K)$	$e_3(K)$	$e_4(K)$...		
SPAEVSKVRVQF	F	$e_1(F)$	$e_2(F)$	$e_3(F)$...			
	V	$e_1(V)$	$e_2(V)$...				
	...	$e_1(\cdot)$...					

The approach is to build a HMM with a repetitive structure of states but different emission probabilities in each position.

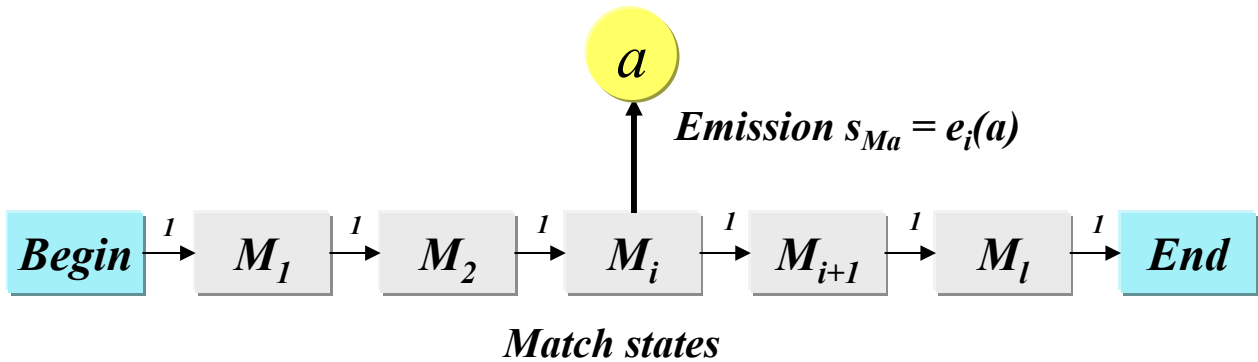
Profile HMM

```

SSAPLRTVKEVQF
SACPLRTIKRVQF
EAKVKKQIKSIQF
SPAEVSKVRVVQF
    
```

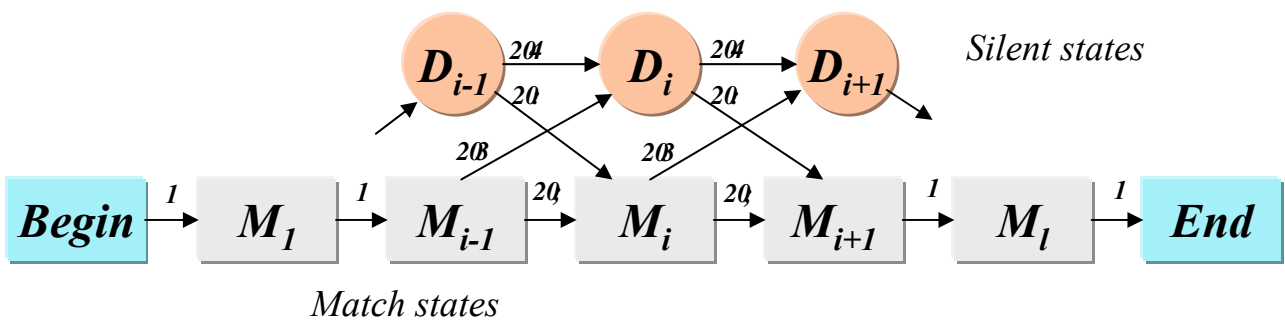
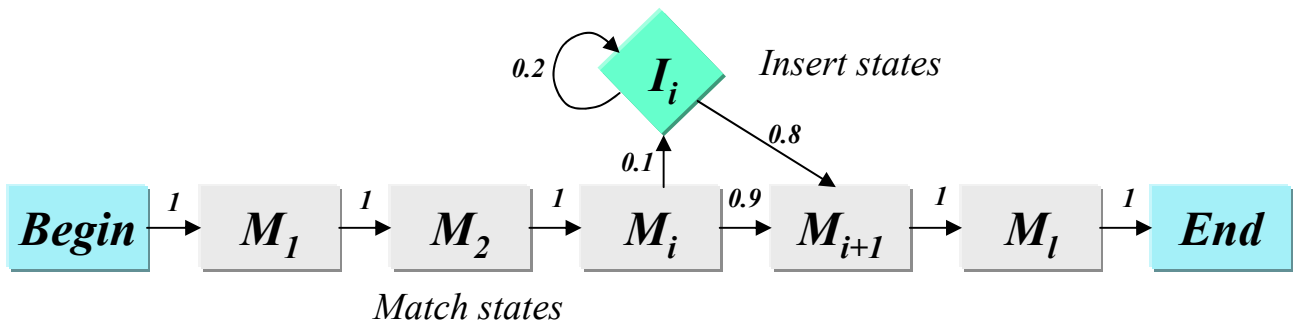
```

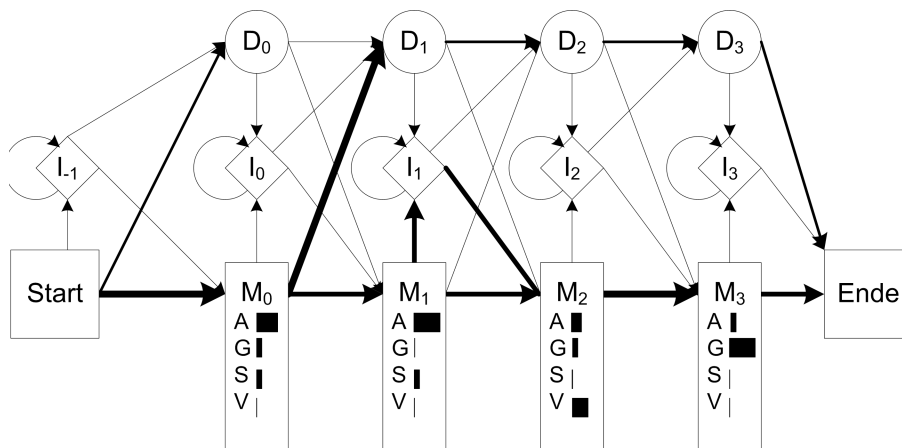
VGQ-Q---YSSAPLRTVKEVQF
HGGGPPSGDSACPLRTIKRVQF
F-----EAKVKKQIKSIQF
DTRFP---FSPAEVSKVRVVQF
    
```



What about gaps? - Insertion/Deletion

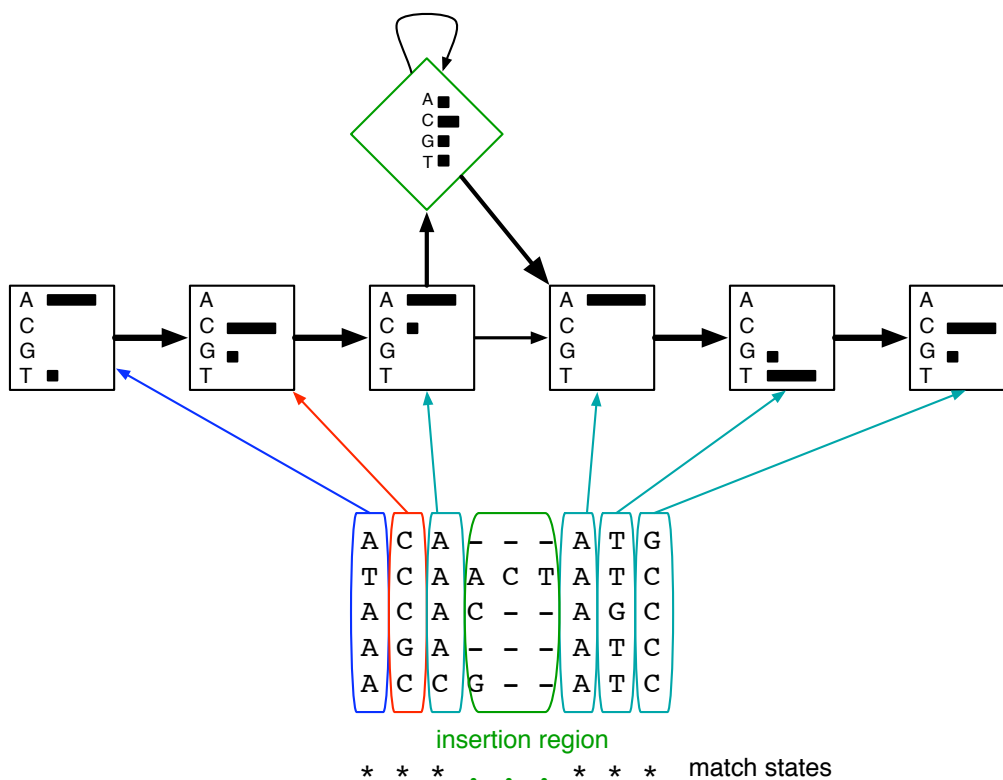
Profile HMM

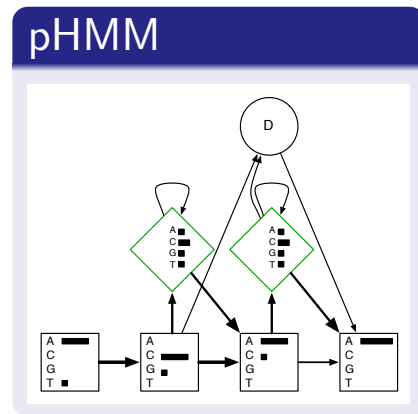
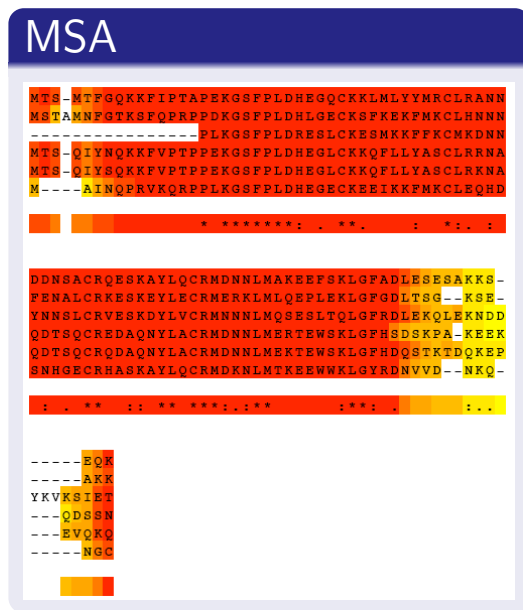
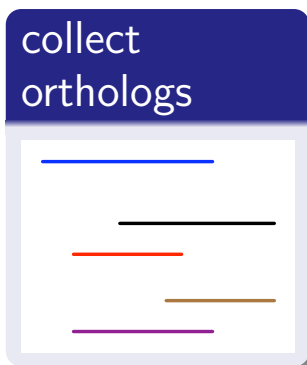




- How would you choose the number of states in the model?
- What states should be chosen?
- How are the model parameters chosen?

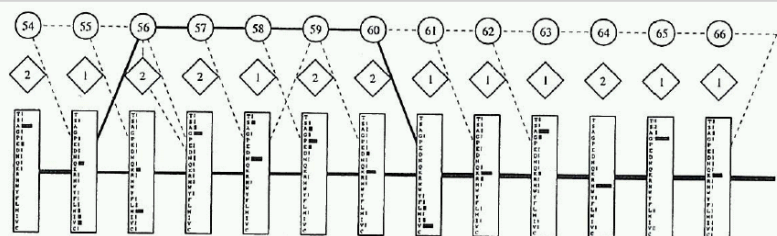
pHMM from aligned sequences





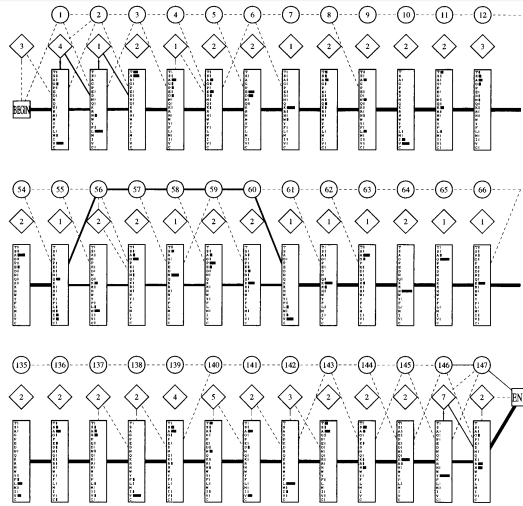
Is a new protein P also part of the group?

(p)HMMs - training



- Usually we do not know the **best structure/probabilities** for an HMM.
- We **cannot evaluate all** different possible paths through the HMM separately, neither for training nor for evaluation.
- Efficient algorithms exist to **train pHMMs with sequence alignments** (Baum-Welch algorithm).
- The training process changes the transition probabilities and, thus, leave a trace of the sequence family.
- Also the **structure of the pHMM** can be changed during training (States not used by at least half of the training set are merged with the insertion state; insertions present in more than half the training set are made a new match state.)
- Unfortunately, **large training sets** (> 20 – 50) are necessary to train HMMs

(p)HMMs - applications



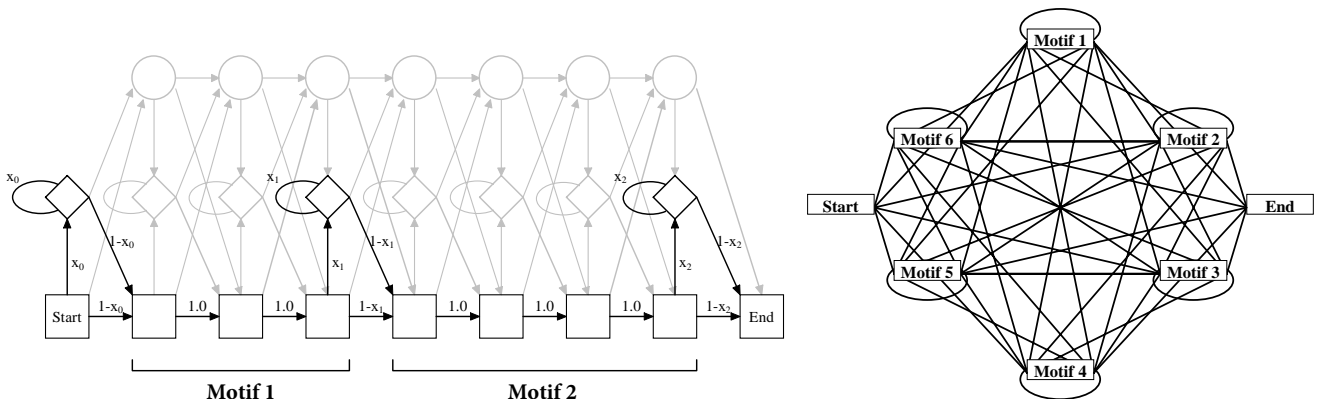
An HMM for globin sequences (Krogh et al. 1994)

The main applications of pHMMs in Bioinformatics are certainly

- to search in databases for **relatives of protein families** with pHMMs generated from alignments of sequences from the respective protein-family
- to detect and annotate **functional domains** with pHMMs generated from alignments of their **sequence motifs**

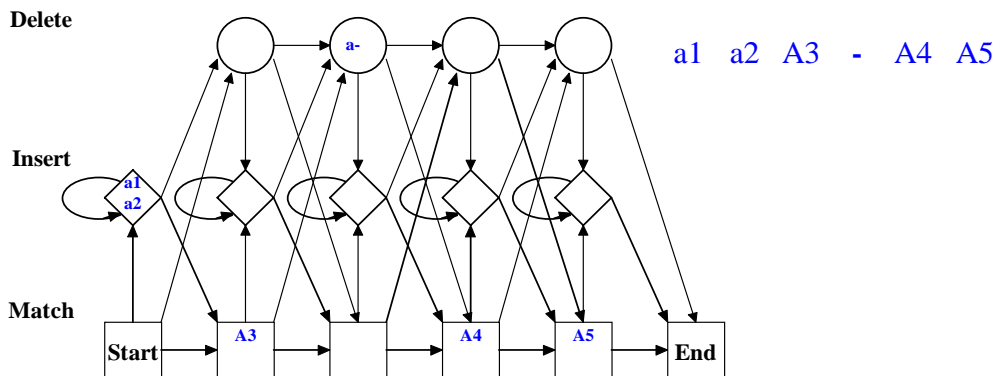
(p)HMMs - application for multi-domain proteins

- One can search for proteins containing several domains
- by joining pHMMs to one linear HMM if the domains occur in a certain order
- but one can also join several domain allowing for unspecific orders



(p)HMMs - application sequence alignment

- Furthermore one can align sequences using HMMs
- by aligning the match states of the Viterbi path.
- E.g. aligning sequences $A_1A_2A_3A_4A_5$ and $B_1B_2B_3B_4B_5$.

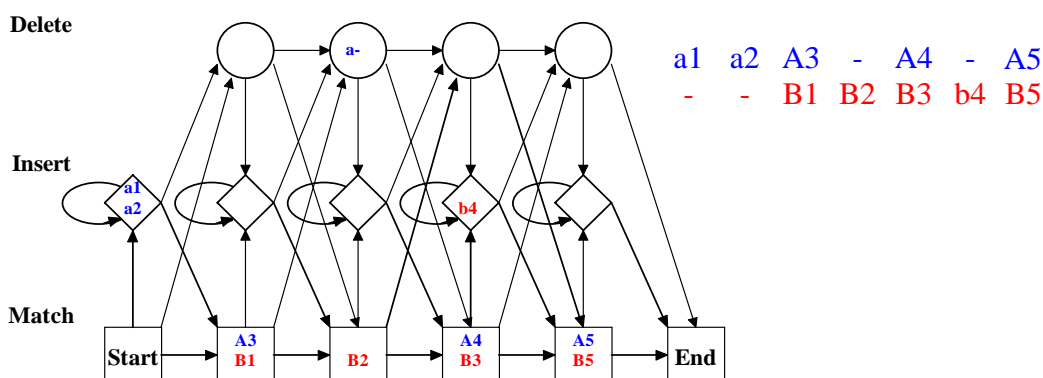


Please note,

- characters in insert/delete states have been marked by lower case letters in this example for distinction.
- a- in the deletion state is not really a character, but it is needed to avoid match state M2.
- characters mapped to the same insert states would be put in separate columns in the alignment.

(p)HMMs - application sequence alignment

- Furthermore one can align sequences using HMMs
- by aligning the match states of the Viterbi path.
- E.g. aligning sequences $A_1A_2A_3A_4A_5$ and $B_1B_2B_3B_4B_5$.



Please note,

- characters in insert/delete states have been marked by lower case letters in this example for distinction.
- a- in the deletion state is not really a character, but it is needed to avoid match state M2.
- characters mapped to the same insert states would be put in separate columns in the alignment.

Family: *Pkinase* (PF00069)

1412 architectures 51174 sequences 21 interactions 3376 species 1323 structures

Summary

- Domain organisation
- Alignments
- HMM logo
- Trees
- Curation & models
- Species
- Interactions
- Structures

Summary

Protein kinase domain [Add annotation](#)

No Pfam abstract.

Literature references

1. Hanks SK, Quinn AM; , Methods Enzymol 1991;200:38-62.: Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members. [PUBMED:1956325](#)
2. Hanks SK, Hunter T; , FASEB J 1995;9:576-596.: Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. [PUBMED:7768349](#)
3. Hunter T, Plowman GD; , Trends Biochem Sci 1997;22:18-22.: The protein kinases of budding yeast: six score and more. [PUBMED:9020587](#)

InterPro entry [IPR017442](#)



```

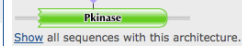
CK1_MOUSE/150-476  Y E I V D T L G E G A F C K V V E C I D H K V G C G . . . . . Y V A V I V K N V . . . . . D R Y C E A A Q S .
CKX21_CHICK/39-324  Y Q L V R K L G R G R T S E V . F E A I N I T N N E . . . . . K V V Y R I L K P V . . . . . K K K K I K .
MKM4_HUMAN/20-312  F V N F Q P L G E G V N G L V . L S A V D S R A C C . . . . . K V A V K K I A L S . . . . . D A R S M K H . A I L
FK1_CANAL/69-371  Y Q I L E I V G E G A Y G I V . C S A I H R P S O Q . . . . . K V A I K K I E P F . . . . . E R S M I C L E T L L .
GSK3A_RAT/115-443  Y T D I R V I G G E F G S V . Y S A I L A E T T E . . . . . L V A I K K I L O D . . . . . K R F R N K .
NAK_RAT/4-284      Y T T M R Q L G D G T Y G S V . L M G K S N E S G E . . . . . L V A I K K M K R K . . . . . F Y S W G E G M N L L .
CDKL1_HUMAN/4-287  Y E K I G K I G E G S Y G V V . F K C R N R D T G Q . . . . . L V A I K K F L E S . . . . . E D D P V I K K I A L L .
CTK1_YEAST/183-469  Y L R I M O V G E G T Y G V V . Y R A R N T N T E K . . . . . L V A L K K L L O . . . . . G E R E G F P I T S I R .
BURL_YEAST/99-366  Y R E D E R L G O G T F G E V . Y K G I H L E T G R . . . . . Q V A M K K I I V S . . . . . V E K S L F P I T A Q R .
CDC21_MESA/1-284   G E N V E K I G E G T Y G V V . Y K A R D V N E K I . . . . . I T A L K K I L E . . . . . Q E K G V P S T A I R .
KIN28_YEAST/7-290  Y T R E K K V G E G T Y A V V . Y L G C H S T G I . . . . . K I A I K E I K T S . . . . . E F K D G L D M S A I R .
TTK_HUMAN/525-791  Y S I L K Q I G S G G S K V . F Q V L N E K Q T I . . . . . Y A I R Y V N L S . . . . . E A D N G T L D S Y R I N .
PKM1_HUMAN/129-393  Y Q V G P L L G S G F G S V . Y S G I R V S D N I . . . . . P V A I K V Y E R D . . . . . R I S D W G E . . . . .
KAR7_YEAST/1096-1354  F V S L Q K M G E G A Y G V V . N L C I H K K N R V . . . . . L V V I K M I F K E . . . . . A I L V D T V V D R K .
PKM1_MIXXA/59-320  F R L V R R L G R G G M G A V . Y L G E H V S I G S . . . . . L V A V K L H A H . . . . . L T M P F E L V Q D F H .
HRR25_YEAST/9-273  F I T G R I R G S G S F G D I . Y R G T N L I S G E . . . . . L V A I K L E S I R . . . . . S R H P Q L D M I .
RPL1_CERFU/1094-1292  Q I T Q S I G S G S S A T V . E K A V W L G T F . . . . . V A K K T I P Q . . . . . N E E I F R K .
AVR2A_HUMAN/152-479  Q L L E V K A R G R F G C V . W K A Q L L N E Y . . . . . V A V K I F P I Q . . . . . D K Q S W N E T I .
ACVR1_HUMAN/208-455  L L L E C V G R G R T G E V . W R G S W Q G E N . . . . . V A V K I F S S R . . . . . D E K S W F N E T I .
MSK9_HUMAN/144-403  L T L E E I I G I G G F G K V . Y R A F W I G D E . . . . . V A V K A A R H D P P E D I . . . . . S Q I L E R N V Y
    
```

Domain organisation

Below is a listing of the unique domain organisations or architectures in which this domain is found. [More...](#)

There are 33933 sequences with the following architecture: Pkinase

[PETK2_HUMAN](#) [Homo sapiens (Human)] Serine/threonine-protein kinase PPTAIRE-2 EC=2.7.11.22 (384 residues)



[Show all sequences with this architecture.](#)

There are 1484 sequences with the following architecture: Pkinase x 2

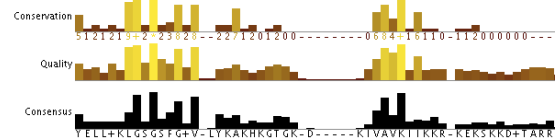
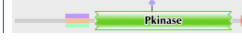
[CDC7_YEAST](#) [Saccharomyces cerevisiae (Baker's yeast)] Cell division control protein 7 EC=2.7.11.1 (507 residues)



[Show all sequences with this architecture.](#)

There are 537 sequences with the following architecture: Pkinase, Pkinase_C

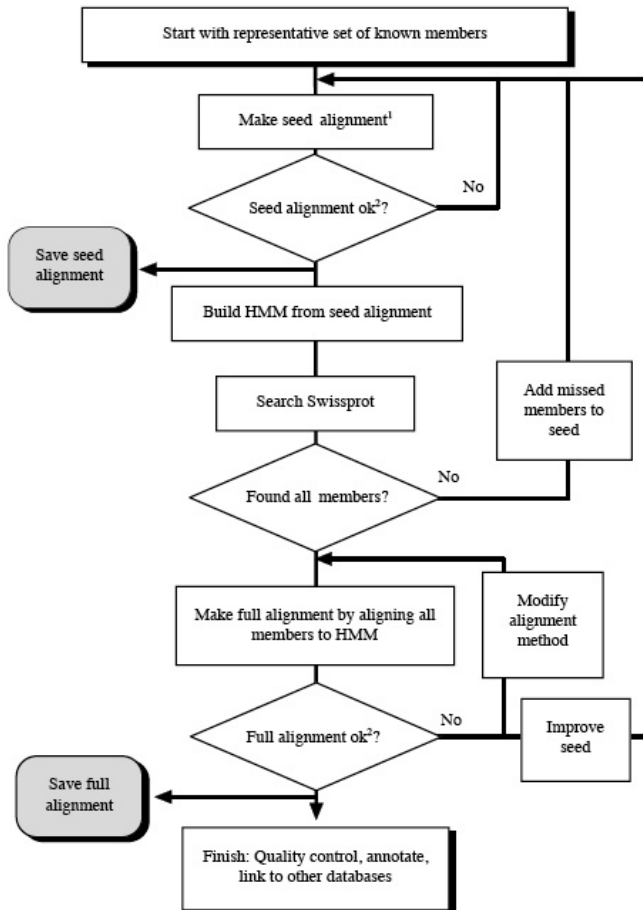
[DBF2_YEAST](#) [Saccharomyces cerevisiae (Baker's yeast)] Cell cycle protein kinase DBF2 EC=2.7.11.1 (572 residues)



Pfam Database (Sonnhammer et al. 1997)

- The Pfam database contains Protein (Domain) Families based on the data in the Uniprot protein databases.
- There are two sections: Pfam (or Pfam-A) and Pfam-B
- **Pfam/Pfam-A:**
 - contains a set of hand curated seed alignments containing data from different sources (Uniprot, Prosite, Prodom, structural alignments, BLAST results, Repeats found with Dotter, published alignments)
 - from the seed alignments (profile) HMMs are created
 - the HMMs are used to collect additional data from Uniprot
 - create a full alignment (and HMM)
- **Pfam-B:** (abandoned 2013)
 - utilizes an automated clustering of all sequences from Uniprot in the ADDA database (without the sequences already used in Pfam-A).

Pfam-A Generation (Sonnhammer et al. 1997)



If the results during the curated generation of alignments/HMMs, one can optimize in different ways:

- seed alignment construction
- HMM construction
- generation of full alignment

Pfam Database Entries

An Entry in the Pfam database consists of:

- Annotation/summary about the protein (domain) family,
- the full alignment,
- the seed alignment (Pfam-A only),
- the (profile) HMM (Pfam-A only),
- background information about curation and HMM creation etc. (Pfam-A only)

Pfam currently (Rel. 31.0, 03/2017) contains

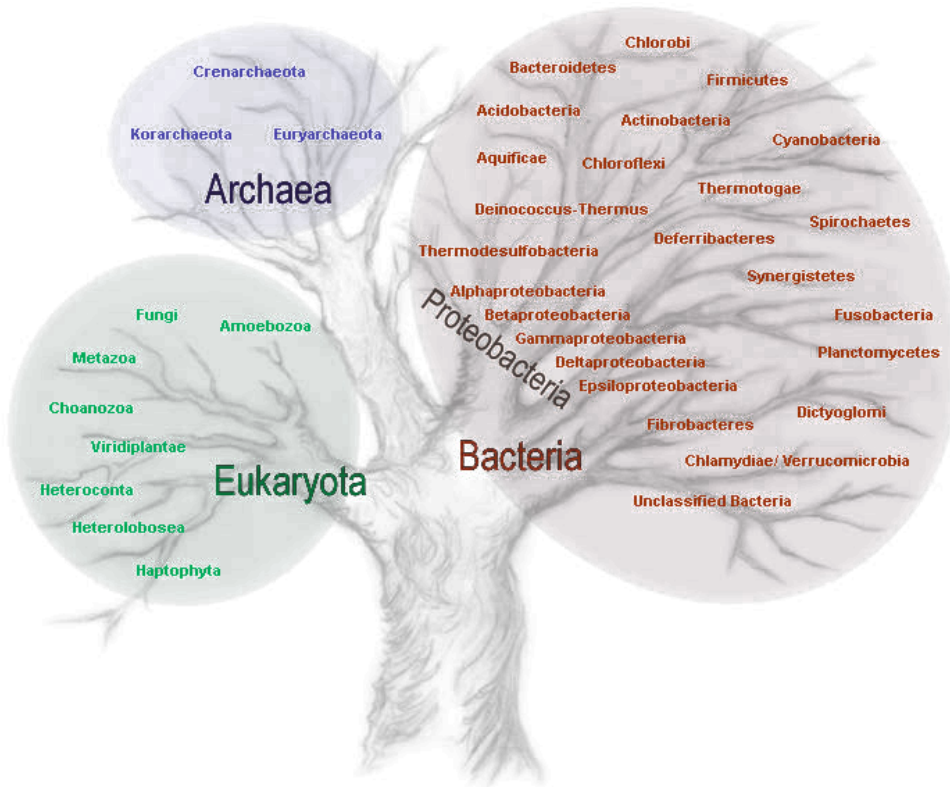
- **16712 Pfam-A** families based on *UniProtKB reference proteomes* (since Rel 29.0)
- last with both **Pfam-A (14831)** and **Pfam-B (544866)** was 27.0 (03/2013) based on SwissProt+SP-TrEMBL
- compared to **100 Pfam-A** and **11763 Pfam-B** families in Release 0.2, 01/1996 (based on SwissProt only).

There are several ways to search in/with the Pfam HMM database:

- search with a **query sequence** against all HMMs in Pfam – e.g., to **classify proteins** or their **domains**
- one can download **an HMM** and search in a set of sequences to find (distant) **homologs**
- search with the **whole set of HMMs** against a set of (unknown) sequences, e.g., to **annotate** and/or **find functional domains**.

Phylogeny Reconstruction

A more recent view of the Tree of Life



Source: genome.jgi-psf.org

Theodosius Dobzhansky: The Light of Evolution

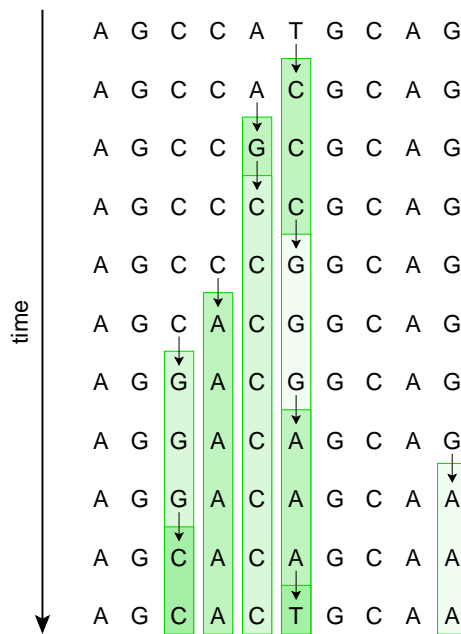


Theodosius Dobzhansky (1900-1975)

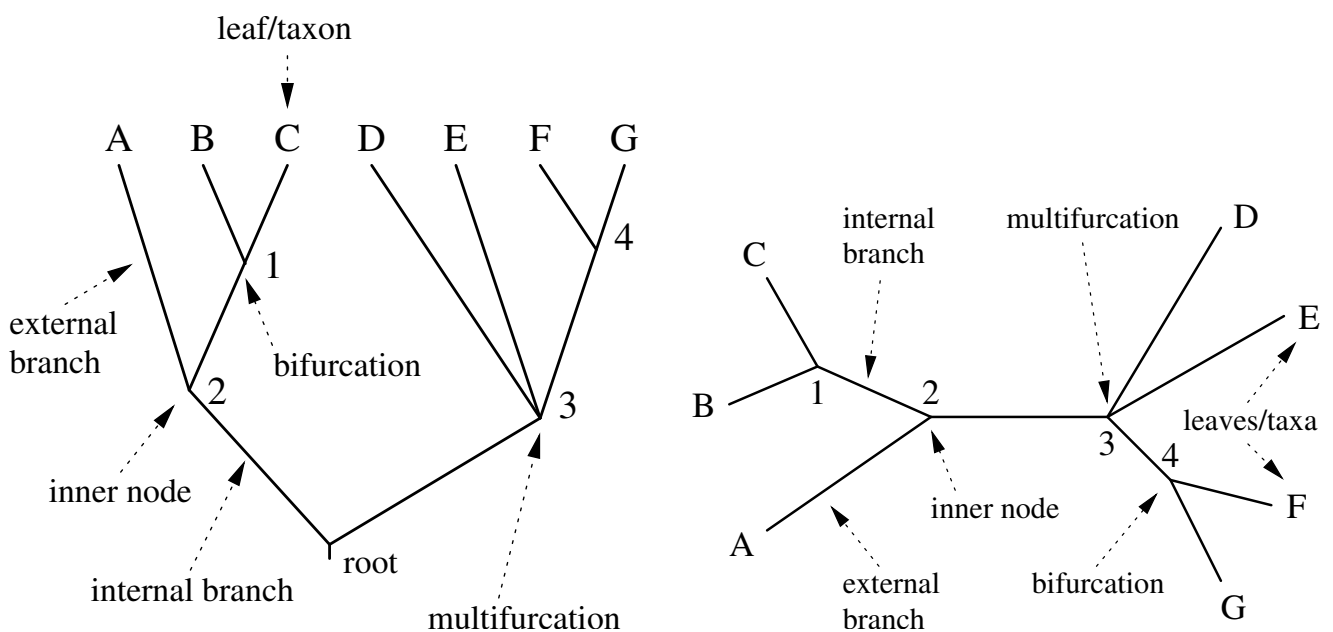
Nothing in Biology Makes Sense Except in the Light of Evolution.

Dobzhansky, 1973

Traces of Sequence Evolution



Some Notation



Note: branch = edge = split, external node = leaf = taxon = sequence are used interchangeably.

Main Types of Phylogenetic Methods

Data	Method	Evaluation Criterion
Characters (Alignment)	Maximum Parsimony	Parsimony
	Statistical Approaches: Likelihood, Bayesian	Evolutionary Models
Distances	Distance Methods	

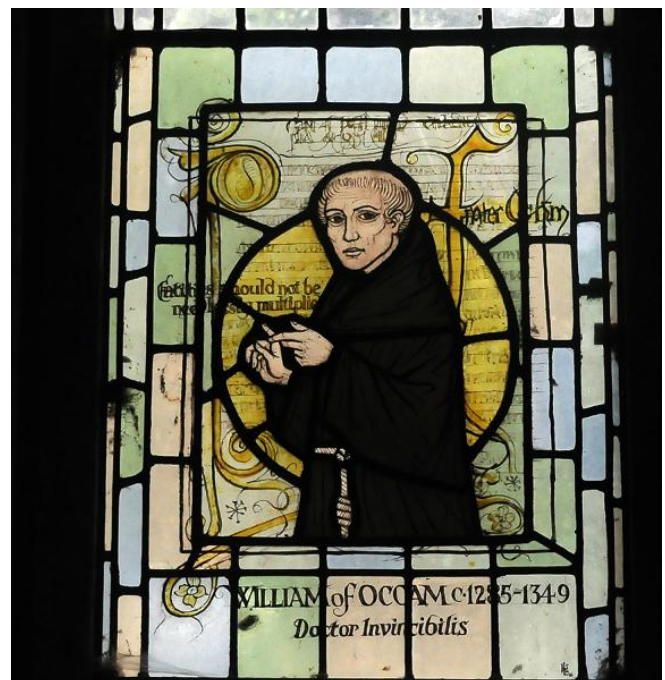
William of Ockham: The Law of Parsimony

Occam's Razor (law of parsimony) states:

Pluralitas non est ponenda sine necessitate.

Plurality should not be posited without necessity.

The principle gives precedence to simplicity; of two competing theories, the simplest explanation of an entity is to be preferred.

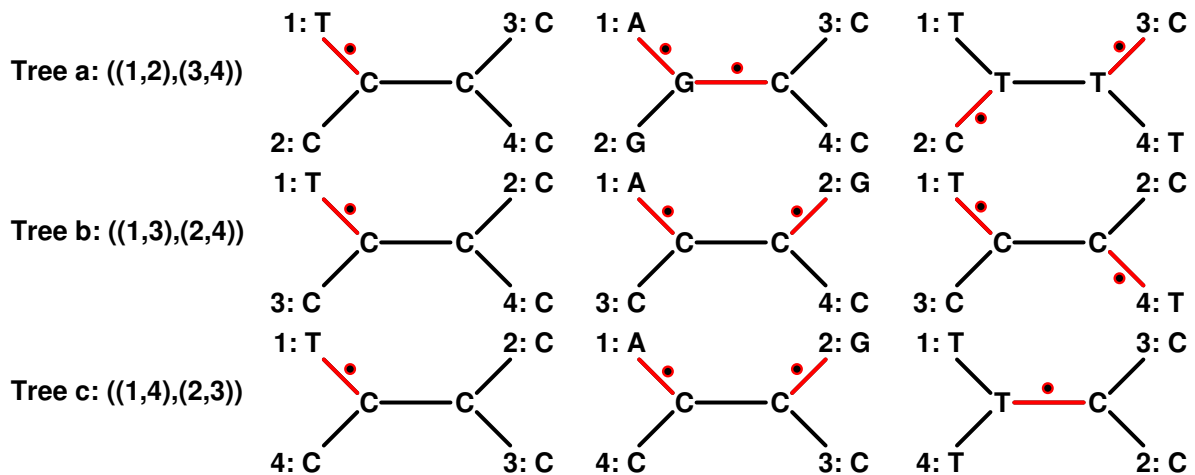


William of Ockham (1285-1347/49)

Maximum Parsimony (MP)

taxon	1	2	3	4	5	6	7	8	9
1:	T	G	A	A	C	T	G	T	T
2:	C	G	G	A	C	T	G	C	T
3:	C	G	C	A	C	T	G	C	T
4:	C	G	C	A	C	T	G	T	T

↑
↑
↑



Parsimony Informative Sites

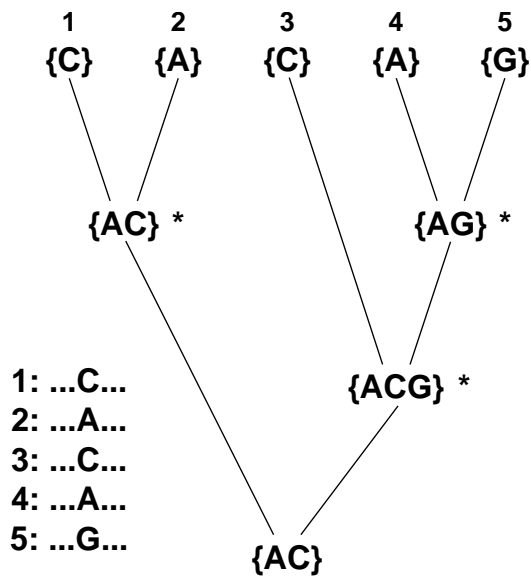
We have seen that not all variable columns are informative for the parsimony reconstruction.

(Parsimony) Informative Sites: 2-2-2-rule

To be an informative site for the parsimony principle the column has to contain **at least two different** character states, and **at least two of these** states have to occur **at least twice**.

Maximum Parsimony: Fitch's (1970) algorithm

What is the minimum of mutations required?



Note: We need 1 substitution per union in tree T (tree-length = substitutions needed).

- 1 Initialize state set S_k at each leaf k with the characters from the alignment.
- 2 Construct the state sets of all internal nodes in a post-order-traversal starting at the root.
- 3 Let k be the current node and i, j its decedents, then build the intersection of S_i and S_j :
 - If $S_i \cap S_j$ non-empty (*): set $S_k = S_i \cap S_j$,
 - if $S_i \cap S_j$ empty: set $S_k = S_i \cup S_j$ and increase the tree-length by 1.
- 4 Continue with the traversal until you have reconstructed the state set S_{root} of the root of T . If we have a sequence for the root, repeat Step 3 for its character and S_{root} .

How to find the Most Parsimonious Tree?

Ideally we would evaluate all trees and take the one(s) with the lowest tree-length.

However, there are too many trees. This problem affects almost every method that aims to find trees with optimal score.

So we need other strategies (which we will see later).

- Parsimony is often considered **model-free**. This is not entirely correct.
- One has no choice of a model, but nevertheless the algorithm assumes a **very simple model**.
- Parsimony assumes that **substitutions are rare** and that **back-mutations do not occur**.
- Although this was often true for morphological data, it is certainly not true for distantly related DNA sequences which only have four character states.

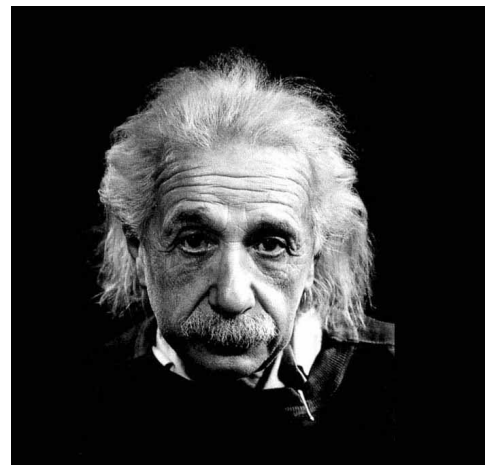
Simplicity with Caution – Einstein's Principle

Everything Should Be Made as Simple as Possible, But Not Simpler!

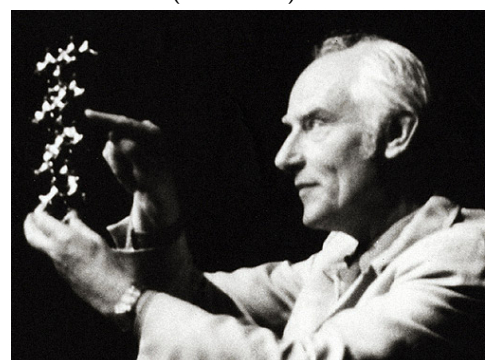
(attributed to) Albert Einstein

While Ockham's razor is a useful tool in the physical sciences, it can be a very dangerous implement in biology. It is thus very rash to use simplicity and elegance as a guide in biological research.

Francis Crick, 1988



Albert Einstein (1879-1955)



Francis Crick (1916-2004)