Automatic Classification of Traced Neurons Using Morphological Features

José D. López-Cabrera, Juan V. Lorenzo-Ginori

Universidad Central “Marta Abreu” de Las Villas,
Centro de Investigaciones en Informática, Santa Clara,
Cuba

josedaniellc@uclv.cu, juanl@uclv.edu.cu

Abstract. The great advances in the field of neuron tracing have made possible a high availability of free-access data in the Internet, which motivates the realization of automatic classifications. The increase of neuronal reconstruction databases makes the manual classification of neurons a time-consuming and tedious task for the analysts. Classification by human experts is also prone to inter- and intra-analyst variability due to the process' inherent subjectivity. In this context, the need arises to find new descriptors having discriminative properties which allow separating the various neuron classes, and this constitutes currently an open problem. Such descriptors would contribute to improve the results of automatic classification. In this study the attention is focused on the use of new morphological features in supervised classification of traced neurons. Furthermore, we present a comparative analysis of different supervised learning algorithms oriented to the classification of reconstructed neurons. The results were validated using non-parametric statistical tests and they show the usefulness of the proposed solution.

Keywords. Neuron tracing, morphological features, feature selection, automatic classification, non-parametric tests.

1 Introduction

The great advances in the field of neuron tracing have made possible a high availability of data in the Internet, which encourages the realization of automatic classifications [21, 3]. Note that although the issue of classifying neurons had its beginnings since the emergence of the neuroscience as a scientific discipline, manual classification is a slow and tedious task for human analysts, and this fact determines the existence of an increasing interest in the use of machine learning techniques for this application [2]. It is worth to mention that manual classification is also subjected to inter- and intra-analyst variability due to the subjectivity inherent to this process. Major efforts have been made in terms of automatic classification of neurons, but most of them have been done by using unsupervised techniques [14, 19]. These have been exploratory techniques aimed at discovering new types of cells or to confirm some known hypotheses about the neurons. Although useful, these classification systems are hampered by at least two deficiencies. First, a distinctive feature can be shared by several cell types. Second, one discriminative feature for certain neuronal types may be irrelevant and highly variable for other types. It is therefore necessary to find new descriptors having discriminative properties in regard of neuron classes, in order to improve the automatic classification.

Today neural structures such as the axon and the dendrites remain the cornerstone in the analysis of neural development; pathology; computation and connections between neurons [21]. Neuron classification, however, has also been treated as a multimodal problem in which in addition to the morphological features, biochemical and electrophysiological features are also employed [13]. There are several software that perform morphometric analysis [22, 17, 7, 1]. The morphometric features are oriented mainly to the density of the branches of neuronal trees, the size of the roads and the relationship between their thickness, tortuosity, angles at bifurcations, volume and area, among others. One of the most commonly used ways to quantify neurons is the Sholl analysis, to which several modifications have

ISSN 1405-5546
doi: 10.13053/CyS-21-3-2495

Table 1. Examples of some features computed with L-Measure for an interneuron cell

Metric	Total Sum	Compartments considered	Compartments discarded	Minimum	Average	Maximum	Std
Soma Surface	43.709	30	2211	0.184	1.457	2.657	0.693
Number of bifurcation	123	123	2118	1	1	1	0
Number of tips	126	126	2115	1	1	1	0
Volume	1112.2	2240	1	0.011	0.496	155.21	3.557
Fractal Dimension	177.9	168	2073	1	1.059	1.394	177.97

already been made [10, 11]. In order to characterize neurons and find common rules in the geometry of the dendrites, a study related with the angles formed by its branches has been carried out in [18]. Another way to represent the structure of neuronal trees is proposed in [12], where bifurcations are encoded as strings.

In this study we focus the attention in the use of morphological features for supervised classification, which has been less treated and has shown better results than unsupervised techniques [14, 23]. In this case, a priori information which is used in unsupervised algorithms only to validate the classification process, allowed us to build our models. This paper presents a comparative analysis among different supervised learning techniques, oriented to the classification of reconstructed neurons using morphological features.

2 Materials and Methods

The data used for classification were extracted from the NeuroMorpho.Org website, which contains a large number of neuronal reconstructions from different species, brain regions and laboratories [15], freely downloadable. Neural reconstructions used in our research came from [14], where the procedure for the used neuron reconstruction process is explained in detail. The neurons reconstructed there belong to laboratory rats. This data set is composed of 318 traced neurons, classified by human experts in 192 interneurons and 126 pyramidal cells. The features were computed from arbor reconstruction files in the standard SWC file format. From each of these cells a set of morphometric features were computed to be used later in the classification process. The way of representing the neurons is by means of a graph, where their branching structures are represented by the directed adjacency matrix. A tree is composed by a set of labelled nodes connected by edges, these edges are also called compartments and they have an associated diameter. Only the compartments associated to a specific feature are to be taken into account, the rest being discarded.

The first data set (called LM) is composed by features extracted using the L-Measure software, which provides 43 morphological features. Each of these features is associated with 7 parameters [22], as shown in Table 1. In many cases, some of these 7 parameters are meaningless, therefore it is necessary to preprocess the data set. The second data set (called NF) is composed by new features proposed in this research which are shown in Table 2.

The 3rd data set is simply the union of the two previous ones (called Union), and it was created in order to determine if the incorporation of the proposed features can improve the performance of classifiers. Three matrices were then formed, having 318 rows and a number of columns equal to the number of features used. Some feature selection techniques were used to reduce the cardinality of the data sets. This was made in order to comply with the recommendations made in [9],

Table 2. Description of the new features proposed

Name	Description
IPmean	Mean value of the in-plane angle IP.
IPstd	Standard deviation of the in-plane angle IP.
TP5mean	Mean value of the torsion angle TP5
TP5std	Standard deviation of angles TP5
D-CM-C	Euclidean distance from C_m to C.
D-S-CM	Euclidean distance from S to C_m
D-S-C	Euclidean distance from S to C

Table 3. Results after applying feature selection techniques for each dataset, showing the number of feature (NuF) obtained

Data Set	Search Method		Evaluator	NuF
LM	Best First	Forward	CFS	8
			Consistency	11
		Backward	CFS	8
			Consistency	16
		Bi-Directional	CFS	8
			Consistency	11
NF		Forward	CFS	4
			Consistency	5
		Backward	CFS	4
			Consistency	5
		Bi-Directional	CFS	4
			Consistency	5
Union		Forward	CFS	9
			Consistency	11
		Backward	CFS	9
			Consistency	13
		Bi-Directional	CFS	9
			Consistency	11

about having an appropriate relationship between the numbers of cases and features to prevent overfitting of the classification algorithms, This procedure tends also to improve the performance of classifiers.

2.1 Feature Selection

A selection was made of a subset of the features included in the original set, with the purpose of obtaining maximum performance with minimum effort. All feature selection algorithm consists of two basic components: evaluation function and search method. As search method was used Best First, with three alternatives: Forward, Backward and Bi-Directional. In the case of the evaluation function, the methods used were Correlation-based Feature Selection (CFS) and Consistency-based Subset Evaluation. The CFS evaluation function tends to produce subsets containing features that are highly correlated with the class and uncorrelated between them [16]. In the case of Consistency, it is characterized by having a strong dependence on the training set, trying to remove the minimum subset that satisfies an acceptable rate of inconsistency, usually set by the user.

2.2 Neuronal Classification

The classification process had two purposes: to determine whether using the new proposed features improved the quality of classification and to compare the performance of several classifiers. The metric selected to quantify the performance of classifiers was the AUC or area under the ROC (receiver operating characteristic) curve. As methods of machine learning classifiers were used Logistic Regression (LR), KNN, Random Forest (RF), C4.5 and Naive Bayes (NB) [8]. Both feature selection techniques and classification algorithms were applied using the widely known open-source data mining software tools named WEKA [6]. The parameters of the classifiers were those who come by default in WEKA. For KNN we decided to use k = 3, since this value led to the best result obtained during the experiments.

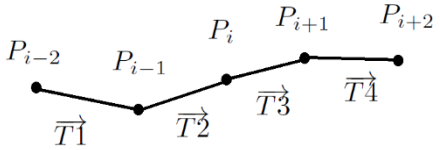

Fig.1. Representation of the vectors used to compute the angles in a neuron

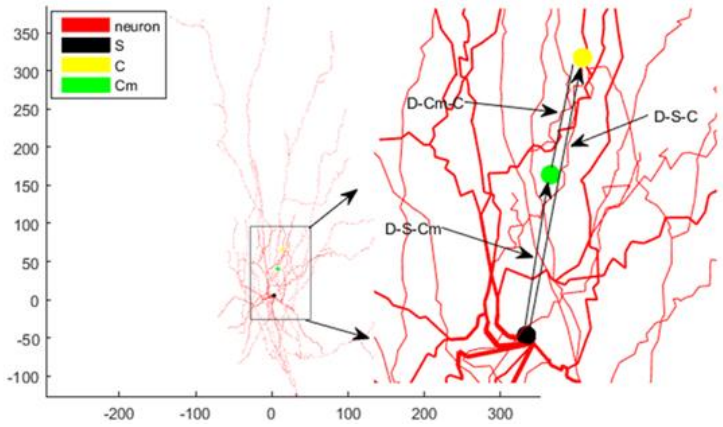

Fig. 2. Representation of an interneuron cell used in the experiments

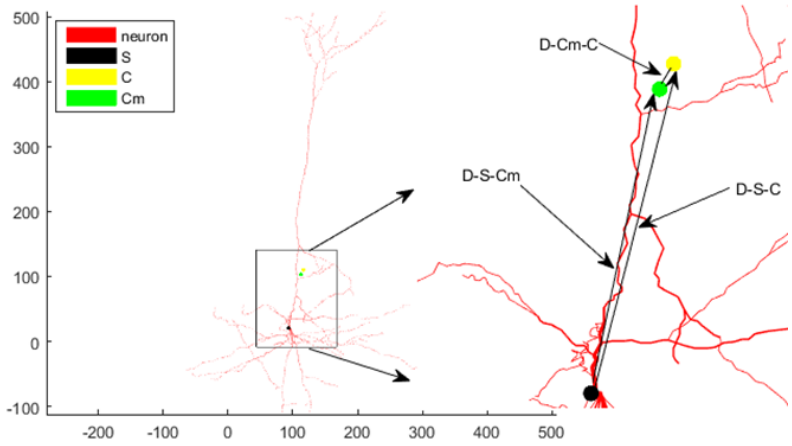

Fig. 3. Representation of a pyramidal neuron used in the experiments and some features computed related with Euclidean distances

The R software was used, specifically the SCMAMP package, in order to apply statistical tests [4]. We applied 5x2cv as described in [5] as well as the Friedman test, to find out if there were significant differences between some of the classifiers analyzed. In the case where such differences were found, the Finnes post hoc test was applied, which is considered generally a good choice because of its simplicity and power.

2.3 New Proposed Features

The first feature implemented was the in-plane angle (IP_i). It is computed using three points of the neuron, as it is shown in equation 1. Figure 1 shows the vectors to compute the angles in the neurons:

$$IP_i = \cos^{-1}\left(\frac{\overrightarrow{T2}}{|\overrightarrow{T2}|} \cdot \frac{\overrightarrow{T3}}{|\overrightarrow{T3}|}\right), \quad (1)$$

where

$$\overrightarrow{T2} = \overrightarrow{P_i} - \overrightarrow{P_{i-1}} \text{ and } \overrightarrow{T3} = \overrightarrow{P_{i+1}} - \overrightarrow{P_i}.$$

Notice that the point P_i has the coordinates $(X_i\ Y_i\ Z_i)$ in a three dimensional space and in this analysis it is represented by a position vector

Table 4. Performance of the Logistic Classifier

CR	SM	Evaluator	mean	std	NuF
LM	Forward	CFS	0.953	0.022	8
		Consistency	0.965	0.009	11
	Backward	Consistency	0.799	0.011	16
NF	Forward	CFS	0.776	0.015	4
		Consistency	0.767	0.020	5
Union	Forward	CFS	0.974	0.012	9

Table 5. Performance of the Random Forest Classifier

CR	SM	Evaluator	mean	std	NuF
LM	Forward	CFS	0.958	0.012	8
		Consistency	0.976	0.009	11
	Backward	Consistency	0.883	0.012	16
NF	Forward	CFS	0.845	0.024	4
		Consistency	0.848	0.026	5
Union	Forward	CFS	0.973	0.009	9

traced from the coordinate axes' origin up to the point itself. The number of IP_i calculated for the whole neuron is computed by equation 2, where Pt is a total of nodes and $Pter$ is a number terminal points. Notice that it is necessary also to subtract one, because the initial point is not counted:

$$PvalIP = Pt - Pter - 1. \quad (2)$$

Another feature implemented was the torsion angle. It is composed of the angle between two consecutive planes whose pivot point is P_i.

Then using the procedure described above for the vectors associated to points P_{i-1} and P_{i-2}, we defined $\overrightarrow{T1}$ and analogously from the vectors associated to P_{i+1} and P_{i+2} is defined $\overrightarrow{T4}$.

The orthogonal vectors to the two planes containing the vectors $\overrightarrow{T1}, \overrightarrow{T2}$ and $\overrightarrow{T3}, \overrightarrow{T4}$ were obtained using the vector cross product and then the angle between these vectors which corresponds to the rotation between the two planes considered is calculated as shown in equation 3:

$$TP5_i = \cos^{-1}\left(\frac{\overrightarrow{T1} \times \overrightarrow{T2}}{\left|\overrightarrow{T1} \times \overrightarrow{T2}\right|} \cdot \frac{\overrightarrow{T3} \times \overrightarrow{T4}}{\left|\overrightarrow{T3} \times \overrightarrow{T4}\right|}\right), \quad (3)$$

where

$\overrightarrow{T1} = \overrightarrow{P_{i-1}} - \overrightarrow{P_{i-2}}$, $\overrightarrow{T2} = \overrightarrow{P_i} - \overrightarrow{P_{i-1}}$, $\overrightarrow{T3} = \overrightarrow{P_{i+1}} - \overrightarrow{P_i}$, $\overrightarrow{T4} = \overrightarrow{P_{i+2}} - \overrightarrow{P_{i+1}}$.

The number of torsion angles calculated for the whole neuron is computed by equation 4, where Pt is the total number of points (nodes) and $Pter$ is the number of terminal points in the neuron:

$$PvalTP5 = Pt - Pter * 2 - 2. \quad (4)$$

For the list obtained with $TP5_i$, we compute global statistics like mean ($\bar{x}$) and standard deviation (σ). Notice that differently to what is proposed in [17], to compute the torsion angle we used here five points instead of four, as shown in Figure 1.

The features calculated afterwards are associated with the Euclidean distances in the three-dimensional space formed between three points. The first is the root of neuron tree (soma, S).

The second point is the centroid (C) of the tree, and the third is the center of mass (C_m). C is defined by equation 5. C_m is computed using equation 6, where r_i is the position vector (X_i, Y_i, Z_i) and V_i is associated with the cylindrical volume formed between two consecutive nodes, which has its radius as prior information.

Its height (h) is calculated as the distance between the two node points, as shown in equation 7. The description of the implemented features is shown in Table 2:

$$C = \frac{\sum_{i=0}^{P_t-1} r_i}{P_t}, \tag{5}$$

$$C_m = \frac{\sum_{i=0}^{P_t-1} V_i r_i}{\sum_{i=0}^{P_t-1} V_i}, \tag{6}$$

$$V = \pi r^2 h. \tag{7}$$

Figure 2 and Figure 3 show examples of the classes of neurons that are being analyzed in this research. In these figures it is observed the graphical representation of some of the new features proposed, notice that the distances between the above mentioned node points are different.

3 Results and Discussion

Once the feature selection method is applied, new subsets of features are obtained, which coincide in many cases, as shown in Table 2.

In the case of the set LM and using the CFS evaluator, the same subset of features is obtained by the three search methods. The new subset contained 8 features. On the other hand, the selection using the Consistency evaluator led to different results when compared to the previous one. The Forward and Bi-Directional search strategies selected the same subset of 11 features and in the case of the Backward search method the subset had 16 features.

For the Union set, the results obtained using the Consistency evaluator were the same obtained for the LM subset, i.e. the feature selection method chose the same subset of features in the case of LM as well as in the case of Union. Something different happened with the CFS evaluator for the Union set where nine features were selected. Three of these 9 features selected belonged to the NF set, these were $IPmean$, *D-S-Cm* and $TP5mean$ i.e. three out of the nine features contained in this set, were among the new features proposed.

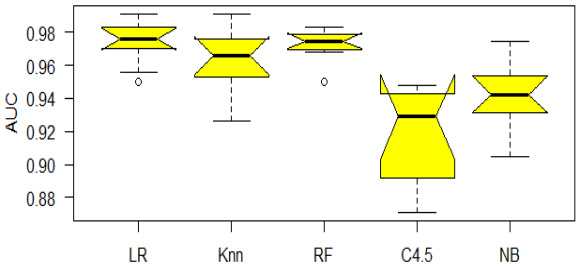

Fig. 4. Boxplot of performance of the five classifiers according to their AUC values

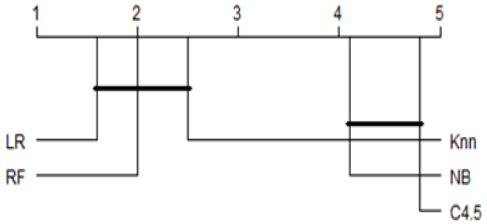

Fig. 5. Results of the Finnes Post Hoc test, to determinate if any of the classifiers showed significant statistical differences when compared to the others

In the case of NF with the CFS evaluator, there is a coincidence among the sets of features obtained for the cases of Forward, Backward and Bi-Directional, resulting in a total of four features. In the case of the Consistency evaluator, the resulting subset contained five features instead of 4.

Table 4 shows the results of the LM classifier for the subsets of features obtained with different feature selection techniques. Given that in some cases the subsets match, the results of the classification were also the same. Hence only the results that were different are shown. For the first subset of features belonging to the set obtained from L-Measure, the highest value was 0.965, using the Consistency evaluator with 11 features. In the case of the second set of features (NF, proposed in this research) AUC values ranged between 0.76 and 0.77, demonstrating its

discriminative power, however they did not reach the values achieved by the features from L-Measure.

In the case of the last set of features that is the union of the previous two sets, relatively higher values were obtained than those previously achieved independently when using features from L-Measure and the AUC for the new features attained the value 0.974, and this was the highest AUC performance for this classifier. It is pointed out that the subset obtained with Union had fewer features and better performance in the classification, making this subset more computationally efficient.

The Random Forest classifier performance is shown in Table 5. It is observed that there was a noticeable increase in AUC for the case of the new features proposed, because in the previous classifiers AUC were around 0.77 while now this value was raised up to 0.847. We also found for this classifier a higher AUC value of 0.976, which was reached with the LM set using the Consistency search strategy and 11 features. However, the Union set also had a good performance, demonstrating that the proposed new features together with those calculated with L-Measure are a good alternative for automatic classification of these classes of reconstructed neurons.

3.1 Comparison between Classifiers

Taking into account that the best performances of the classifiers were observed for the Union set of features, this was used as a basis to determine whether there are significant differences between the performance of the five classifiers that were used in the experiments. Fig shows the distribution for each classifier of the calculated AUCs in each of the classification experiments.

After applying the Friedman aligned rank test, the *p* value was less than 0.05, which means that there were significant differences between the classifiers used. To find out between which of these classifier existed significant differences, the Finnes post hoc test was applied, the results of which are shown in Figure 5. In this figure it is observed that there were neither significant differences between the classifiers forming the first group, LR, RF and KNN, nor differences between those forming the second group, e. g. C4.5 and NB, while there were significant differences between these two groups according to which the first group exhibited a better performance.

4 Conclusions

In this study, various methods to classify reconstructed (traced) neurons were compared, based on the extraction of morphological features by means of the L-Measure software and new features proposed by the authors. Feature selection methods were employed to establish an appropriate relationship between the number of cases and the number of features in order to avoid overfitting of the classification algorithms.

The data used were downloaded from the NeuroMorpho.org website, which offers the largest number of reconstructed neurons freely downloadable in the Internet. There were introduced eight new features with the purpose of increasing the discriminative power of the automatic classification algorithms. These features were based on the in-plane deviation and torsion angles in the neural tree as well as in the distance from S to Cm (*D-S-Cm*). The Union subset of features which contained the new features, showed in many cases an improvement of the classification performance. In addition, this subset had a lower number of features, which made it computationally less expensive. The statistical analysis showed that the best classifiers were LR, RF and KNN.

References

1. **Aguiar, P., Sousa, M., & Szucs, P. (2013).** Versatile morphometric analysis and visualization of the three-dimensional structure of neurons. *Neuroinformatics*, Vol. 11, No. 4, pp. 393–403. DOI: 10.1007/s12021-013-9188-z.
2. **Armañanzas, R. & Ascoli, G.A. (2015).** Towards the automatic classification of neurons. *Trends in neurosciences*, Vol. 38, No. 5, pp. 307–318. DOI:10.1016/j.tins.2015.02.004.
3. **Ascoli, G. A. (2015).** Sharing Neuron Data: Carrots, Sticks, and Digital Records. *PLoS Biol*, Vol. 13, No. 10. DOI:10.1371/journal.pbio.1002275.

4. **Calvo, B. & Santafe, G. (2016).** scmamp: Statistical Comparison of Multiple Algorithms in Multiple Problems. *The R Journal,* Vol. 8, No.1.
5. **Cerpa, N., Bardeen, M., Astudillo, C.A., & Verner, J. (2016).** Evaluating different families of prediction methods for estimating software project outcomes. *Journal of Systems and Software*, Vol. 112, pp. 48–64. DOI:10.1016/j.jss.2015.10.011.
6. **Cunningham, S.J. & Denize, P. (1994).** A tool for model generation and knowledge acquisition. *Selecting Models from Data,* Springer, pp. 471–478.
7. **Cuntz, H., Forstner, F., Borst, A., & Häusser, M. (2010).** One rule to grow them all: a general theory of neuronal branching and its practical application. *PLoS Comput Biol,* Vol. 6, No. 8. DOI:10.1371 /journal.pcbi.1000877.
8. **Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014).** Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research,* Vol. 15, No. 1, pp. 3133–3181.
9. **Foster, K.R., Koprowski, R., & Skufca, J.D. (2014).** Machine learning, medical diagnosis, and biomedical engineering research-commentary. *BioMedical Engineering OnLine*, Vol. 13, No. 1. DOI: 10.1186/1475-925X-13-94.
10. **García-Segura, L.M. & Perez-Marquez, J. (2014).** A new mathematical function to evaluate neuronal morphology using the Sholl analysis. *Journal of neuroscience methods*, Vol. 226, pp. 103–109. DOI: 10.1016/j.jneumeth.2014.01.016.
11. **Gensel, J.C., Schonberg, D.L., Alexander, J.K., McTigue, D.M., & Popovich, P.G. (2010).** Semi-automated Sholl analysis for quantifying changes in growth and differentiation of neurons and glia. *Journal of Neuroscience Methods*, Vol. 190, No. 1, pp. 71–79. DOI: 10.1016/j.jneumeth.2010.04.026.
12. **Gillette, T.A. (2015).** *Comparative Topological Analysis of Neuronal Arbors via Sequence Representation and Alignment*.
13. **Glaser, J.I. & Kording, K.P. (2016).** The Development and Analysis of Integrated Neuroscience Data. *Frontiers in Computational Neuroscience*. DOI:10.3389/fncom.2016.00011.
14. **Guerra, L., McGarry, L.M., Robles, V., Bielza, C., Larrañaga, P., & Yuste, R. (2011).** Comparison between supervised and unsupervised classifications of neuronal cell types: A case study. *Developmental Neurobiology*, Vol. 71, No. 1, pp. 71–82. DOI:10.1002/dneu.20809.
15. **Halavi, M., Polavaram, S., Donohue, D.E., Hamilton, G., Hoyt, J., Smith, K. P., & Ascoli, G. A. (2008).** NeuroMorpho.Org implementation of digital neuroscience: dense coverage and integration with the NIF. *Neuroinformatics,* Vol. 6, No. 3, pp. 241–252. DOI:10.1007/s12021-008-9030-1.00020.
16. **Hall, M.A. (1999).** *Correlation-based feature selection for machine learning.* Ph.D. thesis, The University of Waikato.
17. **Ledderose, J., Sención, L., Salgado, H., Arias-Carrión, O., & Treviño, M. (2014).** A software tool for the analysis of neuronal morphology data. *International Archives of Medicine*, Vol. 7, No. 1, pp. 1–9. DOI: 10.1186/1755-7682-7-6.
18. **Leguey, I., Bielza, C., Larrañaga, P., Kastanauskaite, A., Rojo, C., Benavides-Piccione, R., & De Felipe, J. (2016).** Dendritic branching angles of pyramidal cells across layers of the juvenile rat somatosensory cortex. *Journal of Comparative Neurology*. DOI: 10.1002/cne.23977.
19. **Lu, Y., Carin, L., Coifman, R., Shain, W., & Roysam, B. (2014).** Quantitative arbor analytics: unsupervised harmonic co-clustering of populations of brain cell arbors based on L-measure. *Neuroinformatics,* Vol. 13, No. 1, pp. 47–63.
20. **McGarry, L. M., Packer, A. M., Fino, E., Nikolenko, V., Sippy, T., & Yuste, R. (2010).** Quantitative classification of somatostatin-positive neocortical interneurons identifies three interneuron subtypes. *Frontiers in neural circuits*, Vol. 4, DOI:10.3389/fncir.2010.00012.
21. **Parekh, R. & Ascoli, G. A. (2015).** Quantitative Investigations of Axonal and Dendritic Arbors Development, Structure, Function, and Pathology. *The Neuroscientist*, Vol. 21, No. 3, pp. 241–254. DOI:10.1177/1073858414540216.
22. **Scorcioni, R., Polavaram, S., & Ascoli, G.A. (2008).** L-Measure: a Web-accessible tool for the analysis, comparison, and search of digital reconstructions of neuronal morphologies. *Nature protocols,* Vol. 3, No. 5, pp. 866–876. DOI: 10.1038/nprot.2008.51.
23. **Vasques, X., Vanel, L., Villette, G., & Cif, L. (2016).** Morphological Neuron Classification Using Machine Learning. *Frontiers in Neuroanatomy*. DOI: 10.3389/fnana.2016.00102.

Article received on 02/12/2016; accepted on 14/06/2017.
Corresponding author is José D. López-Cabrera.