Classification of Genomic Regions by Chaos Game Representation Images and Fractal Dimension

Caetano Jimenez Carezzato[1], Junior Barrera[1], Sandro José de Souza[2] and Luciano da Fontoura Costa[3]

[1]DCC-IME-USP, University of São Paulo – Brazil – {caetano, jb}@vision.ime.usp.br

[2]ILPC – Brazil – sandro@compbio.ludwig.org.br, [3]IFSC-USP, University of São Paulo – Brazil – luciano@if.sc.usp.br

Abstract

This work describes a new approach to classify genomic regions by applying the multiscale fractal dimension [2] over images generated by chaos game representation (CGR) of sequences. Since the introduction of chaos game representation of sequences [8], evidences have been found that it is possible to obtain discriminative measures from images produced by this methodology and to feed such features into classifiers in order to identify the origin of gene fragments. Also, it has been showed [3] that it is possible to reconstruct filogenetic trees just using this representation.

In this project, we propose a new feature extractor of sequences that can help the classification of genomic regions. The feature extraction consists of determining the CGR of a sequence and estimating the fractal dimension of the generated image. Such measurements are organized into a feature vector.

For certain genomic regions, as the GC-content (mean percentage of guanine and cytosine) changes, the CGR also changes. We are going to verify if CGR contains more meaningful information than GC-content that can be easily and fastly exploited for genomic regions classification. Our preliminary results based on cluster techniques show that methods based on this approach should be better than GC-content based ones. We compared, for each approach, the ability to distinguish genic regions of the human chromossome 22. Currently, we are finishing the validation of the tests and developing a way to classify genic regions.

CGR Images

The chaos game is an algorithm, generally controlled by a series of random numbers, which allows one to produce images (attractors) of fractal structures. The use of DNA sequences, rather than random numbers, has been proposed to control the chaos game [8]. We call an image generated by the DNA sequence of Chaos Game Representation (**CGR**) of the DNA sequence.

CGR consists of disposing a table of frequencies in one square matrix using a recursive rule.

Given one window w of size s, we build the square matrix m of size $n \times n, n = 2^s$. Each point of the matrix m represents the frequency of one possible configuration of the window w.

In the case where $s = 1$, there are only 4 rows in the frequency table, the alphabet itself. We dispose the following example table

A	10
T	9
C	4
G	13

in the following 2×2 matrix:

C	G
A	T

$\Longrightarrow$

4	13
10	9

In the case where $s = 2$, there are 16 rows in the frequency table as shown in the following example:

AA	10	CA	35
AT	9	CT	44
AC	4	CC	0
AG	13	CG	1
TA	11	GA	22
TT	15	GT	7
TC	24	GC	90
TG	17	GG	19

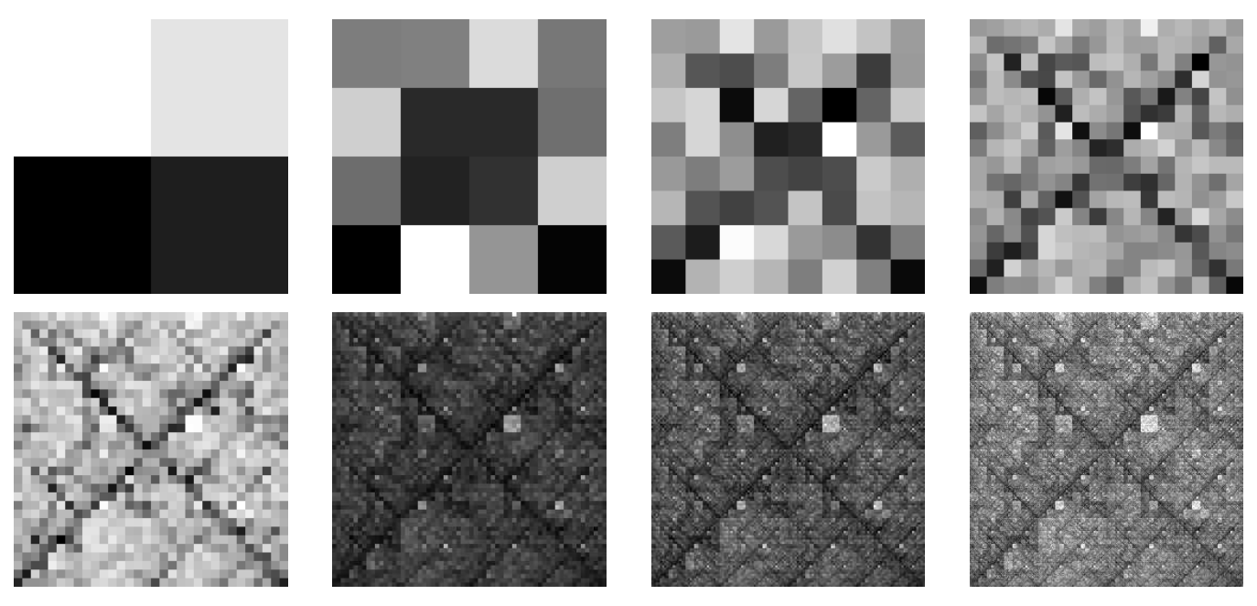
C	G
A	T

CC	GC	CG	GG
AC	TC	AG	TG
CA	GA	CT	GT
AA	TA	AT	TT

$\Downarrow$

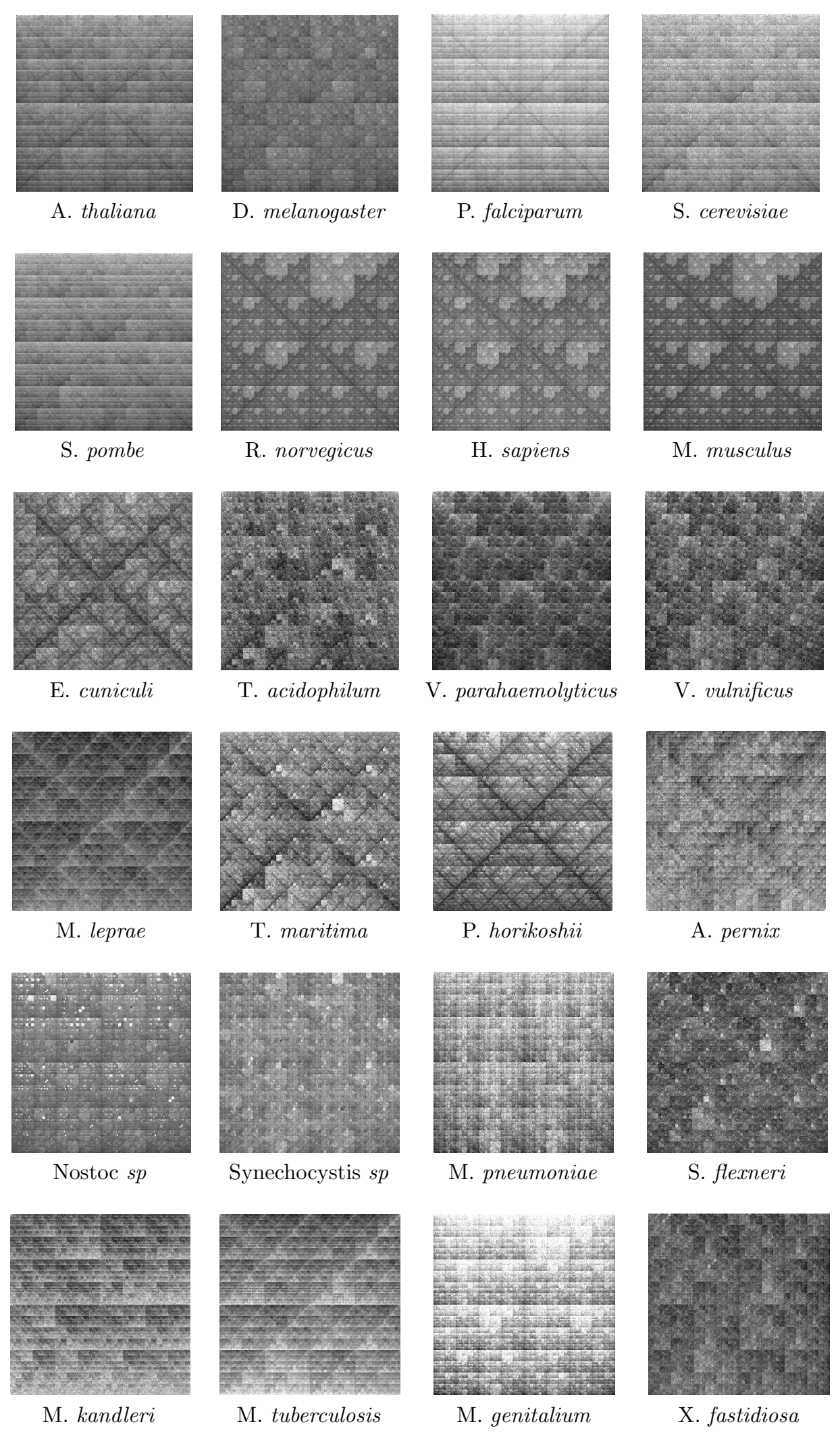
0	1	90	19
35	44	22	7
4	13	24	17
10	9	11	15

We can use gray levels, instead of using numbers, for visualization. In the following example, we have the CGR images for $s = 1 \ldots 8$ of the bacteria A. *fulgidus*. The dark intensity is direct proportional to the frequency. The last three images are log normalized.

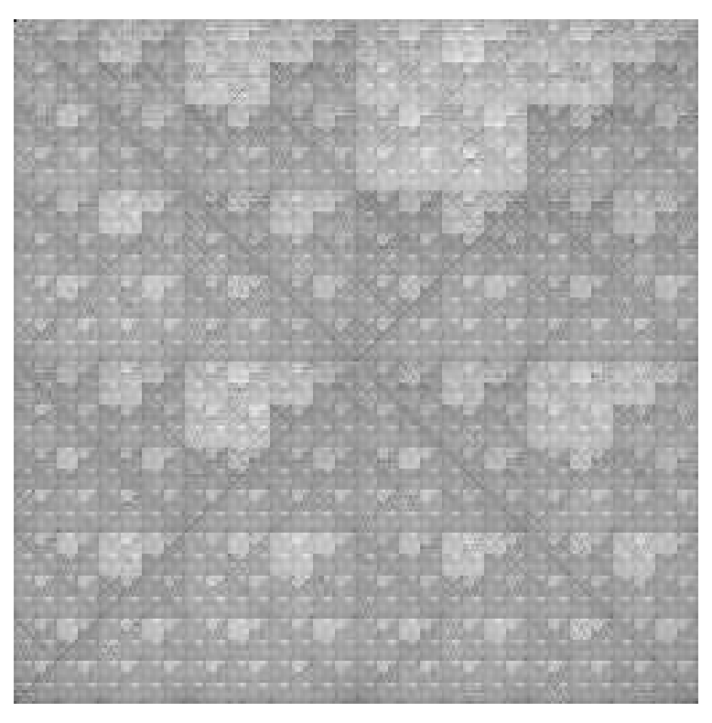


In the next set of imagens we can see the CGR image log normalized, $s = 8$ for some organisms.

Note that there are significative pattern differences between each organism.

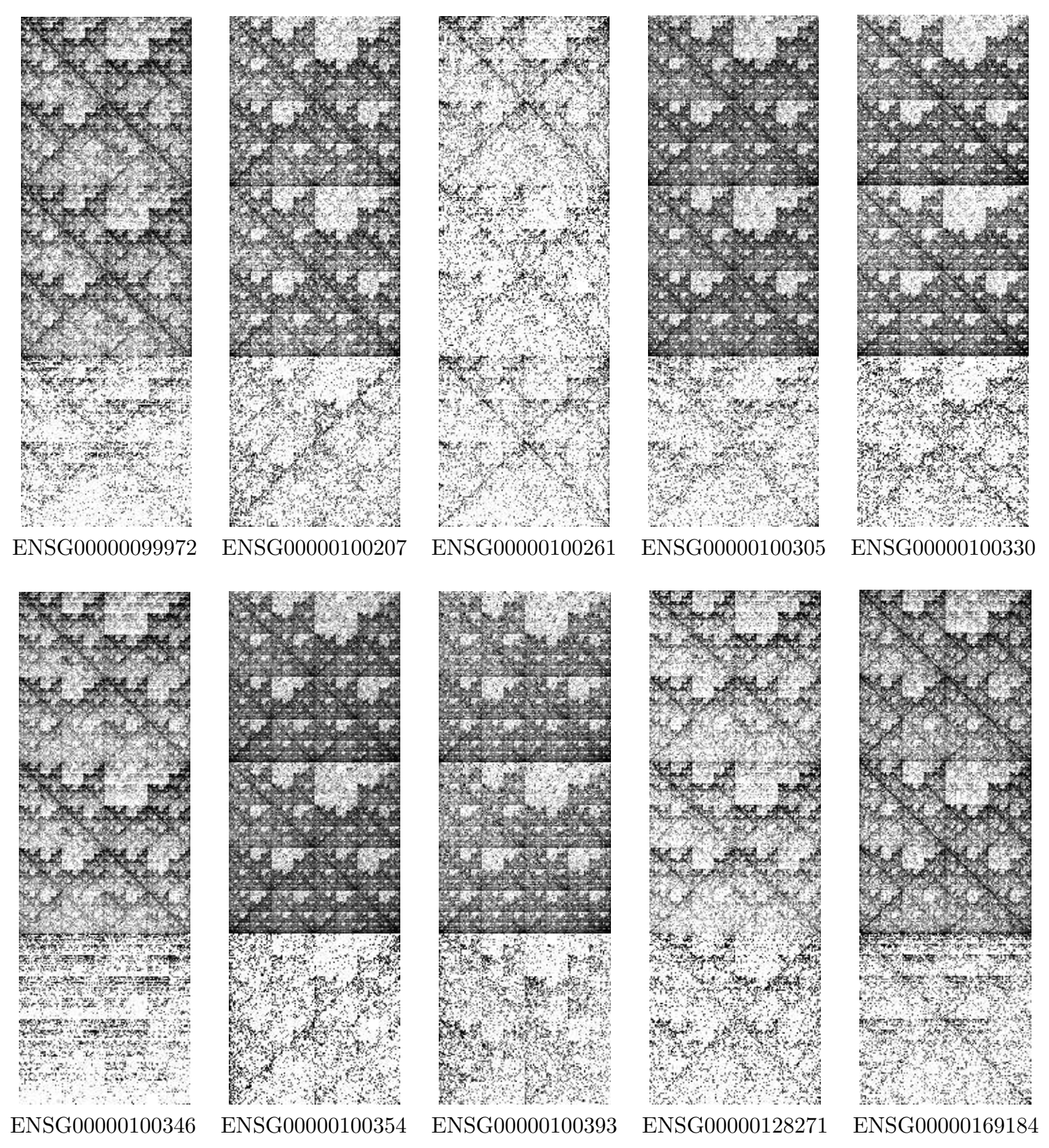


Deschavanne [3] showed that these patterns can be used to build filogenetic trees. Is is also possible to reconstruct the CGR image using just small pieces of the genome. This property can be used to construct a classifier that can decide from which organism is a given piece of DNA [9].



The image above shows the log normalized CGR image of the chromossome 22 where $s = 8$. In the following set of images, we show some CGR images of genes of the chromossome 22. For each image, there are three square sub-images inside. The top one is the CGR of the gene DNA sequence, the medium one is the CGR of the introns sequences and the bottom one is the CGR of the exons sequences.

Note that some genes look like each other and with the CGR image of the chromossome 22 above. Also note that some gene CGR images are very different from each other, specially the exon region images.



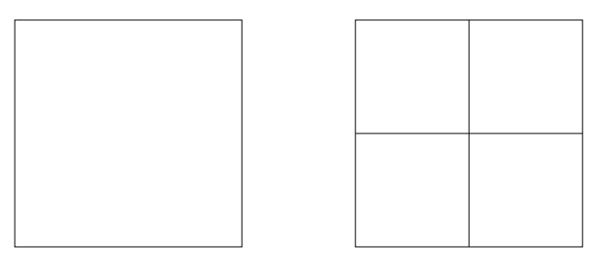
We are trying to explore local variations of the CGR image in the chromosome. Our main goal is to extract some useful information from CGR images that can be explored locally. We are trying to do this using fractal dimension techniques.

Fractal Dimension

To explain the concept of fractal dimension, it is necessary to understand what we mean by dimension. Obviously, a line has dimension 1, a plane dimension 2, and a cube dimension 3.

So why is a line one-dimensional and the plane two-dimensional? Note that both of these objects are self-similar. We may break a line segment into 4 self-similar intervals, each with the same length, and each of which can be magnified by a factor of 4 to yield the original segment. We can also break a line segment into 7 self-similar pieces, each with magnification factor 7, or 20 self-similar pieces with magnification factor 20. In general, we can break a line segment into n self-similar pieces, each with magnification factor n.

A square is different. We can decompose a square into 4 self-similar sub-squares, and the magnification factor here is 2. Alternatively, we can break the square into 9 self-similar pieces with magnification factor 3, or 25 self-similar pieces with magnification factor 5. Clearly, the square may be broken into n^2 self-similar copies of itself, each of which must be magnified by a factor of n to yield the original figure as shown in the figure bellow.



Finally, we can decompose a cube into n^3 self-similar pieces, each of which has magnification factor n.

Fractal dimension is a measure of how "complicated" a self-similar figure is. In a rough sense, it measures "how many points" lie in a given set. A plane is "larger" than a line, while other curves sit somewhere in between these two sets. The fractal dimension provides a quantitative characterization of the complexity of curves as induced by self-similarity.

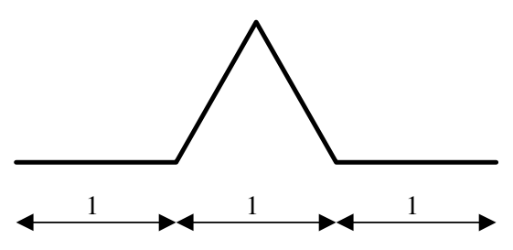
On the other hand, all three of these sets have the same number of points in the sense that each set is uncountable. Somehow, though, fractal dimension captures the notion of "how large a set is" quite nicely. Therefore, the fractal dimension provides a nice indication of how much the curve extends itself through space. As a consequence, more intricate curves will cover the surrounding space more effectively, leading to higher fractal dimensions.

Now we see an alternative way to specify the dimension of a self-similar object: the dimension is simply the exponent of the number of self-similar pieces with magnification factor n into which the figure may be broken. So we can write

$$F = \text{fractal dimension} = \frac{\log(\text{number of self-similar pieces})}{\log(\text{magnification factor})}$$

So, for a straight line, $F = \frac{\log(n)}{\log(n)} = 1$, for a plane, $F = \frac{\log(n^2)}{\log(n)} = 2$ and for a cube, $F = \frac{\log(n^3)}{\log(n)} = 3$.

For the classical Koch Triadic curve, which basic pattern is bellow, the fractal dimension $F = \frac{\log(4)}{\log(3)} \approx 1.26$.

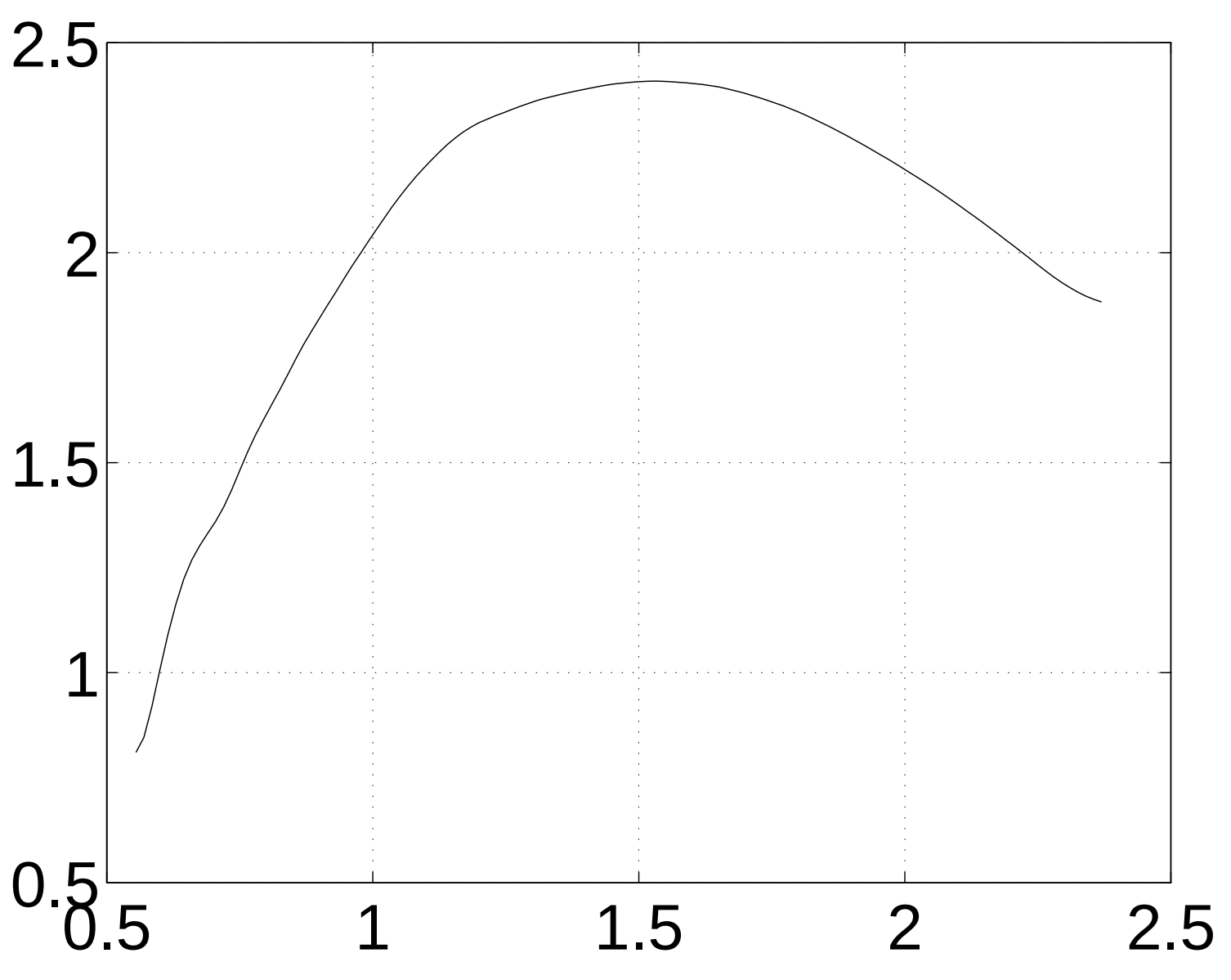


"Despite the good potential of this measure to characterize complexity in a more objective fashion, its extension to real objects is complicated by the fact that the latter are not perfectly self-similar. In fact, only a few orders of similarity are usually found for natural objects, such as the three or four orders found in fern leaves. Indeed, the fractality of such objects, especially when represented in digital images, is limited at both microscopic and macroscopic scales. First, for scales smaller than the image resolution, the fractal dimension tends to zero, the dimension of the image pixels. On the other hand, for scales larger th an the object, the respective dimension tends to zero, as the object tends to behave as a point for large distances. Therefore, real objects will present higher fractal values only along limited intervals of spatial scale. This problem can be suitably addressed by using the multiscale extension of the fractal dimension recently described in [1], which involves the numerical estimation of the first derivative of a log log cumulative function, more specifically the graph of the logarithm of the dilated area in terms of the logarithm of the spatial scale (i.e. radius of the dilating discs).

This extension involves obtaining not a scalar value of fractal dimension as usually done, but expressing a fractal function in terms of the spatial scale that properly reflects the behavior of the object when observed at different magnifications. Therefore, the multiscale fractal dimension represents a less degenerate geometric characterization, in the sense of preserving more information about the geometry of the original object." [2]. We are using this approach to extract information from DNA sequences.

More information on fractal dimension and chaos game can be found in Devaney [4] and Costa [2].

The following image is an example of multiscale fractal dimension graph obtained from the CGR image of the human chromossome 22. The x-axis is the resolution (log of the radius of the ball size in pixels) and the y-axis is the fractal dimension.



Preliminary Results

We did a series of 25 clustering experiments to evaluate the potential of the fractal dimension based approach to searching genes. We compared 5 kinds of characteristics:

1. GC-content
2. CGR image of the sequence
3. CGR image log normalized of the sequence
4. Multiscale Fractal Dimension of CGR image
5. Multiscale Fractal Dimension of CGR image log normalized

We extracted the above characteristics from 526 genes of the chromossome 22 (complete sequence) and 655 continuous regions of the same chromossome where there is a big confidence that there is no genes or genes pieces in them.

The basic idea was to try to separate these 1181 DNA sequences just using one of the above characteristics each time. For each characteristic we calculated a matrix of distances D using, respectively, the following rules where x and y are the two points being considered:

1. GC-Content - Sum of the module of the differences of x and y
2. CGR image - $D = \sum_{i,j} |x_{ij} - y_{ij}|$
3. Multiscale Fractal Dimension - $D = \sum_{i} |x_i - y_i|$

For each distance matrix D, we applied 5 hierarchical clustering algorithms available in Matlab 6 trying to separate the whole set.

For each one of the 25 experiments we calculated 1180 clustering of the same method with the number of clusters from 2 to n.

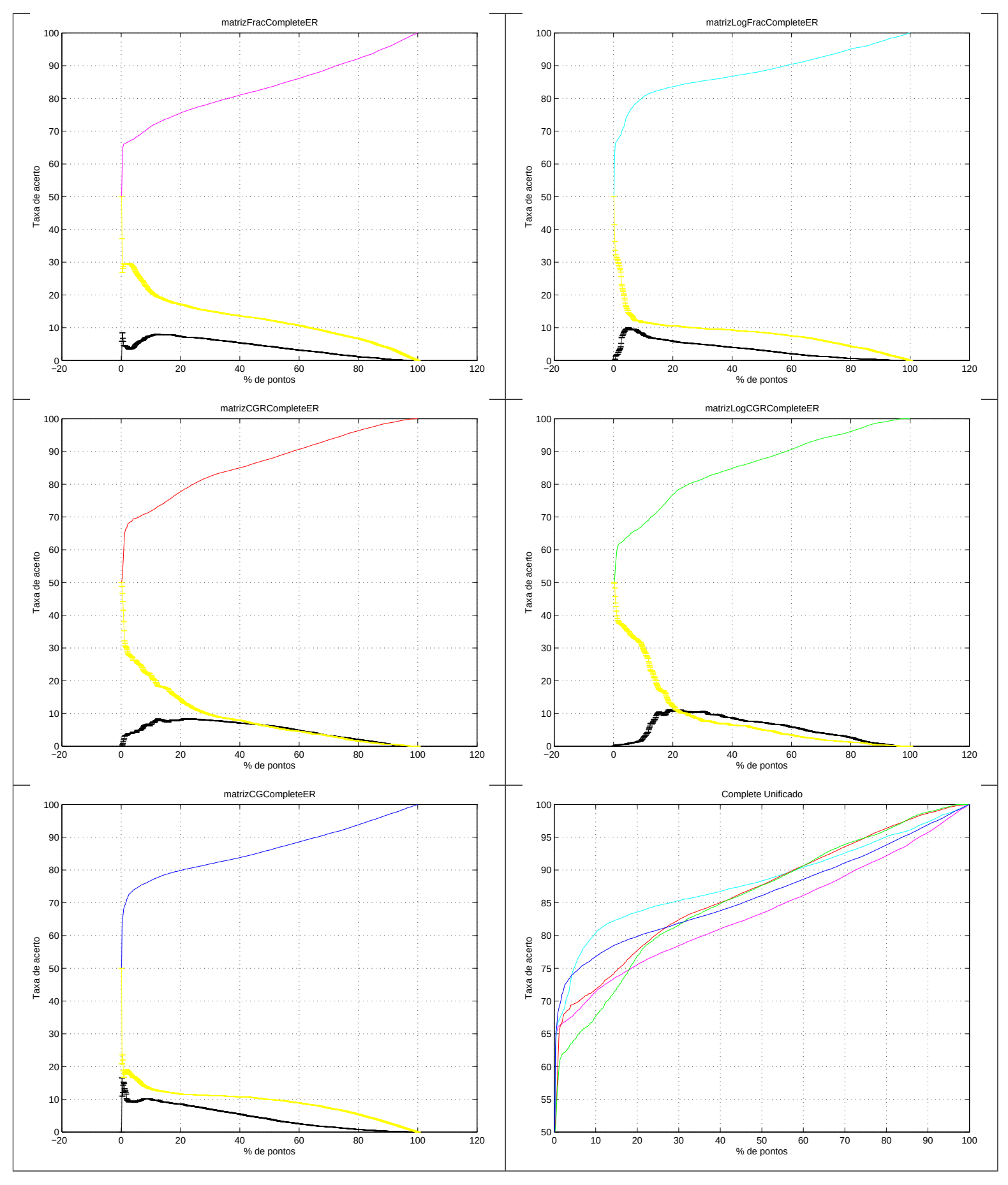
The "error rate" of each experiment is the minimum error possible that we do by labelling each cluster with either "genes" or "not genes". So, for each clustering, the total error will be the sum, for each cluster, of the minimum number between genes and not genes elements in the cluster.

We applied a simple bootstrap method that consists of selecting randomly 350 regions of genes and 350 regions of not genes. We build 100 data sets using this method and applied the clustering procedure for each data set.

In the following set of graphs we can see the results for the "Complete" clustering method. The x-axis represents the number of clusters (in percentage of the total set size) and the y-axis the error rate. Each graph represents a different characteristic.

For the first 5 graphs, in black we have the mean of false positive rates (FP - labelling as gene regions that are not genes) with error bars and in yellow the mean of false negative rates (FN - labelling as not gene regions that are genes) with error bars. In the third color we have the mean of hit rates (100 - mean(FP) - mean(FN)). The last graph shows the mean of hit rates for each characteristic.

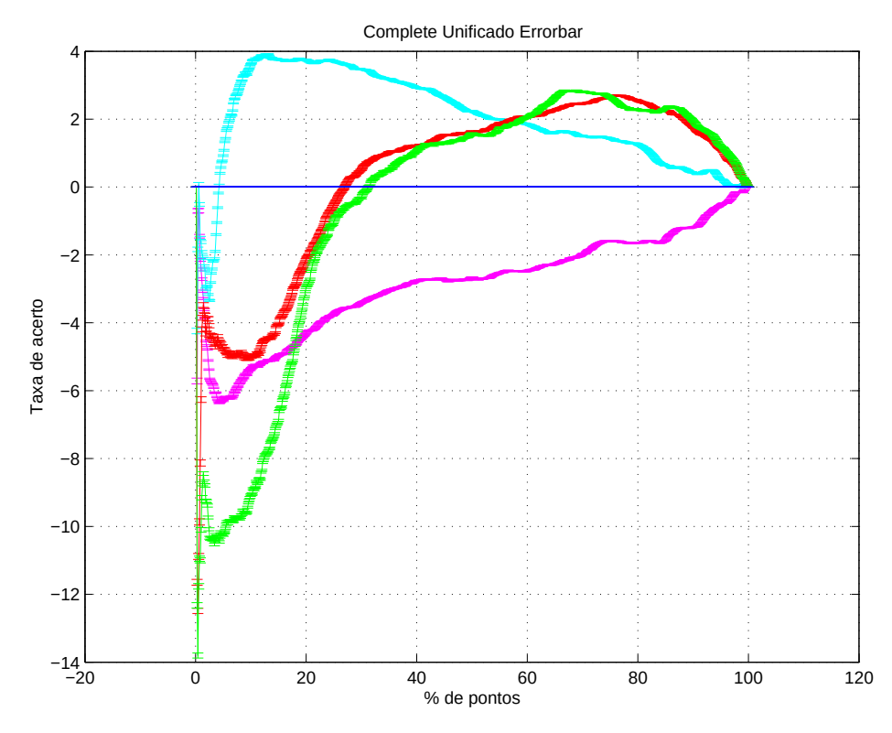
- Magenta - Multiscale Fractal Dimension graph of CGR image
- Cyan - Multiscale Fractal Dimension graph of CGR image log normalized
- Red - CGR image
- Green - CGR image log normalized
- Blue - CG-Content



It is important to observe that for the others clustering methods the results are almost similar. In all experiments the cyan curve (Multiscale Fractal Dimension graph of CGR image log normalized) is better than the others curves. Note that it only make sense to compare the curves in the first 20% part (the beginning of the graph) because values higher than that characterizes overfitting.

Theses preliminary results are interesting because they don't use any additional biological information. In spite of the separation for all methods is not excellent, these results suggests that measures based on fractal dimension of CGR images of sequences should be used in gene searching systems. Note that this is a preliminary result that must be validated through more experiments. In the way they were done, these experiments are equivalent to training a supervised classifier and applying the classifier in the training set.

The following graph shows the mean hit rate for each characteristic being considered, with error bars.



Further Work in Progress

Currently we are doing a series of experiments to validate the application of the multiscale fractal dimension technique. We pretend to test the ability of classification of this measure using randomly selected regions of the human chromossome 22 and to extend these experiments to others chromossomes.

It is important to note that while the Multiscale Fractal Dimension graph of CGR image log normalized has produced nice results, the Multiscale Fractal Dimension graph of CGR image linearly normalized was the worst technique evaluated. This suggests that normalization plays an important role in the process, so we pretend to make experiments to find a nice normalization that should be used.

The biggest problem blocking the realization of these experiments was the algorithm to calculate the fractal dimension. It was very slow. Now, we have just finished a new nice and very faster algorithm so we plan to have final results in some weeks.

This approach needs the definition of a way to build a classifier given a training set. We also pretend to study this problem.

We also may adapt this approach to other problems in Computational Biology such as the determination of intron/exon and exon/intron frontiers.

Acknowledgment

This work is supported by grant 01/03975-5 to C.J.C. from FAPESP of the Brazilian Government. This work is linked to the thematic project **CAGE** http://www.vision.ime.usp.br/∼cage/ FAPESP 99/07390-0.

References

[1] Costa, L. da F., Campos, A. G., and Manoel, E. T. M. An integrated approach to shape analysis: results and perspectives. In *Int. Conf. on Quality Control by Artificial Vision* (2001), pp. 23–34.

[2] Costa, L. da F., Manoel, E. T. M., Faucereau, F., Chelly, J., van Pelt, J., and Ramakers, G. A shape analysis framework for neuromorphometry. *Network 13* (2002), 283–310.

[3] Deschavanne, P. J., Giron, A., Vilain, J., Fagot, G., and Fertil, B. Genomic signature: Characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol. 16*, 10 (1999), 1391–1399.

[4] Devaney, R. L. *Chaos in the Classroom.* Boston University, 1995. http://math.bu.edu/DYSYS/chaos-game/chaos-game.html.

[5] Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Classification.* John Wiley and Sons, 2000.

[6] Jain, A. K., Duin, R. P. W., and Mao, J. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 1 (2000), 4–37.

[7] Jain, A. K., Murty, M. N., and Flynn, P. J. Data clustering: A review. *ACM Computing Surveys 31*, 3 (September 1999).

[8] Oliver, J. L., Bernaola-Galván, P., Guerrero-García, J., and Román-Roldán, R. Entropic profile of DNA sequences trought chaos-game-derived images. *Journal of Theorical Biology 160* (1993), 457–470.

[9] Sandberg, R., Winberg, G., Bränden, C.-I., Kaske, A., Ernberg, I., and JoakimCöster. Capturing whole-genome characteristics in short sequences using a naïve bayesian classifier. *Genome Research 11* (2001), 1404–1409.

[10] Theodoridis, S., and Koutroumbas, K. *Pattern Recognition.* Academic Press, 1999.