Bioinformatics
COMPGI10
Bioinformatics Report
Date: Tuesday 21st March, 2017
Graduate Assignment

Bioinformatics

Protein Classification By feature extraction

Sofiane MAHIOU

Computer Science, UCL, London, WC1E 6BT, UK

Abstract

Aim : The goal of this assignment is to provide an automated system that is able to classify proteins (Amino Acid sequences) into four classes each being a subcellular locations : **[Cytosolic, Secreted, Nuclear, Mitochondrial]**
Results : Using a *Random Forest Classifer* we manage to reache a **67% cross-validation accuracy**.
Improvements: In order to improve the results of the classifier, deepening the feature extraction method seems to be the way to go. Another method would be to use neural network techniques.
Contact: ucabsm1@ucl.ac.uk

1 Introduction

Currently, There is a growing need for fully automated methods to analyse amino acids sequences. One of the process that need to be automated is **the identification of the protein's subcellular location**. This problem can be splitted into two sub problems :

- **feature extraction:** the goal of this task is to choose the features that would allow an efficient classification, to be more precise, the chosen features should allow to easily seperate the sequences into classes or groups which will then be matched with the various subcellular locations
- **clasffication:** once the features obtained, it is then necessary to choose a fitting classification algorithm that will use the various features selected as a *vector representation* of each sequence that will then be fed to the classification algorithm during both training and testing.

It is however, possible to avoid splitting the problem into two sub-problems by using methods that have been designed to classify sequences of variables lengths such as :

- **HMM**
- **Recurrent neural networks & Seq2Seq Models**
- **1D Convolutional Neural Networks**

Although these methods usually yields better results than the methods presented before, the results obtained are far harder to interpret as these systems behave as **"black boxes"** and it's quite difficult to interpret what was *learned*.

Therefore, The first approach was used in order to ease the analysis of the results and the task was splitted into a **feature-extraction** task and **classification** task. Therefore, in order to indentify what features might be useful to this problem, a research phase was realized where several research papers on the same subject have been studied, and used as a reference to select several features.

2 Sequence preprocessing :

It seems important to mention that before proceeding the **feature extraction** phase, it was necessary to preprocess the data due to the presence of unexpected characters : **U, B, X** in the amino acid sequences.

- **X:** Given that "X" refers to "any amino acid" we randomly replace it by a given amino acid among the 20 amino acids
- **B:** "B" refers to either *asparagine* and *aspartic acid*. It is therefore automatically replaced by one or the other
- **U:** the "U" amino acid is simply removed from the sequence for lack of a better solution

This preprocessing step, will only affect 64 sequences out of more than 9000, therefore these changes are unlikely to heavily influence the results but will allow and easier implementation of the various features.

3 Feature Extraction

In terms of Features, three differents types of features can be identified:

- **Amino acids composition:** This refers to count or frequency of each amino acid in the sequence to analyse.
- **Protein's properties:** This referes to the various chimical and biological properties that a given protein can have such as *aromaticity, hydrophobicity, iso-electric point ...etc*
- **Subsequence-based features :** This refers to various existing methods that aims to extract relevant discriminative subsequences which *presence or count* will then be used as features to classify the proteins

The goal will then be combine several features from each type in order to optimize the results of the classification. The process through which the features to extract were selected was mainly based on what several research papers on this field advised and recommended. Once all the features were pre-selected and implemented, the final feature selection process was done while attempting classification, by trying to optmize both results and speed. Below you can find all the features that were considered as well as wether or not they were used in the final model.

3.1 sequence length: **Used**

The sequence found in the training have a varying length, therefore, it seemed fitting the provide the length of a given sequence as a feature.
This was confirmed by the results obtained after classification using various classifier, all of them providing better results when using this feature

3.2 Amino acids Counts:

This feature refers to the number of time each amino acid appears in the sequence. Given the variability of the length of the sequences, given the raw count for each amino acids seemed unfit as each dimension would heavily vary.

3.3 Amino acids frequency: **Used**

In order to fix the problem cited above, we use the frequency instead of the counts was recommended. Indeed, the frequency being a value between 0 and 1, it allows each dimension to be normalized, allowing the model to compare **amino acid composition** of sequences of various length.

3.4 isoelectric point: **Used**

The isoelectronic point or isoionic point is the pH at which the amino acid does not migrate in an electric field.
As presented by Q.-B. Gao *et al*. (2005), **iso-electric point** is among the top features to use when attempting the predict the subcellular location of a protein.

3.5 Presence of specific Sequences : N-Grams **Used**

As explained in Saidi *et al*. (2010), The presence of specific sequences of a specified length as well as the number of appearences can be powerfull features. The goal is then to manage to extract relevant, discriminative features.
to do so two techniques were attempted:

- **tf-idf:** tf-idf is a classic processing alogrithm that allows to extract "relevant" information from a sequence of "words" or "N-Grams" *ie* sequences of characters of length N. For example, for a sequence length or 3, this algorithm allowed us to extract the following sequences to look for : **['WW','MWW', 'CWW', 'WWM', 'WMW', 'WCW', 'CWM', 'WCM', 'MCW', 'WWH', 'FWW']**
- **Descriminative Descriptors:** this is another method that was presented in Saidi *et al*. (2010), it behaves similarily to tf-idf but use the additional information that several classes exist. [———————— TO COMPLETE ————————]

3.6 Nuclear Export Signals: **Used**

This feature that is presented in Xua *et al*. (2012), it describes the following pattern as an efficient discriminative pattern : $\phi_1 - X_3 - \phi_2 - X_2 - \phi_3 - X - \phi_4$. Positions ϕ_3 and ϕ_4 of this prevalent pattern are dominated by the five traditional hydrophobic residues **Leu[L], Ile[I],Val[V], Met[M], and Phe[F]**.
This feature did lead to improvements, however they were as significant as expected.

3.7 Nuclear Localization Signals: **Used**

This feature has also been presented in Xua *et al*. (2012), it refers to the count of subsequences of at least 5 **positively charged amino acids** *.ie* meaning one of the following : **lys[K], arg[R], his[H]**.
Again this feature lead to slight improvement over all models.

3.8 Protein's properties: **Used**

In addition to what was presented above the following properties were attempted as suggested by Q.-B. Gao *et al*. (2005)

- **Hydrophobicity**
- **Aromaticity**
- **Molecular Weight**

3.9 Begning and End of sequences:

Each feature presented above was computed for the **full sequence** as well as the **first 50 amino acids** and the **last 50 amino acids** of each sequence. This aims to identify trends and patterns not only overall but also specific to the begining and the end of the sequences. Indeed, the length of sequences would make it difficult to extract information solely related to the begining of the sequence therefore isolating the most proabably relevant subsequences seems to be a proper way of removing the "noise" due to then length of most amino acid sequences.

4 Classification Methods

Several classification methods are available, using previous knowledge about **Machine Learning - Classifcation** problems as well as classifiers referenced by other research papers of this same field. This lead to try the following classifiers :

- Logistic Regression Classifier
- Random Forest Classifier
- SVM Classifier
- Ridge Regression Classifier

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text

Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

- for bulleted list, use itemize
- for bulleted list, use itemize
- for bulleted list, use itemize

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

4.1 This is subheading

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text. Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

4.1.1 This is subsubheading

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text textText Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text

Table 1. This is table caption

head1	head2	head3	head4
row1	row1	row1	row1
row2	row2	row2	row2
row3	row3	row3	row3
row4	row4	row4	row4

This is a footnote

Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

Fig. 1. Caption, caption.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

4.2 Test1

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

5 Discussion

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text.

Table 1 shows that Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text. Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text. Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text.

6 Conclusion

(Table 1) Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

1. this is item, use enumerate
2. this is item, use enumerate
3. this is item, use enumerate

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the

above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text

Acknowledgements

Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

Funding

This work has been supported by the... Text Text Text Text.

References

R. Saidi, M. Maddouri and EM. Nguifo (2010)
Protein sequences classification by means of feature extraction with substitution matrices, *BMC Bioinformatics*

Q-B Gao, Z-Z Wang, C Yan, Y-H Du (2005)
Prediction of protein subcellular location using a combined feature of sequence, *FEBS Letters*

D Xua, A Farmera, G Colletta, N V. Grishinb, Y M Chooka (2012)
Sequence and structural analyses of nuclear export signals in the NESdb database , *MBoC*