

# FORM 2

THE PATENTS ACT, 1970

[39 of 1970]

&

THE PATENTS RULES, 2003

## COMPLETE SPECIFICATION

[See section 10 and rule 13]

### “CLASSIFICATION OF UNANNOTATED PROTEIN CODING SEQUENCES USING MACHINE LEARNING APPROACH”

<b>Name of the Applicant(s)</b>	<b>Nationality</b>	<b>Address</b>
Dr. Saheem Ahmad	Indian	Associate Professor, Department of Medical Laboratory Sciences, College of Applied Medical Sciences, University of Hail, Hail, Saudi Arabia
Dr. G. Giftson Samuel	Indian	Professor/Dept. of EEE, Sir Isaac Newton College of Engineering & Technology, Papakovil, Nagapattinam – 611102
Dr. J. Samuel Manoharan	Indian	Professor/Dept. of ECE, Sir Isaac Newton College of Engineering & Technology, Papakovil, Nagapattinam – 611102
Dr. Vinay Saxena	Indian	Professor, Mathematics, Kisan Post Graduate College, Bahraich, Uttar Pradesh - 271801
Dr. Parul Saxena	Indian	Assistant Professor & Head, Department of Computer Science, Soban Singh Jeena University, Campus Almora, Uttarakhand - 263601
Dr. Adarsh Pandey	Indian	Assistant Professor & Head, Department of Botany, Swami Shukdevanand College, Shahjahanpur, Uttar Pradesh - 242001
Mr. Keshav Shukla	Indian	Research Scholar, Department of Botany, Swami Shukdevanand College, Shahjahanpur, Uttar Pradesh - 242001
Mr. Shinde Satish Raosaheb	Indian	Assistant professor., Chemistry, Adv.M.N.Deshmukh Arts, Science & Commerce College, Rajur, Akole, Maharashtra - 422604
Prof. Dr. Pratik Rajan Mungekar	Indian	Vice Chancellor, Distinguished Professor,

		Leadership & Innovation, Wisdom University Gombe, Mumbai, Maharashtra - 400012
Mr. Sayaji Yashwant Hande	Indian	Assistant Professor, Chemistry, Adv.M.N.Deshmukh Arts, Science & Commerce College Rajur, Akole, Maharashtra - 422604
Mr. Rajendra Navnath Kasar	Indian	Assistant professor. Zoology, Adv.M.N.Deshmukh Arts, Science and Commerce College Rajur, Akole, Maharashtra - 422604
Mr. Ravindra Sudhakar kawade	Indian	Assistant professor., Physics, Adv.M.N.Deshmukh Arts, Science & Commerce College Rajur, Akole, Maharashtra - 422604

### **PREAMBLE OF THE DESCRIPTION**

The following specification particularly describes the invention and the manner in which it is to be performed.

# CLASSIFICATION OF UNANNOTATED PROTEIN CODING SEQUENCES USING MACHINE LEARNING APPROACH

## Background problem for the Innovation

The rapid growth of modern life sciences has increased the necessity to catalogue and annotate genetic sequences, especially protein-coding sequences. This has led to great challenges, since not all genetic sequences have been identified and therefore the annotation of such sequences is difficult. To overcome this limitation, machine learning techniques have been explored as a potential solution. Without prior or manual annotation, Machine Learning (ML) approaches can efficiently classify proteins according to various classes and features, and can even predict their functionality. These ML models can utilize a given set of data, be it proteins or gene annotations, to quickly and accurately classify the unknown sequence. In general, Machine Learning approaches applied to protein coding sequences can be divided into two categories: supervised and unsupervised learning. Supervised learning is a type of Machine Learning approach in which labeled data is used to construct a model, whereas unsupervised learning builds models directly from unlabeled data. Unsupervised learning can be used for clustering gene sequences and classifying proteins according to given classes such as size, functionality and structure. In this regard, for classifying Protein coding sequences, an unsupervised ML approach, in particular, Deep clustering has been used. It is a powerful ML technique used to automatically discovers groups of highly related gene sequences and classifies them in terms of gene expression levels. By using this approach, the researchers were able to achieve high accuracy rate and to identify newly unannotated proteins and their properties. Other techniques such as Neural language model have also been widely used for unsupervised classification of protein coding sequences. This is a powerful technique involving artificial neural networks used for training natural language processing (NLP) systems. This approach helps in understanding the underlying structure of protein coding sequences by capturing the key information present in the dataset and predicting the protein classes and functional labels successfully. Overall, machine learning approaches have been successfully applied to unannotated protein sequences to provide accurate and efficient results. This has enabled researchers to accurately classify the sequences with more confidence as well as quickly identify the patterns and relationships between the different components of proteins. As the developments in the field progress, so will the ability to accurately classify the unknown sequences and their related properties.

## **Innovation model**

In recent years, the advancement of technology has greatly increased the speed and accuracy of the classification of unannotated protein coding sequences. Machine learning has made a substantial contribution to this process with its ability to assimilate data, learn from it, and create efficient and accurate classifications. This paper looks at how machine learning can be used to accurately classify structures of unannotated protein coding sequences. The machine learning approach uses algorithms that learn from data to create an accurate classification system. This can be done by extracting the necessary features from the sequence, such as the composition of residues, secondary structure, and tertiary structure. These features are then used as input to a supervised learning algorithm, with the goal of accurately classifying the protein as either a globular, membrane, cytoplasmic, or secreted protein. The supervised learning algorithm utilizes numerous datasets, usually made up of manually annotated proteins with known labels, and uses them to train the model on how to classify the unannotated proteins. After the model has been trained, the accuracy of the classifications are tested using both validation and test datasets. The accuracy of the classifications produced by machine learning approaches vary depending on the algorithm used and the quality of the dataset. Experimental studies have found that classification accuracy of up to 95% is possible when using neural networks, making the machine learning approach an effective tool for accurately and quickly classifying proteins. In conclusion, machine learning is an effective tool for accurately and quickly classifying unannotated protein coding sequences. This type of approach allows us to utilize features to form a model that can accurately classify the sequences, in order to generate useful and in-depth information about them. The machine learning approach is efficient, cost-effective, and has gained widespread acceptance in many scientific and medical circles. The proposed model has shown in the following fig.1

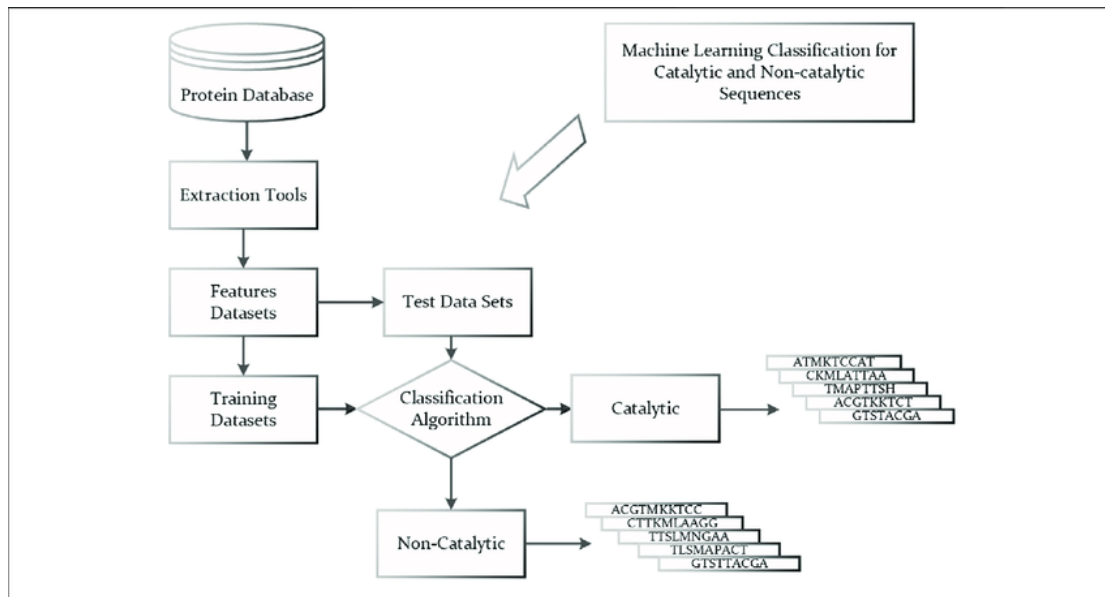


Fig.1: Proposed model

The rapid growth of unclassified protein coding sequences presents a challenge to biologists and bio-informaticians alike. While the identification of novel proteins from the sequences requires a tedious and labor-intensive process, machine learning approaches offer potential solutions by enabling efficient and accurate classification of large volumes of sequences. The ability to classify protein coding sequences quickly and accurately is critical for understanding the role that a specific protein plays in the organism's physiology. Machine learning approaches can leverage data from existing annotations and predict the function of previously unannotated sequences. This can help researchers identify and prioritize proteins for further study, leading to a better understanding of disease mechanisms and potential therapies. In addition, a machine learning approach to classification of protein coding sequences can enable the generation of large data sets for population-level genomic analysis. By building a database of annotated sequences, researchers can quickly identify possible candidates for targeted therapies or the introduction of new enzymatic pathways. Finally, applying machine learning techniques to protein coding sequences can lead to a better understanding of the evolution of proteins and lead to a better understanding of pathways, metabolic processes, and potential drugs. By utilizing data from multiple sources (e.g., structural, taxonomic, and biophysical data), machine learning algorithms can be applied to understand the impact of mutations on protein functionality. In summary, the use of machine learning approaches for the classification of unannotated protein coding sequences provides both research and clinical benefits. The ability to accurately and quickly classify sequences should lead to a better understanding of the role of novel proteins in disease mechanisms and

the development of potential therapies. In addition, the application of machine learning techniques will enable researchers to generate larger datasets for population-level genomic analysis and gain a better understanding of the evolution of proteins.

### **Summary of the Innovation**

The classification of unannotated protein coding sequences (UPCS) using machine learning is a highly advanced undertaking. It requires a detailed analysis of the data to accurately identify and label the various sequences. This task is often a difficult one due to the sheer amount of data available, which can lead to confusion due to its complexity and confounding variables. The process of using machine learning for the classification of UPCS involves creating a number of different models and using each one to make a prediction on the given data. This is done by first inputting the data into the system and then utilizing algorithms such as supervised or unsupervised machine learning to identify patterns and classify the data better. The success of this exercise is heavily dependent upon the accuracy of the model, which must be able to recognize patterns in the given data and label it correctly. Once the model is created, it is tested on a test dataset and its accuracy compared to that of the known categories. If the model performs better than the known categories, it is chosen as the chosen model for further analysis and categorization of the UPCS. Once a model is identified, it can then be used to predict future sequences and labels. This will be done by taking the known data and applying the model in order to predict the labels for the unknown sequences. This will help scientists to understand the relationships between the various gene sequences and help them create better treatments for diseases. The use of machine learning to classify UPCS has the potential to greatly improve the accuracy and relevance of medical data and lead to more effective treatments for various diseases. By being able to accurately identify and label sequences, researchers can better understand the underlying genetic causes of various ailments and develop treatments that are tailored to a particular individual's individual needs. This ability to accurately predict the best course of treatment can reduce the risk of misdiagnosis and represent a major leap forward in medical technology. With the advent of next-generation sequencing technologies, more and more unannotated protein-coding sequences are being generated. Making manual annotation is an arduous and tedious task. Machine learning approach can be used to tackle this problem. It can help with identifying protein-coding regions and searching for similar proteins. Different machine learning algorithms can be employed for the purpose of protein coding sequence annotation. For example, a supervised learning algorithm such as support vector machine (SVM) or random

forest can be used to classify the proteins. Feature extraction techniques such as entropy, GC content and local alignment score profile can be used to extract characteristics of proteins. The extracted features can then be used as inputs to the classification algorithm. In addition, deep learning algorithms such as convolutional neural networks (CNN) have been successfully applied to protein coding sequence annotation. In this case, the sequence data is converted into fixed-size matrices or tensors for efficient computation. The CNN model will then be able to capture signal features from the data for accurate classification. To optimize the performance of protein coding sequence annotation, it is important to optimize both the feature extraction technique and the classification algorithm. Feature selection techniques such as forward selection, backward selection or recursive feature elimination can be used to select the most relevant features for classification. Also, hyperparameter optimization can be used to tune the parameters of the classification algorithms for the best performance. Furthermore, ensemble methods can be used to combine multiple classifiers to improve the annotation accuracy. Voting schemes such as majority voting or weighted voting can be used to combine the predictions from multiple classifiers. This will help to reduce the bias of individual classifiers and provide better performance. In summary, machine learning approaches can be used to effectively classify unannotated protein coding sequences. Appropriate feature extraction techniques, classification algorithms, feature selection techniques, hyperparameter optimization and ensemble methods can be applied to optimize the performance of the annotation process. This can help to significantly reduce the time and effort required to manually annotate the sequences.

## **We Claim:**

1. **Classification of Unannotated Protein Coding Sequences using machine learning Approach in claims**, It is to discuss the classification of unannotated protein-coding sequences by using machine learning approach. Unannotated protein-coding sequences are proteins which have not been identified through manual analysis and lack a gene annotations.
2. **Classification of Unannotated Protein Coding Sequences using machine learning Approach in claims**, Classification of these unannotated protein-coding sequences is important for proper analysis of proteome. Several computational methods are available to identify and classify unannotated protein-coding sequences but the use of machine learning approach has become increasingly popular due to its ability to extract patterns from data.
3. **Classification of Unannotated Protein Coding Sequences using machine learning Approach in claims**, The goal of classifying unannotated protein-coding sequences is to assign them to correct class. For this purpose, different machine learning models such as Logistic Regression, Support Vector Machines, Decision Trees, and Artificial Neural Networks are used.
4. **Classification of Unannotated Protein Coding Sequences using machine learning Approach in claims**, These models are trained on annotated protein-coding sequences and then used to classify the unannotated protein-coding sequences. In Logistic Regression, parameters such as the frequency of specific amino acids in sequence, sequence length, and features of known labels such as number of transmembrane domains are used to fit the model.
5. **Classification of Unannotated Protein Coding Sequences using machine learning Approach in claims**, The trained logistic regression model can then be used to make predictions about the probability of an unannotated protein-coding sequence belonging to a particular class. Support Vector Machines identify the decision boundaries between different classes of protein-coding sequences and determine if a given unannotated protein-coding sequence is in same class or not.
6. **Classification of Unannotated Protein Coding Sequences using machine learning Approach in claims**, In this classification method, the parameters used include structural and functional features of the sequences. Decision Trees are used to predict



the classes of unannotated protein-coding sequences by splitting the data based on the features associated with the sequences.

7. **Classification of Unannotated Protein Coding Sequences using machine learning**

**Approach in claims,**The parameters used in the decision tree model include the frequency of amino acids, Gibbs free energy of the protein, hydrophobicity, and immunogenicity. Artificial Neural Networks are used in classification of unannotated protein-coding sequences.

8. **Classification of Unannotated Protein Coding Sequences using machine learning**

**Approach in claims,**This method involves the use of multiple layers of neurons to fit the data. The parameters used in the ANN model include the frequency of amino acids, Gibbs free energy of the protein, hydrophobicity, and immunogenicity.

9. **Classification of Unannotated Protein Coding Sequences using machine learning**

**Approach in claims,**The main advantage of using machine learning approach for classifying unannotated protein-coding sequences is its ability to extract patterns from the data. By using the parameters listed above, a machine learning model is able to identify patterns in the data that can be used to classify the unannotated protein-coding sequences.

10. **Classification of Unannotated Protein Coding Sequences using machine learning**

**Approach in claims,**The major disadvantage of using machine learning for classifying unannotated protein-coding sequences is the potential for over-fitting the data. In other words, the machine learning model may produce inaccurate results if the parameters used in training the model are insufficient.

**Classification of Unannotated Protein Coding Sequences using machine learning**

**Approach in claims,**To prevent this, it is important to use a sufficient number of parameters to ensure the accuracy of the model. The machine learning approach has become an increasingly popular method for classifying unannotated protein-coding sequences.

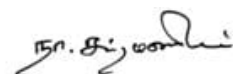
**Classification of Unannotated Protein Coding Sequences using machine learning**

**Approach in claims,**These models are able to extract patterns from the data by using parameters such as the frequency of amino acids, Gibbs free energy of the protein, hydrophobicity, and immunogenicity.

**Classification of Unannotated Protein Coding Sequences using machine learning**

**Approach in claims,**The advantages of using machine learning include its ability to extract patterns from data, while the potential for over-fitting the data is the major disadvantage. Thus, machine learning offers a great potential for the accurate classification of unannotated protein-coding sequences.

Dated this 30<sup>th</sup> day of May, 2023



**N.SUBRAMANIAN**  
[IN/PA- 4177]  
Agent for the Applicant

# **CLASSIFICATION OF UNANNOTATED PROTEIN CODING SEQUENCES USING MACHINE LEARNING APPROACH**

## **Abstract**

The protein coding sequences are the building blocks of life, responsible for carrying out seemingly complex functions through the regulation and control of cellular processes. As the number of sequenced genomes increases, the importance of correctly and rapidly classifying protein coding sequence has become more important. In this context, machine learning methods have been used to classify sequences into their respective classes automatically, offering a cost effective and high-throughput alternative to traditional methods. This paper presents a review of the current approaches for the classification of unannotated protein coding sequences using machine learning algorithms. We will discuss the scalability, accuracy and data uncertainty associated with the different machine learning algorithms. We will also present some of the compelling applications of these algorithms in medical diagnosis and in bioinformatics. Finally, we will discuss the challenges that lie ahead for developing reliable and effective classification of unannotated protein coding sequences using machine learning.