Position-specific Information in HMMs?

- 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

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

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

S	S	A	P	L	R	T	V	K	E	V	Q	F
S	A	C	P	L	R	T	I	K	R	V	Q	F
E	A	K	V	K	K	Q	I	K	S	I	Q	F
S	P	A	E	V	S	K	V	R	V	V	Q	F

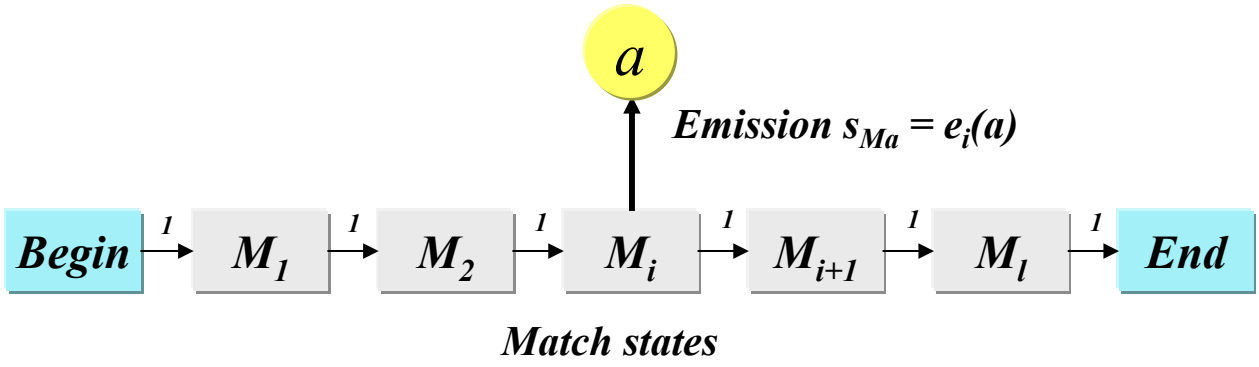
	1	2	3	4	5	6	...
S	$e_1(S)$	$e_2(S)$	$e_3(S)$	$e_4(S)$	$e_5(S)$	$e_6(S)$	...
A	$e_1(A)$	$e_2(A)$	$e_3(A)$	$e_4(A)$	$e_5(A)$	...	
K	$e_1(K)$	$e_2(K)$	$e_3(K)$	$e_4(K)$	...		
F	$e_1(F)$	$e_2(F)$	$e_3(F)$	...			
V	$e_1(V)$	$e_2(V)$	...				
...	$e_1(.)$	...					

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

Profile HMM

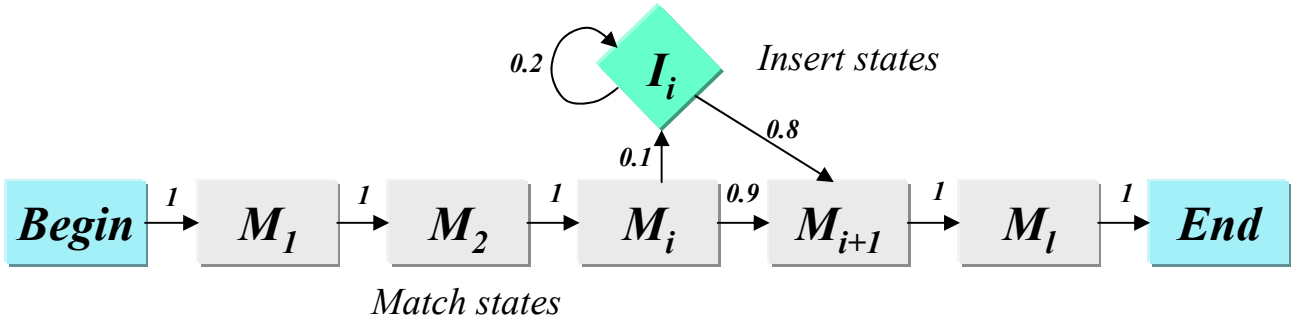
```
SSAPLRTVKEVQF
SACPLRTIKRVQF
EAKVKKQIKSIQF
SPAEVSKVRVVQF
```

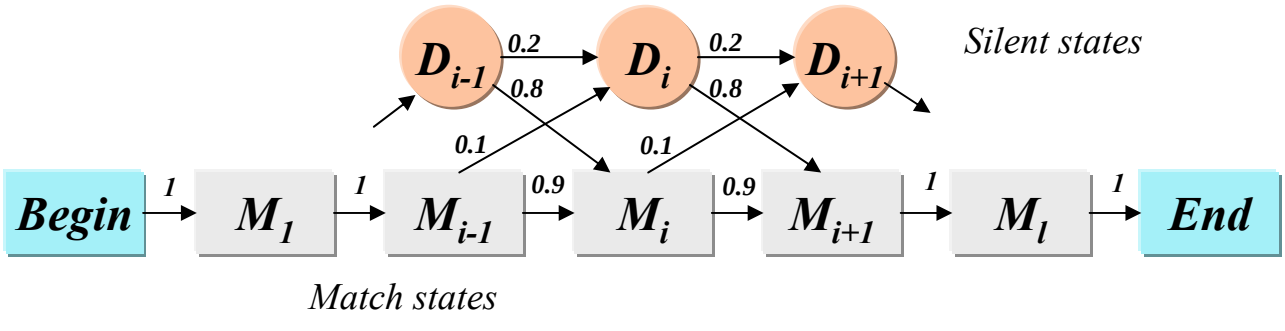
```
VGQ-Q---YSSAPLRTVKEVQF
HGGGPPSGDSACPLRTIKRVQF
F---------EAKVKKQIKSIQF
DTRFP---FSPAEVSKVRVVQF
```



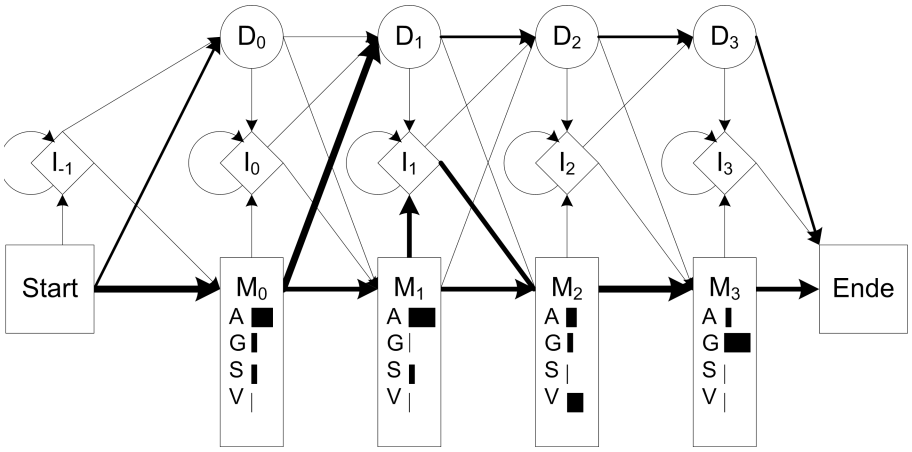
What about gaps? - Insertion/Deletion

Profile HMM



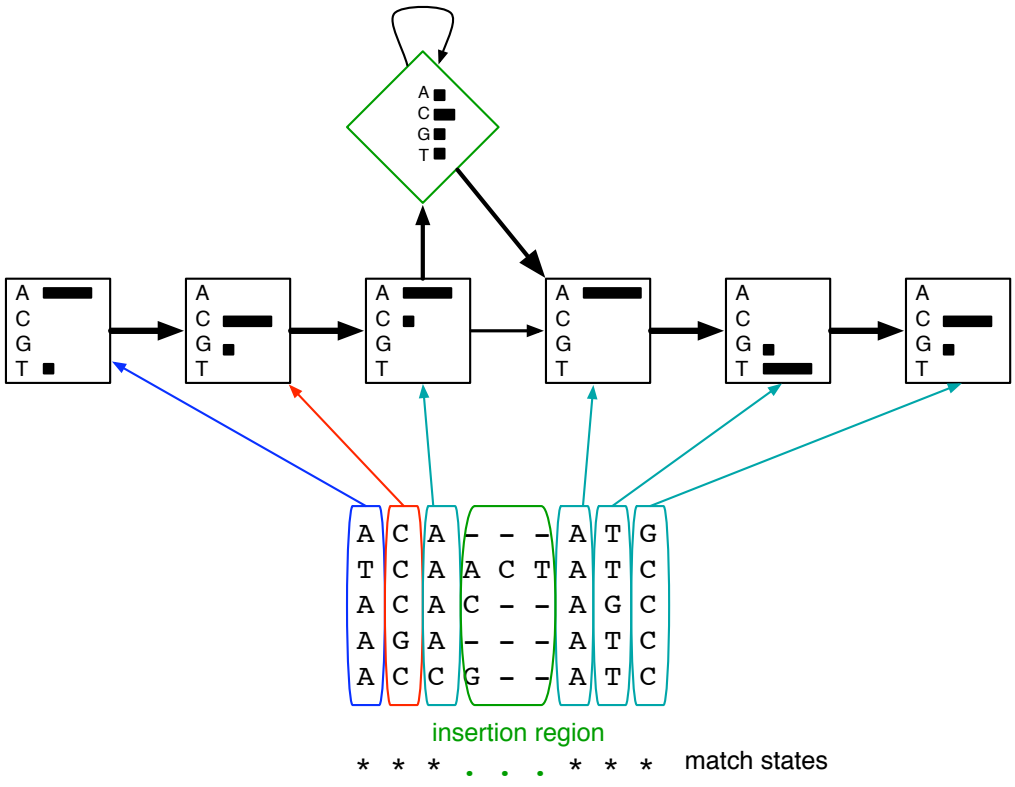


Profile HMM

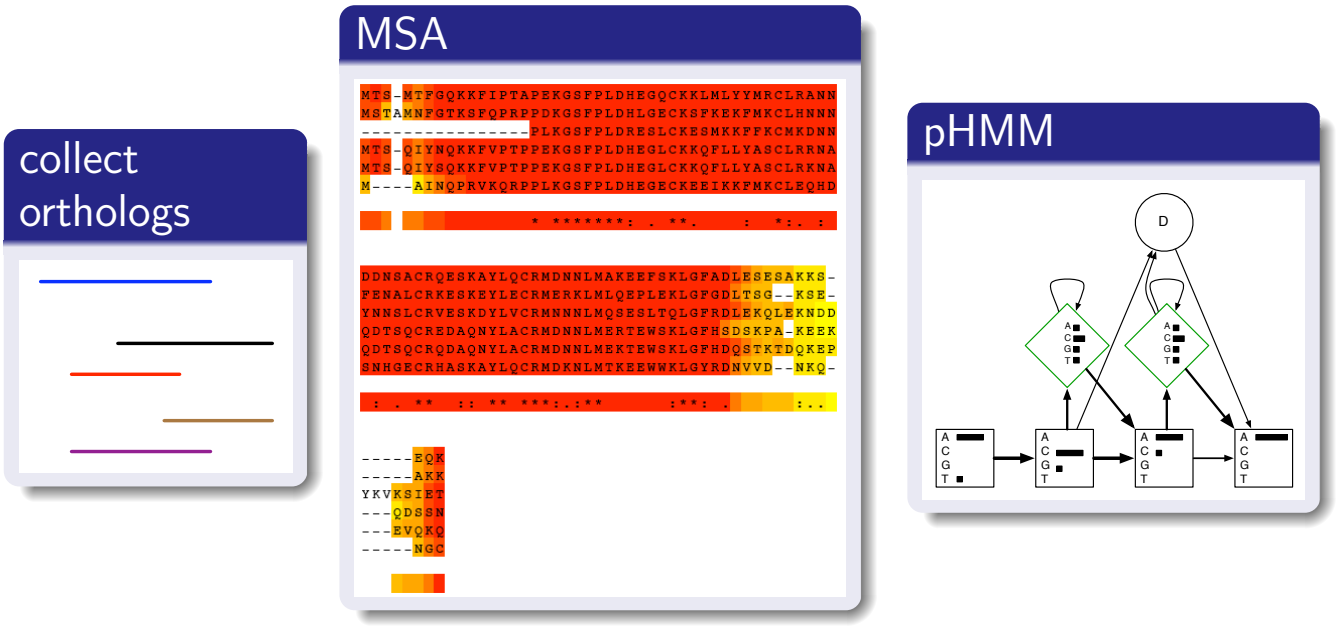


- How would you choose the number of states in the model?
- What states should be chosen?
- How are the model parameter chosen?

pHMM from aligned sequences

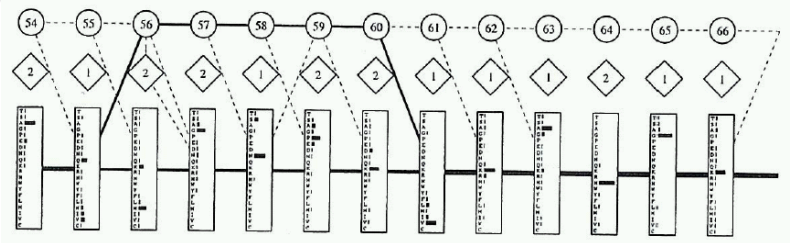


pHMM for related proteins



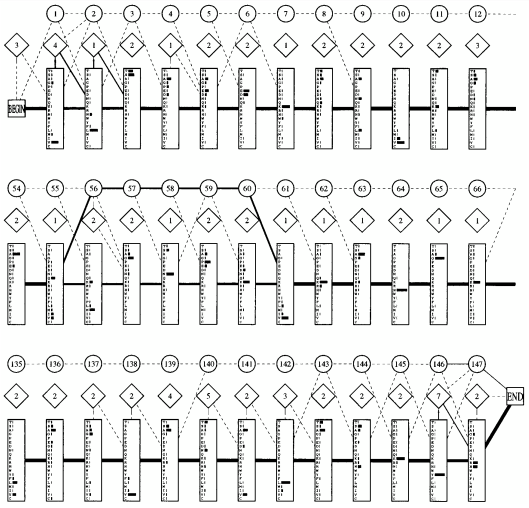
Is a new protein P also part of the group?

(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



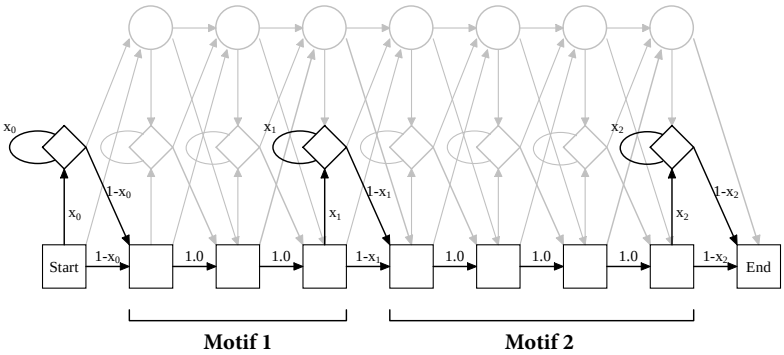
An HMM for globin sequences (Krogh et al. 1994)

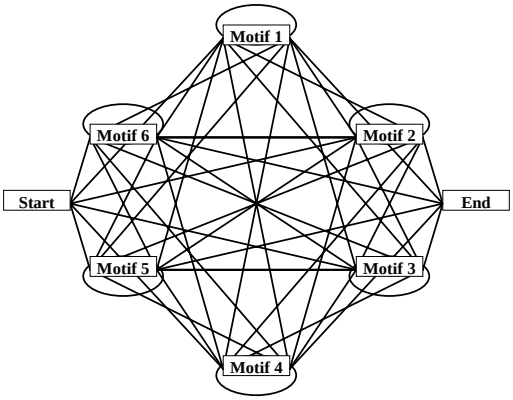
The main applications of pHMMs in Bioinformatics are certainly

- 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

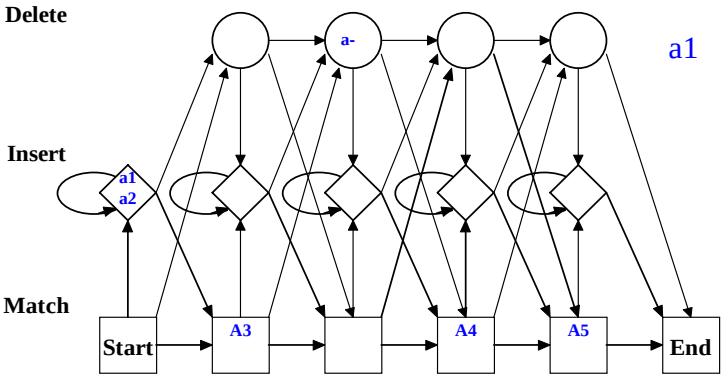
- One can search for proteins containing several domains
- by joining pHHMs to one linear HMM if the domains occur in a certain order
- but on can also join several domain allowing for unspecific orders





(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 $B_1B_2B_3B_4B_5$.



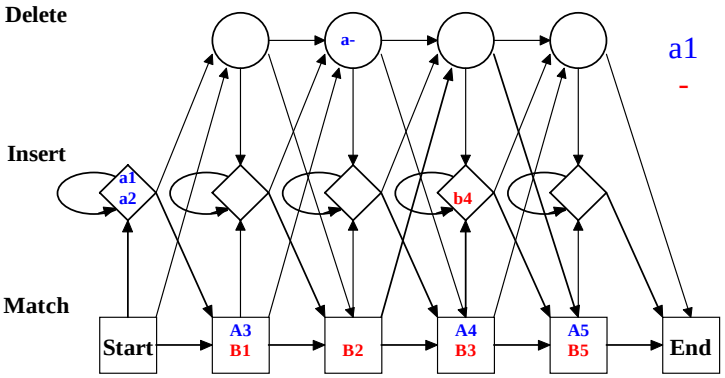
a1 a2 A3 - A4 A5

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 $B_1B_2B_3B_4B_5$.



a1 a2 A3 - A4 - A5
\- - B1 B2 B3 b4 B5

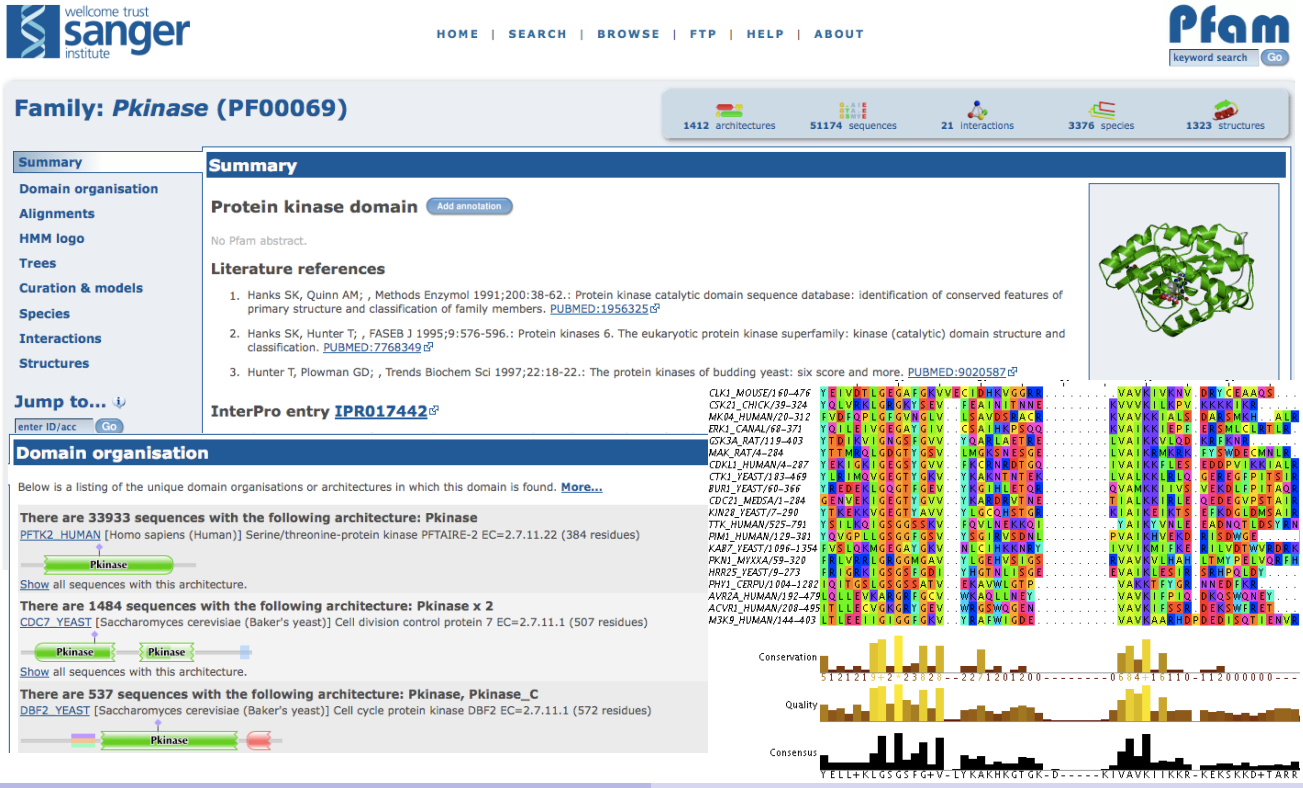
Please note,

- 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.

Pfam

Protein families database of alignments and pHMMs.

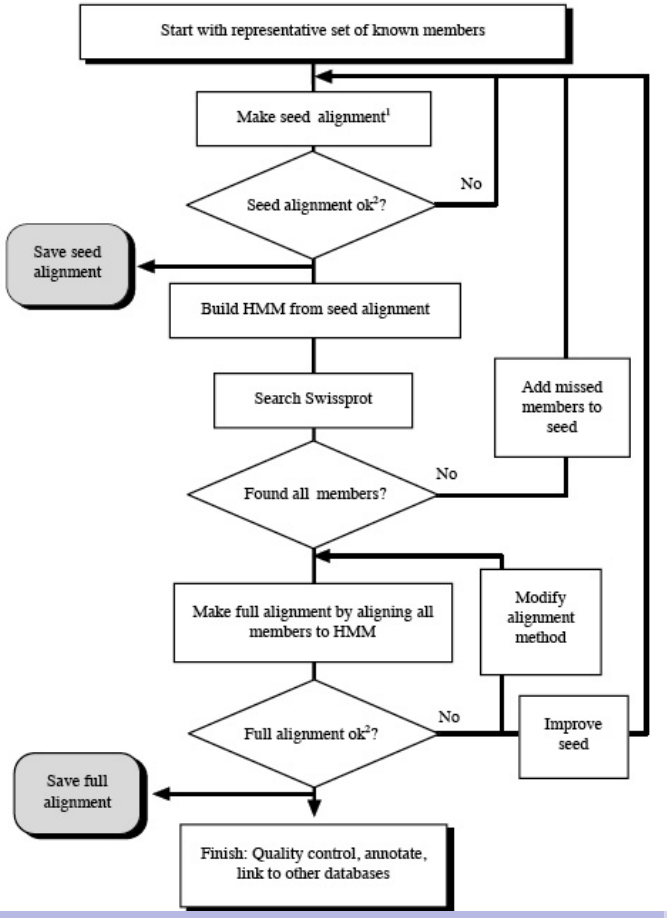
`http://pfam.sanger.ac.uk`



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).

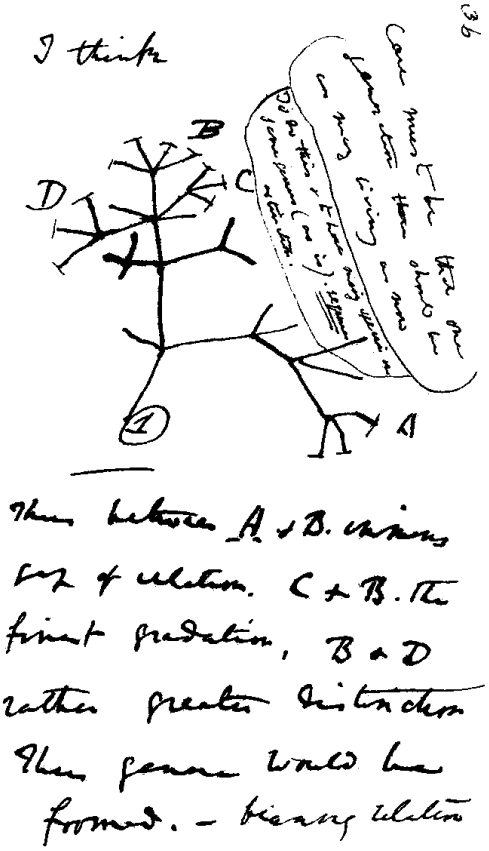
Pfam Database Searching (Sonnhammer et al. 1997)

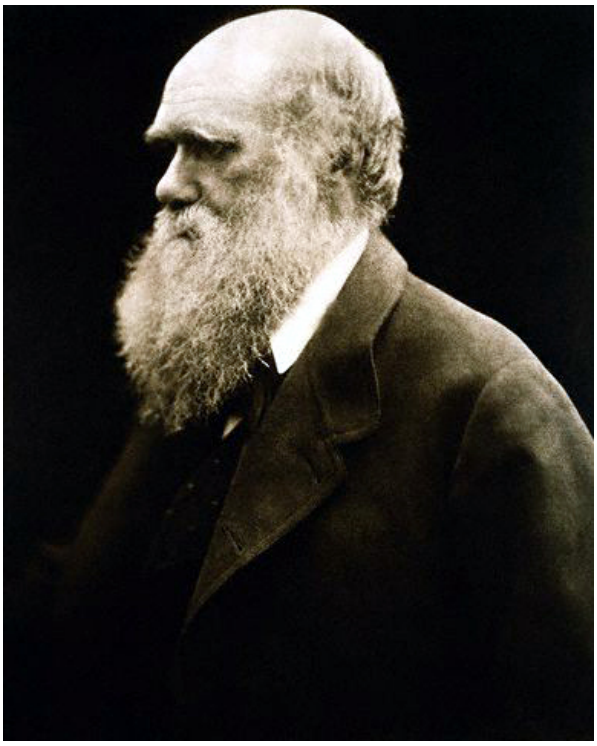
There are several ways to search in/with the Pfam HMM database:

- 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

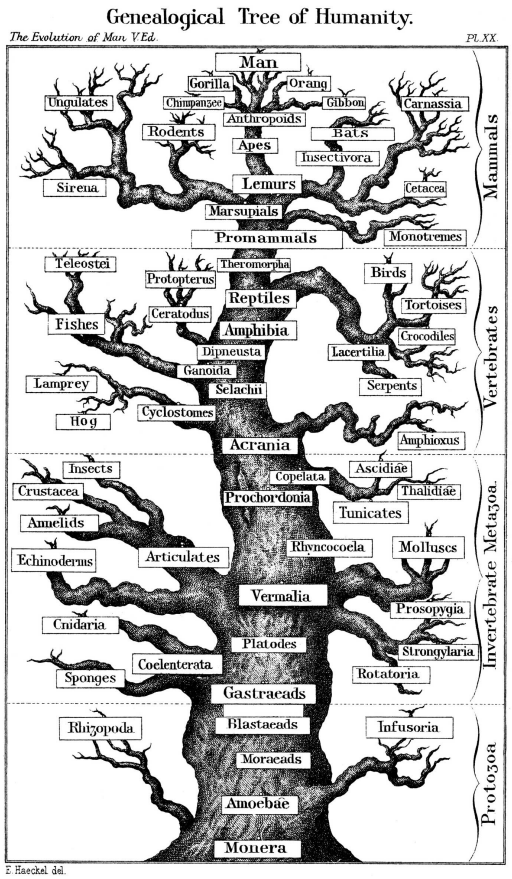
Charles Darwin: Evolutionary Relationships

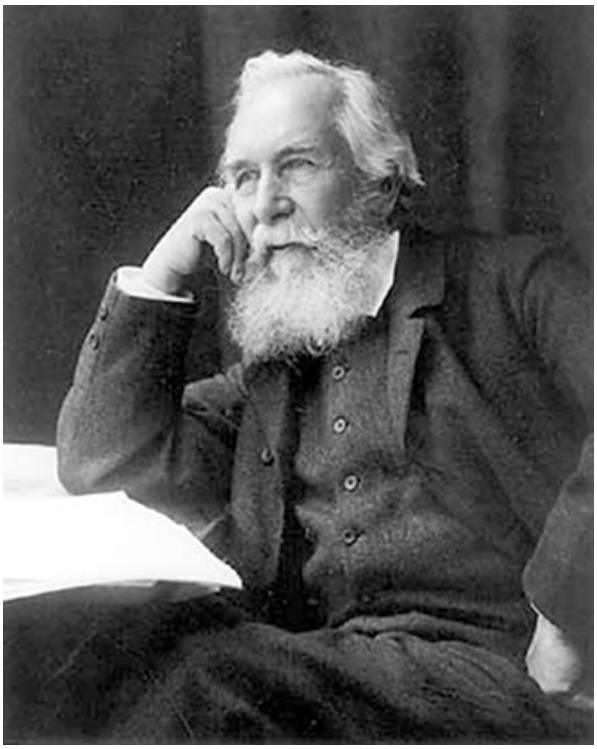




Charles Darwin (1809-1882)

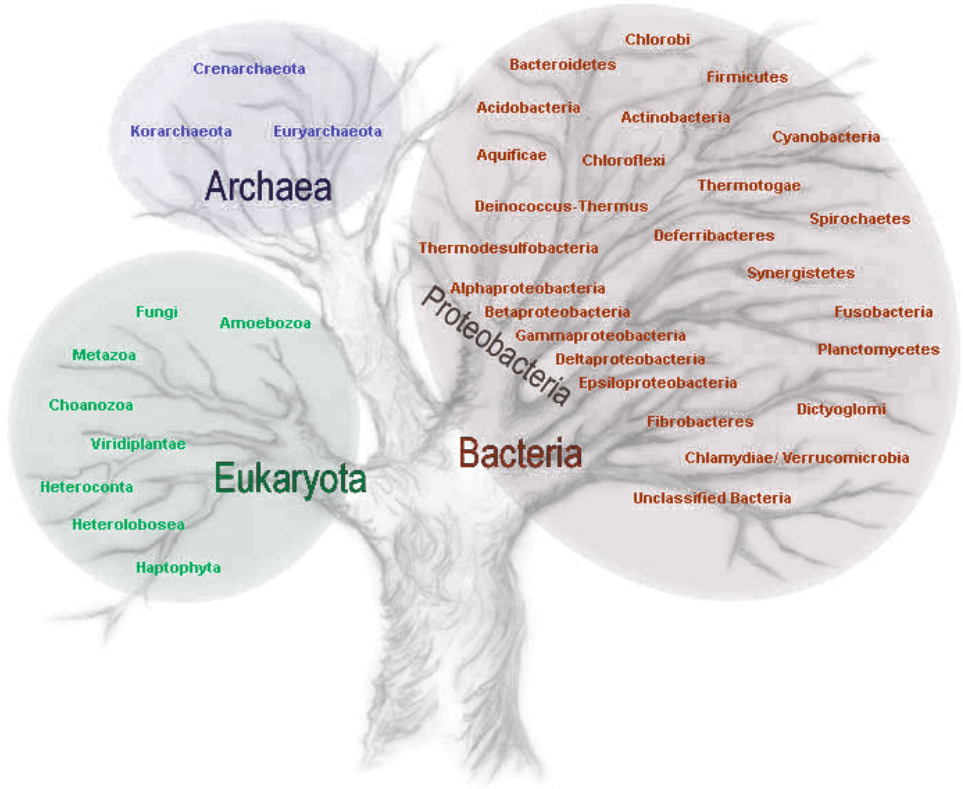
Ernst Haeckel: Evolutionary Trees





Ernst Haeckel (1834-1919)

A more recent view of the Tree of Life



Source: genome.jgi-psf.org

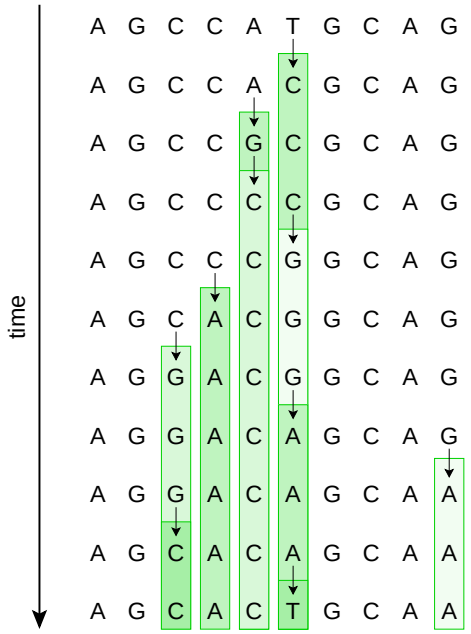
Theodosius Dobzhansky: The Light of Evolution

Theodosius Dobzhansky (1900-1975)

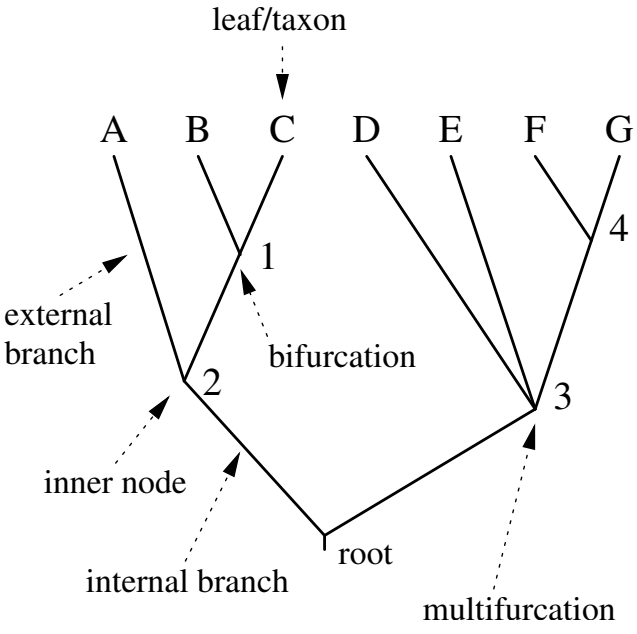
Nothing in Biology Makes Sense Except in the Light of Evolution.

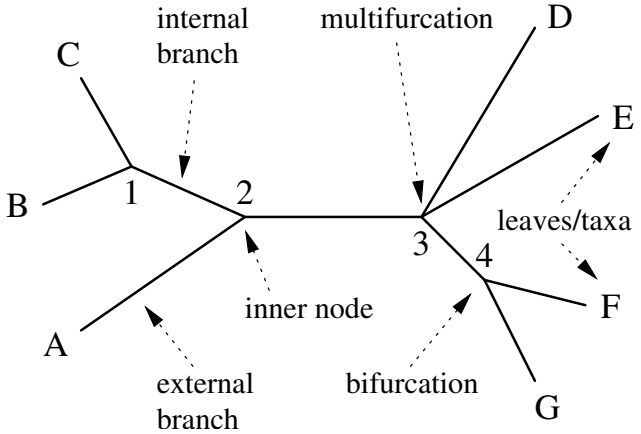
Dobzhansky, 1973

Traces of Sequence Evolution



Some Notation





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

Main Types of Phylogenetic Methods

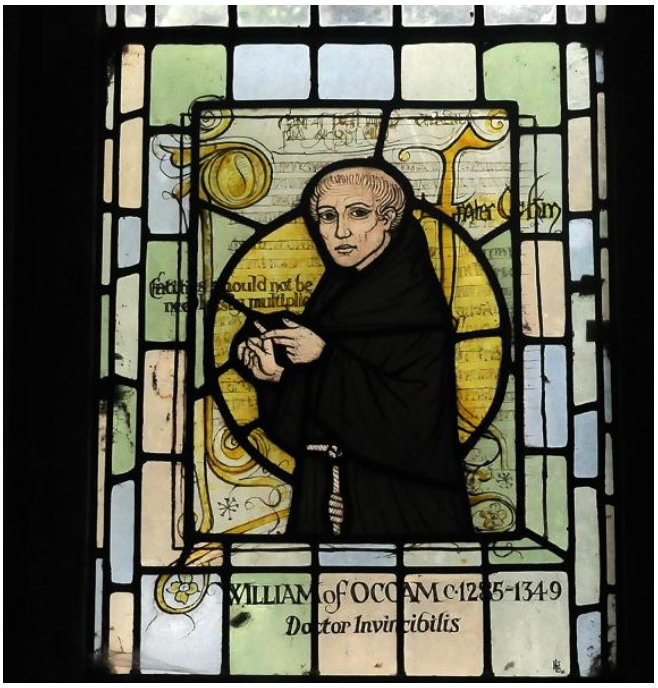
Data	Method	Evaluation Criterion
Characters (Alignment)	**Maximum Parsimony**	Parsimony
	Statistical Approaches: Likelihood, Bayesian	Evolutionary Models
Distances	**Distance Methods**	

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.*

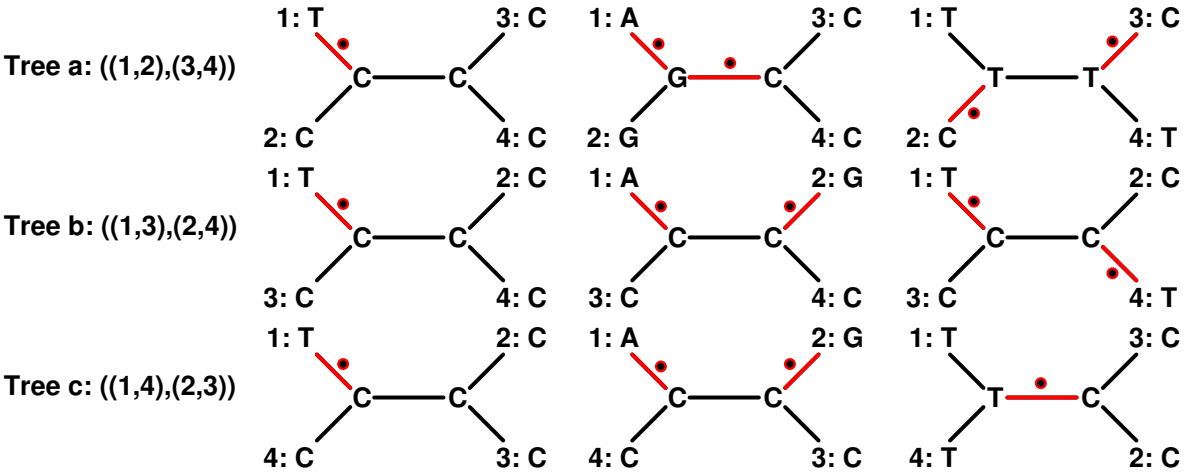
The principle gives precedence to simplicity; of two competing theories, the simplest explanation of an entity is to be preferred.



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

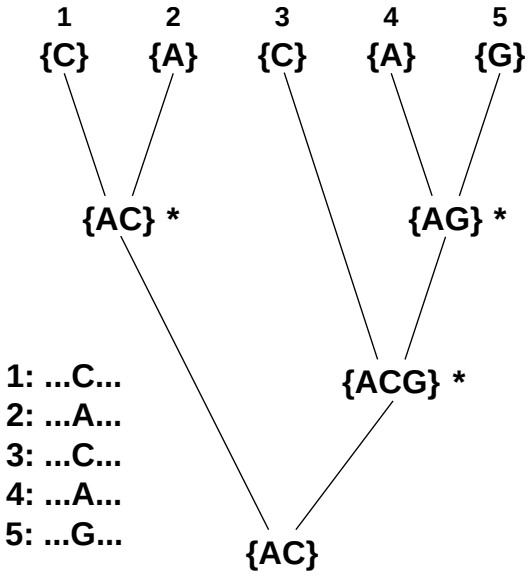
We have seen that not all variable columns are informative for the parsimony reconstruction.

(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 decendents, 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).

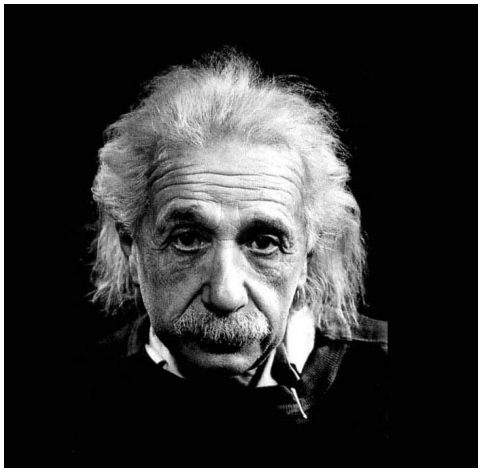
Problems with Parsimony

- 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!

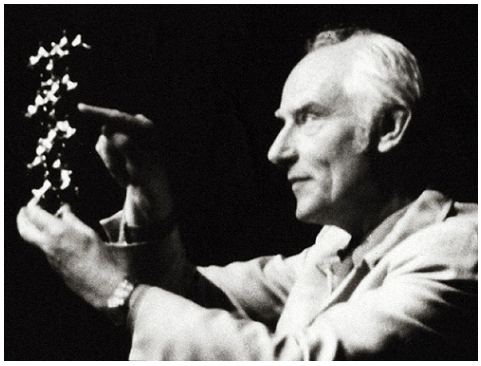
(attributed to) Albert Einstein



Albert Einstein (1879-1955)

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



Francis Crick (1916-2004)