Profile Hidden Markov Models are used to classify Proteins into Families. Proteins are organized into families represented by multiple sequence allignments. Classifying them into families is important, since conserved domains of set families also point to their functions. Paired Similarity Comparisons such as BLAST searches fail to detect the subtle similarities. PHH find those similarities.

→ a proﬁle HMM is a statistical model of a multiple sequence alignment, where

probabilities are assigned to each amino acid at each position in the alignment, and to the transitions between positions.

→ pHHM have formal probabilistic basis and a consis-tent theory behind insertion and gap scores

→ Nature uses protein domains as building blocks, shuffling them to create multi-domain proteins. Domains are conserved part of the protein that have a function and can work independently (often). It makes more sence to calssify the domains not entire proteins into families.

→ MSA: In a multiple sequence alignment, several sequences are ﬁtted on top of each other,such that the amino acids placed above each other in the alignment are as similar as possible. MSAs allow representation inserts and deletions (dashes).

→ Sequence similarties between proteins of the same domain might be really low (sequence of RNA-polymerase domain are highly diverse). So traditional BLASt comparisons will fail here.

→ we construct alignments to break proteins into domains for a given family. We construct profiles for each domain and we compare new sequences against that domain-profile.identify common sequence characteristics

→ We have MSA and theta (fraction of maximum inserts), pseudocount as Input.

→ The used MSA consists of known already known members of the protein family ( for example RNA-polymerase). We construct our profile for that known protein family based on the MSA.

→ GOAL: be used to locate struc-turally or functionally important residues, since they are conserved in the sequences

and consequently receives high probabilities in the HMM.

→ output: transition and emission matrices of the profile HMM.

→ We use pHMM to decide if our new unknown protein belongs to the protein family depicted by the profile (is it an RNA-Polymerase?).

→ We remove columns where the fraction of insertions exceeds a threshold theta. We get a MSA where each column has a relativly low number of insertions.

→ We construct a profile of that allignment.

→ each node has an emission propability corresponding to all possible proteins possible.

→ emission states: We define the emission propability as the count of the emitted aminoacids from a specific state/ column/position.

→ transitions states: We generate hidden paths for each row (msa) in the profile. Transition state „from“ a state should add up to 1, except end state which doesnt have any. Count number of states (number of sequences in msa) and divide them with number of sequences in that state. If I have 10 sequences and 5 go to match, 3 to insert and 2 to delete at position M4, than the propability of match at M4 is 5/10 = 50 %.

→ All Amino acids are represented. If an amino acid did not appear in our MSA, then if will be represented by a pseudocount with a small propability (1/20). Its never 0 propability.

A proﬁle HMM consists of a collection of states of three kinds (Figure 9): match

states, which correspond to the positions in the consensus sequence, insert states,

which model insertions with respect to the consensus, and delete states, which repre-

sent deletions with respect to the consensus.

→ Start and End states also exist. End states are unique to pHMM, since traditional HMM dont have an endstate.

→ Match states: These are transitions between positions where our sequence matches something in our model. The match states represent the conserved regions. Eachz Match state will represent 1 Amino acid, since we can emmit only a single symbol. Match state bildet die Konsensussequenz ab und die Inserts und deletes werden durch andere states abgebildet.

→ Insertionsstates model inserts and are the only once that loop with themselfs. Meaning the pHH is unidirectional otherwise. We always move forward from start state to end state with no returns to an old state. There are other models which allow going back. They allow repeat alignments. We can use the frequency of amino acids in nature for the insert propability if msa is not given for a specific amino acid.loops allow for arbitrary length insertionsrelative to the consensus sequence

→ Deletionsstates also exist. They do not have emission states.

→ We use vitebri Algorithm to find optimal hidden path. If propabilty threshold is satisfied, we classify the unknown protein as belonging to the family. Number of rows in virtebri diagram corresponds to number of emitted states in HMM diagram.

→ Pfam is a commonly used Profile Hidden Markov Model Database. Here we use custom once.

→ disadvantage: We can start wil an insert. We have to start with a match state. Meaning we always start with a conserved region.