Computational Biology
Lecture 12: HMM and beyond

Jo Grundy

ECS Southampton

21-02-2024

Hidden Markov Machines - Recap

- Viterbi algorithm finds most probable path given probabilities and states

$$q^* = \underset{p}{\text{argmax}}\, P(\boldsymbol{x}, q_p)$$

- Forward and Backwards algorithms finding the posterior probability of a state..

$$P(q_t = i|\boldsymbol{x}) = \frac{P(x, q_t = i)}{P(\boldsymbol{x})} = \frac{f_i(t)b_i(t)}{P(\boldsymbol{x})}$$

- are simplifications. Actually need the parameters:

$$q^* = \underset{p}{\text{argmax}}\, P(\boldsymbol{x}, q_p|\boldsymbol{\theta})\ P(q_t = i|\boldsymbol{x}|\boldsymbol{\theta}) = \frac{f_i(t)b_i(t)}{P(\boldsymbol{x}|\boldsymbol{\theta})}$$

.. where $\boldsymbol{\theta}$ are the parameters of the model.

We can sometimes infer these parameters if we know alot about the system generating the sequence.. however usually we don't

Training HMMs - Introduction

Given a sequence and a state path, we can count

- A_{ij} count of $i \to j$ transitions
- $E_i(l)$ count of letter l when in state i

To maximise likelihood of the sequence:

$$e_i(l) = \frac{E_i(l)}{\sum_{l \in \mathcal{A}} E_i(l)} \qquad a_{ij} = \frac{A_{ij}}{\sum_{k \in \mathcal{Q}} A_{ik}} \tag{1}$$

Training HMMs - Introduction

Need to know:

- probabilities to define the state sequence
- state sequence to work out the probabilities

How do we get round it?

Training HMMs - Introduction

Need to know:

- probabilities to define the state sequence
- state sequence to work out the probabilities

How do we get round it?

Expectation Maximisation

Training HMMs - Baum Welch

- make initial guess at transition and emission probabilities
- use this to estimate posterior probabilities - Expectation
- calculate the new transition and emission probabilities given the state posterior probabilities calculated - Maximisation

The second and third steps can be repeated until and good enough estimation is reached

Training HMMs - Baum Welch

One more point of detail is Pseudocounts

$$e_i(l) = \frac{E_i(l)}{\sum_{l \in \mathcal{A}} E_i(l)} \qquad a_{ij} = \frac{A_{ij}}{\sum_{k \in \mathcal{Q}} A_{ik}} \tag{2}$$

Problem:

- HMM may want to allow transitions that we don't see much in the training data
- Can solve by adding *pseudo counts*

Training HMMs - Baum Welch

One more point of detail is Pseudocounts

$$e_i(l) = \frac{E_i(l)}{\sum_{l \in \mathcal{A}} E_i(l)} \quad a_{ij} = \frac{A_{ij}}{\sum_{k \in \mathcal{Q}} A_{ik}} \tag{2}$$

Problem:

- HMM may want to allow transitions that we don't see much in the training data
- Can solve by adding *pseudo counts*

$$e_i(l) = \frac{E_i(l) + r_i(l)}{\sum_{l \in \mathcal{A}} E_i(l) + r_i(l)} \quad a_{ij} = \frac{A_{ij} + r_i(l)}{\sum_{k \in \mathcal{Q}} A_{ik} + r_i(l)} \tag{3}$$

There is more detail but I will not cover it here. For more see Durbin *et al*, Biological Sequence Analysis

HMM and Beyond - HMM Structure

The structure of an HMM is usually designed by an expert.
Need to decide on:

- number of states
- what transitions are allowed
- what order

HMM and Beyond - HMM Structure

Number of states:

Too few states:

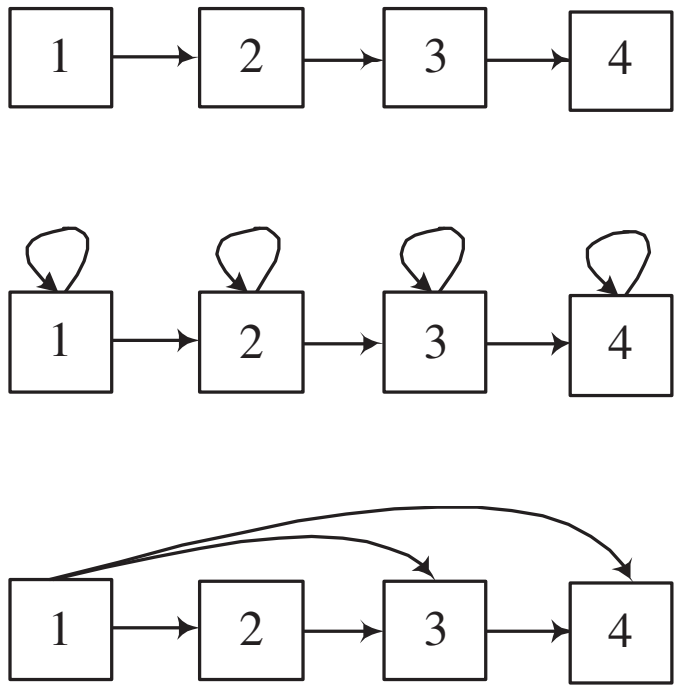
- fail to model significant differences
- miss detail
- underfit

Too many states:

- model noise
- poor generalisation
- overfit

HMM and Beyond - HMM Structure

In Biology HMMs are usually built from blocks
These represent common motifs in biological sequences



Given a set of motifs, a genetic algorithm can also generate good structures
For more complex problems, they are better than experts

HMM and Beyond - HMM Structure

Transitions between states:

- should represent real physical structure
- fully connected gives best likelihood but can overfit

HMM and Beyond - HMM Structure

Order

The Markov property of HMM imposes too little short range structure

They can be more generalised to second order:

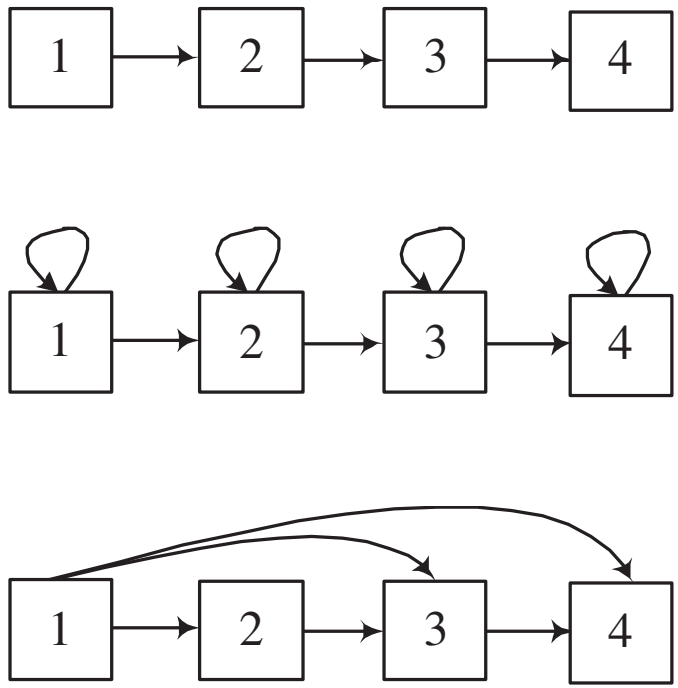
$$P(x_i|x_{i-1}, x_{1-2}, ..., x_1) = P(x_i|x_{i-1}, x_{i-2})$$

and to nth order

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

HMM and Beyond - HMM Structure

In Biology HMMs are usually built from blocks
These represent common motifs in biological sequences



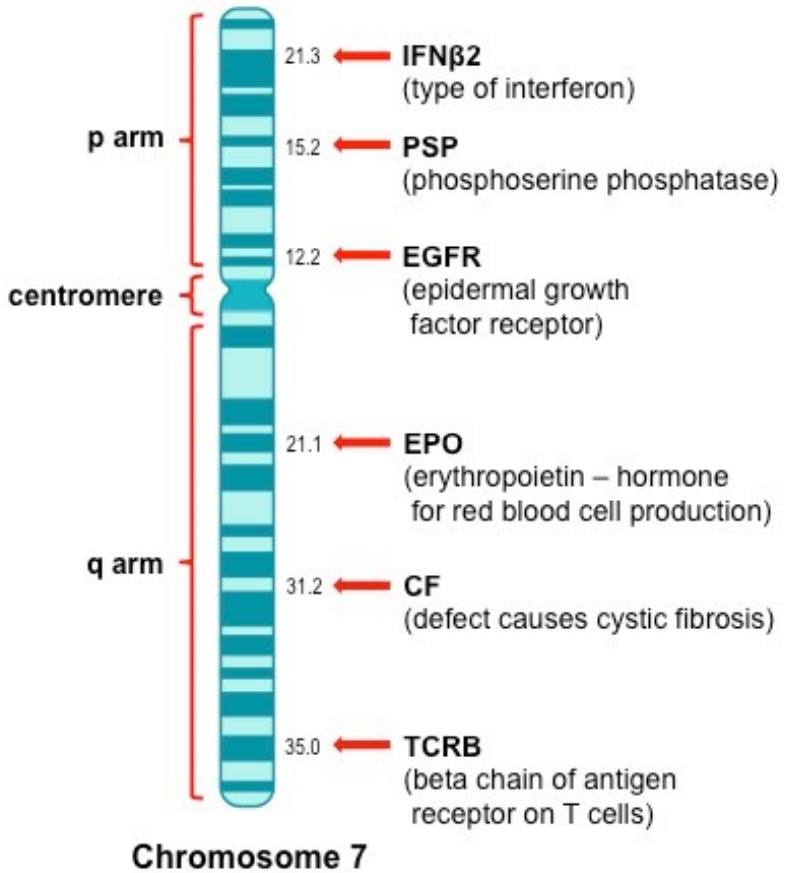
Given a set of motifs, a genetic algorithm can also generate good structures
For more complex problems, they are better than experts

HMM and Beyond - Finding Genes

- There are about 20,000 genes in the genome
- 25% of the genome is genes

Genes contain:

- promoters
- enhancers
- silencers
- insulators
- coding - 1%

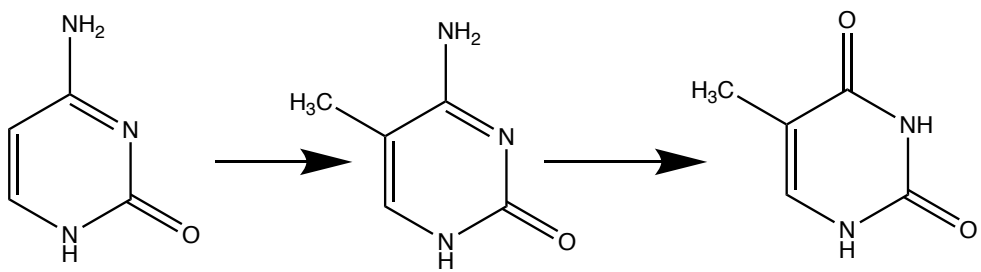


Finding Genes on the genome

HMM and Beyond - CpG islands

The 'CG' base dinucleotide sequence in DNA is less common than expected

- methylation of the Cytosine base is common
- more likely to mutate in to a Thymine base

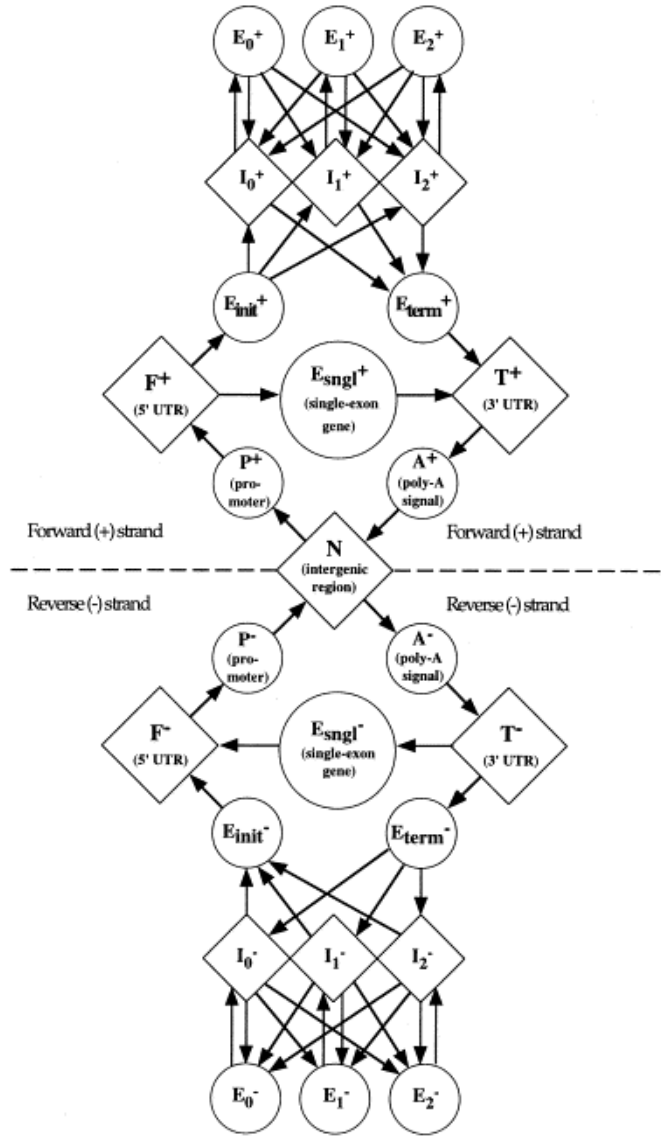


Methylation is supressed in promoter regions of the genome

HMM and Beyond - HMM Example

GENESCAN

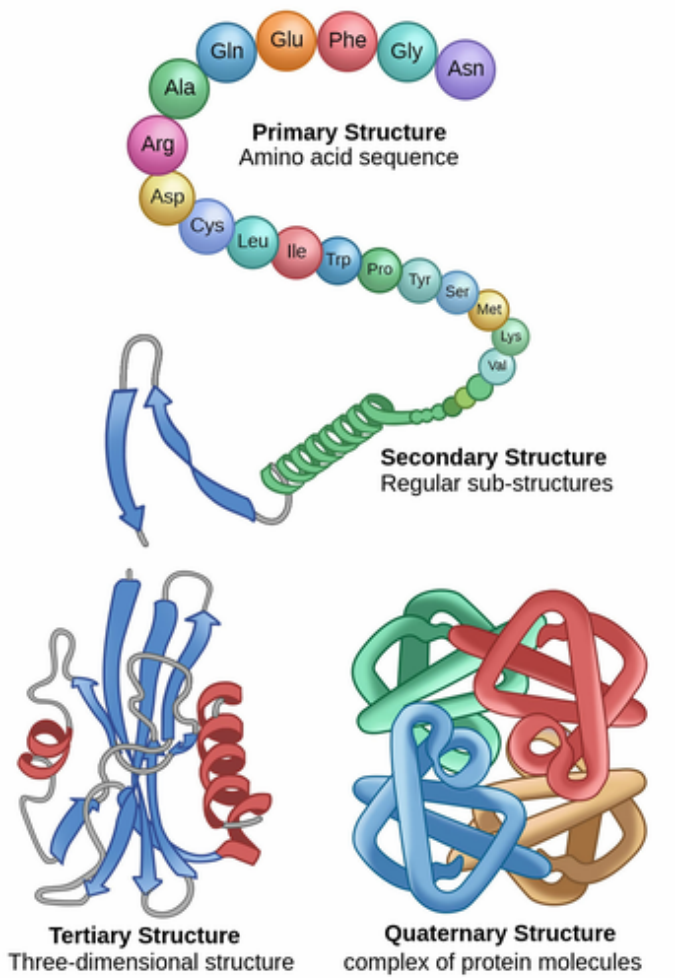
- found genes
- fifth order Markov model
- Burge and Karlin 1997
- looks for areas of high 'CpG' content



HMM and Beyond - Protein Structure Prediction

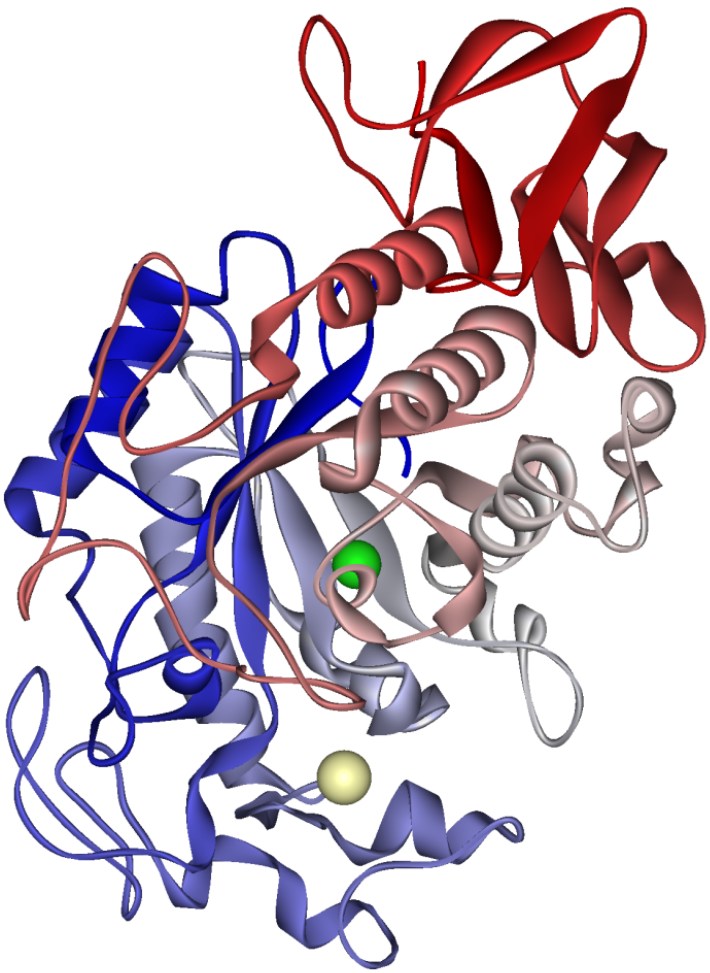
Proteins structure:

- Primary - amino acid sequence
- Secondary - α helices, β sheets or coil
- Tertiary - 3D structure of polypeptide
- Quaternary - protein complex

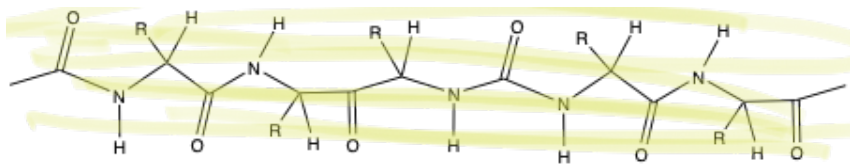


HMM and Beyond - Protein Structure Prediction

- Amylase enzyme protein



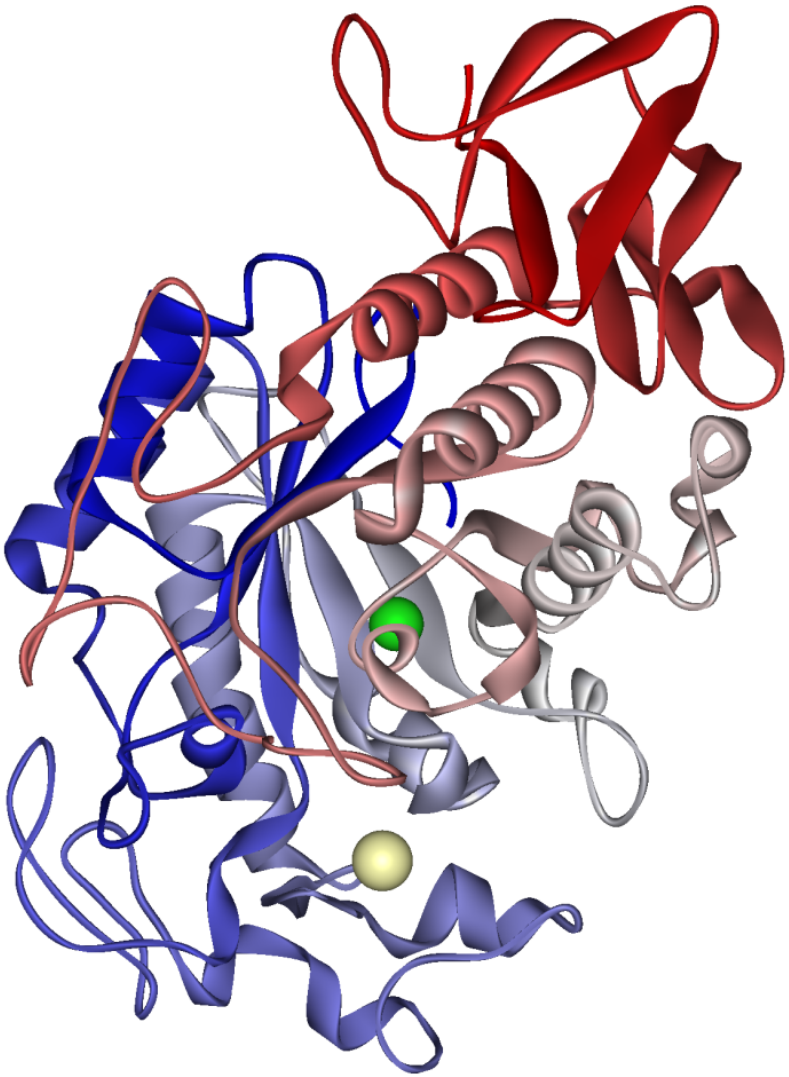
This is the amylase in your saliva, breaks down starch in to sugars. If you chew bread or plain pasta for a while, it goes sweet because of this enzyme.

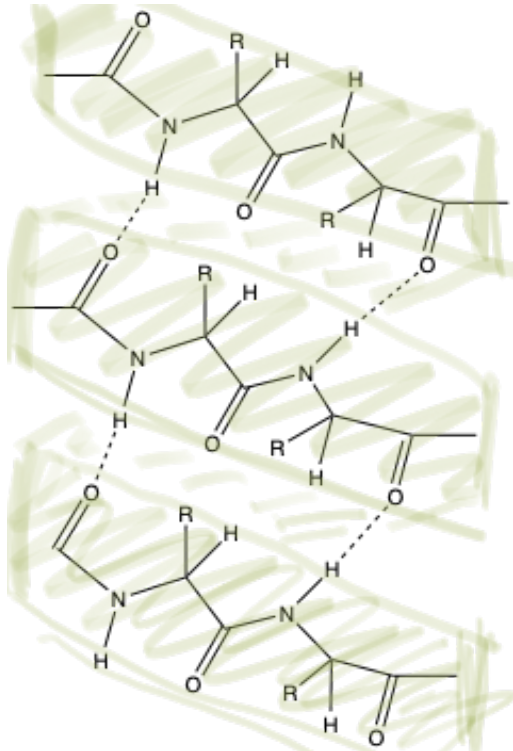


Each strand is made of a sequence of amino acids

HMM and Beyond - Protein Structure Prediction

- Amylase enzyme protein

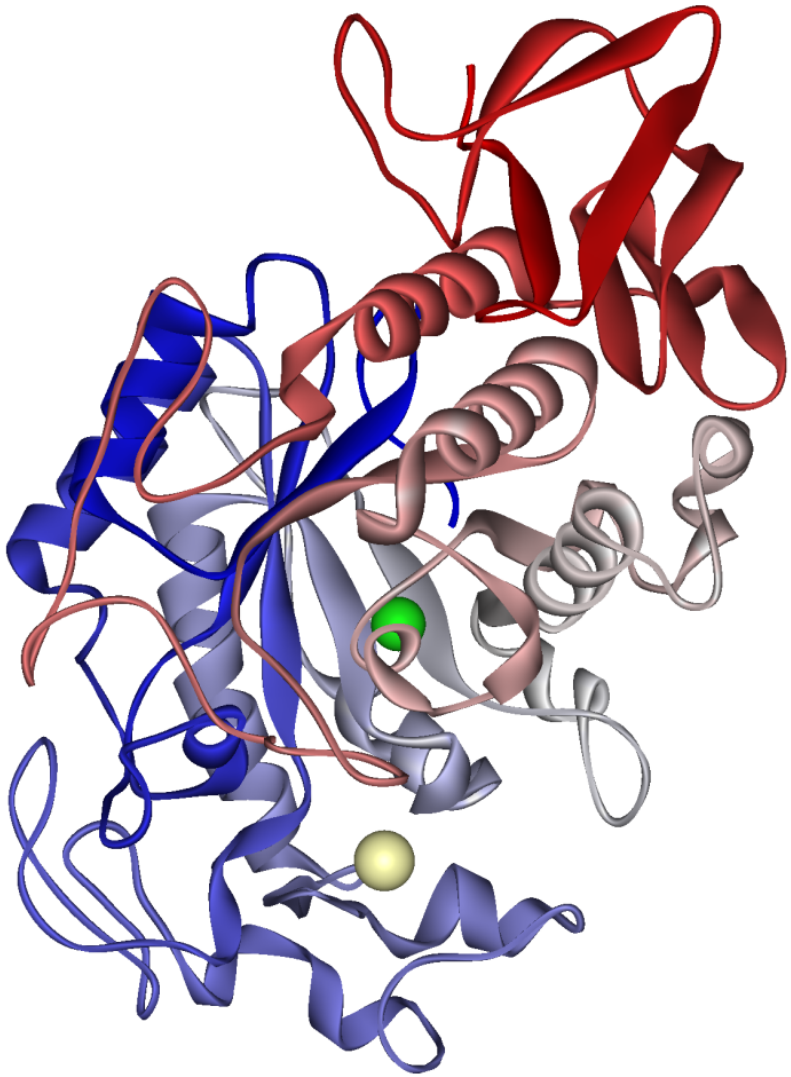


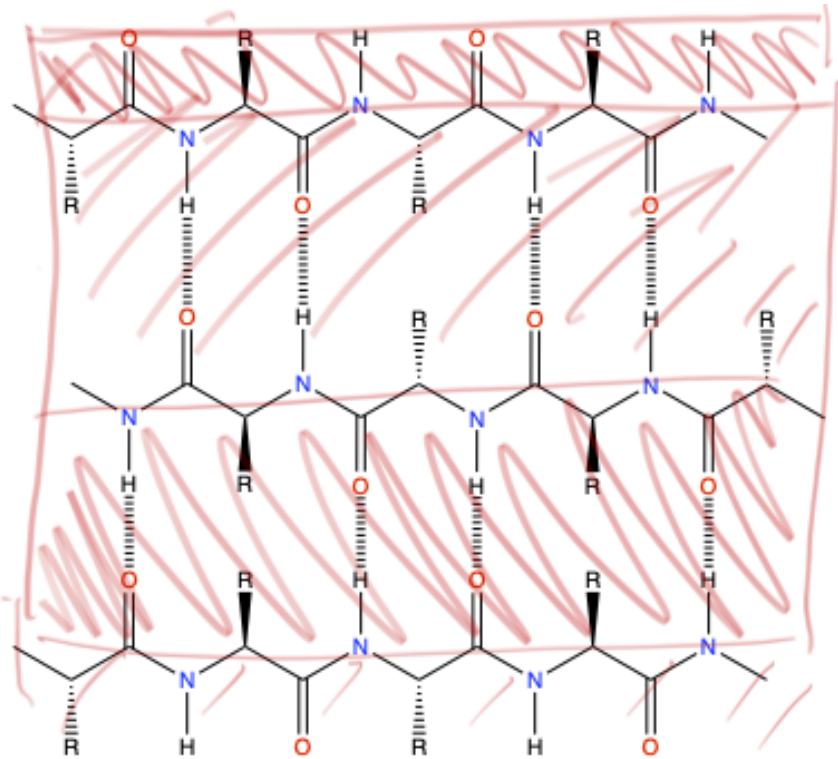


The spiral shapes are called 'alpha helix', and are the amino acid chains stacked up together, held together with hydrogen bonds

HMM and Beyond - Protein Structure Prediction

- Amylase enzyme protein





The flat ribbons are called 'beta sheets', and have the amino acid chain forming a flat structure, held together with hydrogen bonds

HMM and Beyond - Protein Structure Prediction

Can the amino acid sequence be used to generate the *secondary structure?*

Difficulties:

- *Non local*
- *Different lengths of input*
- *Representation*

However:

- *There is now good training data*
- *Can use windows to get a fixed input*

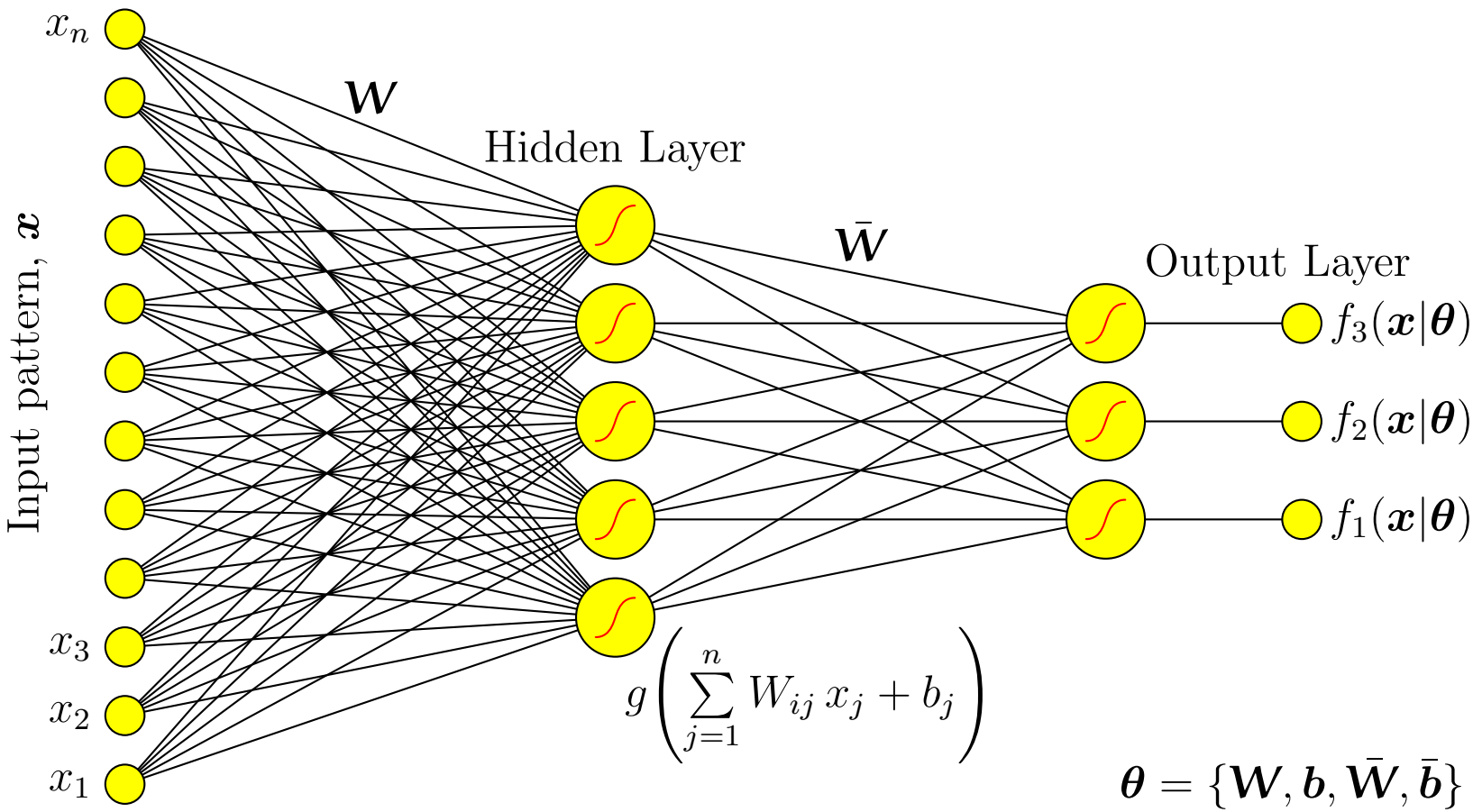
Yes.

HMM and Beyond - Secondary Structure Prediction

Qian and Sejnowski 1988:
Adapted their 'NETtalk' system from a word sequence to an amino acid sequence

- mean squared error loss
- one hot encoding representation
- windowed input

HMM and Beyond - Protein Structure Prediction



$$\boldsymbol{\theta} = \{\boldsymbol{W}, \boldsymbol{b}, \bar{\boldsymbol{W}}, \bar{\boldsymbol{b}}\}$$

$$f_i(\boldsymbol{x}|\boldsymbol{\theta}) = g\left(\sum_{j=1}^{n_h} \bar{W}_{i,j}\, g\left(\sum_{k=1}^{n} W_{ik}\, x_k + b_k\right) + \bar{b}_i\right)$$

HMM and Beyond - Protein Structure Prediction

Loss function - mean squared error

$$L(\theta, \mathcal{D}) = \frac{1}{m} \sum_{i=1}^{m} ||y_i - f(x_i|\theta)||^2$$

Follow negative gradient wrt. parameters θ
Make a step in the direction of negative gradient

$$\theta(t+1) = \theta(t) - r \nabla L(\theta, \mathcal{D})$$

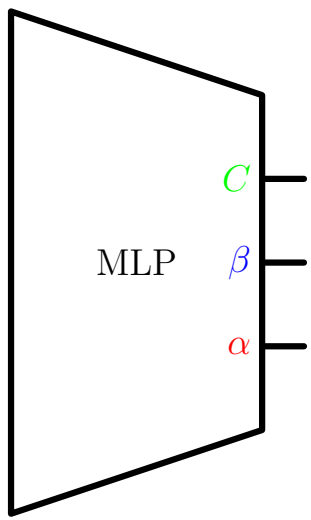
r is learning rate

HMM and Beyond - Protein Structure Prediction

How to make a variable length sequence a fixed length input?

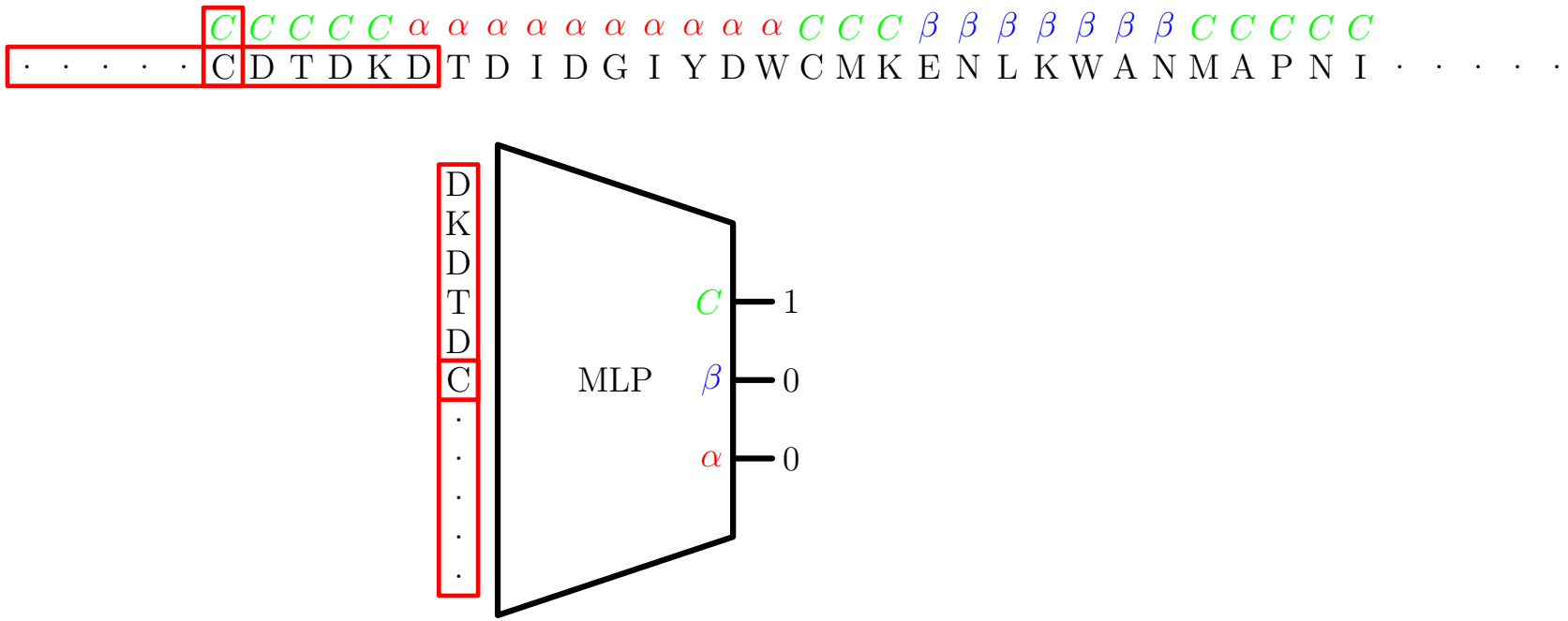
$C\ C\ C\ C\ C\ \alpha\ \alpha\ \alpha\ \alpha\ \alpha\ \alpha\ \alpha\ \alpha\ \alpha\ \alpha\ C\ C\ C\ \beta\ \beta\ \beta\ \beta\ \beta\ \beta\ \beta\ C\ C\ C\ C\ C$

$\cdot\ \cdot\ \cdot\ \cdot\ \cdot$ C D T D K D T D I D G I Y D W C M K E N L K W A N M A P N I $\cdot\ \cdot\ \cdot\ \cdot\ \cdot$



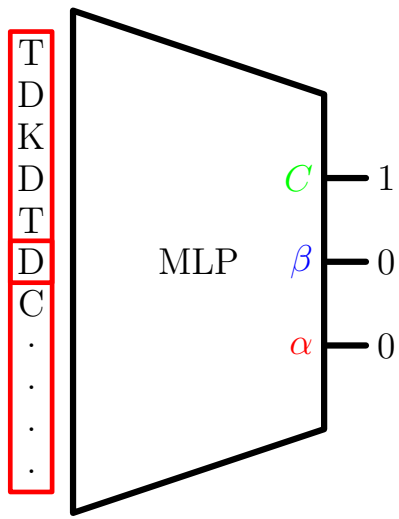
HMM and Beyond - Protein Structure Prediction

Windowed input



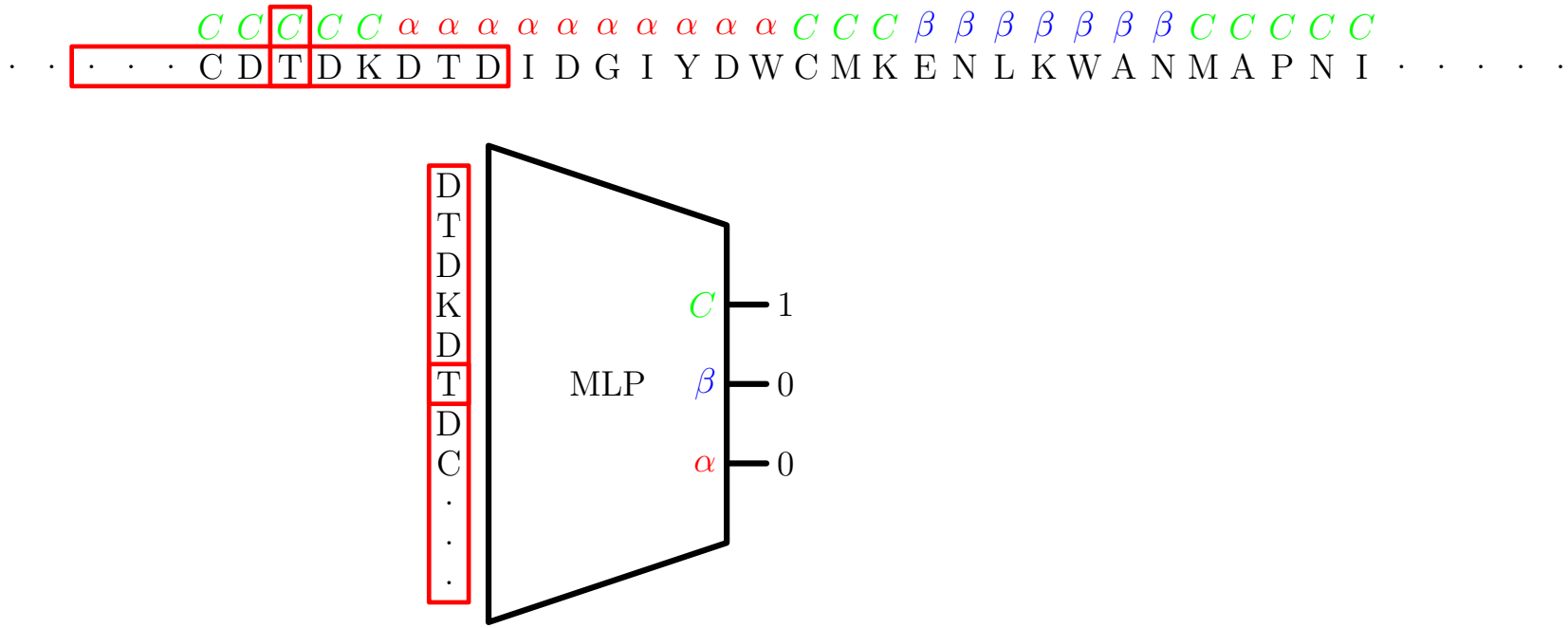
HMM and Beyond - Protein Structure Prediction

Windowed input



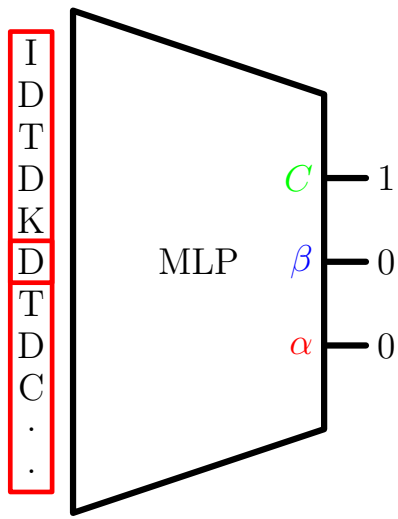
HMM and Beyond - Protein Structure Prediction

Windowed input



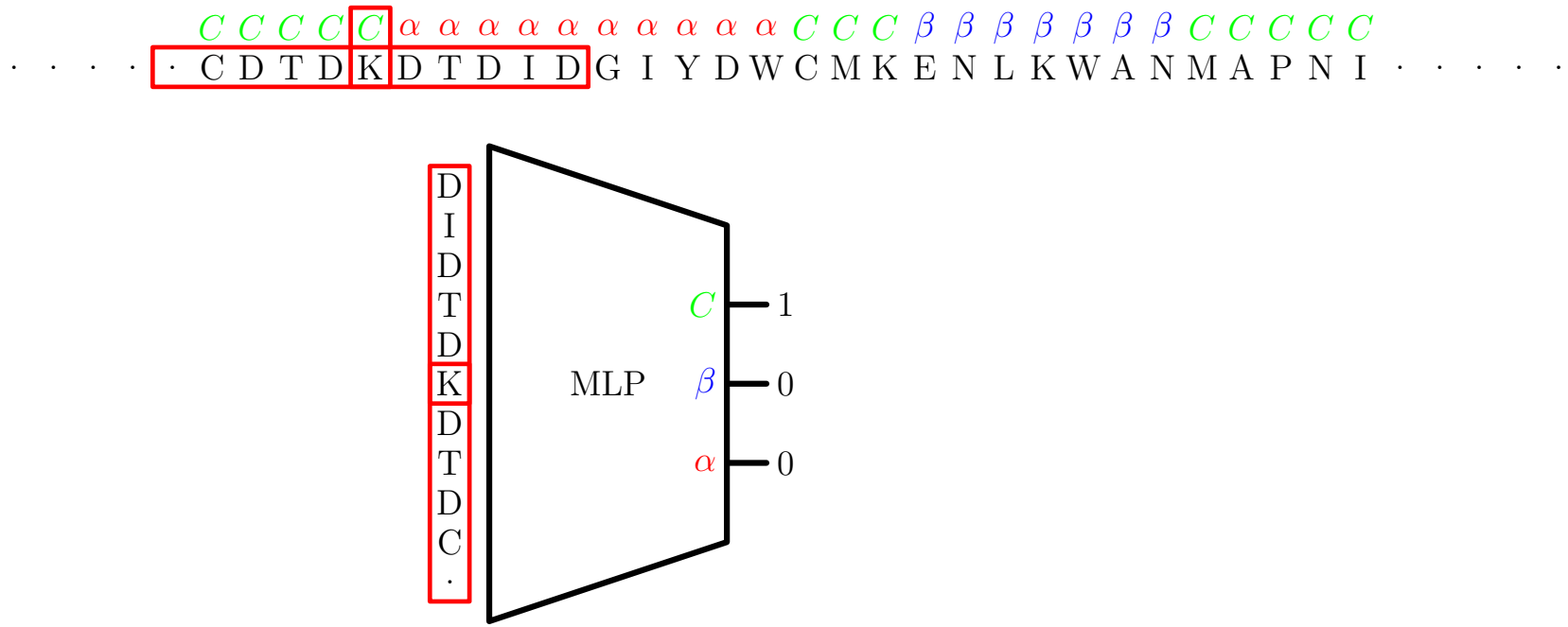
HMM and Beyond - Protein Structure Prediction

Windowed input



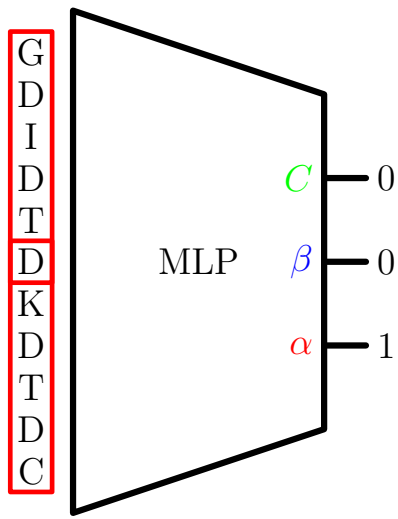
HMM and Beyond - Protein Structure Prediction

Windowed input



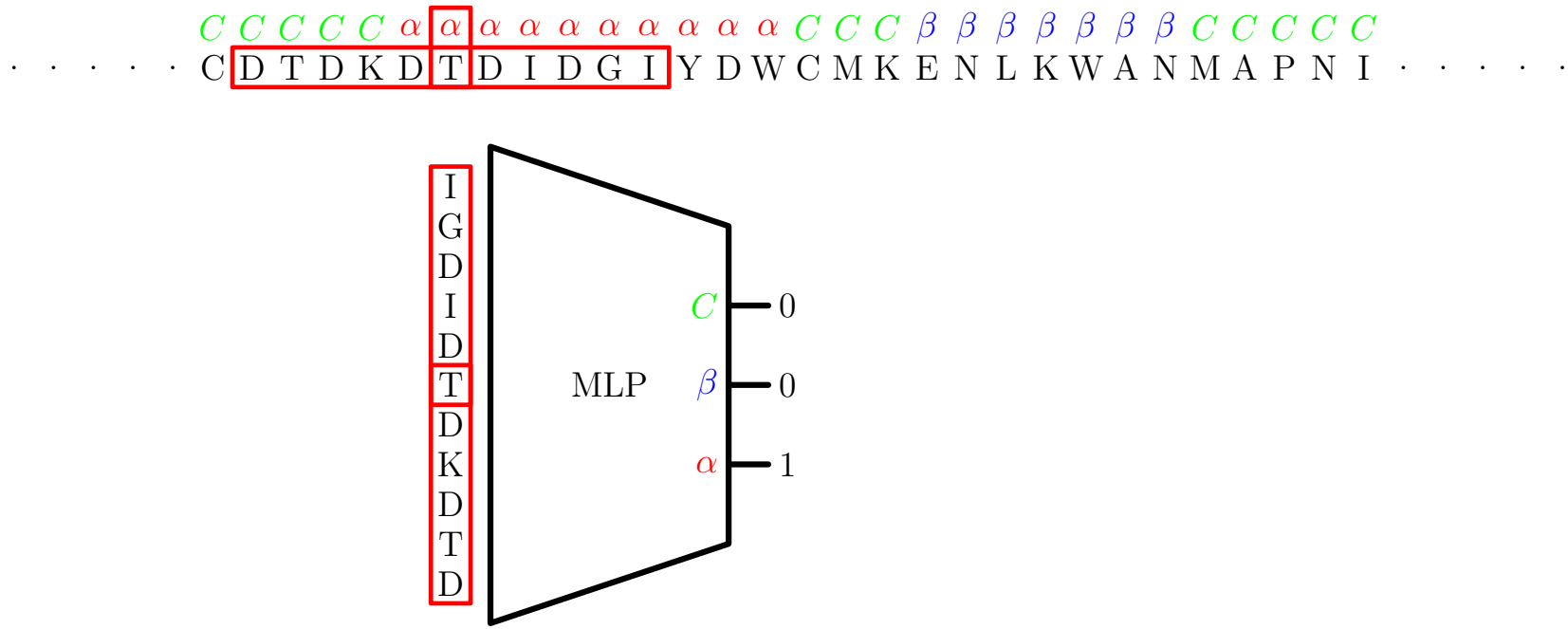
HMM and Beyond - Protein Structure Prediction

Windowed input



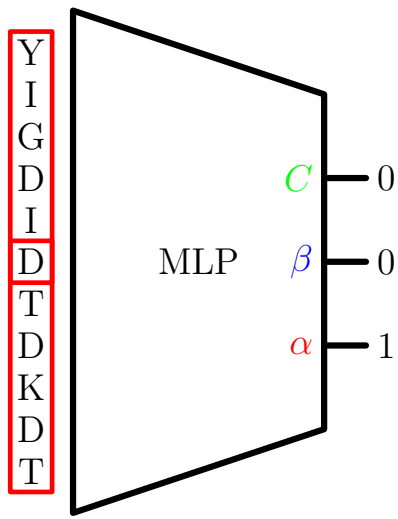
HMM and Beyond - Protein Structure Prediction

Windowed input



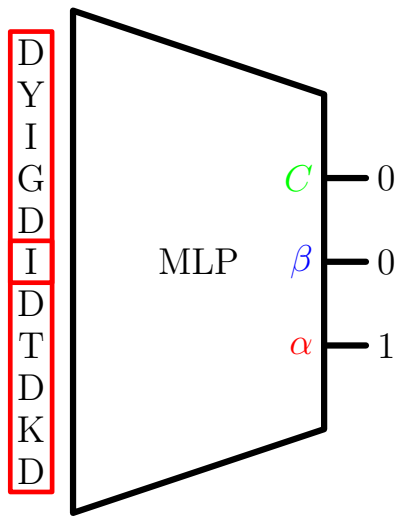
HMM and Beyond - Protein Structure Prediction

Windowed input



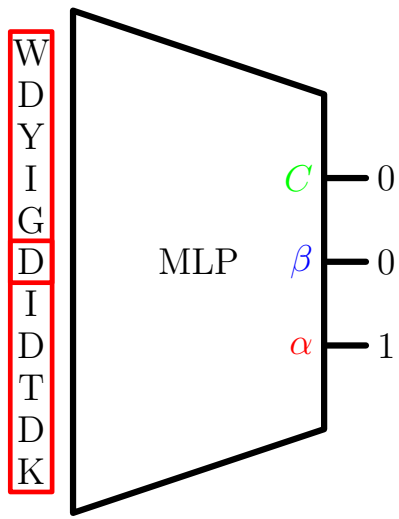
HMM and Beyond - Protein Structure Prediction

Windowed input



HMM and Beyond - Protein Structure Prediction

Windowed input

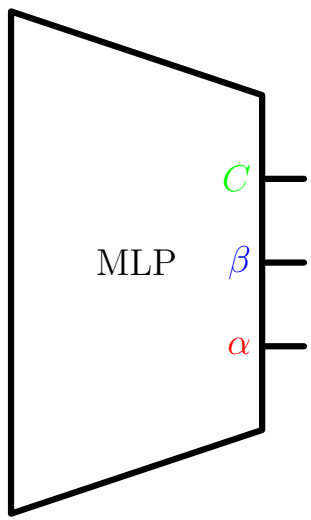


HMM and Beyond - Protein Structure Prediction

Need to shuffle:

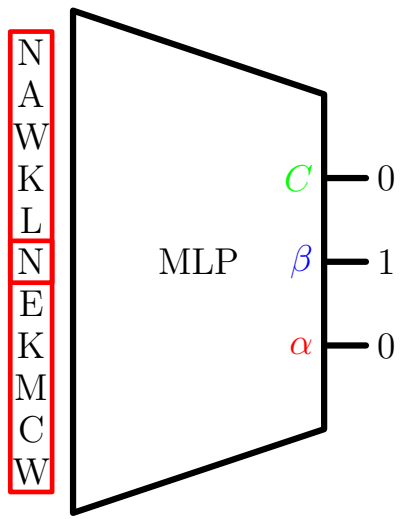
$C\ C\ C\ C\ C\ \alpha\ \alpha\ \alpha\ \alpha\ \alpha\ \alpha\ \alpha\ \alpha\ \alpha\ \alpha\ C\ C\ C\ \beta\ \beta\ \beta\ \beta\ \beta\ \beta\ \beta\ C\ C\ C\ C\ C$

$\cdot\ \cdot\ \cdot\ \cdot\ \cdot$ C D T D K D T D I D G I Y D W C M K E N L K W A N M A P N I $\cdot\ \cdot\ \cdot\ \cdot\ \cdot$



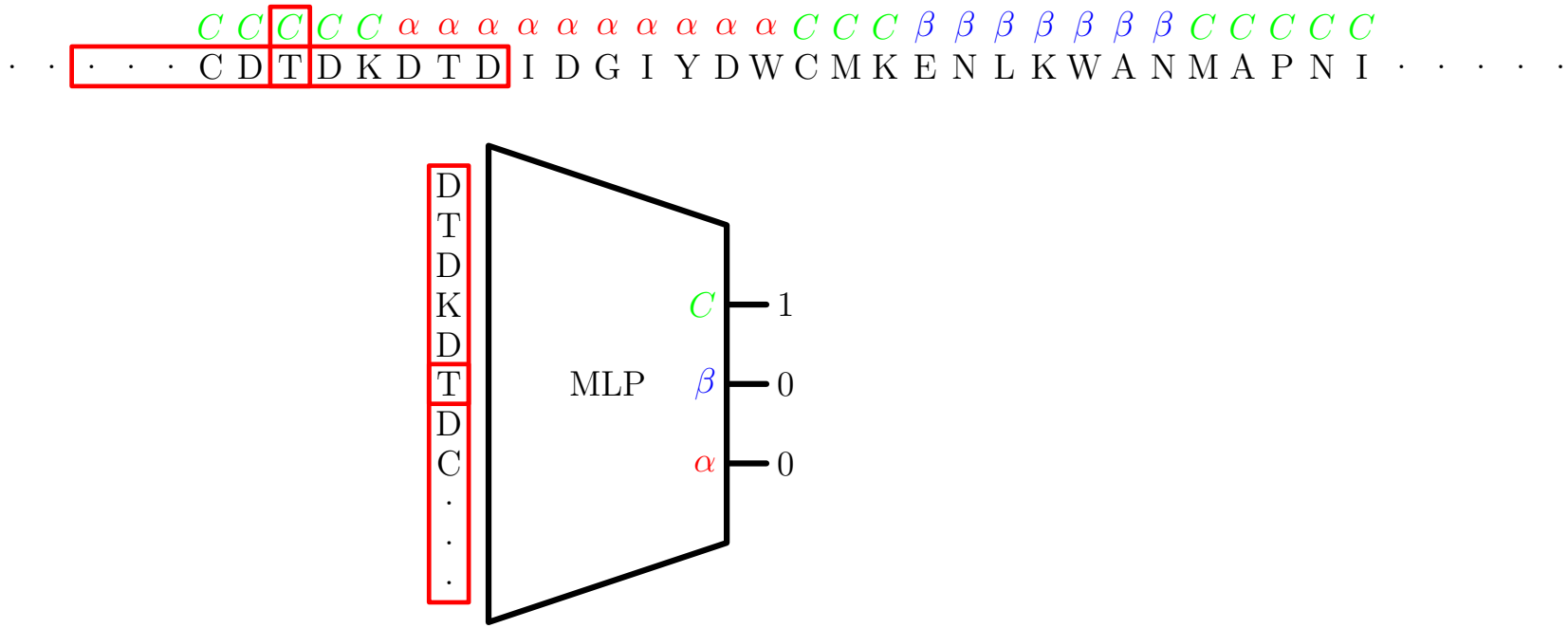
HMM and Beyond - Protein Structure Prediction

Randomised windowed input



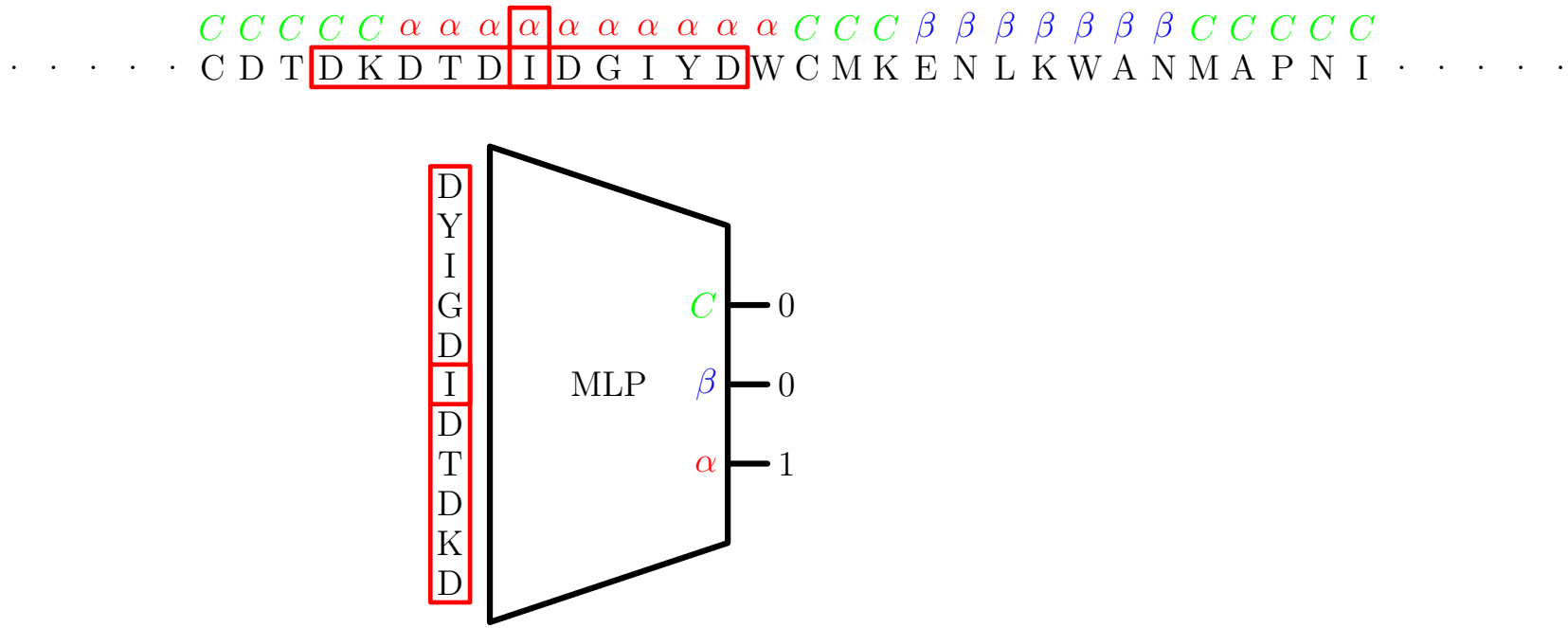
HMM and Beyond - Protein Structure Prediction

Randomised windowed input



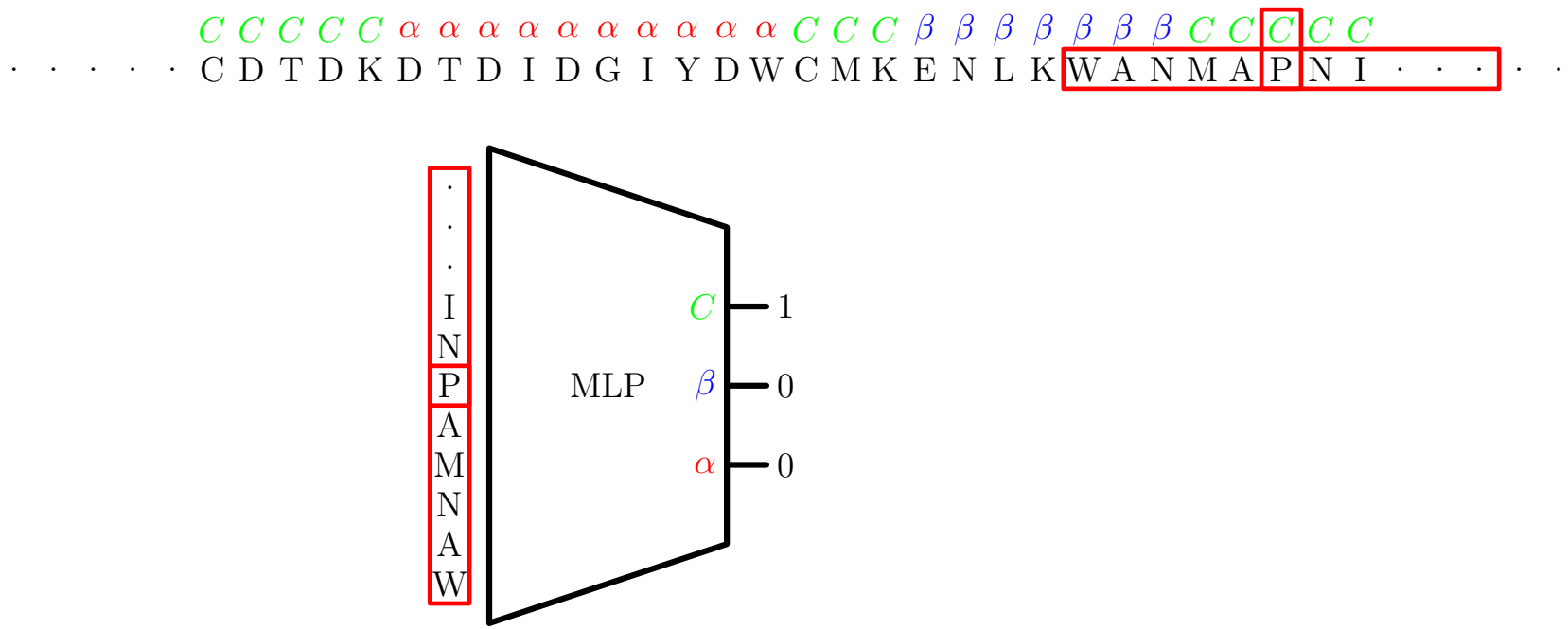
HMM and Beyond - Protein Structure Prediction

Randomised windowed input



HMM and Beyond - Protein Structure Prediction

Randomised windowed input



HMM and Beyond - Protein Structure Prediction

Qian Sejnowski - 1988 - 64.3%
'Coil' more common, 47%, easier for network to recognise.
Improved Rost and Sander - 1993

- profiling
- balanced training
- structural context
- jury of networks

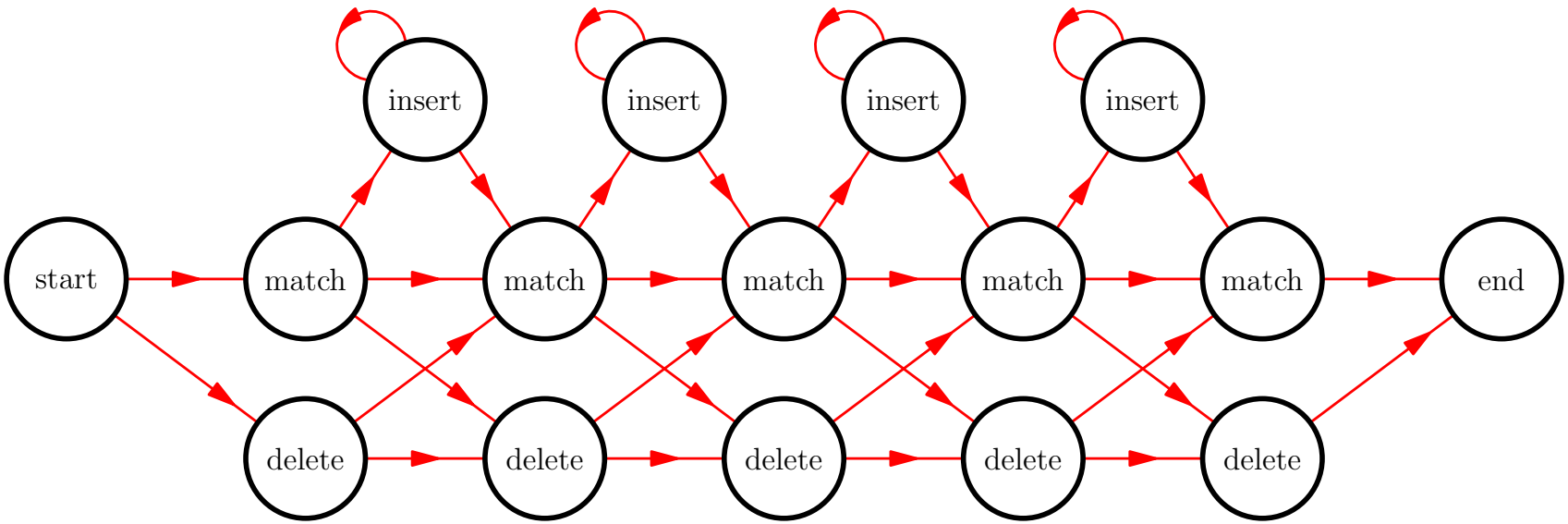
HMM and Beyond - Protein Structure Prediction

Profiling Homologous proteins

- same tertiary structure
- approximately same secondary structure

Add each aligned protein coding up

Using sequence alignments of a family of proteins - HMM



Allows relationships between different proteins to be expressed in the coding

Increases accuracy by about 5%

HMM and Beyond - Protein Structure Prediction

Coding

Coding amino acids is problematic

- Use a number 1-20

HMM and Beyond - Protein Structure Prediction

Coding

Coding amino acids is problematic

- Use a number 1-20 Gives a false representation 19 is not similar to 20

HMM and Beyond - Protein Structure Prediction

Coding

Coding amino acids is problematic

- Use a number 1-20 Gives a false representation 19 is not similar to 20
- Use chemical features

HMM and Beyond - Protein Structure Prediction

Coding

Coding amino acids is problematic

- Use a number 1-20 Gives a false representation 19 is not similar to 20
- Use chemical features Should work better than it does.

HMM and Beyond - Protein Structure Prediction

Coding

Coding amino acids is problematic

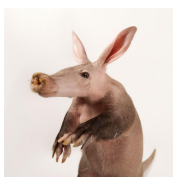
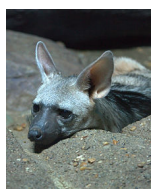
- Use a number 1-20 Gives a false representation 19 is not similar to 20
- Use chemical features Should work better than it does.
- One hot encoding works very well

HMM and Beyond - Protein Structure Prediction

One hot encoding:

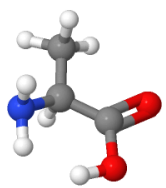
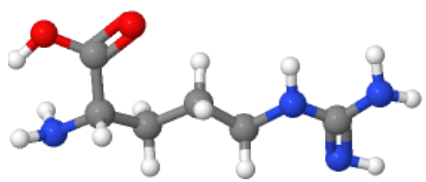
For words,

- a $\rightarrow$ $[1, 0, 0, 0, 0, 0, 0, 0, \ldots, 0]$
- aa $\rightarrow$ $[0, 1, 0, 0, 0, 0, 0, 0, \ldots, 0]$
- aardvark $\rightarrow$ $[0, 0, 1, 0, 0, 0, 0, 0, \ldots, 0]$
- aardwolf $\rightarrow$ $[0, 0, 0, 1, 0, 0, 0, 0, \ldots, 0]$

Similarly for amino acids, however only 20 amino acids makes this easier.

- Ala $\rightarrow$ $[1, 0, 0, 0, 0, 0, 0, 0, \ldots, 0]$
- Arg $\rightarrow$ $[0, 1, 0, 0, 0, 0, 0, 0, \ldots, 0]$
- Asn $\rightarrow$ $[0, 0, 1, 0, 0, 0, 0, 0, \ldots, 0]$
- Asp $\rightarrow$ $[0, 0, 0, 1, 0, 0, 0, 0, \ldots, 0]$

HMM and Beyond - Protein Structure Prediction

Profiling - using one hot encodings and alignments of similar proteins

Protein	G	Y	I	Y	D	P	A	V	G	D	...
Alignments	G	Y	I	Y	D	P	E	V	G	D	...
	G	Y	I	Y	D	P	A	V	G	D	...
	G	Y	E	Y	D	P	A	E	G	D	...
	G	Y	E	Y	D	P	A	E	G	D	...
G	5	0	0	0	0	0	0	0	5	0	
A	0	0	0	0	0	0	4	0	0	0	
P	0	0	0	0	0	5	0	0	0	0	
D	0	0	0	0	5	0	0	0	0	5	
E	0	0	2	0	0	0	1	2	0	0	
V	0	0	0	0	0	0	0	3	0	0	
I	0	0	3	0	0	0	0	0	0	0	
Y	0	5	0	5	0	0	0	0	0	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

HMM and Beyond - Protein Structure Prediction

Balanced Training

The distribution of secondary structure types is unbalanced:

- 32% α helix
- 21% β sheet
- 47% coil

These are sampled in equal quantities so in the training data:

- $33.\dot{3}\%$ α helix
- $33.\dot{3}\%$ β sheet
- $33.\dot{3}\%$ coil

This reduces performance over all slightly, but does increase accuracy on the α helix and coil structures.

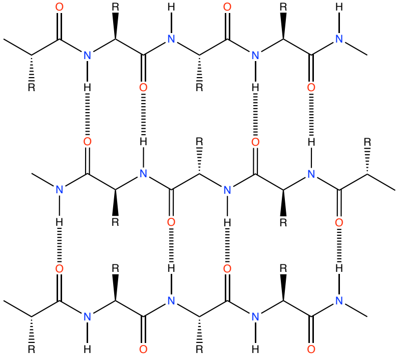
HMM and Beyond - Protein Structure Prediction

Structural context:

- α helices are typically 10 amino acids long
- β sheets are typically 6 amino acids long

The prediction: $\alpha\ \alpha\ \alpha\ \beta\ \beta\ \ \alpha\ \alpha\ \alpha$ makes no sense.
Should be $\alpha\ \alpha\ \alpha\ \alpha\ \alpha\ \alpha\ \alpha\ \alpha$
A structure to structure MLP is attached at the end to propagate this training signal back.
This is known as *stacking.*
Gives $2-3\%$ *improvement*

For example, β sheet



HMM and Beyond - Protein Structure Prediction

Secondary Structure prediction:

- Qian Sejnowski - 1988 - 64.3%
- Rost and Sander - 1993 - 69.7%

and..?

HMM and Beyond - PSI PRED

PSI BLAST:

- Position Specific Interative - Basic Local Alignment Search Tool - PSI BLAST
- Derives a Position Specific Scoring Matrix - PSSM
- This matrix is used to search for new matches
- PSSM is then updated
- repeated to give both near and distant relationships between proteins

PSI PRED:

- Predicts secondary structure based on PSI BLAST profiles
- uses MLP

Gets 77% accuracy

HMM and Beyond - Protein Structure Prediction

Secondary Structure prediction:

- Qian Sejnowski - 1988 - 64.3%
- Rost and Sander - 1993 - 69.7%
- PSI PRED - 1999 - 77%

Little progress until 2010..
Next Lecture: Deep Learning

HMM and Beyond - Protein Structure Prediction

Multi class classification is better with cross entropy loss and softmax. Deals better with minority classes.
Why? - We want to maximise the likelihood of the correct classification given the data
Posterior probability for class k is:

$$P(C_k|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|C_k)P(C_k)}{\sum_j P(\boldsymbol{x}|C_j)P(C_j)}$$

Given:

$$a_k = \ln P(\boldsymbol{x}|C_k)P(C_k) \approx f_{C_i}(\boldsymbol{x}_i|\boldsymbol{\theta})$$

$$P(C_k|\boldsymbol{x}) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

Softmax: Allows output to be interpreted as the probability of the classification

HMM and Beyond - Protein Structure Prediction

Cross Entropy Error

Maximising the relative entropy is maximising the log-likelihood of the data (over N patterns)

$$P(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{N} P(C_i|\boldsymbol{x}_i, \boldsymbol{\theta})$$

Take logs:

$$\log P(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log P(C_i|\boldsymbol{x}_i, \boldsymbol{\theta})$$

We approximate that probability using the output

$$\log P(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log f_{C_i}(\boldsymbol{x}_i|\boldsymbol{\theta})$$

Maximising the cross entropy we maximise the log probability. Gives more and better information to the machine than squared error loss.

HMM and Beyond - Summary

Machine learning tools used on language have been adapted for biological sequence analysis

- HMM - still used today
- MLP - need to tune correctly

tools to use:

- Cross entropy loss
- Soft max
- windowing
- profiling

Now deep learning predominates - next lecture