UCS2735 BIOINFORMATICS TECHNOLOGIES

Unit 4

Modelling for Bioinformatics

Modelling for Bioinformatics : Hidden Markov Modelling for Biological Data Analysis - Comparative modelling - Probabilistic Modelling - Molecular Modelling

* Modelling for Bioinformatics

→ Modeling - creating representations, often mathematical, to understand specific scientific phenomena.

→ In bioinformatics, modelling simulates biological processes.

→ Important modelling approaches used are :

1. Hidden Markov Model (HMM)
2. Comparitive Modelling
3. Probabilistic Modelling
4. Molecular Modelling

* Hidden Markov Model for Biological Data Analysis

→ HMM is a dynamic statistical profile built from the analysis of a training dataset.

→ In an HMM, some information is hidden. For example, while analyzing DNA sequences, one may not know where a gene

starts or ends. HMM can help infer these hidden parts.

→ HMM models sequences as states and the transitions between them, which can be visualized as a finite state machine.

→ Probabilities are then assigned to each state (emissions) and between states (transitions).

→ There are 3 primary roles of HMM in biological sequence analysis. These are:

(i) Sequence Identification

(ii) Sequence Classification

(iii) Generation of multiple Alignments

A. Sequence Identification

Terminology

1. Consensus - A consensus sequence is derived from multiple sequence alignments and summarizes the most frequently occurring nucleotides / amino acids.

Example of Consensus

Consider 3 aligned sequences:

Seq 1 : A T G (C) G (T)

Seq 2 : A T G A G (T)

Seq 3 : A T G (C) G C

consensus =

A T G C G T

2. Conserved Elements : DNA sequences that remain mostly unchanged across different species, indicating they have important functions (eg. TATA Box)

3. Non-Conserved Elements : DNA sequences that vary across species, often related to evolutionary changes or gene regulation.

4. Codon : A sequence of three nucleotides (DNA or RNA bases: A, T/U, and G) that together code for a specific amino acid. Each codon instructs the cell to add a specific amino acid to a growing protein chain or to stop building the protein

5. Start Codon : The first codon in mRNA that signals the beginning of protein synthesis, usually AUG (methionine)

6. Exon : Parts of a gene that contain information to make proteins.

7. Intron : Non-coding regions in a gene that are not involved in protein creation.

8. Stop Codon : Codons like UAA, UAG, UGA that signal the end of protein synthesis.

9. Intergenic Region : DNA sequences between genes that do not code for proteins, sometimes involved in gene regulation.

10. TATA Box : A conserved DNA sequence, that is a consistent marker found in the promoters of many genes.

→ In HMMs, the TATA box is a key feature for identifying gene start sites, helping the model recognize where transcription likely begins. (3'- TATAA - 5')

11. Promoter : A specific region of DNA located just before a gene. Its main role is to control the start of transcription, which is the process where DNA is copied into messenger RNA (mRNA), so the gene can eventually be made into a protein.

* Key characteristics of HMM Models for Sequence Identification

→ HMM can be used to identify sequences (target) from the background of biological data.

→ HMM model aims to match the sequences in the DNA, that are identical to the consensus

→ This is done by assigning probabilities to each DNA base, and then calculating how likely a given sequence is to match our target.

→ Biological data can vary, so HMMs include insert and delete states to handle gaps. This means that the model can ignore extra or missing parts in the DNA sequence.

Main states - capture the core, consistent parts

Insert states - allow for extra bits that may not always match

Delete states - handle missing parts of the sequence.

→ To improve accuracy, HMMs used scoring to determine how closely a sequence matches a target. One scoring method is Log-odds:

$$\text{Positional score} = \ln\left(\frac{\text{weighting}}{0.25}\right)$$

* Types of Target Sequences that HMMs attempt to identify

→ The form, length and location of the target sequence will define the structure of the HMM used for its identification.

→ There are 2 kinds of target sequences.

ID relatively large regions of genetic code

eg. gene promoter region

ID a small subsequence

eg. basal promoter element within a gene sequence.

Regional Identification & Specific Short Sequence Identification

Example for Regional Identification - Identify the occurrence of a specific short sequence in the DNA motif 'ATG' within a set of DNA sequences using HMM. The sample sequences are:

GCTAGTACATGCGTA , TACGATGCTAG, ATGCCGTA

Step1 : Encode the sequences

A = 0

T = 1

G = 2

C = 3

Step2 : Encode the sample sequences:

S_1 : G C T A G T A C A T G C G T A

[2 3 1 0 2 1 0 3 0 1 2 3 2 1 0]

S_2 : T A C G A T C G C T A G

[1 0 3 2 0 1 3 2 3 1 0 2]

S_3 : A T G C C G T A

[0 1 2 3 3 2 1 0]

Step 3 : Define the HMM structures

S_1: not in target motif

S_2 : in target motif

Step 4 : Define the transition probabilities and emission probabilities

Transition $P(S_1 \to S_2)$ $P(S_2 \to S_1)$ $P(S_1 \to S_1)$ $P(S_2 \to S_2)$

Emission $P(A|S_1)$ $P(T|S_1)$ $P(G|S_1)$ $P(C|S_1)$

$P(A|S_2)$ $P(T|S_2)$ $P(G|S_2)$ $P(C|S_2)$

Step 5 : Train the HMM

→ use the Baum-Welch algorithm

Step 6 : Model Evaluation

→ Use a separate validation set of sequences to evaluate the model.

→ Check accuracy of identified motif, compared to known occurrences

Step 7 Sequence Identification

→ Apply the Viterbi algorithm to the test sequences to determine the most likely state sequence, identifying where 'ATG' occurs

Step 8 - Interpret Results

eg. S_1 : ATG at index 6 S_3: ATG at index 0

S_2 : ATG at index 3

* Scoring Scheme in HMM for Sequence Identification

→ Used to identify the most likely sequence of 8 bases from a longer DNA sequence using a scoring scheme based on a HMM.

Algorithm

START:

While not at end of file

For each base

Get substring of 8 bases

(i) Calculate 5-base cumulative score

(ii) Calculate highest score of last 3 bases

(iii) Add the two scores together

Return the highest scoring sequence

END.

Example on Scoring Schemes

Given the input sequence:

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

and the scoring scheme

A = 1 T = 2 G = 3 C = 4

calculate the score of the most likely sequence

① A T G C G T A C
1 2 3 4 3 2 1 4

First 5 : 13

Last 3 : 4 score = 17

② T G C A G T A C
2 3 4 1 3 2 1 4

First 5 : 13

Last 3 : 4 score = 17

③ G C G T A G C T
3 4 3 2 1 3 4 2

First 5 : 13

Last 3 : 4 score = 17

: continue for all substrings

ID sequence w/ max. score.

If a seq length < 8

eg. T G C G T A

First 5 : T G C G T

Last 3 : G T A

⇒ then compute scores

* Co-Location Algorithm for Short Sequence Identification

→ The co-location algorithm can be applied if the target sequence is located in known proximity of another highly conserved sequence.

Steps

1. Gain a clear understanding of the target for identification in terms of:

   (i) Length

   (ii) Composition

   (iii) Location

   (iv) Deviation Trends

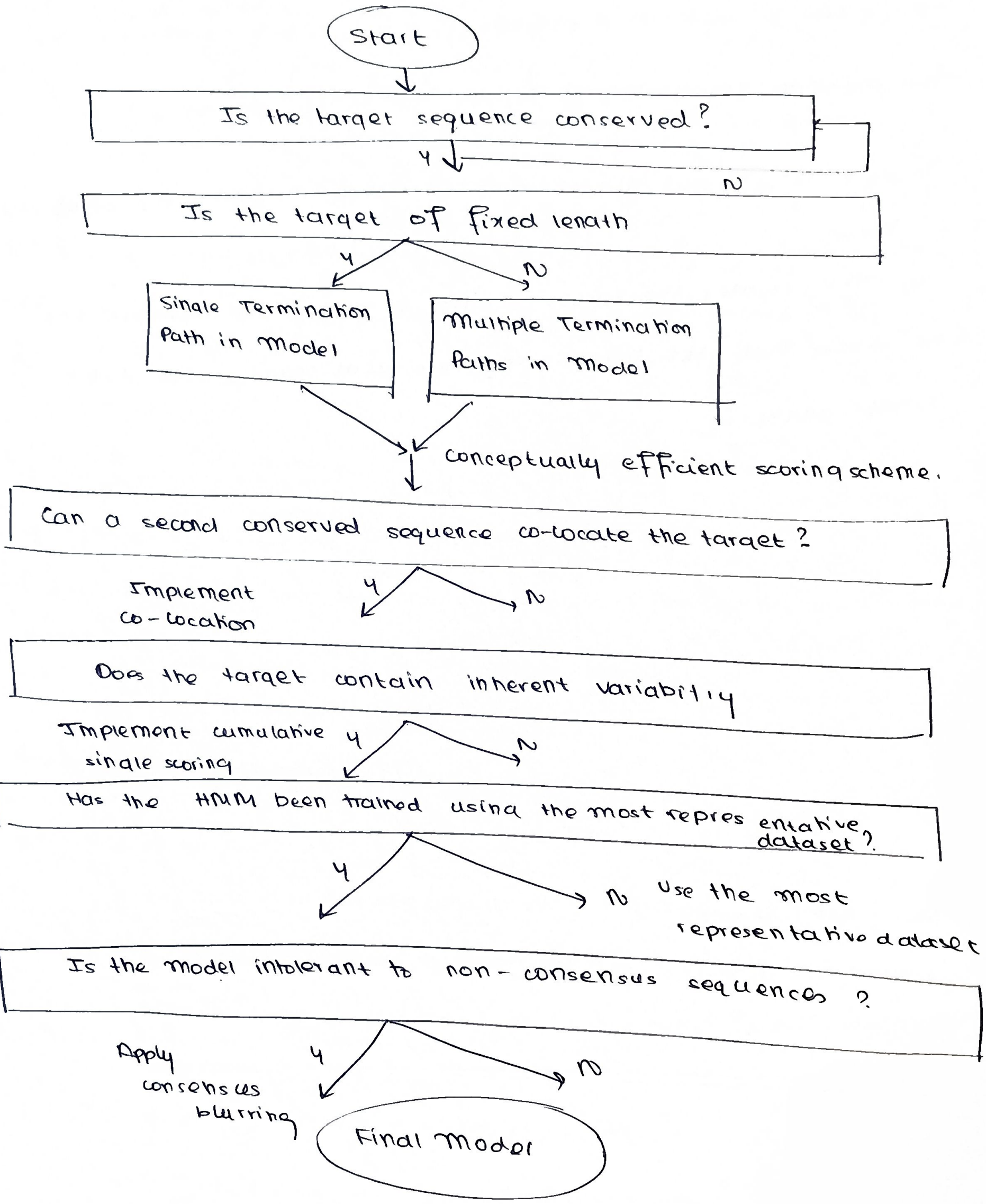
2. Use the most representative dataset possible

3. Account for model length variability

4. Implement cumulative single-scoring - calculate a score for each base in the context of its neighboring bases.

5. Implement consensus blurring - allow for some variability in the consensus sequence - accomodating natural biological variation

Flowchart for co-location algorithm



* Limitations of HMM for Sequence Identification

→ limited to a small alphabet {A, C, T, G} → more need paths for accurate recognition of target matching sequences

→ 25%. Likelihood of random occurrence matching the consensus base

→ The greater the target length & larger the alphabet, the less likely that the target sequence will occur in the data under analysis.