Mutual Information Based Distance Measures for Classification and Content Recognition with Applications to Genetics

Zaher Dawy[1,3], Joachim Hagenauer[1], Pavol Hanus[1], and Jakob C. Mueller[2]

[1]Munich University of Technology, Institute for Communications Engineering (LNT), Arcisstr. 21, 80290 Munich, Germany

[2]National Research Center for Environment and Health (GSF), Ingolstaedter Landstr. 1, 85764 Neuherberg, Germany

[3]American University of Beirut, Department of Electrical and Computer Engineering, Beirut, Lebanon

Email: zaher.dawy@aub.edu.lb, hagenauer@tum.de, pavol.hanus@mytum.de, jakob.mueller@gsf.de

***Abstract*— Possibilities of using mutual information for classification and content recognition are exploited. Two different mutual information based distance measures are proposed, one for classification and one for content recognition. The measure proposed for classification is shown to be a metric. The influence of compression based estimation methods on the proposed measures is investigated. Several examples of successful applications in the field of genetics are presented.**

I. Introduction

Mutual information describes the amount of information shared by stochastic processes. It can thus be used to derive distance measures quantifying the similarity of the processes. Such distance measures are needed for classification and content recognition. A source generating messages, e.g. a person writing articles or nature generating genetic sequences for individuals from a particular species, can be regarded as a stochastic process. Mutual information based distance measure can thus be used to compare texts written by different authors or to build phylogenies of animal species. The calculation of mutual information requires knowledge about the entropies of the compared sources. Often the entropy has to be approximated based on one representative message generated by a given source. According to Shannon's source coding theorem an approximate value for the entropy of a source can be derived from the compression ratio achieved by an ideal lossless compressor for that source. The idea of using compression based distance measures for classification and content recognition is not new. Different distance measures have been proposed for this purpose [1] [2].

The main aim of this paper is to propose two universal well-founded distance measures based on mutual information for both classification and content recognition. An analysis of their theoretical performance, as well as their behavior when compression based approximations of the entropy rates are being used, is given. The presented results focus on applications from the field of genetics. However, the proposed measures are universal and applicable to all kinds of sources, e.g. text based sources. The problems of classification and content recognition are formally defined in Section II. A derivation of two mutual information distance measures follows in Section III. In Section IV these measures are approximated using lossless compression algorithms. Some results are presented in Section V, conclusions in Section VI.

II. Classification and Content Recognition

Let S be a discrete source emitting symbols X_i where $i \in \mathbb{N}$. The symbols X_i are modeled as random variables with realizations in the finite alphabet $\mathcal{X}$ of size $|\mathcal{X}| = L$. The source's output is a sequence of random variables $X_1, X_2, \ldots, X_n$ and can be modeled as a stochastic process. The source is characterized by the joint probability mass function

$$p_{X_1,X_2,\ldots,X_n}(X_1, X_2, \ldots, X_n), \qquad n \in \mathbb{N}. \tag{1}$$

The sources entropy rate is calculated as follows:

$$H(S) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n). \tag{2}$$

An upper bound for the entropy rate can be estimated by assuming a stationary memoryless source with a uniformly distributed symbol alphabet. For a DNA sequence using a four symbol alphabet the upper limit would be 2 bit, but since life represents order a lower entropy rate can be expected for the DNA.

Shannon's fundamental theorem on data compression states that every source can be losslessly compressed up to its entropy rate $H(S)$. Thus, the compression ratio achieved by an optimal compression algorithm designed for a given source S when compressing a message s generated by this source is a good approximation of the sources actual entropy rate

$$H(S) \approx \frac{|\mathrm{comp}(s)|}{|s|}. \tag{3}$$

It is impossible to design one optimal compression algorithm for all sources. For practical reasons, *universal* compressors are made for a *class of sources*. They use a compressor *model* that is based on statistical properties common to all sources in the class. The model corresponds to the way the algorithm organizes and fills its memory. The actual content of the compressor's memory represents the *parameters* of the compressor model. The investigated compression algorithms work sequentially and stepwise. In each step they compress a fragment of the message and adjust their parameters based on the part of the message compressed so far. DNACompress is currently the best compression algorithm for DNA class of sources [3]. The fact that DNA sequences contain many approximate repeats is used as compressor model. The parameter

of the model is the buffer containing the sequence searched for the repeats.

Classification is used to build clusters of sources $S_i, i \in \{1 \dots |\mathcal{M}|\}$ from a set $\mathcal{M}$ based on chosen criteria. A distance function $d(S_i, S_j)$ quantifying the similarity of the sources needs to be defined. The distance function has to span a bounded metric space. Formally, a metric space $\mathcal{M}$ is a set of points with an associated distance function $d(S_i, S_j) : \mathcal{M} \times \mathcal{M} \mapsto \mathbb{R}^+$ satisfying the following conditions [4]:

1) $d(S_i, S_j) = 0$ iff $S_i = S_j$ (identity of indiscernibles),
2) $d(S_i, S_j) = d(S_j, S_i)$ (symmetry),
3) $d(S_i, S_j) \leq d(S_i, S_k) + d(S_k, S_j)$ (triangle inequality).

The metric space is bounded if there exists some number r, such that $d(S_i, S_j) \leq r, \forall S_i, S_j \in \mathcal{M}$. In a bounded metric space it is possible to build unambiguous clusters based on the distances $d(S_i, S_j)$.

Content recognition serves a different purpose. Here a set $\mathcal{C}$ of known *content* sources $S_i^{\mathcal{C}}, i \in \{1 \dots |\mathcal{C}|\}$ is provided together with a set $\mathcal{U}$ of *unknown* sources $S_j^{\mathcal{U}}, j \in \{1 \dots |\mathcal{U}|\}$. The goal is to find the best matching content source $S_b^{\mathcal{C}}$ to every unknown source $S_j^{\mathcal{U}}$, which leads to the minimum distance $b = \arg\min_i(d(S_i^{\mathcal{C}}, S_j^{\mathcal{U}}))$. The distance measure for content recognition does not have to be a metric. The quality of content recognition can be quantified as the difference in percent between the minimal and the actual distance

$$q(i,j) = \frac{d(S_i^{\mathcal{C}}, S_j^{\mathcal{U}}) - d(S_b^{\mathcal{C}}, S_j^{\mathcal{U}})}{d(S_b^{\mathcal{C}}, S_j^{\mathcal{U}})}. \quad (4)$$

Distance measures leading to higher values of $q(i,j)$ are more robust for content recognition.

III. Mutual Information Distance Measures

Information theory describes the relatedness of two sources S_i and S_j as the mutual information $I(S_i; S_j)$ shared by these sources

$$\begin{aligned} I(S_i; S_j) &= H(S_i) - H(S_i|S_j) \\ &= H(S_j) - H(S_j|S_i) = I(S_j; S_i). \end{aligned} \quad (5)$$

From this, one can easily show that

$$0 \leq I(S_i; S_j) \leq \min(H(S_i), H(S_j)). \quad (6)$$

Mutual information is an absolute measure of information common to both sources. For independent sources, it converges towards zero. However, mutual information by itself would not be a suitable measure, because it is not a distance and it is not bounded. We can make it a bounded distance by normalizing it and subtracting it from one. The normalization can be done in two different ways.

One way is to normalize by the maximum possible mutual information the two sources can share, see (6), resulting in

$$d_{CR}(S_i, S_j) = 1 - \frac{I(S_i; S_j)}{\min(H(S_i), H(S_j))}. \quad (7)$$

Thus,

$$0 \leq d_{CR}(S_i, S_j) \leq 1. \quad (8)$$

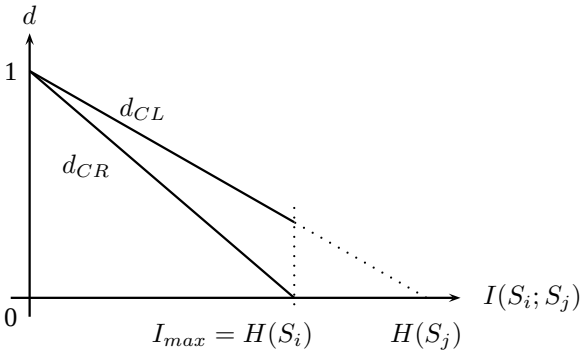

Fig. 1. Plot of d_{CR} and d_{CL} depending on $I(S_i; S_j)$ assuming $H(S_i) < H(S_j)$.

The lower bound is reached for sources that share the maximum possible mutual information given their entropies.

The other possibility is to normalize by the maximum entropy of both sources resulting in

$$d_{CL}(S_i, S_j) = 1 - \frac{I(S_i; S_j)}{\max(H(S_i), H(S_j))}. \quad (9)$$

Thus,

$$1 - \frac{\min(H(S_i), H(S_j))}{\max(H(S_i), H(S_j))} \leq d_{CL}(S_i, S_j) \leq 1. \quad (10)$$

The distance measure in (9) evaluates to zero only for identical sources $S_i = S_j$, because only identical sources share the maximum possible information and have identical entropies at the same time $H(S_i) = H(S_j) = I(S_i; S_j)$. This distance additionally reflects whether the sources are identical or not.

Fig. 1 shows the behavior of d_{CR} and d_{CL} depending on the mutual information $I(S_i; S_j)$, assuming $H(S_i) < H(S_j)$. The distance measure d_{CL} can reach zero only for sources with identical entropies $H(S_i) = H(S_j)$. The distance measure d_{CR} however can reach zero even if the entropies $H(S_i)$ and $H(S_j)$ differ and the sources are thus not identical. This means that d_{CR} unlike d_{CL} does not fulfill the identity of indiscernibles axiom, but only the identity axiom $d(S_i, S_i) = 0$. The symmetry axiom is fulfilled by both measures, as can be seen from the measure definitions (7) and (9). Fig. 1 additionally shows that both measures depend linearly on the mutual information $I(S_i; S_j)$. While d_{CR} uses the whole spectrum of values from 0 to 1 for every entropy combination, d_{CL} only for sources with equal entropies. Therefore d_{CR} can be considered more robust which makes it at least a better candidate for content recognition. The steeper slope of d_{CR} leads to higher values for $q(i,j)$ for d_{CR} than for d_{CL} given the same sources $S_i^{\mathcal{C}}$, $S_b^{\mathcal{C}}$ and $S_j^{\mathcal{U}}$, see (4) and Fig. 1 .

For classification purposes however, the triangle inequality must be fulfilled as well for the distance measure to be a metric (necessary for forming unambiguous clusters). Unfortunately, this is not the case for d_{CR}. Just imagine a situation where $d_{CR}(S_i, S_k) + d_{CR}(S_k, S_j) = 0$, then $d(S_i, S_j)$ must be equal to zero to satisfy the triangle inequality. This can however not be guaranteed, if $H(S_k) = \min(H(S_i), H(S_j), H(S_k))$.

The distance $d_{CL}(S_i, S_j)$ can also be written as:

$$d_{CL}(S_i, S_j) = \frac{\max(H(S_i|S_j), H(S_j|S_i))}{\max(H(S_i), H(S_j))}. \quad (11)$$

This resembles the similarity metric based on Kolmogorov complexity proposed in [2], proven to satisfy the triangle inequality up to an additive constant term. By conducting a similar proof it can be shown that the d_{CL} measure satisfies perfectly the triangle inequality and is thus a metric. A prerequisite resulting from the chain rule for uncertainty is necessary to be able to conduct the proof. By applying the chain rule twice for three sources S_i, S_j and S_k it can be written:

$$H(S_i|S_j) = H(S_k|S_j) + H(S_i|S_j, S_k) - H(S_k|S_i, S_j). \quad (12)$$

Thus following inequality must be true

$$H(S_i|S_j) \leq H(S_k|S_j) + H(S_i|S_k). \quad (13)$$

In summary, it can be said that the distance function d_{CR} should be the better choice for content recognition as it is more robust and d_{CL} should be the better choice for classification as it satisfies the triangle inequality. The presented measures are universal and can be used with any kind of sources.

IV. The Measures and Compression Algorithms

In the following we will analyze the performance of the measures, when real-life compression algorithms are used to estimate the entropy rates of the sources. For this purpose d_{CR} will like d_{CL} in (11) be first reformulated using conditional entropies to give

$$d_{CR}(S_i, S_j) = \frac{\min(H(S_i|S_j), H(S_j|S_i))}{\min(H(S_i), H(S_j))}. \quad (14)$$

We do this because the resulting terms are very compact and for non-ideal approximations of entropy this is an advantage.

The following approximations will be used

$$\begin{aligned} H(S_i|S_j) &= \frac{|\mathrm{comp}(s_j, s_i)| - |\mathrm{comp}(s_j)|}{|s_i|}, \\ H(S_i) &= \frac{|\mathrm{comp}(s_i)|}{|s_i|}, \end{aligned} \quad (15)$$

where s_i and s_j are the messages generated by the sources S_i and S_j, respectively. Because we use one compressor designed for a class of sources, we approximate the conditional entropy as the compression ratio achieved for s_i when the compressors parameters are first trained on s_j. This can be done by first concatenating s_j and s_i in the mentioned order and compressing them together resulting in $|\mathrm{comp}(s_j, s_i)|$. The result of the compression up to the start of s_i is identical to the compression of s_j by itself, $|\mathrm{comp}(s_j)|$, because the algorithm works sequentially. Therefore $(|\mathrm{comp}(s_j, s_i)| - |\mathrm{comp}(s_j)|)/|s_i|$ is the most suitable approximation for the conditional entropy rate $H(S_i|S_j)$. Plugged into (11) for $|\mathrm{comp}(s_i)| > |\mathrm{comp}(s_j)|$, this results in

$$d_{CL} = \frac{|\mathrm{comp}(s_j, s_i)| - |\mathrm{comp}(s_j)|}{|\mathrm{comp}(s_i)|}. \quad (16)$$

The same expression as in (16) applies to d_{CR} except that for $|\mathrm{comp}(s_i)| < |\mathrm{comp}(s_j)|$. What sequence to compress first, is decided based on the *sizes* of the compressed sequences, not the compression *ratios* achieved for the sequences, when the approximations in (15) are used. The distance can be seen as the ratio of the compressed size of s_i ($|\mathrm{comp}_{s_j}(s_i)| = |\mathrm{comp}(s_j, s_i)| - |\mathrm{comp}(s_j)|$), when the compressors model parameters are first tuned using s_j, to the compressed size of s_i ($|\mathrm{comp}(s_i)|$), when the compressor starts using the default parameters. The compressed size $|\mathrm{comp}(s_i)|$ has now become the normalizing term.

The presented approximation works best for compression algorithms that greatly suffer from a wrong initial parameter set and perform very well with a good initial parameter set for the given class of sources. These are not necessarily the algorithms achieving the best compression ratios for the investigated class of sources. An example would be the PPM (Prediction by Partial Matching) compressor [5]. When applied to genetic sequences, it performs quite badly with respect to compression, however it works well for classification and content recognition. Due to the denominator term in (16) compressors complying with the above property and performing well when compressing the class of sources by itself are expected to achieve the best results.

V. Simulation results

The performance of the presented distance measures d_{CR} and d_{CL} for classification and content recognition is tested on genetic data. The simulation results confirm that d_{CL} is a better choice for classification and d_{CR} for content recognition. Performance of different compression algorithms with the proposed compression based entropy rate approximation is evaluated. Following algorithms were tested: Lempel-Ziv (LZ), Context Tree Weighting (CTW), Burrows Wheeler Transform (BWT), Prediction by Partial Matching (PPM) and DNACompress. In general PPM and DNACompress performed best for the classification and content recognition of genetic sequences.

First the compression performance of the investigated algorithms on data used later for classification and content recognition is analyzed. All considered sequences are taken from the NCBI database [6]. The upper part of Table I shows the compression performance using mtDNA (mitochnodrial DNA) sequences of different high level animals. The sequences are each about 16,000 nucleotides long and consist mainly of coding DNA. The mtDNA is particularly suitable for phylogenetic research as it mutates 6 to 17 times faster than nuclear DNA and is inherited only maternally preserving information about ancestry. It can be seen that the compression ratios for mtDNA are relatively high and that BWT, LZ and PPM expand the sequences instead of compressing them. CTW and DNACompress manage to compress the mtDNA.

The lower part of Table I shows compression ratios achieved for three major types of DNA. The first 300,000 nucleotides of each type were cut out from the human chromosome 1 and concatenated for this purpose. *Extra-gene* (eg) DNA are

Sequence	BWT	LZ	CTW	PPM	DNAc
mtDNA-human	2.2106	2.4303	1.9345	2.0566	1.9306
mtDNA-horse	2.1825	2.4058	1.9280	2.0303	1.9116
mtDNA-rat	2.2037	2.4245	1.9225	2.0299	1.9171
c1eg-300kb	2.0749	2.1533	1.8841	1.8718	1.4294
c1in-300kb	2.0614	2.1569	1.8803	1.7637	1.5295
c1ex-300kb	2.1084	2.1566	1.8840	1.8992	1.8300

TABLE I

COMPRESSION RATIOS IN bit/nucleotide.

regions between genes not participating in protein synthesis. Inside genes *exons* (ex) and *introns* (in) are found, whereas only exons are coding regions actually translated into proteins. Exons experience a relatively low mutation rate compared to non-coding regions. Triplets of exon nucleotides code for amino acids. Therefore a stronger local dependency can be expected from an exon source, however it contains only little redundancy and its compressibility is thus limited. Non-coding DNA contains many approximate repeats and palindromes and should therefore be better compressible by a suitable compressor model. The results indicate that CTW, PPM and DNACompress are capable of actually compressing these sequences, whereas DNACompress achieves the best results.

For classification and content recognition the absolute compression performance is secondary . It is more important that the approximation used for conditional entropy in (15) is very sensitive to dissimilarities. For this purpose the compressors parameters must be difficult to unlearn once trained on a dissimilar sequence to the one being compressed and they must lead to good compression ratios when a similar sequence is compressed. Such algorithm must consider distant dependencies equally important to local ones, so that the training sequence does not gradually loose its influence on the overall compression performance. The LZ type compressors are strongly biased in favor of local dependencies. The CTW algorithm gives gradually decreasing importance to distant dependencies. Finally PPM and DNACompress treat distant repeats about equally important to local ones. Table II demonstrates this behavior. For this purpose six different mtDNA sequences were concatenated. One time two close sequences of two chimpanzees s_1 and s_2 were put next to each other after the other four sequences and compressed. The other time the other four sequences were placed between them in the concatenation. Additionally, it can be seen that PPM and DNACompress were able to use the similar sequence to achieve better compression (compare Table I with Table II).

Sequence	BWT	LZ	CTW	PPM	DNAc
4x-s_1-s_2	2.0487	2.0041	1.8849	1.7430	1.5660
s_1-4x-s_2	2.0487	2.2126	1.8837	1.7397	1.5692

TABLE II

COMPRESSION PERFORMANCE: LOCAL VS. DISTANT DEPENDENCIES.

Compression based classification relying on mutual information can be successfully applied e.g. to phylogenetic research. The evolutionary model assumes a common ancestor for all living organisms. A new species evolves when one part of the members of a species stops interbreeding with the other. The exchange of genetic information stops and the two new species continue developing separately. They become different sources with independently mutating statistics. Gradually they share less and less of the originally identical information. Thus, by taking DNA samples of currently existing species and measuring the amount of shared information with the other species, we should be able to reconstruct the phylogeny. Fig. 2 shows an evolutionary tree constructed from distances calculated using the mutual information distance metric d_{CL} and DNACompress as compression algorithm. Please note, that the length of the branches does not correspond to the actual distances. The used sequences are complete mtDNA reference sequences of the organisms obtained from the NCBI database [6] using the following accessions (V00654-cow, X61145-finback whale, X72204-blue whale, X63726-harbor seal, X72004-gray seal, U20753-cat, X79547-horse, X97336-rhinoceros, AJ428577-bear, AJ222767-guinea pig, AY172335-mouse, X14848-rat, AF347015-human, D38113-chimpanzee, D38116-pygmy chimpanzee, D38114-gorilla, D38115-orangutan, X99256-gibbon, Z29573-opossum, Y10524-wallaroo, X83427-platypus, AJ303116-echidna, AY694420-carp). The chosen tree generation method is the quartet method presented in [7], with carp used as outgroup. It can be seen that using d_{CL} we are able to group *marsupials=(metatheria, prototheria)*, *eutheria=((primates,ferungulates),rodents)* and *mammals=(marsupials,eutheria)* complying with the currently accepted ordering. The origin of guinea pig is widely disputed.

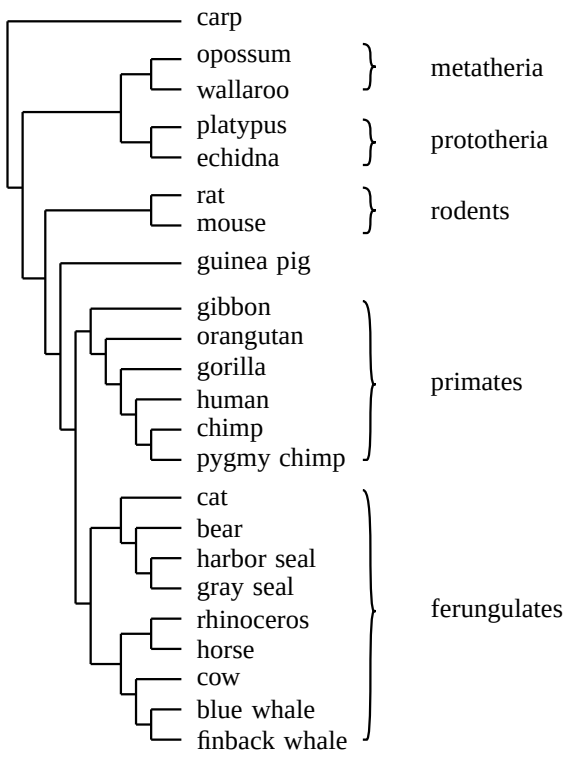

Fig. 2. Evolutionary tree based on mtDNA. Built using the distance metric d_{CL}, DNACompress for compression, and quartet method for tree generation.

Our result reflects the hypothesis that it is a mammal that separated at an early stage from the remaining mammals. The detailed branches of the subtrees are biologically correct as well. Using PPM and d_{CL} results in a slightly different tree, where the group (rhinoceros+horse) is misplaced. The same happens when d_{CR} is used in combination with PPM or DNACompress. This verifies the expected superior behavior of d_{CL} for classification.

To demonstrate the content recognition performance of the measures, we present the results for content recognition of extra-genetic regions (eg), exons (ex) and introns (in). As content sequences the first 50,000 nucleotides of concatenated sequences of each type from the human chromosome 19 were taken. Sequences of different sizes of each type taken from the beginning of chromosome 1 were used as unknown sequences. For each unknown sequence j the distance $d_{CR}(S_i^{\mathcal{C}}, S_j^{\mathcal{U}})$ and $d_{CL}(S_i^{\mathcal{C}}, S_j^{\mathcal{U}})$ to every content sequence i was calculated together with the difference in percent to the best matching content sequence $q(i,j)$, see (4). The best matching sequence is indicated with the string 'best' instead of the value $q(i,j)$. Results obtained using DNACompress and PPM are shown in Table III (notation: c19eg-50kb means an extra-gene sequence from chromosome 19 with size 50 kilo bases). It can be seen that the values of $q(i,j)$ are higher for d_{CR} than for d_{CL}. This confirms our expectation and emphasizes that d_{CR} is better suited for content recognition. Additionally using d_{CR} as distance measure both PPM and DNACompress were able to recognize c1in-13kb correctly as an intron, unlike with d_{CL}, confirming our hypothesis about better robustness of d_{CR}. A closer look reveals that PPM failed to recognize c1ex-13kb (sequence of 13,000 nucleotides from chromosome 1) while DNACompress did not. This can be explained by the superior performance of DNACompress when compressing extra-genetic and intrinsic regions, which allows it to easily distinguish them from exons. Looking at the very low distance values obtained for d(c1eg-300kb,c19eg-50kb) it seems that the two chromosomes share some part of the sequence. In general, the content recognition of exon, intron and extra-gene sequences worked quite well using both DNACompress and PPM. As introns and extra-genetic regions are non-coding regions they are not as sensitive to mutations as exons. Their higher mutation rate lets us assume quite a different statistics from the exons. The obtained results confirm this by indicating that extra-genetic regions are closer to introns, see the lower distances for eg-in combinations.

	DNACompress		
$S_j^{\mathcal{U}} \backslash S_i^{\mathcal{C}}$	c19eg-50kb	c19in-50kb	c19ex-50kb
	$d_{CR}(S_i^{\mathcal{C}}, S_j^{\mathcal{U}}) - q(i,j)$		
c1eg-300kb	0.041-best	0.842-1940%	1.025-2383%
c1eg-13kb	0.651-best	1.013-55.6%	1.009-55.0%
c1in-300kb	0.933-59.5%	0.585-best	1.014-73.3%
c1in-13kb	1.000-1839%	0.052-best	1.068-1970%
c1ex-300kb	1.017-5.6%	1.006-4.5%	0.963-best
c1ex-13kb	0.985-18.7%	0.944-13.7%	0.830-best
	$d_{CL}(S_i^{\mathcal{C}}, S_j^{\mathcal{U}}) - q(i,j)$		
c1eg-300kb	0.786-best	0.971-23.5%	1.006-28.0%
c1eg-13kb	0.912-best	1.004-10.1%	1.003-9.9%
c1in-300kb	0.988-5.8%	0.933-best	1.007-7.8%
c1in-13kb	0.935-best	0.994-6.3%	1.012-8.2%
c1ex-300kb	1.006-1.2%	1.003-0.9%	0.994-best
c1ex-13kb	0.997-3.6%	0.986-2.6%	0.962-best
	PPM		
$S_j^{\mathcal{U}} \backslash S_i^{\mathcal{C}}$	c19eg-50	c19in-50	c19ex-50
	$d_{CR}(S_i^{\mathcal{C}}, S_j^{\mathcal{U}}) - q(i,j)$		
c1eg-300kb	0.067-best	0.803-1093%	1.005-1394%
c1eg-13kb	0.630-best	0.969-53.8%	1.001-58.8%
c1in-300kb	0.929-73.1%	0.536-best	1.013-88.8%
c1in-13kb	0.955-964%	0.090-best	1.021-1038%
c1ex-300kb	1.026-2.5%	1.026-2.6%	1.000-best
c1ex-13kb	0.951-best	0.967-1.7%	1.009-6.0%
	$d_{CL}(S_i^{\mathcal{C}}, S_j^{\mathcal{U}}) - q(i,j)$		
c1eg-300kb	0.832-best	0.971-16.7%	1.000-20.2%
c1eg-13kb	0.904-best	0.996-10.2%	1.001-10.7%
c1in-300kb	0.986-6.6%	0.926-best	1.003-8.4%
c1in-13kb	0.929-best	0.977-5.2%	1.002-7.9%
c1ex-300kb	1.001-0.0%	1.003-0.2%	1.001-best
c1ex-13kb	0.991-best	0.994-0.4%	1.001-1.0%

TABLE III
CONTENT RECOGNITION USING DNACOMPRESS AND PPM WITH d_{CR} AND d_{CL}.

VI. CONCLUSIONS

Two universal distance measures for comparison of sources based on the mutual information they share were presented. One of the measures (7) has been shown to be superior for content recognition applications. On the other hand, the second measure (9) has been shown to be a metric, thus, superior for classification. An analysis of the performance of both measures was provided for entropy rate approximations using non-ideal lossless compression algorithms. Successful applications from genetics were presented confirming the assumed properties for each measure. Different compression algorithms were examined with respect to their classification and content recognition performance when applied to genetic data. Finally, very similar results and trends have been obtained from applications involving other classes of sources, namely the applications of language recognition and authorship attribution.

REFERENCES

[1] S. Grumbach and F. Tahi, "A new challenge for compression algorithms: Genetic sequences," *Journal on Information Processing and Management*, vol. 30, no. 6, pp. 875–886, 1994.

[2] M. Li, X. Chen, X. Li, B. Ma, and P. Vitanyi, "The similarity metric," in *Proc. of the 14th annual ACM-SIAM symposium on Discrete algorithms (SODA)*, (Baltimore, Maryland), pp. 863–872, 2003.

[3] X. Chen, M. Li, B. Ma, and J. Tromp, "DNACompress: fast and effective DNA sequence compression," *Bioinformatics*, vol. 18, no. 12, pp. 1696–1698, 2002.

[4] M. Hagedoorn, *Pattern matching using similarity measures*. PhD thesis, Universiteit Utrecht, Netherlands, Sept. 2000.

[5] J. G. Cleary and I. H. Witten, "Data compression using adaptive coding and partial string matching," *IEEE Transactions on Communications*, vol. COM-32, pp. 396–402, April 1984.

[6] NCBI, "National center for biotechnology information."

[7] R. Cilibrasi and P. M. B. Vitányi, "Clustering by compression," 2003.