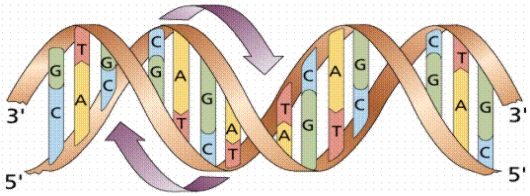
HIDDEN MARKOV MODELS IN COMPUTATIONAL BIOLOGY TO ANALYSIS A PROTEIN

BY

N.R.Jegannathbabu

Relationship Between DNA, RNA And Proteins



DNA

CCTGAGCCAACTATTGATGAA

transcription

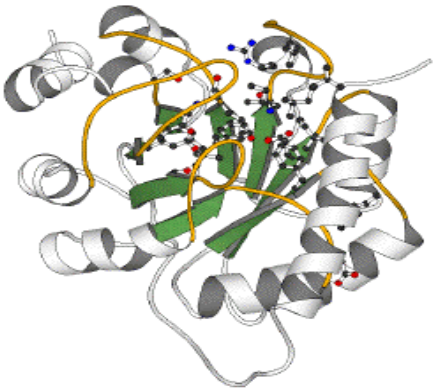
mRNA

CCUGAGCCAACUAUUGAUGAA

translation

Protein

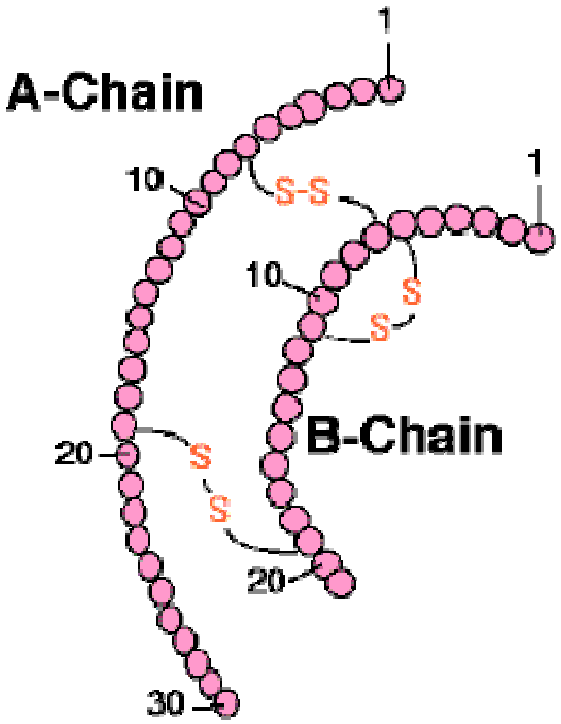
PEPTIDE



Protein Structure

Primary Structure of Proteins

The primary structure of peptides and proteins refers to the linear number and order of the amino acids present.

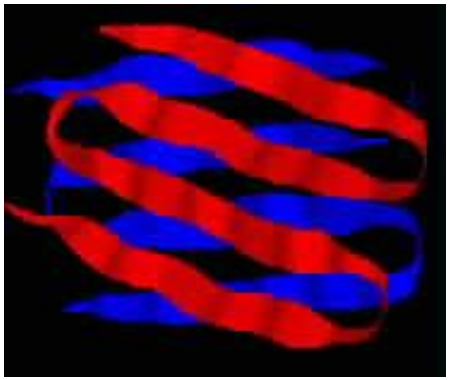


Protein Structure

Secondary Structure

Protein secondary structure refers to regular, repeated patters of folding of the protein backbone. How a protein folds is large by the primary sequence of amino acids

Alpha Helix

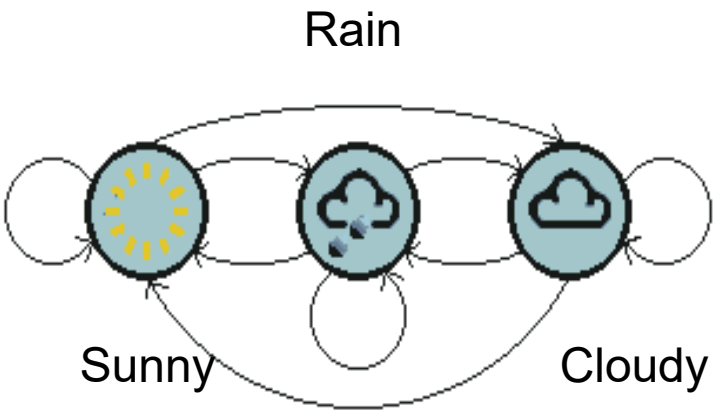


Beta Sheet

Multiple Alignment Process

- Process of aligning three or more sequences with each other
- Generalization of the algorithm to align two sequences
- Local multiple alignment uses Sum of pairs scoring scheme

Markov Chains



weather today

weather yesterday	Sun	Cloud	Rain
Sun	0.5	0.25	0.25
Cloud	0.375	0.125	0.375
Rain	0.125	0.625	0.375

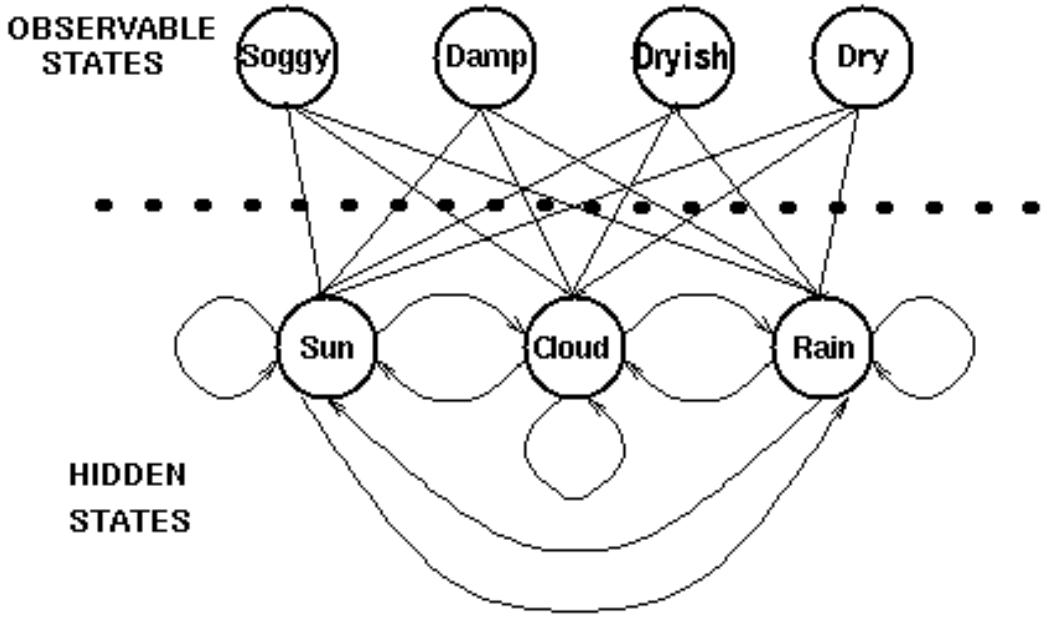
States : Three states - sunny, cloudy, rainy.

State transition matrix : The probability of the weather given the previous day's weather.

Sun	**Cloud**	**Rain**
1.0	0.0	0.0

Initial Distribution : Defining the probability of the system being in each of the states at time 0.

Hidden Markov Models



Hidden states : the (TRUE) states of a system that may be described by a Markov process (e.g., the weather).

Observable states : the states of the process that are `visible' (e.g., seaweed).

Components Of HMM

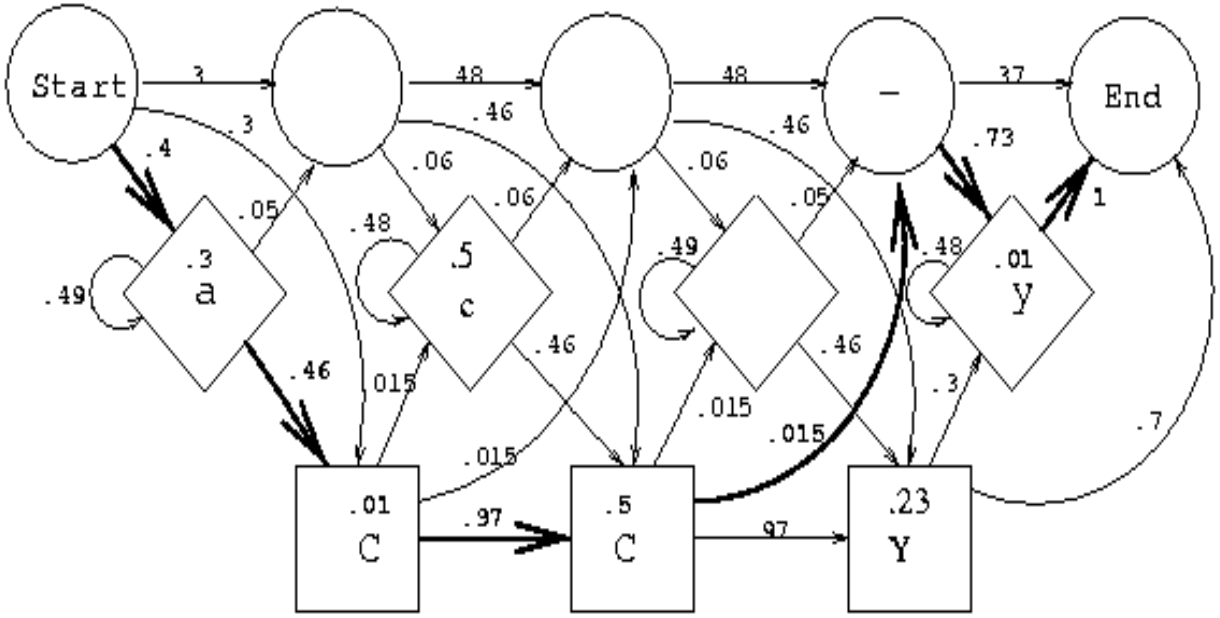
		Seaweed			
		Dry	**Dryish**	**Damp**	**Soggy**
weather	**Sun**	0.60	0.20	0.15	0.05
	Cloud	0.25	0.25	0.25	0.25
	Rain	0.05	0.10	0.35	0.50

Output matrix : containing the probability of observing a particular observable state given that the hidden model is in a particular hidden state.

Initial Distribution : contains the probability of the (hidden) model being in a particular hidden state at time t = 1.

State transition matrix : holding the probability of a hidden state given the previous hidden state.

Problems With HMM



Scoring problem:

Given an existing HMM and observed sequence , what is the probability that the HMM can generate the sequence

Problems With HMM

Training Problem

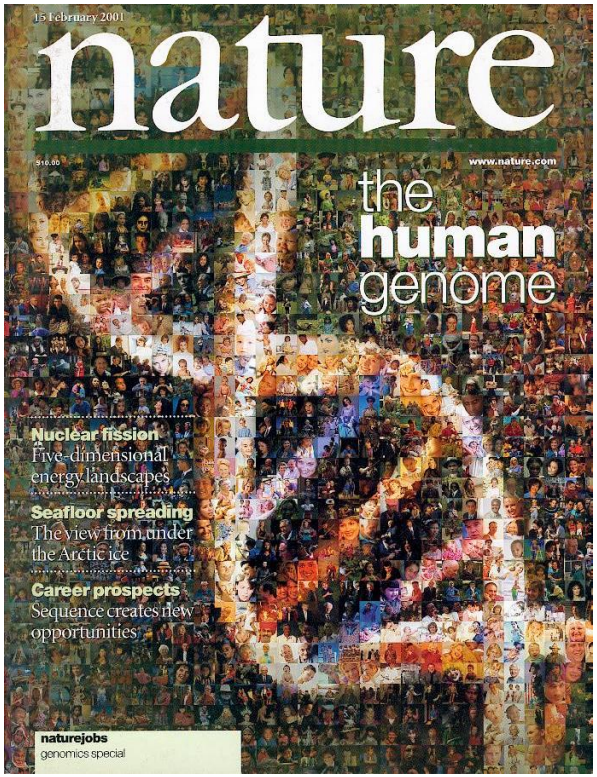
Given a large amount of data estimate the structure and the parameters of the HMM that best accounts for the data

HMMs in Biology

- Gene finding and prediction
- Protein-Profile Analysis
- Secondary Structure prediction

Finding genes in DNA sequence

This is one of the most challenging and interesting problems in computational biology at the moment. With so many genomes being sequenced so rapidly, it remains important to begin by identifying genes computationally.

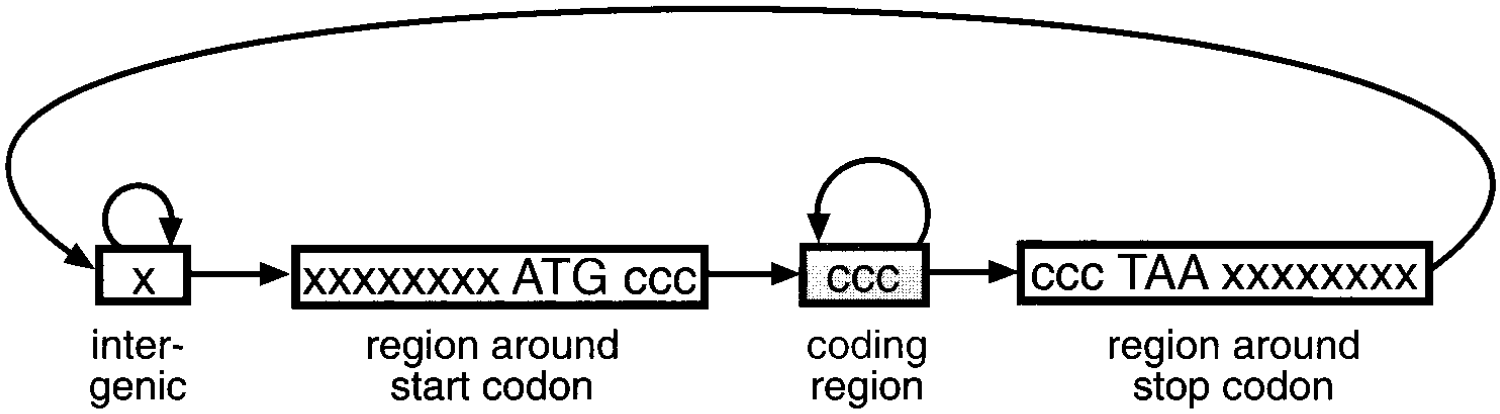


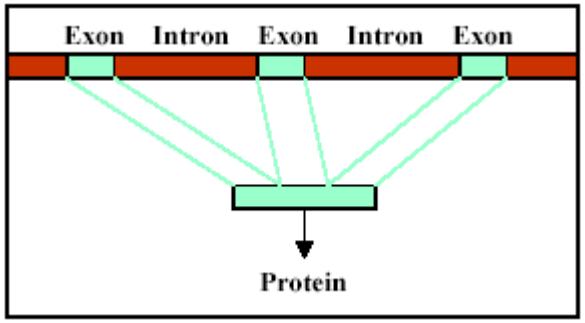
Gene Finding HMMs

- To find the coding and non-coding regions of an unlabeled string of DNA nucleotides
- Assist in the annotation of genomic data produced by genome sequencing methods
- Gain insight into the mechanisms involved in transcription, splicing and other processes

HMMs for gene finding

- Algorithms are used to find genes , by computational methods





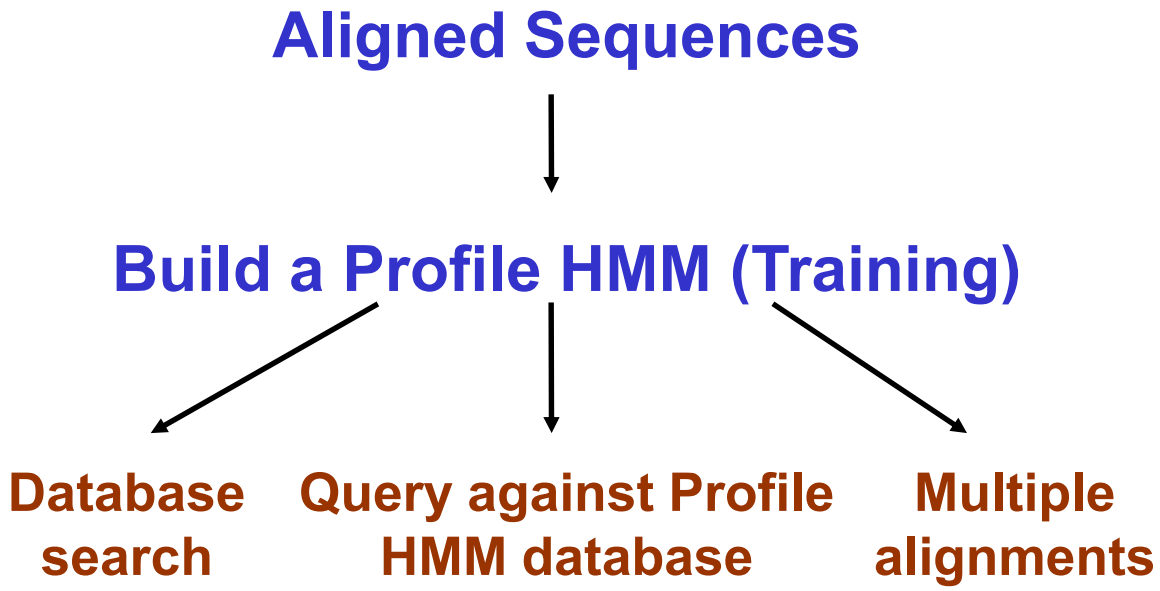
An HMM for unspliced genes.
x : non-coding DNA
c : coding state

Protein Profile HMMs

- Protein Profile Given a single amino acid target sequence of unknown structure

- Profile
  - Proteins families of related sequences and structures
  - Same function
  - Clear evolutionary relationship
  - Patterns of conservation, some positions are more conserved than the others

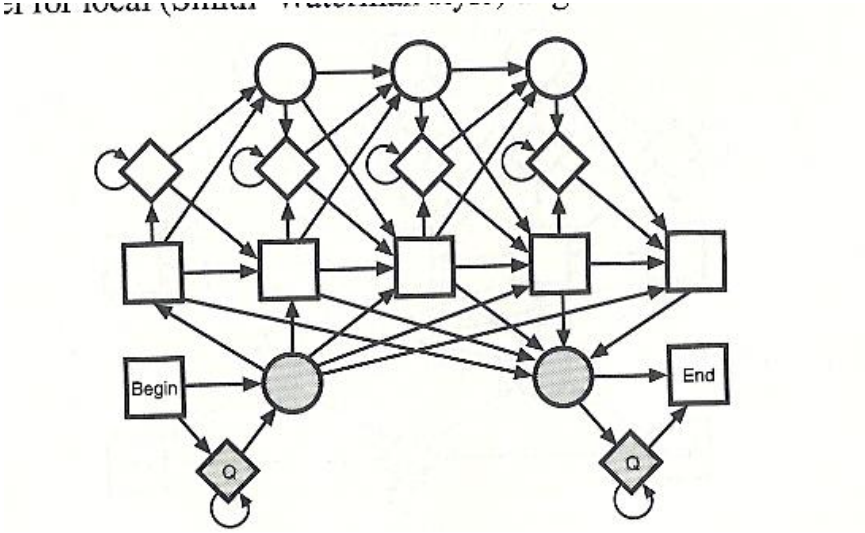
HMM Process


Aligned Sequences
Build a Profile HMM (Training)
Database search
Query against Profile HMM database
Multiple alignments

Database Searching

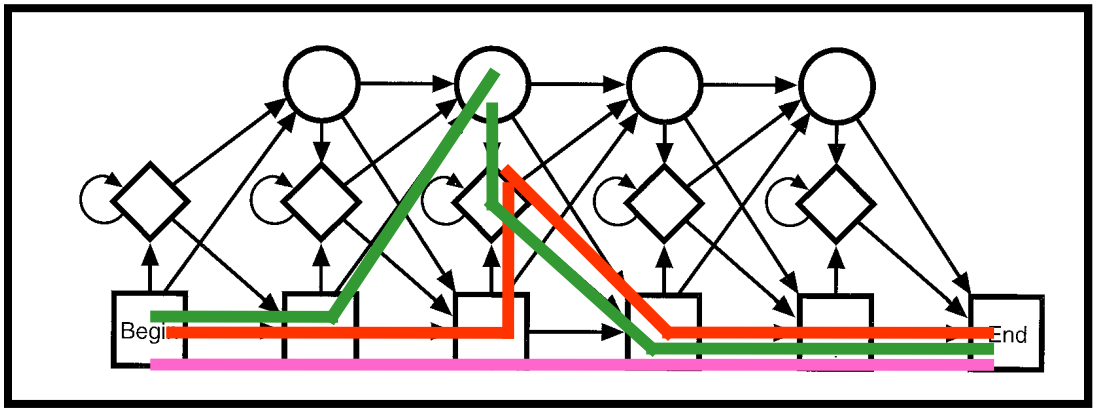
- Given HMM, M, for a sequence family, find all members of the family in data base.
- LL – score LL(x) = log P(x|M)

  (LL score is length dependent – must normalize or use Z-score)

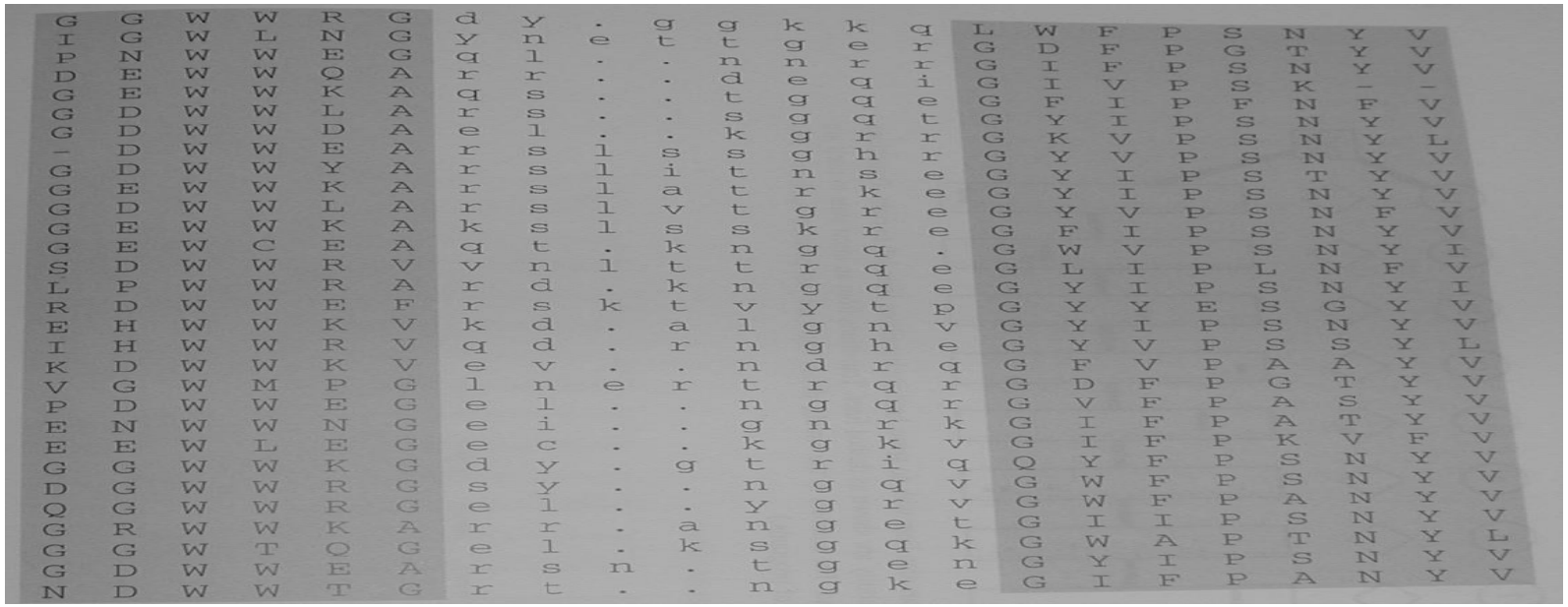


Multiple Alignments

- Try every possible path through the model that would produce the target sequences
  - Keep the best one and its probability.
  - Output : Sequence of match, insert and delete states
- alg. Dynamic Programming



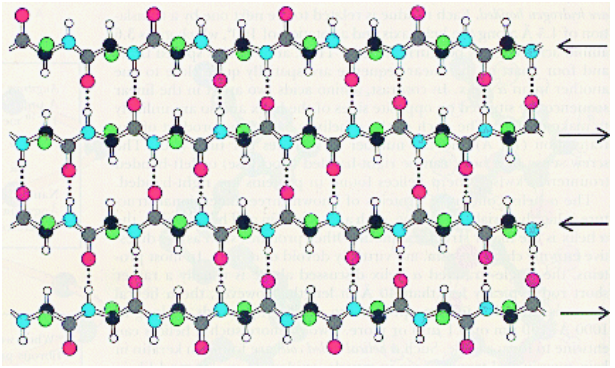
PHMM Example



An alignment of 30 short amino acid sequences chopped out of a alignment. The shaded area are the most conserved and were represented by the main states in the HMM.

Prediction of Protein Secondary structures

- Prediction of secondary structures is needed for the prediction of protein function.
- Analyze the amino-acid sequences of proteins
- Learn secondary structures
  - helix, sheet and turn



- Predict the secondary structures of sequences

Advantages

- Batch an entire family of sequences.
- Position-dependent character distributions and position-dependent insertion and deletion gap penalties.
- Built on a formal on probability basis
- Can make libraries of hundreds of profile HMMs and apply them on a large scale (**whole genome**)