**Implementing Gaussian Naïve Bayes for protein Secondary Structure prediction**

**Rahul Alapati**1,\*,+**, Sritika Chakladar**1,+**, Anuj Gupta**1,+**, Vineet Nayak**1,+ **and Austin Ream**2,+

1Graduate Student, Computer Science and Software Engineering, Auburn University, 36830, USA 2Undergraduate Student, Computer Science and Software Engineering, Auburn University, 36830, USA

\* rza0037@auburn.edu

+these authors contributed equally to this work

**ABSTRACT**

Protein secondary structure predictions from its amino acid sequence is useful in the study of protein functions. Therefore, it is vital to accurately predict the secondary structure of a protein. Gaussian Naïve Bayes learning algorithm is a well-known algorithm for classification. Based on Bayes theorem, it predicts conditional probabilities for each class. The class with the highest probability is considered as the most likely class by the classifier. In this project, we explore the probabilistic nature for protein secondary structure using Naïve Bayes algorithm. Our approach models the Gaussian Naïve Bayes learning algorithm to train and model a classifier that predicts the 3 class labels for secondary structure. We use this model to predict the class labels on any given test dataset by selecting the most probable class assignment.

**1. Introduction**

Protein structure is the three-dimensional arrangement of atoms in an amino acid-chain molecule1. There are different distinct levels of protein structure. The [primary structure](https://en.wikipedia.org/wiki/Primary_structure) of a protein refers to the sequence of amino acids in the polypeptide chain. Tertiary structure refers to the three-dimensional structure of monomeric and multimeric protein molecules. Secondary structure refers to highly regular local sub-structures on the actual polypeptide backbone chain.

Protein folding is the physical process by which a polypeptide folds into its characteristic and functional three-dimensional structure from random coil3. The primary structure of a protein, its linear amino-acid sequence, determines its native conformation4. The specific amino acid residues and their position in the polypeptide chain are the determining factors for which portions of the protein fold closely together and form its three dimensional conformation. Formation of a secondary structure is the first step in the folding process that a protein takes to assume its native structure. Characteristic of secondary structure are the structures known as α-helices and β-sheets2 that fold rapidly because they are stabilized by intramolecular hydrogen bonds. α-helices are formed by hydrogen bonding of the backbone to form a spiral shape. The β pleated sheet is a structure that forms with the backbone bending over itself to form the hydrogen bonds.

Dictionary of Protein Secondary Structure (DSSP) is the most widely used tool to standardize secondary structure assignment. The DSSP program was designed by Wolfgang Kabsch and Chris Sander. It is a database of secondary structure assignments for all protein entries in the Protein Data Bank (PDB). It also calculates DSSP entries from PDB entries.

Prediction of a secondary structure of a protein from its amino acid sequence is an important and challenging work. Not only can successful predictions provide a starting point for direct tertiary structure modelling, but they can also significantly improve sequence analysis and sequence-structure threading for aiding in structure and function determination7.

Naive Bayes algorithm is one of the most efficient and effective inductive learning algorithms for machine learning and data mining. Its competitive performance in classification is surprising, because of the conditional independence assumption on which it is based5. It is based on Bayes theorem and it predicts conditional probabilities for each class such as the probability that given record or data point belongs to a particular class6.  The class with the highest probability is considered as the most likely class.

In this project, we explore the probabilistic nature for protein secondary structure using Naïve Bayes algorithm. Our approach models the Gaussian Naïve Bayes learning algorithm to train and model a classifier that predicts the 3 class labels for secondary structure. We use this model to predict the class labels on any given test dataset by selecting the most probable class assignment.

**2. Methods**

The protein sequences in Multi-FASTA format and the true 3-class secondary structure (SS) labels (‘H’, ‘E‘, ‘C’) of the proteins have been used to implement the Gaussian Naïve Bayes learning algorithm on training set and predict the class labels on test dataset by selecting the most probable class assignment.

The input protein sequences and their respective RSAs have been curated into non-overlapping sets Training (75%) and Test (25%) datasets using simple random sampling.

**2.1. Subsection**

Example text under a subsection. Bulleted lists may be used where appropriate, e.g.

• First item

• Second item

***2.1.1. Second-level section***

Here goes the text.

**3. Results**

Topical subheadings are allowed. Authors must ensure that their Methods section includes adequate experimental and characterization data necessary for others to reproduce their work.

**Table 1.** Legend (350 words max). Example legend text.

|  |  |  |
| --- | --- | --- |
| Condition | n | p |
| A | 5 | 0.1 |
| B | 10 | 0.01 |



**Figure 1.** Legend (350 words max). Example legend text.

**4. Discussion**

The Discussion should be succinct and must not contain subheadings.

**5. References**

1. Wikipedia contributors. (2018, March 21). Protein structure. In Wikipedia, The Free Encyclopedia. Retrieved 19:23, March 26, 2018, from <https://en.wikipedia.org/w/index.php?title=Protein_structure&oldid=831674619>
2. Pauling, L., Corey, R. B., & Branson, H. R. (1951). The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proceedings of the National Academy of Sciences, 37(4), 205-211.
3. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). The shape and structure of proteins.
4. Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. Science, 181(4096), 223-230.
5. Zhang, H. (2004). The optimality of naive Bayes. AA, 1(2), 3.
6. Schmidler, S. C., Liu, J. S., & Brutlag, D. L. (2000). Bayesian segmentation of protein secondary structure. Journal of computational biology, 7(1-2), 233-248.
7. Robles, V., Larrañaga, P., Peña, J. M., Menasalvas, E., Pérez, M. S., Herves, V., & Wasilewska, A. (2004). Bayesian network multi-classifiers for protein secondary structure prediction. Artificial Intelligence in Medicine, 31(2), 117-136.