NORMALIZED DISTANCE MATRIX METHOD FOR CONSTRUCTION OF PHYLOGENETIC TREES USING NEW COMPRESSOR - DNABIT COMPRESS.

P. RAJARAJESWARI 1, ALLAM APPARAO 2
1DMSSSVH College of Engineering, Machilipatnam.  
2Jawaharlal Nehru Technological University, Kakinada.

ABSTRACT
We define a compression distance, based on a normal compressor to show it is an admissible distance. The first theme concerns the statistical significance of compressed file sizes. Only in recent years have scientists begun to appreciate the fact that compression ratios signify a great deal of important statistical information. In applying the approach, we have used a new DNA sequence compressor “DNABIT compress” C. A compressor C approximates the information distance EC(x,y) based on Kolmogorov complexity, by the compression distance EC(x,y). Compression algorithms can be used to approximate the Kolmogorov complexity. The normalized compression distance, an efficiently computable, and thus practically applicable form of the normalized information distance is used to calculate Distance Matrix. In this paper this new distance matrix is proposed to reconstruct Phylogenetic tree. Phylogenies are the main tool for representing the relationship among biological entities. Phylogenetic reconstruction methods attempt to find the evolutionary history of given set of species. This history is usually described by an edge weighted tree, where edges correspond to different branches of evolution, and the weight of an edge corresponds to the amount of evolutionary change on that particular branch. We constructed a phylogenetic tree with BChE DNA sequences of mammals giving new proposed distance matrix by DNABIT compressor to NJ (Neighbor-Joining algorithm) tree.

Keywords: Normalized Compression Distance, kolmogorov complexity, DNABIT compress, Phylogeny, Bioinformatics, Distance matrix, Phylogenetic tree, neighbor-joining algorithm.

I: INTRODUCTION
DNA sequences seem ideally suited for the compression distance approach. A DNA sequence is a finite string over a four-letter alphabet {A,C,G,T}. Phylogenies are reconstructed using data of all kinds, from molecular data, metabolic data, morphological data to geographical and geological data[1]. Phylogenetic analysis elucidate functional relationship within living cells [2-4]. The problem of inferring the evolutionary history and constructing the phylogenetic tree is a major task in computational biology. There are three major methods for performing a phylogenetic analysis, distance method, maximum parsimony, and maximum likelihood methods.

The alignment methods seem inadequate for post-genomic studies since they do not perform well with data set size and they seem to be confined only to genomic and proteomic sequences. Therefore, alignment-free similarity measures are actively pursued. Ferragina et al.[5] experimentally tested the normalized information distance using 25 compressors to obtain the NCD, and six data sets of relevance to molecular biology. They compared the methodology with methods based on alignments. They assessed the intrinsic ability of the methodology to discriminate and classify biological sequences and structures. The NCD has been put to numerous tests. Keogh et al.[6,7] have tested a closely related metric as a parameter-free and feature-free data mining tool on a large variety of sequence benchmarks. They established clear superiority of the NCD method for
NORMALIZED DISTANCE MATRIX METHOD FOR CONSTRUCTION OF PHYLOGENETIC TREES

clustering heterogeneous data, and for anomaly detection, and competitiveness in clustering domain data and phylogeny construction.

In this proposed method we have used Normalized Compression Distance to calculate Distance Matrix required to construct Phylogeny tree[8]. DNA sequences are compressed with our new proposed “DNABIT compress algorithm” The Normalized distance Matrix focuses on Similarity Metric.

1.1: Similarity Metric
In mathematics, different distances arise in all sorts of contexts, and one usually requires these to be a “metric”. We give a precise formal meaning to the distance notion of “degree of similarity”. Metric: Let Ω be a nonempty set and R + be the set of nonnegative real numbers. A metric on Ω is a function D : Ω * Ω → R + satisfying the metric (in)equalities:
- • D(x, y) = 0 iff x = y,
- • D(x, y) = D(y, x) (symmetry), and
- • D(x, y) ≤ D(x, z)+D(z, y) (triangle inequality).
The value D(x, y) is called the distance between x, y.

A familiar example of a metric is the Euclidean metric, the everyday distance e(a,b) between two objects a,b expressed in, say, meters. Clearly, this distance satisfies the properties e(a,a) = 0, e(a,b) = e(b,a), and e(a,b) ≤ e(a, c)+e(c,b).

1.2: Normalized Compression Distance
The normalized version of the admissible distance EC(x,y), the compressor C based approximation distance is called the Normalized Compression Distance or NCD[9].

NCD(x,y)=C(xy)−min{C(x),C(y)}/max{C(x),C(y)}

Here C(xy) denotes the compressed size of the concatenation of x and y, C(x) denotes the compressed size of x, and C(y) denotes the compressed size of y.

The numerator and denominator of NCD(x,y) is calculated with the above formula subsequently using the DNABIT compressor C, we obtain the distance NCD. The distance NCD is actually a family of distances parameterized with the compressor C. The better C is, the closer NCD approaches the normalized information distance, the better the results are expected to be. Under normal conditions the compressor DNABIT gives values in [0,1] and is a metric[10]. More formally, a compressor C is normal if it satisfies the axioms:
- • C(xx) = C(x) and C(ε ) = 0, (identity)
- • C(xy) ≥ C(x),(monotonicity)
- • C(xy) = C(yx), (symmetry)
- • C(xy)+C(z)≤C(xz)+C(yz), (distributivity)
up to an additive O(log n) term, with n the maximal binary length of a string involved in the equality concerned.

This NCD is the main concept of this work. In practice, the NCD is a non-negative number 0 ≤ D ≤ 1 representing how different the two DNA sequences are. Smaller numbers represent more similar files. Most standard compression algorithms gzip and bzip2 achieved NCD above 1 and compressors PPMZ always had NCD at most 1. [5] OUR NEW PROPOSED DNABIT compressor achieved NCD well between [0,1] which satisfies the similarity metric distance values.

II: MATERIALS AND METHODS

2.1: New Proposed DNABIT COMPRESSOR.

OUR proposed DNABIT algorithm to compress DNA sequences, is used to compress sequences used in distance matrix calculation. Normalized Compression Distance uses compressed sequences length of DNA sequences. In each case, the individual compressed sizes of each sequence are calculated. Then, sum of possible pairs of sequences are combined and compressed to yield pairwise compressed sizes.
This compressor satisfies axioms determining a large family of compressors that include most axioms (if not all) of real-world compressors and ensure the desired properties of the NCD.

DNABIT compressor $C$ is normal as it satisfies, up to an additive $O(\log n)$ term, with $n$ the maximal binary length of an element involved in the inequality concerned, the following:

1. **Idempotency**: $C(xx) = C(x)$, and $C(\Omega) = 0$, where $\Omega$ is the empty string.
2. **Monotonicity**: $C(xy) \geq C(x)$.
3. **Symmetry**: $C(xy) = C(yx)$.
4. **Distributivity**: $C(xy) + C(z) \leq C(xz) + C(yz)$.

**Idempotency**: This compressor $C$ sees exact repetitions and compress the empty string to the empty string.

**Monotonicity**: This compressor satisfies the monotonicity property, at least up to the required precision.

**Symmetry**: Stream-based compressors of the Lempel-Ziv family, like gzip and pkzip, and the predictive PPM family, like PPMZ, are possibly not precisely symmetric. This is related to the stream-based property. Our DNABIT compressor to a great extent is symmetrical, and real experiments show no departure from symmetry.

**Distributivity**: The distributivity property is also satisfied, $C(xy) + C(z) \leq C(xz) + C(yz)$.

### 2.2: DNA COMPRESS JAVA TOOL

The proposed DNABIT compress tool screens are displayed below. The tool has the option of selecting the type as either encrypt or decrypt. This tool is used to compress the Bche DNA sequences of nearly 3 lakhs base pairs within seconds.

The DNA genomes of any length can be given as Input in the input column selecting the encryption option. The compression ratio is displayed in the output column with the time taken for computing.

The encrypted text can be **decrypted** back to its original DNA sequence by using the decryption option. The original DNA text is displayed in the output option.
The proposed new DNABIT compressor performs better than all the current existing DNA compressors like Gencompress, DNA compress[11,12].

Table 1. Comparison of Compression Ratios for different algorithms (bits/base)

<table>
<thead>
<tr>
<th>SEQUENCENAME</th>
<th>SEQ LENGTH</th>
<th>Normal CTW</th>
<th>CTW +LZ</th>
<th>BIOCOMPRESS2</th>
<th>GenCompress</th>
<th>DNA Compress</th>
<th>DNA PACK</th>
<th>DNABIT Compress</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHMPXX</td>
<td>121024</td>
<td>1.838</td>
<td>1.669</td>
<td>1.684</td>
<td>1.673</td>
<td>1.671</td>
<td>1.660</td>
<td>1.5170</td>
</tr>
<tr>
<td>CHNTXX</td>
<td>155844</td>
<td>1.933</td>
<td>1.612</td>
<td>1.617</td>
<td>1.614</td>
<td>1.612</td>
<td>1.610</td>
<td>1.5843</td>
</tr>
<tr>
<td>HEHCMV CG</td>
<td>229354</td>
<td>1.958</td>
<td>1.841</td>
<td>1.848</td>
<td>1.847</td>
<td>1.849</td>
<td>1.834</td>
<td>1.5731</td>
</tr>
<tr>
<td>HUMDY STROP</td>
<td>33770</td>
<td>1.920</td>
<td>1.917</td>
<td>1.926</td>
<td>1.922</td>
<td>1.911</td>
<td>1.908</td>
<td>1.5721</td>
</tr>
<tr>
<td>HUMHB B</td>
<td>73308</td>
<td>1.892</td>
<td>1.808</td>
<td>1.88</td>
<td>1.820</td>
<td>1.789</td>
<td>1.777</td>
<td>1.606</td>
</tr>
<tr>
<td>HUMHD ABCD</td>
<td>58864</td>
<td>1.897</td>
<td>1.821</td>
<td>1.877</td>
<td>1.819</td>
<td>1.795</td>
<td>1.739</td>
<td>1.606</td>
</tr>
</tbody>
</table>

Since the compression ratio is very less compared to other compressors, its yields better distances for phylogeny construction.

III. CONSTRUCTION OF PHYLOGENY (USING NCD).

We used the Bche DNA sequences of 10 mammals, each of about more than 1000 base pairs, to test our methodology. The Bche sequences x and y, are evaluated with the Normalised Compression Distance using DNABIT DNA sequence compressor. The resulting distances are the entries in a 10x10 distance matrix. Constructing a phylogeny tree from the distance matrix, using common tree reconstruction software NJ method[13] gives the tree in Fig.1.

Similarity of sequences in biology is currently primarily handled using multiple sequence alignment methods ClustalW [14]. However, the alignment methods seem inadequate for post-genomic studies since they do not scale well with large data set size. Therefore, alignment-free similarity measures are actively pursued ,Ferragina et al[5]. This proposed new method with DNABIT compressor yields a better distance matrix for reconstruction of phylogenetic tree.

IV. ANALYSIS FOR VARIOUS CASES OF DNABIT COMPRESSOR.

1) MONOTONICITY: \( C(xy) \geq C(x) \)
Example:
Let x be a DNA Bche sequence of Human(homosapiens).
Let y be a DNA Bche sequence of Rat(Rattus Norvegicus).

Length of x = 1809
Compressed bits of x = 2858.58
Length of y = 1794
Compressed bits of y = 2851.74

C(xy) = compressed bits of concatenation of x and y.
C(xy) = 5548.98
C(x) = 2858.58
C(xy) > C(x) , 5548.98 > 2858.58 , which satisfies the Monotonicity property.

2) SYMMETRY : C(xy) = C(yx)

Example:
Let x be a DNA Bche sequence of Human(homosapiens).
Let y be a DNA Bche sequence of Rat(Rattus Norvegicus).

C(xy) = 5548.98
C(yx) = 5548.98
Hence C(xy) = C(yx) , which satisfies the Symmetry property.

3) DISTRIBUTIVITY : C(xy)+C(z) <= C(xz)+C(yz)

Example:
Let x be the DNA Bche sequence of Human = 1809.
Let y be the DNA Bche sequence of Rat = 1794.
Let z be the DNA Bche sequence of Cat = 1809.

C(xy) = 5548
C(xz) = 5546
C(yz) = 5532
C(z) = 2847

5548+2847 = 8395
5546+5532 = 11078
8395<11078, which satisfies the Distributive property of a normal Compressor.

V: ANALYSIS FOR VARIOUS TEST CASES OF COMPRESSED DISTANCE

1) Test Case LEMMA 1: If C is a normal compressor, then EC(x, y)+O(1) is an admissible distance.

Case 1: Assume C(x) = C(y). Then EC(x, y) = C(xy)−C(x). Then, given x and a prefix-program of length EC(x, y) consisting of the suffix of the C-compressed version of xy, and the compressor C in O(1)bits, we can run the compressor C on all xz’s, the candidate strings z in length-increasing lexicographical order. When we find a z so that the suffix of the compressed version of xz matches the given suffix, then z = y by the unique decompression property.

PROOF. Case 1: Assume C(x) = C(y). Then EC(x, y) = C(xy)−C(x).

Let x be the Bche DNA sequence of Rat = 1794
Let y be the Bche DNA sequence of Human = 1809

C(x) = 2851.74; C(y) = 2858.58;
Hence C(xy) = C(x) = 5549.34 − 2851.74 = 2697.6
Then
If z = Length of Bche DNA sequence of Cat = 1809
Compressed bits of xz = C(xz) = 5532.
Therefore z = y , (1809 = 1809). Which satisfies the lemma.

2) Test case LEMMA 2: If C is a normal compressor, then EC(x, y) satisfies the metric (in)equalities up to logarithmic additive precision.

PROOF. Only the triangular inequality is non-obvious. By Distributive property: C(xy)+C(z) <= C(xz)+C(yz) up to logarithmic additive precision.
For the three possible possibilities we verify the correctness of the triangular inequality in turn for each of them.

1. Assume \( C(x) \leq C(y) \leq C(z) \): Then \( C(xy) - C(x) \leq C(xz) - C(x) + C(yz) - C(y) \)

2. Assume \( C(y) \leq C(x) \leq C(z) \):
\[
\text{Then } C(xy) - C(y) \leq C(xz) - C(y) + C(yz) - C(x).
\]

3. Assume \( C(z) \leq C(y) \leq C(x) \): Then \( C(xy) - C(y) \leq C(xz) - C(z) + C(yz) - C(z) \).

Proof:

Let \( x \) be the Bche DNA sequence of Pig \( = 1050; \)
\( C(x) = 1627.60 \)
Let \( y \) be the Bche DNA sequence of Cat \( = 1809; \)
\( C(y) = 2847.72 \)
Let \( z \) be the Bche DNA sequence of Chicken \( = 1812; \)
\( C(z) = 2906.81 \)
\( C(xy) - C(x) = 4471.76 - 1627.6 = 2844.16 \)
\( C(xz) - C(x) + C(yz) - C(y) = (2425 - 1627) + (5750 - 2847) = 2618 + 2903 = 5521 \)
\( 2844.16 < 5521, \) hence the property is satisfied.

2) Assume \( C(y) \leq C(x) \leq C(z) \):
\[
\text{Then } C(xy) - C(y) \leq C(xz) - C(y) + C(yz) - C(x).
\]

Proof:

Let \( x \) be the Bche DNA sequence of Mouse \( = 1812; \)
\( C(x) = 2885.06 \)
Let \( y \) be the Bche DNA sequence of Rat \( = 1794; \)
\( C(y) = 2851.74 \)
Let \( z \) be the Bche DNA sequence of Chicken \( = 1812; \) C(z) = 2906.81
\( C(z) \leq C(x) \leq C(z) \)
\( C(y) \leq C(x) \leq C(z) \)
\( C(x) \leq C(xz) \leq C(yz) \)
\( C(yz) \leq C(y) \leq C(x) \)
\( C(z) \leq C(z) \leq C(z) \)
\( C(xy) - C(y) = 5735-2851 = 2884 \)
\( C(xz) - C(y) + C(yz) - C(x) = (5788-2851)+(5607-2885) = 2937 + 2722 = 5659 \)
\( 2884 < 5659, \) hence the property is satisfied with the proposed new Distance method.

3) Assume \( C(z) \leq C(y) \leq C(x) \):
\[
\text{Then } C(xy) - C(y) \leq C(xz) - C(z) + C(yz) - C(z).
\]

Proof:

Let \( x \) be the Bche DNA sequence of Chicken \( = 1812; \)
\( C(x) = 2906.81 \)
Let \( y \) be the Bche DNA sequence of Cat \( = 1809; \)
\( C(y) = 2847.72 \)
Let \( z \) be the Bche DNA sequence of Pig \( = 1050; \)
\( C(z) = 1627.60 \)
\( C(z) \leq C(y) \leq C(x) \)
\( C(y) \leq C(x) \leq C(z) \)
\( C(z) \leq C(y) \leq C(x) \)
\( C(1627.60) \leq 2847.72 \leq 2906.81 \)
\( C(1627.60) < 2847.72 < 2906.81 \)
\( 2884 < 5659, \) hence the property is satisfied with the proposed new Distance method.

6: Results

The normalized distance of NCD with DNABIT Compressor, has values in \([0,1]\) and satisfies the metric equalities up to additive \(O((\log n)/n)\) terms, with \(n\) the length of a string involved in the equality concerned. Informal experiments [10] have shown that these axioms are in various degrees satisfied by good real world compressors like bzip2, and PPMZ. The maximum usable length of the arguments \(x\) and \(y\) (32KB for gzip, 450KB for bzip2, unlimited for PPMZ) Cebrian et al [15] systematically investigated how far the performance of real-world compressors gzip, bzip2, and PPMZ are. The Normalized Compression Distance NCD is intended to be universally applicable.

We tested our method on BChE DNA sequences of 10 mammals. The trees are generated using the Neighbor Joining (NJ) method [13]. All the experiments in this paper were performed on a PC with
Pentium IV CPU (ZGHZ), 512KB Cache, and 256MB RAM. We chose our group of sequences from [http://www.ncbi.nlm.nih.gov/] and European Molecular Biology Laboratory (EMBL) [http://www.ebi.ac.uk/].

Mammals: AAH18141.1 Homo sapiens (human), EDM00889.1 Rattus norvegicus (Norway rat), AAC06261.1 Felis catus (domestic cat), AAH99977.1 Mus musculus (house mouse), AAF61480.1 Equus caballus (horse), AA123601.1 Bos taurus (cattle), AAC06262.1 Panthera tigris tigris (Bengal tiger), CAC37792.1 Gallus gallus (chicken), AAG41127.1 Sus scrofa (pig), CAA36308.1 Oryctolagus cuniculus (rabbit). We applied the new distance measure to the above data sets. Fig. 1 shows the tree generated by proposed method. The data set is applied to normalized distance matrix tool calculation and the distance matrix [table 3] is analyzed by the NJ program for phylogeny tree construction.

### TABLE :2 Compressed Bits and Compression Ratio of Bche DNA sequences.

<table>
<thead>
<tr>
<th>S.N o</th>
<th>Name of Species</th>
<th>Sequenc e Length (Bases)</th>
<th>Compressed Bits (Bits)</th>
<th>Compression Ratio(bpb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Human</td>
<td>1809</td>
<td>2858.58</td>
<td>1.5802</td>
</tr>
<tr>
<td>2.</td>
<td>Rat</td>
<td>1794</td>
<td>2851.74</td>
<td>1.5896</td>
</tr>
<tr>
<td>3.</td>
<td>Cat</td>
<td>1809</td>
<td>2847.72</td>
<td>1.5742</td>
</tr>
<tr>
<td>4.</td>
<td>Mouse</td>
<td>1812</td>
<td>2885.06</td>
<td>1.5922</td>
</tr>
<tr>
<td>5.</td>
<td>Horse</td>
<td>1809</td>
<td>2864.73</td>
<td>1.5836</td>
</tr>
<tr>
<td>6.</td>
<td>Cattle</td>
<td>1809</td>
<td>2857.85</td>
<td>1.5798</td>
</tr>
<tr>
<td>7.</td>
<td>Tiger</td>
<td>1809</td>
<td>2848.63</td>
<td>1.5747</td>
</tr>
<tr>
<td>8.</td>
<td>Chicken</td>
<td>1812</td>
<td>2906.81</td>
<td>1.6042</td>
</tr>
<tr>
<td>9.</td>
<td>Pig</td>
<td>1050</td>
<td>1627.60</td>
<td>1.5501</td>
</tr>
<tr>
<td>10.</td>
<td>Rabbit</td>
<td>1746</td>
<td>2753.26</td>
<td>1.5769</td>
</tr>
</tbody>
</table>

### TABLE :3 Distance matrix calculated from NCD.

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Rat</th>
<th>Cat</th>
<th>Mouse</th>
<th>Horse</th>
<th>Cattle</th>
<th>Tiger</th>
<th>Chicken</th>
<th>Pig</th>
<th>Rabbit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.000</td>
<td>0.943</td>
<td>0.944</td>
<td>0.933</td>
<td>0.946</td>
<td>0.948</td>
<td>0.943</td>
<td>0.950</td>
<td>0.949</td>
<td>0.946</td>
</tr>
<tr>
<td>Rat</td>
<td>0.943</td>
<td>0.000</td>
<td>0.941</td>
<td>0.942</td>
<td>0.944</td>
<td>0.946</td>
<td>0.970</td>
<td>0.947</td>
<td>0.947</td>
<td>0.943</td>
</tr>
<tr>
<td>Cat</td>
<td>0.944</td>
<td>0.941</td>
<td>0.000</td>
<td>0.998</td>
<td>0.999</td>
<td>0.999</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>Mouse</td>
<td>0.933</td>
<td>0.942</td>
<td>0.998</td>
<td>0.000</td>
<td>1.000</td>
<td>0.999</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>Horse</td>
<td>0.946</td>
<td>0.944</td>
<td>0.999</td>
<td>1.000</td>
<td>0.000</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>Cattle</td>
<td>0.948</td>
<td>0.946</td>
<td>0.999</td>
<td>0.999</td>
<td>0.000</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>Tiger</td>
<td>0.943</td>
<td>0.970</td>
<td>0.998</td>
<td>0.998</td>
<td>0.999</td>
<td>0.999</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>Chicken</td>
<td>0.950</td>
<td>0.947</td>
<td>0.998</td>
<td>0.998</td>
<td>0.999</td>
<td>0.999</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>Pig</td>
<td>0.949</td>
<td>0.947</td>
<td>0.998</td>
<td>0.998</td>
<td>0.999</td>
<td>0.999</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>Rabbit</td>
<td>0.946</td>
<td>0.943</td>
<td>0.998</td>
<td>0.998</td>
<td>0.999</td>
<td>0.999</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.000</td>
</tr>
</tbody>
</table>
We proposed a new sequence distance measure using Normalised Compression Distance measure and used it to generate distance matrix for constructing phylogenetic tree. Unlike most existing phylogeny construction methods, the proposed method does not require multiple alignments and is fully automatic. We tested our method on a dataset and applied it to analyse the evolutionary relationship among BChE sequences. It is to be noted that we use no approximations and assumptions in calculating the distances between sequences. The key concept used here in DNABIT compress is that if two files are very similar in their contents, then they will compress much better when combined together prior to compression, as compared to the sum of the size of each separately compressed file. Although the principle is simple, it has surprising breadth of applicability. By using simplest DNA compressor it is possible to construct evolutionary trees for animals fully automatically using files containing their mitochondrial [DNA] gene sequence. The tree is constructed so that species with “similar” DNA are “close by” in the tree. In this way we may lend support to certain evolutionary theories.

Our DNABIT compress algorithm compression ratios are better when compared to other existing DNA compressors hence these results can often be more accurate than simpler compressors in a variety of particular circumstances or domains.

Throughout the proposed method there is a focus on coding theory and data compression, both as a theoretical construct as well as practical approximations. The quantitative analysis supports the conclusion that the normalized compression method is worth using because of its flexibility, scalability, and competitiveness with existing techniques. In particular, the methodology applies to all biological data in textual format.

The broad range of successful applications of NCD can be demonstrated in various other applications like bioinformatics, linguistics and plagiarism detection.

ACKNOWLEDGEMENT
Authors are thankful for the support rendered by T.Babu AND Kiran during its development phase.

REFERENCES
3. X. Gu. Maximum-likelihood approach for gene family evolution under functional
NORMALIZED DISTANCE MATRIX METHOD FOR CONSTRUCTION OF PHYLOGENETIC TREES

8. Statistical Inference Through Data Compression BY Rudi Cilibrasi
9. Normalized Information Distance by Paul M. B. Vitányi, Frank J. Balbach, Rudi L. Cilibrasi, and Ming Li.

P. RAJARAJESWARI et al.

Dr. Allam Appa Rao has received PhD in Computer Engineering from Andhra University, Visakhapatnam, Andhra Pradesh, India. He has worked as the Professor in Bioinformatics & Computational Biology, Department of Computer Science and Systems Engineering & Principal, Andhra University College of Engineering (AUTONOMOUS). Currently he is Vice Chancellor to Jawaharlal Nehru Technological University, Kakinada. His research interest includes Bioinformatics, Software Engineering and Network Security. He is a member of professional societies like IEEE, ACM and a life member of CSI and ISTE. www.allamapprao.net.

Associate Professor in DMSSVH college of Engineering, Machilipatnam since 2000 to till date. She is pursuing her Ph.D from Acharya Nagarjuna University in Computer Science under the guidance of Dr. Allam Appa Rao. Her research interests includes Bioinformatics, compression techniques, design and analysis of Algorithms, development of software tools.

P. RAJARAJESWARI et al.