Clustering Digital Data by Compression: Applications to Biology, Medical Images and Remote Sensing

Bruno Carpentieri

Abstract— A new, "blind", approach to clustering by compression that classifies digital objects depending on how they pair-wise compress has been recently proposed. This clustering is based on the Normalized Compression Distance (NCD) distance metric that considers only the pairwise compressibility of data but does not include any explicit semantic knowledge. In this paper we review this clustering method and we show how this approach can be used in bio-sequences clustering, medical images clustering, and remote sensing applications.

***Keywords*—** Clustering, Data Compression, Prediction Coding, Remote Sensing.

I. Introduction

Compression is the coding of data to minimize its representation. In compressed form data can be stored more compactly and transmitted more rapidly. Recent advances in compression span a wide range of applications and new general compression methods are always being developed, in particular those that allow indexing over compressed data or error resilience. Compression also inspires information theoretic tools for pattern discovery and classification. Today we know that data compression, data prediction, data classification, learning and data mining are all facets of the same (multidimensional) coin.

Paul Vitányi and his Ph.D..student Rudi Cilibrasi have recently proposed a new strategy for clustering that is based on compression algorithms (see Vitányi and Cilibrasi [2] and Cilibrasi [3]). This approach leads to an interesting clustering algorithm that does not use any "semantic" information on the data to be classified but does a "blind" and effective classification of digital data that is based only on the data compressibility and not on its "meaning".

They have introduced a new distance metric, called NCD (Normalized Compression Distance). NCD is based on data compression and it can be used as a metric to cluster digital data.

In this paper we successfully apply this clustering by compression in different domains: biology, medical images, and remote sensing.

B. Carpentieri is with the Dipartimento di Informatica of the Università di Salerno, 84084 Fisciano (SA), ITALY, phone: +39 089969500 (email bc@di.unisa.it).

In the next Section we review the clustering by compression approach. Section 3 presents the results obtained by applying the clustering by compression approach to biological digital data. Section 4 is devoted to the results obtained on medical images and in Section 5 we discuss the clustering by compression applications on remote sensing data. Finally, in Section 6, we present our conclusions and outline new research directions.

II. The "Clustering by Compression" Approach to Classification

Clustering is the process of organizing objects into groups based on similarity. Clustering is performed by assigning a set of objects to homogenous groups with respect to a given distance metric.

It is an unsupervised learning problem but generally we embed information about the objects that have to be clustered in the distance metric that therefore includes our knowledge of the data.

In clustering by compression this is not the case: the NCD distance metric is based only on compressibility and it does not include any explicit semantic knowledge.

Compression can be used as a distance metric: compression ratios signify a great deal of important statistical information. In fact let's suppose that we have two digital files A and B; if we compress A and B, by using gzip or bzip or with any general-purpose, lossless, data compressor, we can indicate with L(A) and L(B) the compressed lengths of A and B.

If we need to compress both A and B then we have two choices. The first is to compress first A and then B (or vice-versa), so we have as resulting length of the two compressed files L(A) + L(B). The second is to append file B to file A and then to compress the resulting file AB by obtaining length L(AB).

Experimentally it is possible to show that, if and only if A and B are "similar", then:

$L(AB) << L(A) + L(B)$

Therefore if we want to cluster digital files we might be able to do this by considering how well they compress together in pairs.

Vitányi and Cilibrasi, in [2] and [3], have introduced the concept of Normalized Compression Distance (NCD) that measures how close (or different), from the compression point

of view, two files are one from another.

Given a general purpose lossless compressor with length function L, the Normalized Compression Distance between two files x and y, can be defined as:

$$\mathrm{NCD}(x,y) = \frac{L(xy) - \min\{L(x), L(y)\}}{\max\{L(x), L(y)\}}$$

where L( ) is the length, in bits, of the compressed file.

The Complearn software ([3]) is a powerful software tool that takes as two inputs a set of digital files and a general purpose data compressor and outputs a clustering of the data objects that can be visualized as an un-rooted binary tree.

Complearn, at the moment of writing, is freely available from complearn.org. The software works by building a distance matrix obtained by computing the NCDs between each pair of files in the data set that we want to cluster. This matrix is then given as input to a classification algorithm based on the quartet method: the final output is an un-rooted binary tree where each digital object is now represented as a leaf of the tree.

Complearn requires no background knowledge about data. There are no domain-specific parameters, and only a few general settings, to set.

The clustering results depend on the choice of the compressor: different compressors lead to different clustering trees.

III. Clustering biological data

The computational analysis of biological data is today a challenge of considerable interest

Complearn has not been designed specifically to cluster biological data. In particular the main obstacles for Complearn are the large amount of data (for example the size of the human DNA consists of about three billion elements) and the complex informational content.

Our objective is to verify the behavior of CompLearn in identifying relationships between protein sequences contained in the various digital files by clustering them in appropriate groups.

The content of the first dataset consists of one hundred and one protein sequence files. The files are semantically homogeneous, in fact they contain information related to proteins of the same type, i.e. that represents transmembrane receptors, namely integral membrane proteins localized mainly at the level of the cytoplasmic membrane.

The transmembrane proteins differ by the membrane proteins because, unlike membrane, they stretch in the hydrophobic interior of the lipid bilayer, while the membrane proteins remain adherent to only one of the faces of the membrane. Tying with one specific molecule, defined ligand, the transmembrane receptors mediate an intracellular biochemical response, acting as a fundamental role in the process of signal transduction.

The tests were carried out by considering all the one hundred and one files that are named with the scientific nomenclature of the species to which they refer.

We have begun to perform our biological test by using the standard BZLIB compressor that is included in the Complearn suite.

Fig. 1 shows the un-rooted clustering tree obtained by Complearn. In this tree we notice for example that files that refer to mammals, as for instance Homo Sapiens, Mus Musculus, Rattus Norvegicus, etc., are close.

Fig 2 zooms on three particulars of Fig. 1 to show how organisms that are close in nature are still close in the clustering.

The quality of the hierarchical clustering tree can be measured by the normalized tree benefit score S(T) that is associated with the clustering.

For this clustering tree the value of S(T) is 0.915592 (where 1 is the maximum) so the clustering tree is good.

We speculated that the clustering in Figure 1 could be improved by using a more appropriate compressor. In fact in Figure 3 we show a new clustering tree for the same data set of Figure 1 obtained this time by using the LZMA compressor: an improved and optimized version of the LZ77 algorithm that was designed specifically for biological data.

With this new compressor the clustering tree is improved, and the S(T) values increases to 0.96.

IV. Clustering medical images

Medical images are an important source of digital data. There is an increasing interest in this data because of new medical applications, such as telemedicine, tele-radiology, real time tele-consultation, PACS (Picture Archiving and Communication Systems), etc..

Some of these digital imaging technologies, such as magnetic resonance (MR) and computed tomography (CT), produce three-dimensional images.

In the case of MR and CT images each examination produces multiple slices. A slice is the graphical representation of a cross section of the part of the human body that is currently analyzed. The collection of all these slices composes a 3-D medical image.

We have experimented the clustering strategy on several CT and MR images by using the same test set that is generally used in the data compression of this kind of data and that is described in Table I.

Each slice has 256 columns, 256 lines and 8-bit per sample.

From the compression point of view, the 3-D medical images show a strong correlation among consecutive slices (inter-slice) and an high relation in the spatial context (intra-slice).

We have clustered this set of medical images by using the complearn approach and Figure 4 shows the clustering tree obtained.

The result is almost optimal. In fact the images belonging to the type CT (computed tomography) are grouped together while those belonging to the type MR (magnetic resonance imaging) are grouped together.

Type	*History* (Age / Sex / # of Slices)	Image
CT	*Tripod fracture* (16 / M / 192)	CT_skull
	Healing scaphoid dissection (20 / M / 176)	CT_wrist
	Internal carotid dissection (41 / F / 64)	CT_carotid
	Apert's syndrome (2 / M / 96)	CT_Aperts
MR	*Normal* (38 / F / 48)	MR_liver_t1
	Normal (38 / F / 48)	MR_liver_t2e1
	Left exophthalmos (42 / M / 48)	MR_sag_head
	Congenital heart disease (1 / M / 64)	MR_ped_chest

Table 1: Medical Images

V. Clustering remote sensing data

Hyperspectral data are usually generated by using sensors installed on airplanes: for example those used by NASA, known as the Airborne Visible \ Infrared Imaging Spectrometers (AVIRIS), or by using special satellites.

These sensors, through the observation of an object, are able to capture a large portion of the object's electromagnetic spectrum.

Given that any object has an unambiguous fingerprint on the electromagnetic spectrum, hyperspectral data allows the identification of different types of materials: for example, the spectral signature of oil can help mineralogists to find new oil wells.

Hyperspectral data need to be efficiently compressed to be stored and / or transmitted.

Dynamic range and noise level of AVIRIS data (instrument noise, reflection interference, vibrations, etc.) are higher than those in photographic images. This is why a spatial predictor like the median predictor in JPEG LS tends to fail on this kind of data. Nevertheless, the speed and efficiency of JPEG-LS would be highly attractive in the context of on-board, hardware implementation.

The Spectral oriented Least SQuares (SLSQ) optimized linear predictor is today the state of the art for the low complexity compression of hyperspectral images (see [4], [5] and [6]).

SLSQ determines for each sample the coefficients of a linear predictor that is optimal (i.e., that minimizes the energy of the prediction error) with respect to a three dimensional subset of past data.

From the analysis of the intra band correlation, it can be observed that in each AVIRIS hyperspectral image there is a sub-set of bands that have a very low correlation with respect to any other bands.

We have named these bands as NR-bands (that can be grouped into NR-sets).

The basic idea behind our approach is that the NR-bands are not efficiently predicted by a three-dimensional predictive model. Thus, our approach is based on the identification of the NR bands for which it will be used a bi-dimensional predictive model which shall exploit only the spatial correlation.

Since the identification of the NR-bands is substantially a data clustering problem, we can use the CompLearn Toolkit as the data clustering tool in order to classify the AVIRS bands.

Therefore we will be able to distinguish the NR-bands from the other bands.

By analyzing the resulting trees, we can observe that the NR-bands are grouped in at most two sub-trees.

If we define a cut as the elimination of a sub-tree from the clustering tree, we need at most of two cuts in order to obtain the initial NR-sets.

Figure 5 shows as example the two cuts on the resulting tree produced by the analysis of the first scene of Lunar Lake.

By identifying the NR-sets through the clustering by compression approach we have been able to improve significantly the performances of the SLSQ algorithm on AVIRIS images. Therefore given a set of sensors it is possible to improve the SLSQ performance by considering a few hyperspectral images obtained by those sensors and by identifying the NR-sets on those images through the clustering by compression approach. Those NR-sets will be used for all the other images acquired by those sensors.

VI. Conclusion

Compression inspires information theoretic tools for clustering, pattern discovery and classification. Complearn is a powerful tool for clustering by compression. It is a blind clustering but it can be a valid method in many domains.

Here we have experimented this approach with success on biological data, medical images and hyperspectral images.

Future work will include a deeper analysis of the complearn dependence from the compression algorithm, a wider experimentation of the clustering method and new applications.

Acknowledgments

We wish to thank Marco Pastena, Annarita Leone, Raffaele Pizzolante and Giovanni Murano for performing some of the experiments described in this paper.

References

[1] B. Carpentieri, "A "Blind" Approach to Clustering Through Data Compression". International Journal of Matematics and Computers in Simulation, Vol.7, No.2, pp. 162-170, 2013.

[2] R. Cilibrasi and P. Vitányi. "Clustering by Compression". IEEE Transactions on Information Theory, 51(4):1523-1545, 2005.

[3] R. Cilibrasi, Statistical Inference through Data Compression. Ph.D. Dissertation, University of Amsterdam, 2007.

[4] B. Carpentieri, J.A. Storer, G. Motta, F. Rizzo, "Compression of Hyperspectral Imagery". In Proceedings of IEEE Data Compression Conference (DCC '03), Snowbird, UT, USA, 25–27 March 2003; pp. 317–324.

[5] F. Rizzo, B. Carpentieri, G. Motta, J. A. Storer, "Low-complexity lossless compression of hyperspectral imagery via linear prediction". IEEE Signal Process. Lett., 12, 138–141, 2005.

[6] [6] R. Pizzolante, B. Carpentieri, "Visualization, Band Ordering and Compression of Hyperspectral Images". Algorithms, V. 5, pp, 76-97, 2012.

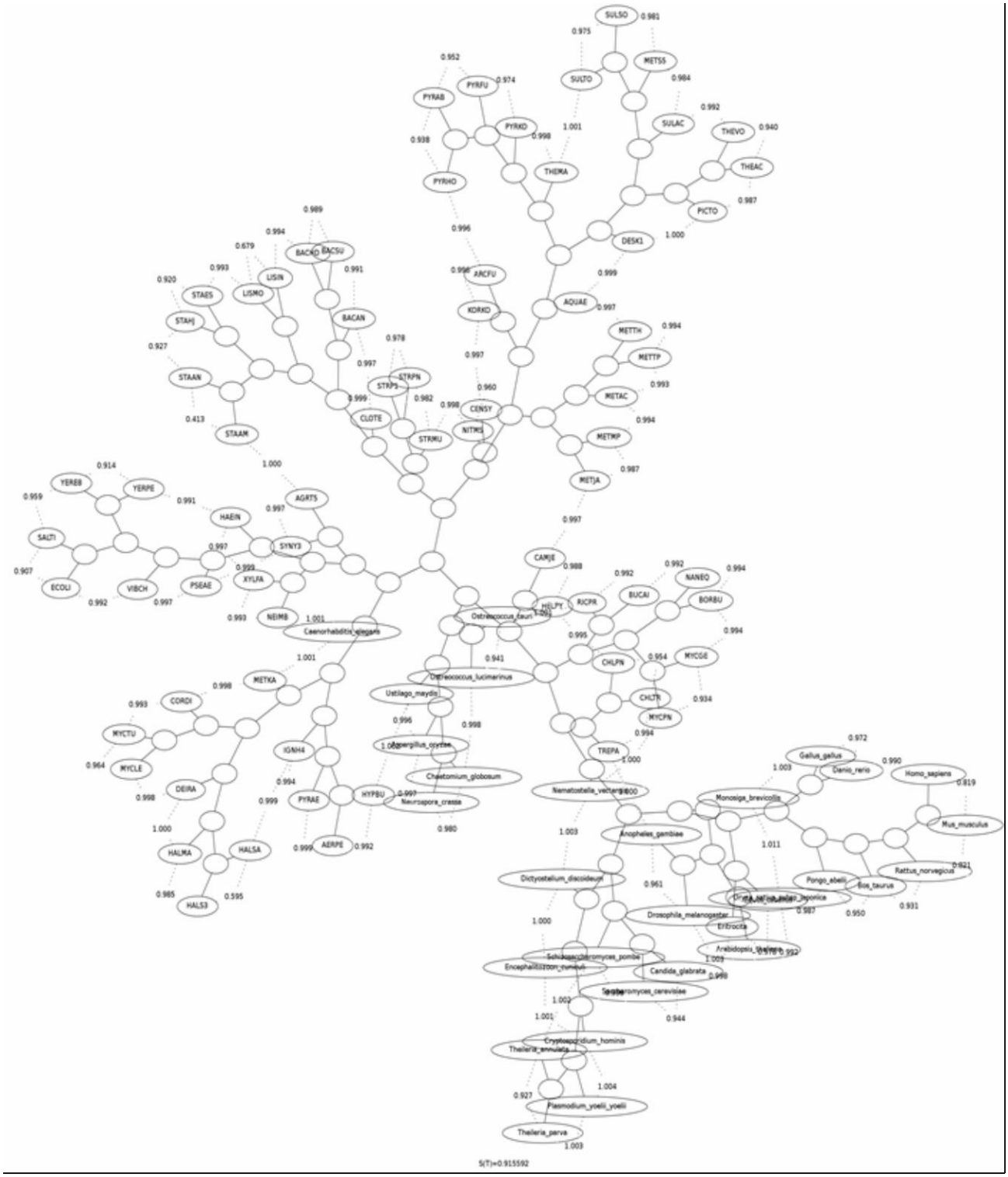

Figure 1: Clustering obtained on the 101 transmembrane receptors by the BZLIB compressor

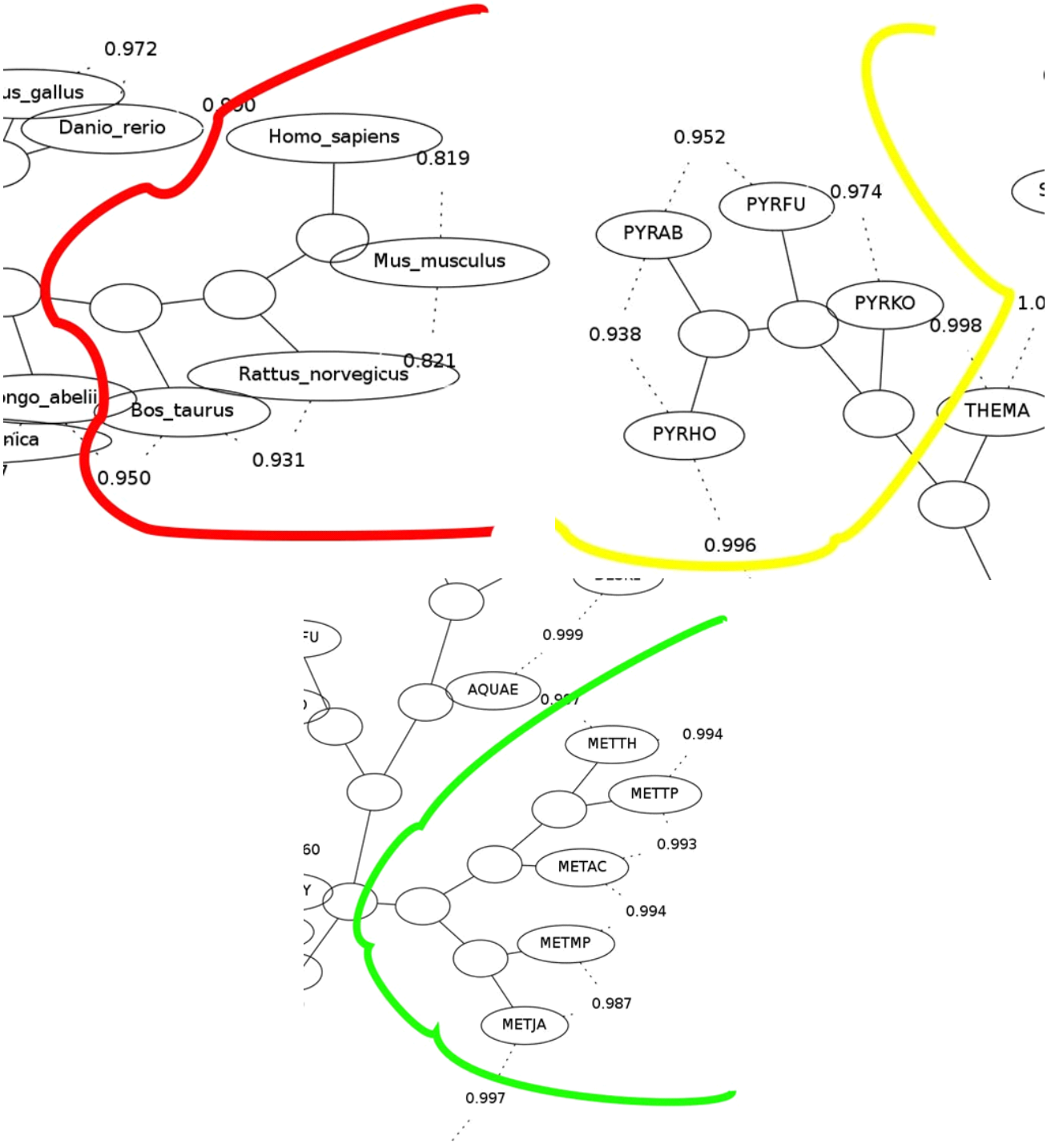

Figure 2: Three zooms of the clustering in Fig.1.

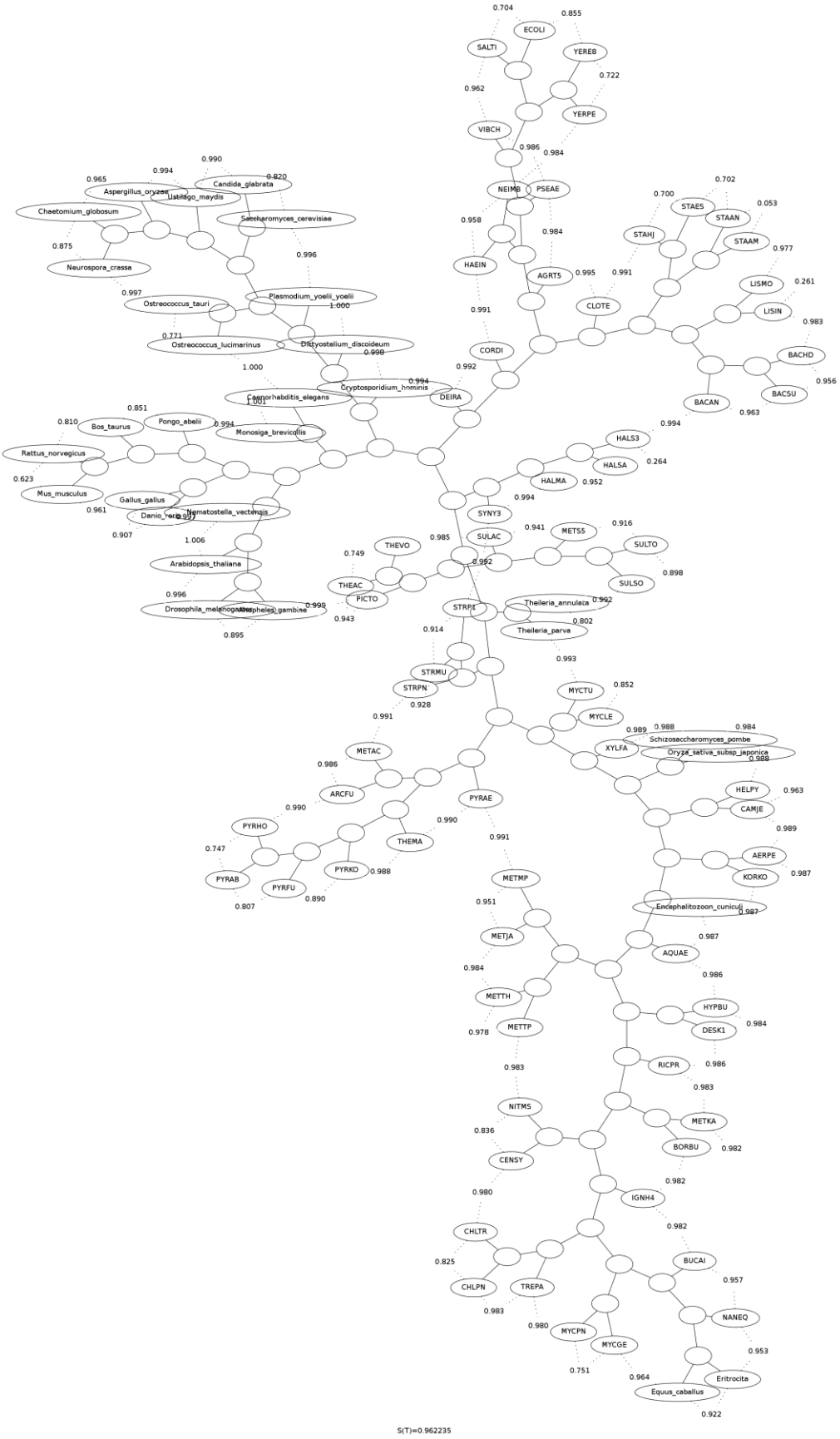

Figure 3: Clustering obtained on the 101 transmembrane receptors by the LZMA compressor

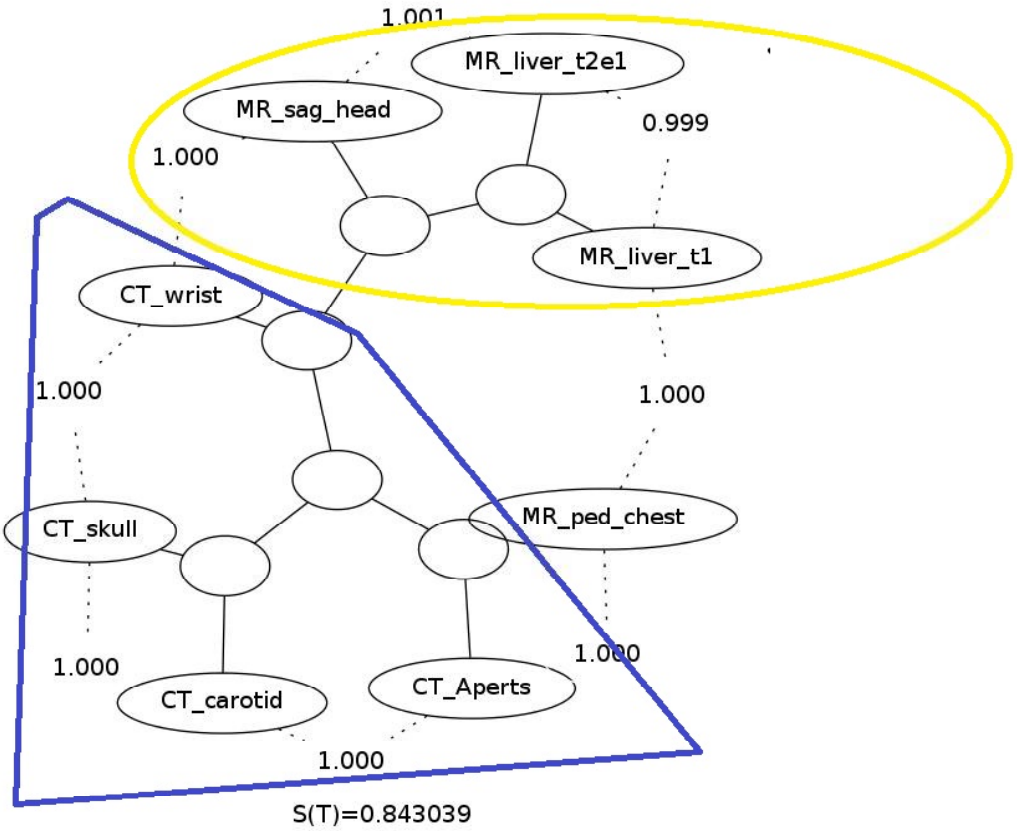

Figure 4: Clustering obtained on medical images

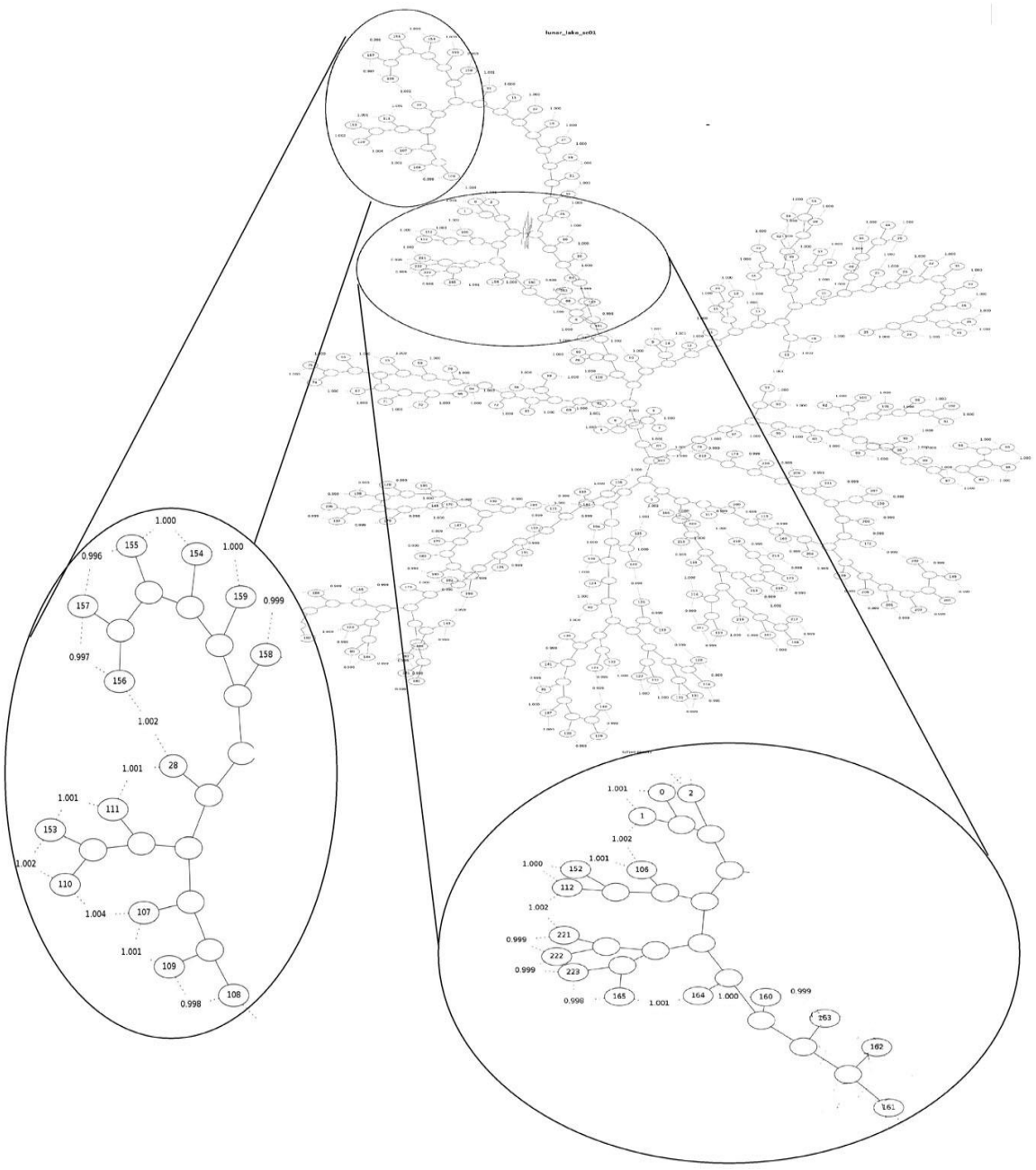

Figure 5: The two cuts on the resulting tree produced by the analysis of the first scene of the Lunar Lake AVIRIS image