WEEK 7 ASSIGNMENT

Data Systems in the Life Sciences – BIOL 51000 | Fall 2 2020

Christina Morgenstern

MARKOV CHAINS AND THEIR USE IN BIOINFORMATICS

1. Introduction to Markov chains

In probability theory individual trials are processed independently from each other with a sequence of chance experiments occurring with the same probability. Consider rolling a die: if it is a fair die, each of the six numbers is equally likely with a probability of 1/6. Knowing which numbers came up in previous experiments does not influence the outcomes of the next experiment. This is a simplification of the random process and might not reflect the natural dependencies.

Russian mathematician Andrey Markov challenged this assumption at the beginning of the 20th century by claiming that when we observe a sequence of chance experiments the past outcomes could influence the predictions for the next experiment [1]. Proposed by Andrey Markov, Markov chains are stochastic models that describe a sequence of possible events by only considering the previous event [2]. These models incorporate conditional probabilities for a sequence of events which assigns a probability given the case another event has occurred. Markov chains are used to study a variety of real-world phenomena such as queues or lines of customers arriving at an airport, exchange rates of currencies and weather. They are applied in economics, game theory, communication theory, genetics and finance. Google's PageRank algorithm for searching the web is also based on a Markov process.

A Markov chain is characterized by a set of possible states $S=\{s_1, s_2, ..., s_r\}$ that represent all possibilities of events that the chain can advance in. A Markov process is the movement along the chain of possible states where the process can either stay in the same state or move to some other state in in S Figure 1. Figure 1 shows a simple 2-state Markov chain with states s_1 and s_2, respectively.

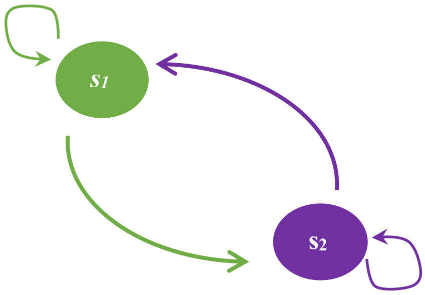


Figure 1. Diagram of Markov chain with two states (S1 and S2).

A Markov process starts in one state and moves from one state to the next in steps. The step p_{ij} denotes the probability of moving from state s_i to state s_j and is called a transition probability. The way transitions between states occur are stochastic and characterized as memoryless. All possible transitions can be visualized in a transition matrix where all transition probabilities are listed. Table 1 shows an example of a transition matrix for all possible transitions in a two state Markov chain. Every state is represented once as a row and once as a column with each cell in the matrix denoting the probability of transitioning from row state to column state. As states are added to the Markov chain, the number of cells in the transition matrix grows quadratically.

Table 1. Transition matrix for a two state Markov chain.

		Next State	
Current State		S_1	S_2
	S_1	$P(S_1 \mid S_1)$	$P(S_2 \mid S_1)$
	S_2	$P(S_1 \mid S_2)$	$P(S_2 \mid S_2)$

An example for a two-state Markov chain as depicted in Figure 1 could be a rolling die example, where one state comprises a fair die and the second state a loaded die. While the fair die produces numbers with equal probabilities i.e., 1/6, the loaded die is biased towards a number, e.g. the number 6 occurs with a probability of 1/2 while all other numbers occur with a probability of 1/10. Both dice have a certain probability of staying a fair or loaded die as well as a certain probability of switching states from fair to loaded or from loaded to fair. Once these conditional probabilities have been determined from experimental observation, the resulting Markov model can be challenged in defined settings where the probability for the occurrence of a certain sequence is calculated based on the given model.

Usually, when we talk about Markov chains, we consider discrete time because of the fixed states for the trials of the chain. Also, the conditional probabilities don´t vary with time and thus such Markov chains are described as discrete-time homogeneous Markov chains. However, there are also other types of Markov chains, that can consider continuous-time processes where the steps advance within an interval.

A Markov chain is the simplest Markov model consisting of a system with a random variable changing through time. The distribution of this variable is only dependent on the previous state which denotes the characteristic Markov property. A hidden Markov model (HMM) is a Markov chain for which the state is not directly observable (and thus hidden) but produces an observable random output according to a given stationary probability law [3]. In order to compute the probability that a given model generates a specified series of observations, several algorithms can be used. The forward algorithm, finds the most probable sequence, given a model, by starting at the beginning of the sequence. The backward algorithm, does the same but starting from the end of the sequence [3]. Both algorithms can also be used in conjunction in the forward-backward algorithm where it is used to estimate all probabilities of hidden states in a hidden Markov model for any sequence of observations. This is a dynamic programming approach and originates from Bayes' rule.

The Viterbi algorithm is another useful algorithm in decoding Markov where it can find the sequence of hidden states that as a whole has the highest probability. Proposed by Andrew Viterbi in 1967 the Viterbi algorithm is an example of a dynamic programming algorithm which can minimize the amount of computation by not computing all possible routes but the best ones for each position [4].

Markov chains and HMMs have been considered a valuable approach of modeling and studying biological sequences, such as protein and DNA sequences [5]. The technological revolution in the life sciences sector, also termed omics revolution, has generated a tremendous amount of sequence data. Computational models are needed to infer new knowledge from those data. HMMs have been applied to a variety of biological problems, such as modeling DNA sequencing errors, pairwise and multiple sequence alignment, base-calling, RNA structural alignment, protein secondary structure prediction and many more [5]. Biological sequences, nucleic acids and proteins, lend themselves for analysis with Markov chains and HMMs because they are structured as sequences with a defined set of states where common patterns, motifs and domains can be represented based on statistical properties.

2. Database search approaches using Markov chains and Hidden Markov Models

Several databases exist that use HMMs to find genes or perform protein family characterization. Pfam [6], SAM [7] and SUPERFAMILY [8] are some of the biological databases that use HMMs.

Proteins can be separated into domains which are functional regions, and a relatively small number of these domains are used by many different proteins. The diversity of proteins found in nature is dependent on a unique combination of those domains. Proteins can share similar domains and thus functionality. Traditional database searches using the BLAST algorithm e.g., on GenBank or Swissprot, can generate a vast amount of hits that are difficult to interpret. Using protein families in the database search can aide in producing more satisfactory results. Databases based on protein families use multiple sequence alignments of known family members whereby conserved features are more easily recognized and awarded with a higher weight than distant similarities.

The protein families database Pfam constitutes a large collection of protein families that are represented by multiple sequence alignments and hidden Markov models (HMM) [9][10]. Within the database, similar proteins are grouped into profile-HMMs which are used for searching and should facilitate the interpretation of the results as well as perform more prone to conservation and insertions/deletions [6].

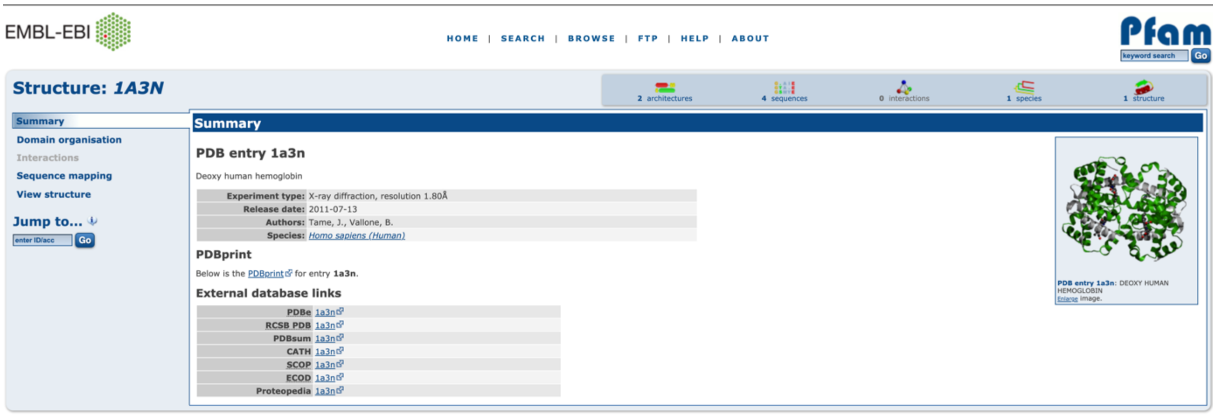
The data within Pfam comprises of three files: seed alignment, full alignment and HMM-profile. The seed alignment is a multiple alignment of a set of sequences that has been manually verified. To perform the database searching, an HMM-profile is built based on the seed alignment. By searching Swissprot, a full alignment is produced by aligning the HMM-profile to all detectable members. Having these different kind of alignments makes sure that the database is updated. For any new protein sequence, full alignments and HMM-profiles are generated automatically, whereas seed alignments provide stable resources within the database [6]. When searching Pfam with a protein sequence, the result displays the family name, a permanent accession number and a record of methods that have been used to create the alignments for family identification. Furthermore, descriptions of the function and the domain structure and links to other databases are given. The generation of Pfam families is an iterative process that makes sure that at each step the quality requirements are met. When generating Pfam families the search is cross-referenced to other databases, such as Prosite and Prints entries. Prosite is a database of protein domains, families and functional sites [11] and Prints is a compendium of protein fingerprints [12]. Pfam entries can be grouped together into Clans, which are large and divergent superfamilies where a single profile HMM can’t capture the diversity of the sequence members [13].

The Pfam database is available via the website https://pfam.xfam.org and in the release version of 32.0 covers a total of 17,929 different protein families [9]. Each Pfam entry has the following tags that are characteristic of the functional unit: domain, family, repeat, motif, coiled coil or disordered. A domain is characterized as a collection of related sequence regions forming a structural unit. A family denotes a collection or related sequence regions containing one or more domains. Repeats are short units that by itself is unstable but in multiple arrangement can form powerful structures. Motifs are sequences that have a distinct role e.g. transcription factor binding. Coiled-coil regions are structures based on alpha-helices whereas disordered regions don't have any specific structure but are conserved in their sequence [14]. Clans denote related Pfam entries that have been grouped together based on sequence similarity, three-dimensional structure similarity, functional similarity of similarity based on the profile HMMs [14].

Using the Pfam database one can search for protein or DNA sequences against the Pfam models, browse the families and clans, retrieve text annotation about any given family/entry, view multiple sequence alignments of a family or clan, view relationships between families in a clan, see protein structure information in the context of a family, view families according to their taxonomic spread, search the database by keywords [15].

On the Pfam homepage, the database can be searched in different ways. These include, searches using a protein name or accession number, a Pfam family name of Pfam accession number, a clan name or accession, a PDB accession or certain keywords [16].

In the following, I am giving an example for a Pfam search using human hemoglobin. Entering the PDB accession number of human hemoglobin (1A3N) into the JUMP TO box on the Pfam start page and pressing Go leads to the summary result shown in Figure 2 with links to external databases.



Structure: *1A3N*

Summary

PDB entry 1a3n

Deoxy human hemoglobin

Experiment type:	X-ray diffraction, resolution 1.80Å
Release date:	2011-07-13
Authors:	Tame, J., Vallone, B.
Species:	*Homo sapiens (Human)*

PDBprint

Below is the PDBprint for entry **1a3n**.

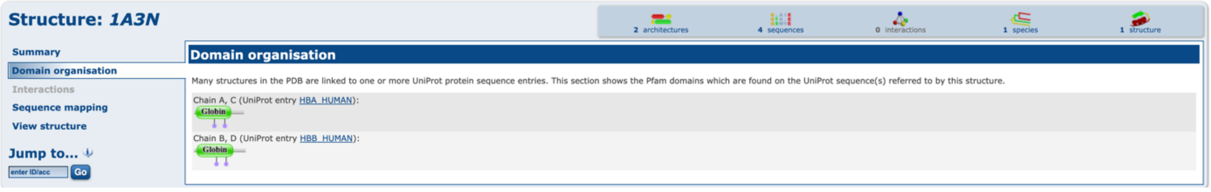
External database links

PDBe	1a3n
RCSB PDB	1a3n
PDBsum	1a3n
CATH	1a3n
SCOP	1a3n
ECOD	1a3n
Proteopedia	1a3n

PDB entry 1a3n: DEOXY HUMAN HEMOGLOBIN

Figure 2. Pfam search result for human hemoglobin.

The result of the domain organization highlights two domains, referring to the globin domain architecture characteristic to the four chains within hemoglobin Figure 3.



Structure: *1A3N*

Domain organisation

Many structures in the PDB are linked to one or more UniProt protein sequence entries. This section shows the Pfam domains which are found on the UniProt sequence(s) referred to by this structure.

Chain A, C (UniProt entry HBA_HUMAN):

Chain B, D (UniProt entry HBB_HUMAN):

Figure 3. Result of domain organization for human hemoglobin.

Further information on the globin structure can be found when clicking on the domain organization entry. Globins belong to the superfamily of heme-containing globular proteins that are essential in binding oxygen. Investigating the nature of the clan, shows that the globin chains of hemoglobin are a member of the clan Globin (CL0090) which is an evolutionary conserved six helical fold found in bacteria and eukaryotes. Seven other members belong to this clan. The profile HMM can also be visualized under the HMM logo tab on the left. HMM logos give an overview of the properties of an HMM in a graphical manner with the height of the letter stack at a particular position denoting the degree of conservation at that position and the height of an individual letter within the stack demonstrating the frequency of that letter at that position [17]. Evolutionary relationships of the globin family's seed alignment can be viewed under the Tree section which displays the phylogenetic tree. Details of the Globin family (PF00042) can be extracted from the Curation & model tab.

3. Markov chains and Hidden Markov Models (HMMs) for multiple sequence alignment

Another major bioinformatics application for Markov chains and HMMs is multiple sequence alignment which can assess if a novel sequence is part of a family of homologous sequences. The goal of multiple sequence alignment is finding a consensus sequence or a profile. Whereas the consensus sequence finds for each position the residue that occurs most frequently for the alignment, the profile assigns for each position a set of scores for any residue to occur. The use of profile HMMs are a way of implementing HMMs together with the notion of a profile. As in part 2 of this paper described the profile HMM provides a good technique to deal with multiple sequence alignment. While in normal profile analysis, the scores are given heuristically, HMMs derive the scores from statistical models [3].

Figure 4 compares different consensus modeling methods from simple patterns to an HMM [18]. The top part of the figure shows a consensus sequence derived from a multiple sequence alignment with the conserved columns and insertions. A sequence profile is an extension of this alignment taking into account variable amino acids scores and variable gap penalties at each consensus position. An HMM is an extension to the profile by replacing the arbitrary scores with probabilities giving rise to a probabilistic model that considers mismatches, insertions and deletions with respect to the consensus [18].

The HMMER database performs sequence searching and sequence alignments based on profile HMMs [18][19]. This tool is often used together with Pfam as stated in part 2 of this paper but can also work in isolation like BLAST. While BLAST and FASTA use position-independent substitution score matrices such as BLOSUM and PAM, HMMER implements HMMs. HMMER can address problems of single sequence queries as well as multiple sequence alignments using probabilistic inference methods based on profile HMMs. Position-specific probabilistic modelling of the alignment which incorporates residue conservation and rates for insertions and deletions account for the sensitivity of profile HMMs. The third generation of the HMMER tool suit has achieved a great reduction in computation overload owing for the growing success of this tool for building profile HMMs [18].

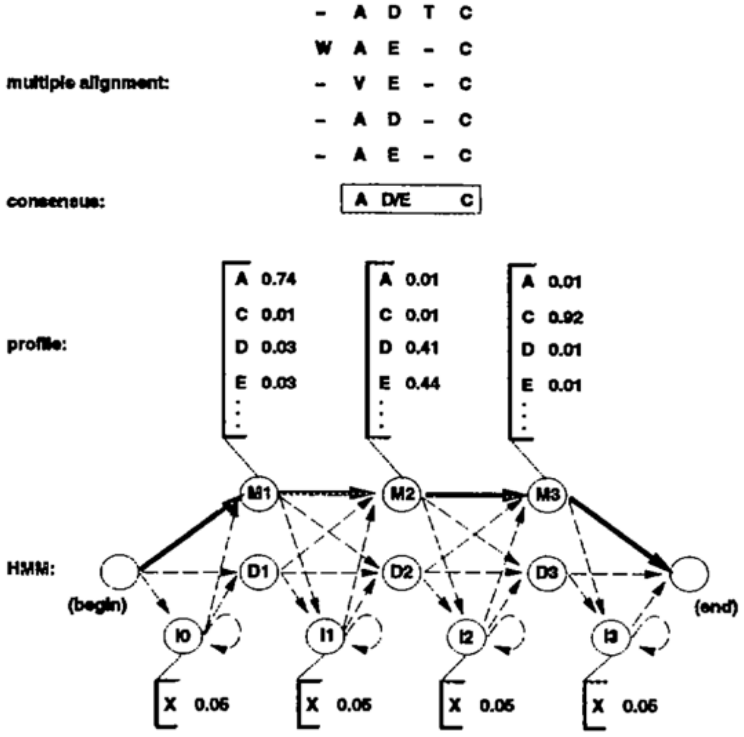


Figure 4. Consensus modeling from simple patterns to an HMM (taken from [18])

Bibliography

[1] A. A. Markov, ‘An Example of Statistical Investigation of the Text *Eugene Onegin* Concerning the Connection of Samples in Chains’, *Sci Context*, vol. 19, no. 4, pp. 591–600, Dec. 2006, doi: 10.1017/S0269889706001074.

[2] ‘Markov chain’, *Wikipedia*. Nov. 29, 2020, Accessed: Dec. 11, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Markov_chain&oldid=991285685.

[3] V. De Fonzo, F. Aluffi-Pentini, and V. Parisi, ‘Hidden Markov Models in Bioinformatics’, *CBIO*, vol. 2, no. 1, pp. 49–61, Jan. 2007, doi: 10.2174/157489307779314348.

[4] A. Viterbi, ‘Error bounds for convolutional codes and an asymptotically optimum decoding algorithm’, *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967, doi: 10.1109/TIT.1967.1054010.

[5] B.-J. Yoon, ‘Hidden Markov Models and their Applications in Biological Sequence Analysis’, *Curr Genomics*, vol. 10, no. 6, pp. 402–415, Sep. 2009, doi: 10.2174/138920209789177575.

[6] E. L. Sonnhammer, S. R. Eddy, E. Birney, A. Bateman, and R. Durbin, ‘Pfam: multiple sequence alignments and HMM-profiles of protein domains.’, *Nucleic Acids Res*, vol. 26, no. 1, pp. 320–322, Jan. 1998.

[7] K. Karplus, C. Barrett, and R. Hughey, ‘Hidden Markov models for detecting remote protein homologies’, *Bioinformatics*, vol. 14, no. 10, pp. 846–856, 1998, doi: 10.1093/bioinformatics/14.10.846.

[8] J. Gough, ‘The SUPERFAMILY database in structural genomics’, *Acta Crystallogr D Biol Crystallogr*, vol. 58, no. Pt 11, pp. 1897–1900, Nov. 2002, doi: 10.1107/s0907444902015160.

[9] S. El-Gebali *et al.*, ‘The Pfam protein families database in 2019’, *Nucleic Acids Res*, vol. 47, no. D1, pp. D427–D432, Jan. 2019, doi: 10.1093/nar/gky995.

[10] E. L. L. Sonnhammer, S. R. Eddy, and R. Durbin, ‘Pfam: A comprehensive database of protein domain families based on seed alignments’, *Proteins: Structure, Function, and Bioinformatics*, vol. 28, no. 3, pp. 405–420, 1997, doi: https://doi.org/10.1002/(SICI)1097-0134(199707)28:3<405::AID-PROT10>3.0.CO;2-L.

[11] ‘ExPASy - PROSITE’. https://prosite.expasy.org/ (accessed Dec. 13, 2020).

[12] ‘PRINTS’. http://130.88.97.239/PRINTS/index.php (accessed Dec. 13, 2020).

[13] EMBL-EBI, ‘What is Pfam? | Pfam’. https://www.ebi.ac.uk/training-beta/online/courses/pfam-quick-tour/what-is-pfam/ (accessed Dec. 13, 2020).

[14] EMBL-EBI, ‘Pfam families and clans | Pfam’. https://www.ebi.ac.uk/training-beta/online/courses/pfam-quick-tour/what-is-pfam/pfam-families-and-clans/ (accessed Dec. 13, 2020).

[15] EMBL-EBI, ‘What can I do with Pfam? | Pfam’. https://www.ebi.ac.uk/training-beta/online/courses/pfam-quick-tour/what-is-pfam/what-can-i-do-with-resource-name/ (accessed Dec. 13, 2020).

[16] EMBL-EBI, ‘Getting started with Pfam | Pfam’. https://www.ebi.ac.uk/training-beta/online/courses/pfam-quick-tour/getting-started-with-pfam/ (accessed Dec. 13, 2020).

[17] T. J. Wheeler, J. Clements, and R. D. Finn, ‘Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models’, *BMC Bioinformatics*, vol. 15, no. 1, p. 7, Jan. 2014, doi: 10.1186/1471-2105-15-7.

[18] S. R. Eddy, ‘Multiple Alignment Using Hidden Markov Models’, p. 7.

[19] S. C. Potter, A. Luciani, S. R. Eddy, Y. Park, R. Lopez, and R. D. Finn, ‘HMMER web server: 2018 update’, *Nucleic Acids Res*, vol. 46, no. W1, pp. W200–W204, Jul. 2018, doi: 10.1093/nar/gky448.

[20] ‘HMMER’. http://hmmer.org/ (accessed Dec. 13, 2020).