The General Regression Neural Network to Classify Barcode and mini-barcode DNA

Riccardo Rizzo, Antonino Fiannaca, Massimo La Rosa, and Alfonso Urso

ICAR-CNR, National Research Council of Italy,
viale delle Scienze Ed.11, 90128 Palermo, Italy
{ricrizzo,fiannaca,larosa,urso}@pa.icar.cnr.it

Abstract. In the identification of living species through the analysis of their DNA sequences, the mitochondrial "cytochrome c oxidase subunit 1" (COI) gene has proved to be a good DNA barcode. Nevertheless, the quality of the full length barcode sequences often can not be guaranteed because of the DNA degradation in biological samples, so that only short sequences (mini-barcode) are available. In this paper, a prototype-based classification approach for the analysis of DNA barcode, exploiting a spectral representation of DNA sequences and a memory-based neural network, is proposed. The neural network is a modified version of General Regression Neural Network (GRNN) used as a classification tool. Furthermore, the relationship between the characteristics of different species and their spectral distribution is investigated. Namely, a subset of the whole spectrum of a DNA sequence, composed by very high frequency DNA k-mers, is considered providing a robust system for the classification of barcode sequences. The proposed approach is compared with standard classification algorithms, like Support Vector Machine (SVM), obtaining better results specially when applied to mini-barcode sequences.

Keywords: DNA barcode, Memory-based Neural Networks, GRNN, Classification.

1 Scientific Background

The identification of living species through the analysis of their DNA sequences is an open challenge. Because a massive comparison of a large collection of full genome sequences is not feasible, a bioinformatics approach to this problem is the analysis of some standard gene regions, containing enough information for the assignment to the proper taxa. The mitochondrial "cytochrome c oxidase subunit 1" (COI) gene is a comprehensive species-specific sequence library for all eukaryotes and it has proved to be a good marker for DNA sequences [8,13]; for this reason, it is considered as a DNA barcode for metazoan genomic.

Anyway, even though DNA barcode approach has proven to be useful for the identification and taxonomic rank assignment of very different species [6,12,11], its use can still be difficult if the biological samples are degraded. This is the case of archival specimen where biological samples can not guarantee the quality of

C. di Serio et al. (Eds.): CIBB 2014, LNCS 8623, pp. 142–155, 2015.
DOI: 10.1007/978-3-319-24462-4_13

the full length barcode sequence (650 bp) recovery. In fact, in many cases, only short sequences, also known as mini-barcode, are available (about 200 bp) [14].

In this paper, we propose a novel prototype-based classification approach based on the analysis of DNA barcode. Our method exploits a spectral representation of DNA sequences and a memory-based neural network for taxa estimation: spectral representation uses fixed-length DNA k-mers, whereas the neural network can store a set of prototypes (groups of k-mers) representing all the elements of the learning dataset.

In order to perform the barcode sequences classification, we introduce a modified version of General Regression Neural Network (GRNN) [19] that use, alternatively, a function derived from Jaccard distance and fractional distance (instead of the euclidean one) to compare learned prototypes against test sequences. The proposed approach implements these two kinds of distances and it is able to perform the classification task, even using only short fragments (200 bp) of the complete barcode sequence.

Finally, we compared our approach with the Support Vector Machine (SVM) [17] classification algorithm. Results show our method, implementing both Jaccard and fractional distances, is directly comparable with SVM in terms of classification metrics (accuracy, precision and recall) when considering full length sequences, whereas it overcomes SVM classifier when applied to short fragments of DNA barcode sequences.

2 Materials and Methods

The proposed method is based on two modified versions of the General Regression Neural Network. In the first subsection the basic principles of the GRNN are explained; the following two subsections present the proposed modifications, based on the Jaccard distance and the fractional distance; the last subsection describes the data sets used.

2.1 The General Regression Neural Network

The General Regression Neural Network [19] is a neural network created for regression i.e. the approximation of a dependent variable y given a set of sample $(\mathbf{x}, y)$, where $\mathbf{x}$ is the independent variable. In the following we will discuss the single output case, the extension to an output vector $\mathbf{y}$ is straightforward and can be found in [19]. In order to implement our classification tool for DNA sequences, we obtained the vector representation of the DNA sequences using a k-mer decomposition [10]. In this representation, sequences are coded by using fixed size vectors whose components are the number of occurrences of DNA snippets of k fixed-length, called k-mers. Considering $k = 5$, as proposed in [10], we have representing vectors $\mathbf{x} \in \Re^{1024}$.

The GRNN network has a one–pass training phase, it is just the memorization of all the training couples $(\mathbf{x_i}, y_i)$ each one in a neural unit i of the hidden layer. Fig. 1 shows a representation of the network: input layer has one neuron for each

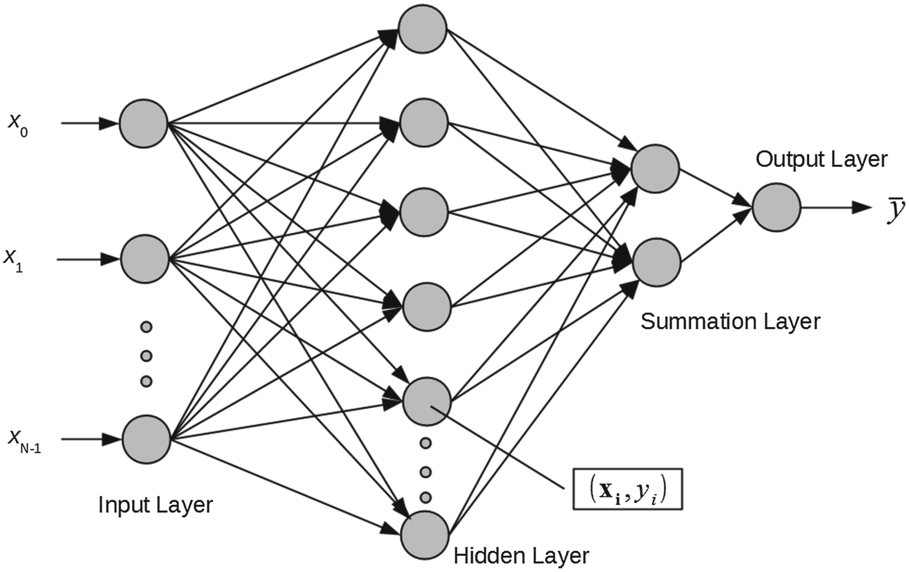


Fig. 1. The representation of the GRNN neural network.

component x_j in the input vector $\mathbf{x_i}$ and the hidden layer has one unit for each training sample.

During the test phase, when an unknown pattern $\mathbf{x}'$ is presented to the network, each hidden unit is excited according to the similarity of the pattern to the memorized input sample $\mathbf{x_i}$. The excitation level of the neural unit i is given by:

$$w_i = \exp\left\{-\frac{d(\mathbf{x}', \mathbf{x_i})}{2\sigma^2}\right\} \tag{1}$$

where $d(\mathbf{x}', \mathbf{x_i})$ is the distance between $\mathbf{x}'$ and $\mathbf{x_i}$, usually euclidean distance, and σ is the spread factor, representing the only parameter of the GRNN network. The hidden units have two outputs, $w_i * y_i$ and w_i, that are collected by the summation units. All the contributions of the hidden units are summed and normalized by the unit in the output layer, in order to obtain the network output y':

$$y' = \frac{\sum w_i * y_i}{\sum w_i}. \tag{2}$$

If we want to use this network as a classification tool we have to change our point of view: first of all the training couples are of the kind $(\mathbf{x_i}, c_h)$ where $c_h \in C$ is the class that is associated to $\mathbf{x_i}$ and C is the set of h classes. This means that all the classes must be coded in a real value vector $\mathbf{y_h}$ were each component y_h^i is given by:

$$y_h^i = \begin{cases} 0 \text{ if } i \neq h \\ 1 \text{ if } i = h \end{cases} \tag{3}$$

and requires a multiple output network. The σ value is the only parameter of the GRNN network. There are some studies on the optimal value of σ that can be a single value for the whole network or a specific value for each hidden unit. In [7] it is suggested a formula that depends on the maximum distance and number of patterns in the training set.

2.2 Jaccard Function

During experiments (see Section 3) we found that the euclidean distance used in the GRNN calculations was not enough "strong" for our purposes: in these kind of problems we have found that the presence or absence of a k-mer mean a lot and the euclidean distance does not emphasize this aspect [3].

Jaccard distance is defined among two sets A and B as

$$J = \frac{\|A \cup B\| - \|A \cap B\|}{\|A \cup B\|} \tag{4}$$

where $\|A\|$ represents the number of elements of the set A; $J \in [0, 1]$ and if $J = 1$ the two sets have not elements in common.

In this work, we redefine the Jaccard distance as a new function between two vectors that considers we will define between two vectors is computed considering the number of components in common between them, so that A and B in Eq. 4 are the set of the indexes of non-zero components in the vectors. This distance, however, has still a low contrast for our purposes because the information related to the magnitude of each component in the vectors is discarded. The component magnitude can be taken into account again if we do not consider all the components but only the m biggest components in the vectors.

More formally the sets A and B are defined as the sets of indexes:

$$s = \{s_0, s_1, ..., s_m\} \tag{5}$$

where s_j is the index of the x components that satisfies the ordering $x_{s_{j-1}} > x_{s_j} > x_{s_{j+1}}$ where $x_0 = max_{l=0,1,...N}\{x_l\}$

These sets can be compared using the Jaccard distance in Eq. 4.

Since Jaccard distance ranges from 0 to 1, it is necessary to map it in the interval $[0, \infty)$, using, for example, the following definition that we call J-function:

$$Jf(\mathbf{x}', \mathbf{x_i}) = \begin{cases} 0 & \text{if } J = 0 \\ \|\frac{1}{logJ}\| & \text{if } J \in (0, 1) \\ \infty & \text{if } J = 1 \end{cases} \tag{6}$$

This is necessary because w_i in eq. 1 should tend to zero if the distance is large. Moreover changing the distance method from euclidean to this normalised J-function distance stretches the original theory of the GRNN network; we leave a formal study of this problem to a future work.

2.3 Fractional Distances

High-dimensional spaces, such as the one defined by the sized vectors representing DNA sequences, are affected by the so called *curse of dimensionality.* In those spaces, in fact, the euclidean norm used to define the distance tend to *concentrate* [4]. That means all pairwise distances between high-dimensional objects appear to be very similar. In order to overcome this phenomenon, fractional norms can be used in place of euclidean norm [9,1]. Fractional norms are obtained from the Minkowski family norms defined as:

$$\|\mathbf{X}\|_p = \left(\sum_i |X_i|^p\right)^{\frac{1}{p}} . \tag{7}$$

With $p = 2$, the euclidean norm is obtained; whereas with $0 < p < 1$ Minkowski norms are called fractional norms, which induce fractional distances. In this work we adopted fractional norms, considering different values of p, in order to compute Eq. 1 and to limit the effects of the curse of dimensionality.

2.4 Barcode Dataset

We downloaded barcode sequences from the Barcode of Life Database (BOLD) [15]. In our study, we considered 10 barcode datasets belonging to different BOLD projects and living organisms. These datasets have been selected according to some criteria: we chose only *barcode compliant* dataset, i.e certified by BOLD as true barcode sequences, with sequence length not shorter than 500 bp and not longer than 800 bp. Following these criteria, we collected 2212 sequences. A full description of our barcode dataset can be retrieved in [3].

As discussed earlier, it is important to find a subset of the barcode gene in order to provide an effective identification mechanism for various animal or bacterial groups [6]. In fact, the recovery of full-length barcode sequence can be a problem in many cases: for example considering archival specimen, due to DNA degradation [5], or environmental samples [14]. There are studies, however, that tries to identify a specific location in the barcode gene, location that are called mini-barcode [14]. Our work is focused on the same idea but, instead of trying to identify a specific location in the gene, we explore the possibility of identifying species using small gene chunks. So that we fixed an amount of genetic material (200 bp) that could be enough to identify 95% of the species [14] and tried to understand what happens if this material comes not from a specific location of the gene, but it is scattered in two chunks of 100 bp (100x2), or in four subsequences of 50 bp (50x4). In both cases we do not check if these subsequences are overlapping or not, trying to reproduce laboratory conditions.

3 Results

Classification results obtained through our GRNN approach have been evaluated in terms of accuracy, precision and recall scores. We implemented both

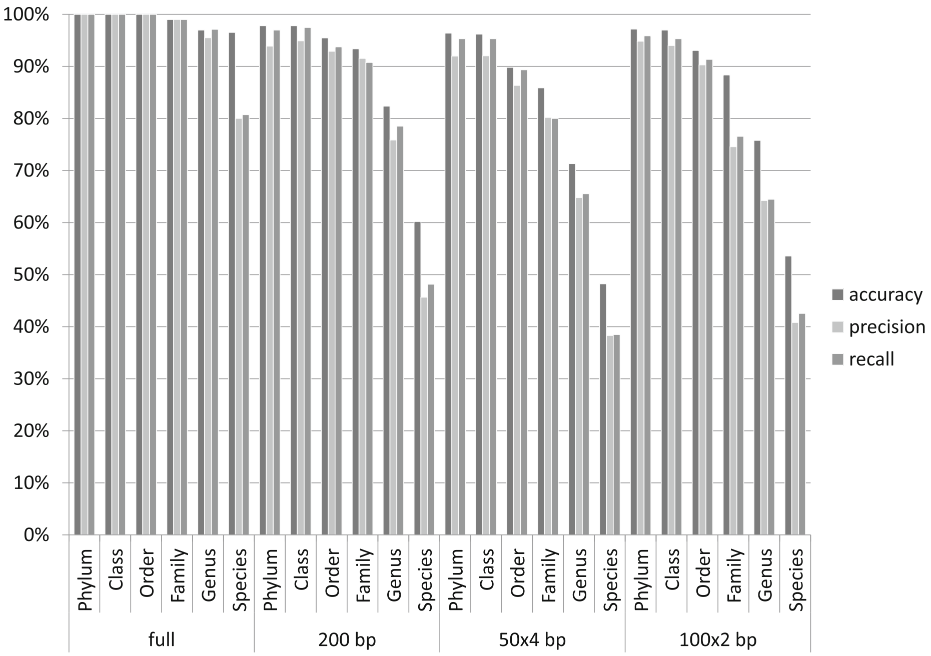


Fig. 2. Classification scores, in terms of accuracy, precision and recall, for the proposed approach based on the GRNN algorithm. The scores are arranged with regards to the taxonomic ranks and sequence sizes.

modified versions of GRNN algorithm using J-function and fractional distances, as explained in Sections 2.2 and 2.3, respectively. We performed two types of training/testing procedures. In the first scenario, we trained our classifiers with the whole full length dataset and then we tested it with all the sequence fragments. Our aim was in fact to assess if the GRNN classifier, trained with the full length sequences, is able to correctly classify the sequence fragments. It is important to underline that although in the training set (full length) and in the test set (fragments) there are the same number of sequences, their vector representations are completely different. In the second scenario, we adopted a ten fold cross validation scheme, considering as training set the full length sequences and as test set corresponding sequence fragments that did not belong to the training set. In this situation we wanted to assess if the GRNN classifier is able to classify sequence fragments even if it did not learned the corresponding full length patterns. Moreover we compared our method with another classifier used for nucleotide sequences classification [18]: the Support Vector Machine (SVM). We adopted the SVM implementation provided by the R package *e1071*, which is an interface to the *libsvm* library [2]. We used a Gaussian Radial Basis kernel, $k(\mathbf{x}, \mathbf{x}') = exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2)$. The parameter C and γ of the Gaussian kernel has been tuned through a grid search over a set of parameters values: γ ranging from 10^{-6} to 10^3; C ranging from 1 to 10^3, as suggested by the authors of *libsvm*.

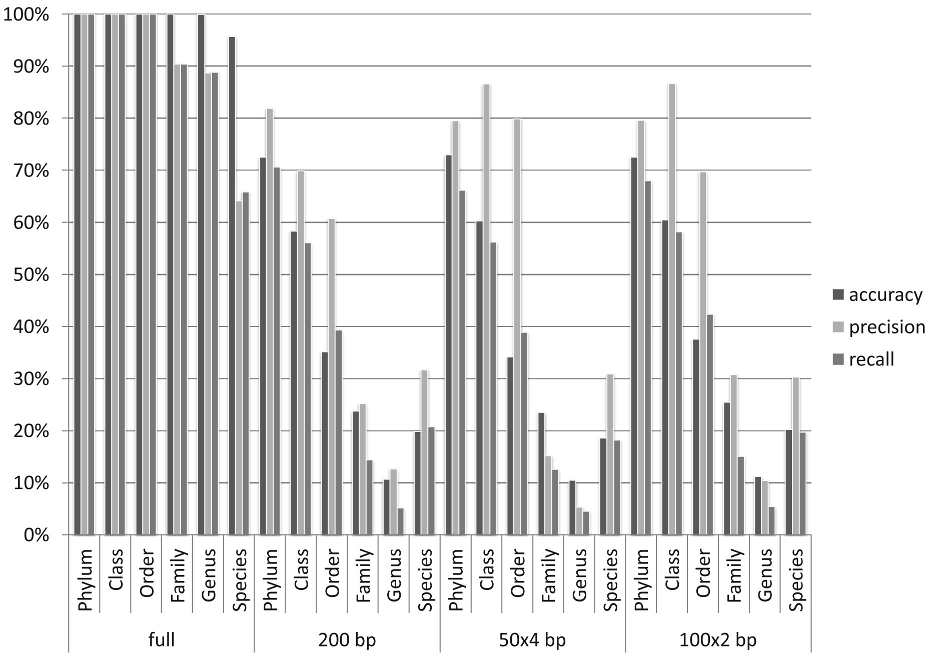


Fig. 3. Classification scores, in terms of accuracy, precision and recall, for the SVM classifier. The scores are arranged with regards to the taxonomic ranks and sequence sizes.

The best parameter values have been computed minimising the error measure using a 10-fold cross validation on the training set. The results obtained with the SVM classifiers have been carried out by means of the same two way training/testing procedure. In other experiments, not shown here, we also adopted Classification Tree, from the R package *rpart*, algorithm and we obtained very similar results to the ones obtained through SVM. For this reason, we do not present those results in this paper.

The GRNN outputs obtained using J-function have been obtained comparing the $m = 30$ biggest components between the vector prototypes and the test vectors. This value has been selected after a series of experiments, not presented here for lack of available space, using for comparison a number of components ranging from 20 to 50 and a σ value during the training phase ranging from 0.5 to 0.8. We reached a trade off between the best results and the smallest number of elements by using 30 components and $\sigma = 0.7$. A number of components $m > 30$ does not give a meaningful improvement in the results.

In the previous version of our work [16], we focused our attention on the comparison of classification results between the GRNN algorithm with J-function and the SVM classifier. Those results, arranged according to the test sequence sizes (full, 200 bp, 50x4 bp, 100x2 bp) and taxonomic ranks (from phylum down to species), are summarized in the charts shown in Fig. 2 and 3. We demonstrated

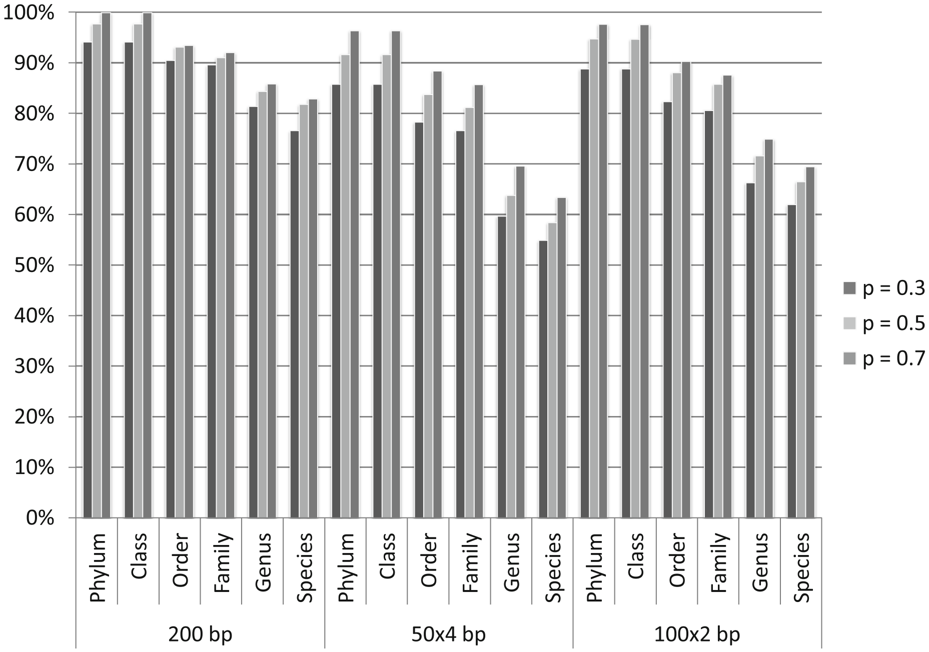


Fig. 4. Classification scores, in terms of accuracy, for the proposed approach based on the GRNN algorithm implementing fractional distances with different values of parameter p. The scores are arranged with regards to the taxonomic ranks and sequence sizes.

that our approach clearly outperforms SVM for the classification of sequence fragments; considering full length sequences, on the other hand, both GRNN and SVM classifiers reached very similar high scores, ranging between 100% and 95% of accuracy score.

In this work we analysed the performances of the GRNN algorithm implementing fractional distances in order to classify short barcode sequences of size 200 bp, 100x2 bp and 50x4 bp. Classification results, in terms of accuracy, precision and recall scores, have been compared with both the GRNN algorithm using J-function and the SVM classifier.

First of all, considering the first training/testing scenario, we studied how classification results change with regard to the parameter p of fractional distances (see Eq. 7). We carried out experiments with $p = 0.3$, $p = 0.5$ and $p = 0.7$ and the classification results, in terms of accuracy score are shown in Fig. 4. The most interesting result is that the best scores, ranging from 100% at phylum level to about 82% at species level, were obtained with $p = 0.7$ considering short fragments of 200 consecutive base pairs. With 50x4 bp and 100x2 bp sequence lengths, we obtained slightly lower scores. The analysis of precision and recall showed very similar scores to the accuracy one, not providing any further meaningful information: for this reason we did not report those results.

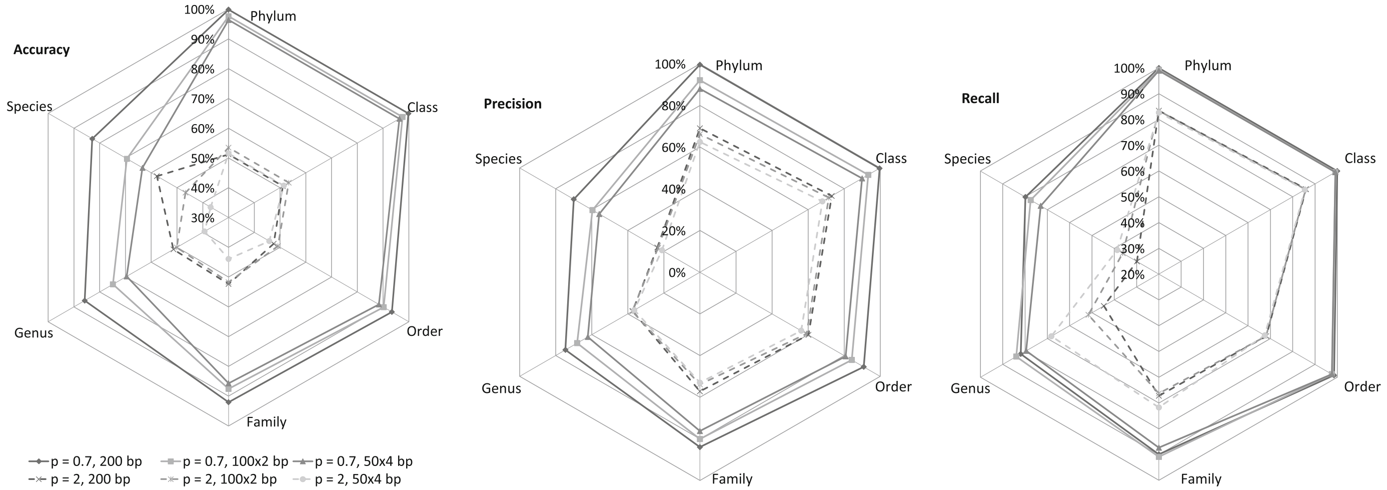


Fig. 5. Classification scores, in terms of accuracy, precision and recall, of the comparison between GRNN implementing fractional distance ($p = 0.7$) and euclidean distance ($p = 2$).

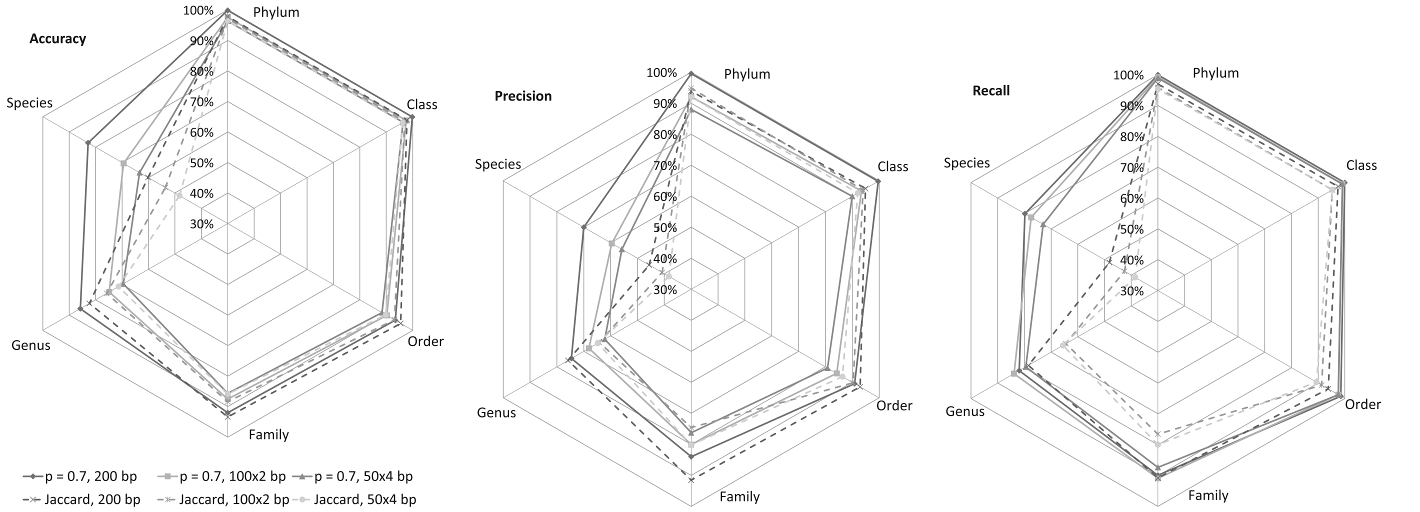


Fig. 6. Classification scores, in terms of accuracy, precision and recall, of the comparison between GRNN implementing fractional distance ($p = 0.7$) and Jaccard distance.

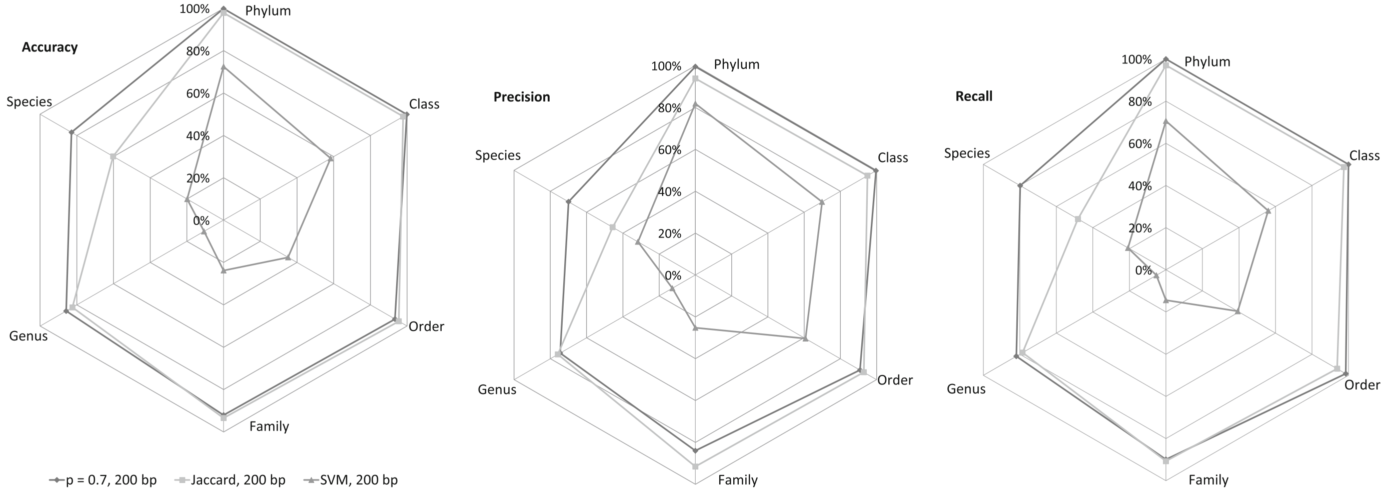
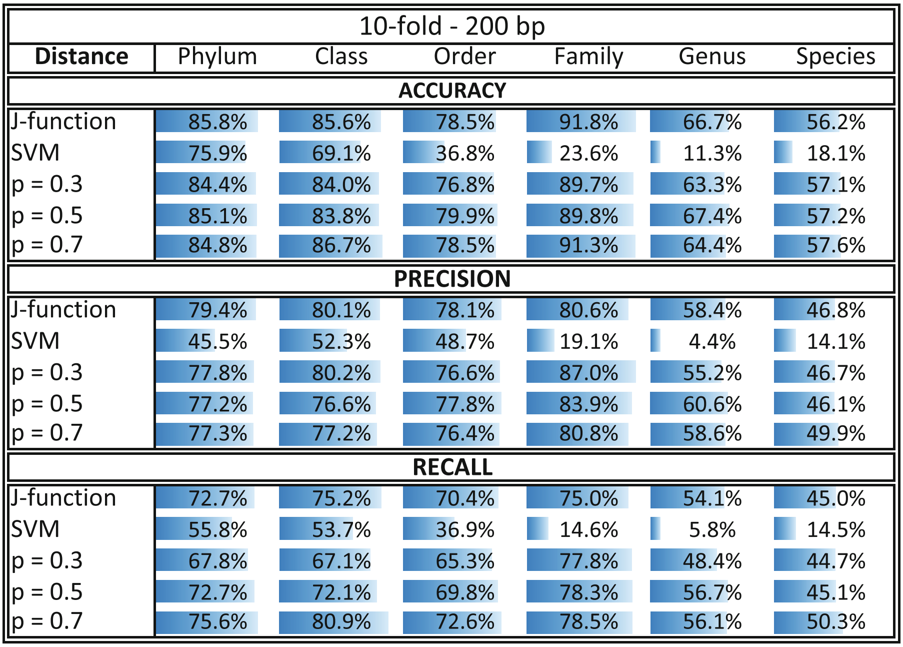


Fig. 7. Classification scores, in terms of accuracy, precision and recall, of the comparison among GRNN implementing both fractional distance ($p = 0.7$) and Jaccard distance, and SVM.

Table 1. Classification scores, in terms of accuracy, precision and recall, for the proposed approach based on the GRNN algorithm against the SVM classifier. The training/testing procedure refers to the second scenario, i.e. ten fold cross validation with full-length sequences as training set and test set composed of the sequence fragments whose corresponding complete sequences do not belong to the training set.



10-fold - 200 bp						
Distance	Phylum	Class	Order	Family	Genus	Species
ACCURACY						
J-function	85.8%	85.6%	78.5%	91.8%	66.7%	56.2%
SVM	75.9%	69.1%	36.8%	23.6%	11.3%	18.1%
p = 0.3	84.4%	84.0%	76.8%	89.7%	63.3%	57.1%
p = 0.5	85.1%	83.8%	79.9%	89.8%	67.4%	57.2%
p = 0.7	84.8%	86.7%	78.5%	91.3%	64.4%	57.6%
PRECISION						
J-function	79.4%	80.1%	78.1%	80.6%	58.4%	46.8%
SVM	45.5%	52.3%	48.7%	19.1%	4.4%	14.1%
p = 0.3	77.8%	80.2%	76.6%	87.0%	55.2%	46.7%
p = 0.5	77.2%	76.6%	77.8%	83.9%	60.6%	46.1%
p = 0.7	77.3%	77.2%	76.4%	80.8%	58.6%	49.9%
RECALL						
J-function	72.7%	75.2%	70.4%	75.0%	54.1%	45.0%
SVM	55.8%	53.7%	36.9%	14.6%	5.8%	14.5%
p = 0.3	67.8%	67.1%	65.3%	77.8%	48.4%	44.7%
p = 0.5	72.7%	72.1%	69.8%	78.3%	56.7%	45.1%
p = 0.7	75.6%	80.9%	72.6%	78.5%	56.1%	50.3%

Experiment trials with fractional distance with $p = 0.7$, therefore, is able to overcome the distance concentration phenomenon, as described in Section 2.3. In order to validate that thesis, we compared classification results obtained with GRNN using fractional distance ($p = 0.7$) and euclidean distance ($p = 2$). This comparison is presented by means of radar charts in Fig. 5. There it is clear how, at all taxonomic level and for each sequence length, euclidean distance is unable to provide acceptable results (accuracy score about 50%).

The comparison between classification results obtained using GRNN with fractional distance ($p = 0.7$) and Jaccard distance is summarized in the radar charts of Fig. 6. Once again the best results, at each taxonomic rank and for all sequence sizes, were reached by means of fractional distance. The most evident difference of the performances between those two approaches is at Species level, where with fractional distance we reached an accuracy score of about 80% against 60% with Jaccard distance.

The last comparison, showed in Fig. 7, was done considering GRNN with fractional distance ($p = 0.7$) and Jaccard distance against SVM. As discussed earlier in this Section, Fig. 3, both approaches implementing GRNN clearly outperforms the SVM classifier.

From these results it is evident that SVM is not able to deal with sequence fragments. In fact, the sequence fragments (mini-barcode) have a vector representation that is very different from the one computed for the original sequences, therefore SVM, from this point of view, can not correctly classify those fragments. Otherwise, our approach, considering both the Jaccard distance and the fractional distance, demonstrate the ability to to provide very reliable classification results when dealing with sequence fragments.

Finally, with regards to the second training/testing scenario, we obtained the classification results summarized in Table 1. We reported only the scores related to the 200 bp fragments because they are very similar to the ones obtained with 100x2 and 50x4 bp fragments. In this situation, although with lower scores, the GRNN algorithm, especially in the case of fractional distances, provides consistent classification performances and it outperforms the SVM classifier. Classification scores are lower with respect to the first scenario because the GRNN has never learned the full length sequences corresponding to the test fragments. In spite of that, our GRNN approach turned out to be robust enough to keep on providing acceptable classification scores.

4 Conclusion

In this work we presented a classification methodology for barcode DNA sequences based on the General Regression Neural Network algorithm. We introduced two modified versions of GRNN in order to overcome limitations of the standard euclidean approach: the first one implements the J-function derived from the Jaccard distance, the second one adopts fractional distances. We obtained very accurate and very robust classifiers with respect to sequence sizes. We tested our approaches, in fact, considering the so-called mini-barcode, that is a sequence fragment 200 bp long extracted from the original sequences. Classification results demonstrate that using fractional distances (with parameter $p = 0.7$) allows to reach the best scores in terms of accuracy, precision and recall. We compared also our methods with SVM classifier. Classification results at all taxonomic levels and for each sequence sizes clearly state that our classifiers outperforms SVM when applied to sequence fragments.

References

1. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional space. In: Van den Bussche, J., Vianu, V. (eds.) ICDT 2001. LNCS, vol. 1973, pp. 420–434. Springer, Heidelberg (2000), doi:10.1007/3-540-44503-X
2. Chang, C.-C., Lin, C.-J.: Libsvm: A library for support vector machines. ACM Trans. Intell. Syst. Technol. 2(3), 1–27 (2011)
3. Fiannaca, A., La Rosa, M., Rizzo, R., Urso, A.: Analysis of DNA barcode sequences using neural gas and spectral representation. In: van Zee, G.A., van de Vorst, J.G.G. (eds.) EANN 2013. LNCS, vol. 384, pp. 215–224. Springer, Heidelberg (1989)

4. Francois, D., Wertz, V., Verleysen, M.: The Concentration of Fractional Distances. IEEE Transactions on Knowledge and Data Engineering 19(7), 873–886 (2007)
5. Hajibabaei, M., Smith, M.A., Janzen, D.H., Rodriguez, J.J., Whitfield, J.B., Hebert, P.D.N.: A minimalist barcode can identify a specimen whose DNA is degraded. Molecular Ecology Notes 6(4), 959–964 (2006)
6. Hajibabaei, M., Singer, G.A.C., Hebert, P.D.N., Hickey, D.A.: DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics.. Trends in Genetics 23(4), 167–172 (2007)
7. Haykin, S.: Neural networks: a comprehensive foundation, 2nd edn. Prentice-Hall (1998)
8. Hebert, P.D.N., Ratnasingham, S., DeWaard, J.R.: Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings of the Royal Society. Series B, Biological Sciences 270(suppl.), S96–S99 (2003)
9. Hinnenburg, A., Aggarwal, C., Keim, D.: What is the nearest neighbor in high dimensional spaces?. In: Proceedings of the 26th International Conference on Very Large Data Bases, VLDB 2000, pp. 506–515. Morgan Kaufmann Publishers Inc. (2000)
10. Kuksa, P., Pavlovic, V.: Efficient alignment-free DNA barcode analytics. BMC Bioinformatics 10(suppl. 14), 9 (2009)
11. La Rosa, M., Fiannaca, A., Rizzo, R., Urso, A.: Alignment-free Analysis of Barcode Sequences by means of Compression-Based Methods. BMC Bioinformatics 14, S4 (2013)
12. La Rosa, M., Fiannaca, A., Rizzo, R., Urso, A.: A study of compression–based methods for the analysis of barcode sequences. In: Peterson, L.E., Masulli, F., Russo, G. (eds.) CIBB 2012. LNCS, vol. 7845, pp. 105–116. Springer, Heidelberg (2013)
13. Marshall, E.: Taxonomy. Will DNA bar codes breathe life into classification? Science 307(5712), 1037 (2005)
14. Meusnier, I., Singer, G.A.C., Landry, J.-F., Hickey, D.A., Hebert, P.D.N., Hajibabaei, M.: A universal DNA mini-barcode for biodiversity analysis.. BMC Genomics 9, 214 (2008)
15. Ratnasingham, S., Hebert, P.D.N.: bold: The Barcode of Life Data System (http://www.barcodinglife.org).. Molecular Ecology Notes 7(3), 355–364 (2007)
16. Rizzo, R., Fiannaca, A., La Rosa, M., Urso, A.: The General Regression Neural Network to Classify Barcode and mini-barcode DNA. In: Proceedings of CIBB (2014)
17. Scholkopf, B., Smola, A.: Learning with kernels. MIT Press, Cambridge (2002)
18. Seo, T.K.: Classification of nucleotide sequences using support vector machines. Journal of Molecular Evolution 71(4), 250–267 (2010)
19. Specht, D.F.: A general regression neural network. IEEE Transactions on Neural Networks 2(6), 568–576 (1991)