

Hidden Markov Models for Database Similarity Search

Higgins&Talyor Ch4 (Overview)

1. Introduction
2. Using profiles and profile-HMM databases
3. Using PSI-Blast
4. Using HMMER
 - (a) Making the first alignment
 - (b) Making a profile-HMM from an alignment
 - (c) Finding homologues and extending the alignment
5. False positives and Validating a profile-HMM search
6. Practical issues of the theories behind-HMMs
 - (a) Overview of profile-HMMs
 - (b) Statistics for profile-HMM
 - (c) Profile-HMM construction
 - (d) Priors and evolutionary information

Introduction

A HMM is used to model a set of sequences that belongs to the same families. We have already seen a method called *profile* where a prob. for each AA at a different position is modeled. A profile-HMM is a generalized, more mathematically convenient model which can effectively model a set of seqs of different length.

There are packages for HMM such as HAMMER, PSI-BALST, SAM, Meta-MEME, PROBE, etc. This chapter explains profile-HMM with HAMMER.

Hidden Markov Model

A Hidden Markov Model (HMM) is defined by an alphabet Σ , a set of hidden states Q , a matrix of state transition prob A , and a matrix if emission prob E , where

- Σ is an alphabet of symbols,
- Q is a set of states that emit symbols from Σ ,
- $A = (a_{kl})$ is a $|Q| \times |Q|$ matrix of state transition prob.,
- $E = (e_k(b))$ is a $|Q| \times \Sigma$ matrix of emission prob.

Using Profiles and Profile-HMM Databases

There are several pre-made profile-HMMs, previously constructed by another group to highlight regions of your protein, which have homology to already well characterized domains.

Among them are Pfam, Prosite profiles, SMART, PRINT, and BLOCKS.

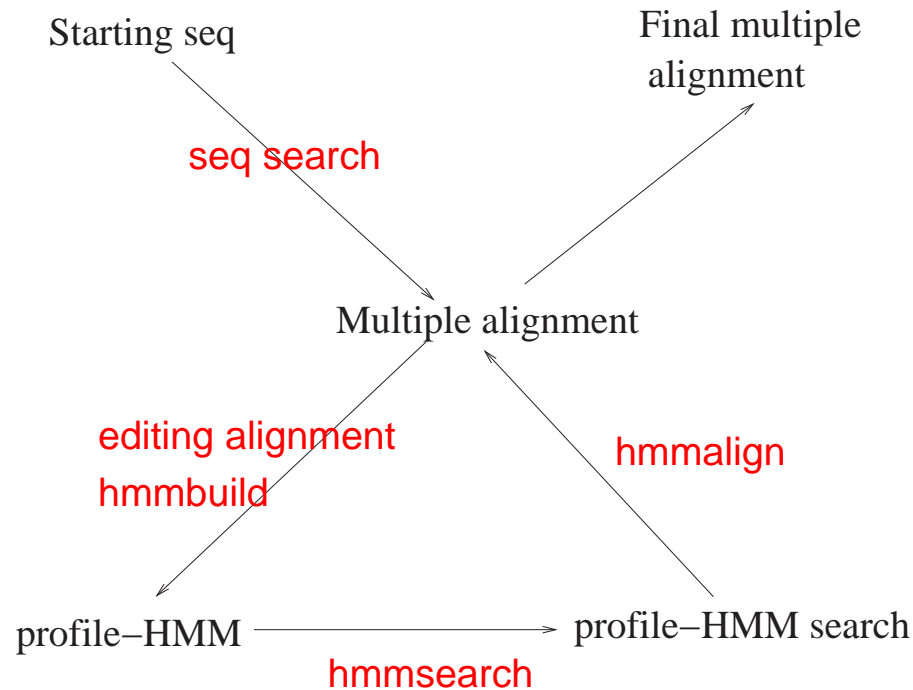
Limitations of profile-HMM databases

1. The domains in your protein should be present in the db.
2. It is possible that the start/end positions or misclassified regions can be reported.

Using HMMER

We will go thru how to use the Hammer package strating from a query seq.

As we discussed several times, it is genrally assumed that the multiple seq. alignment already has been generated. While generating and editing the multiple alignment, human expert knowledge can make a significant difference.



Making a Profile-HMM from An Alignment

You can build a profile-HMM by typing `hmmbuild`, which will execute the following.

1. Tree weighting
2. Dirchlet mixtures
3. The esimation of the effective number of seqs
4. Inertion or align state

Adjusting parameters is difficult even for experts. However, there are options for local/global HMMs. Global mode is a better choice but, at the start of your analysis, you may be not sure of boundaries foa domain. Use `hmmbuild -f` for local alignment.

Having made the profile-HMM, you need to calibrate it by using `hmmcalibrate`.

Finding homologues and extending the alignment.

Once you build a HMM (using `hmmbuild`, you can search a set of seqs with the HMM using `hmmsearch`.

There are two different matches, *per-domain* and **per-sequence**. In case that a seq has multiple domains, it is possible that each of the domain scores is not significant but, when combined, the match as a whole can be significant.

Once you find new sequences that belong to the current HMM, you need to include the sequences into the HMM, generating a new HMM. This can be done with `hmmalign`.

False Positives and Validating a profile-HMM Search

Iterative procedures like PSI-BLAST and the HMM procedure (that we just described) can lead to produce false positives.

1. See whether all the seqs can be connected together by significant scores.
2. Quote the significance of the *new* seqs for the first time they provided a significance score against the profile-HMM.
3. To show that A is related to B, show that by starting either A or B one can generate a profile which finds the other seq using criterion in the above step.
4. Biological justification.

Overview of Profile-HMMs

Two types of gaps: one should neither assume that a particular seq has all the conserved positions in a profile, nor that a particular seq will not introduce additional AAs between two conserved positions in the profile.

Because of these possible gaps, HMM has three different types of states which model the match, insertion, and deletion. Thus a profile-HMM can be seen as an extended version of a profile.

Statistics for profile HMM

There are two different scores that denote the significance of a seq match given a HMM.

- **The bit score** is one to compute the log-likelihood ratio of HMM to a random AA model with the log of 2. HMM need to incorporate *prior information*. The author suggest the minimum significance score of 25 bits.
- **E-value** provides a frequentist statistic. BLAST assumes that for a particular way of making a profile, all profiles have the same EVD parameters, regardless of content. HMMER uses a separate calibration step, where the profile-HMM is compared to a large random database, and an EVD is fitted to the resulting distribution.

Prior and Evolutionary Info

One of our goal is find a new subfamily which is not present in the current HMM. To do this, we need to incorporate extra knowledge into the process of estimating an HMM, *prior information* indicating that it is known before any seq data is seen.

- Sequece weighting
- The undersampling problem and Dirchlet mixtures
- The number of effective seqs.
This is about how to balance the info. from the Dirchlet mixtures (the prior) with the info. from the observed multiple alignment.