The Similarity Metric

Ming Li, Xin Chen, Xin Li, Bin Ma, Paul Vitányi*

Abstract

A new class of similarity measures appropriate for measuring relations between sequences is studied. The "normalized information distance", based on the noncomputable notion of Kolmogorov complexity is shown to be universal in this class (discovers all effective similarities). We demonstrate that it is a metric and takes values in $[0, 1]$. This is a theory foundation for a new general practical tool. We give two distinctive applications in widely divergent areas (the experiments by necessity use just computable approximations to the target notions). First, we computationally compare whole mitochondrial genomes and infer their evolutionary history. This results in a first completely automatic computed whole mitochondrial phylogeny tree. Secondly, we give fully automatically computed language tree of 52 different language based on translated versions of the "Universal Declaration of Human Rights".

1 Introduction

How do we measure similarity—for example to determine an evolutionary distance—between two sequences, such as internet documents, different language text corpora in the same language, among different languages based on example text corpora, computer programs, or chain letters? How do we detect plagiarism of student source code in assignments? The fast advance of worldwide genome sequencing projects has raised the following fundamental question to prominence in contemporary biological science: how do we compare two genomes? This problem figured on two recent lists of major open problems in bioinformatics [20, 39].

Our aim here is not to define a similarity measure for each application field; we develop a general mathematical theory of similarity. To obtain evidence that the theory significantly addresses the question (and is not only theoretically satisfactory), we test it on real-world applications in a wide range of fields. Thus, we first present a new theoretical approach to a wide class of similarity measures; show that among them the "normalized information distance" is a metric, and prove that it is universal in the sense that this single measure uncovers all similarities simultaneously that the other measures in the class uncover a single similarity apiece. It is well-known that when a pure mathematical theory is applied to the real world, for example in hydrodynamics or in physics in general, we can in applications only approximate the theoretical ideal. But the theory gives a framework for the applied science. With this in mind, we demonstrate that the new similarity measure works exemplary on concrete examples in very different application fields—the first completely automatic construction of the phylogeny tree based on whole mitochondrial genomes, and a completely automatic construction of a language tree for over 50 Euro-Asian languages. Other applications we have performed, not reported here, are detecting plagiarism in student programming assignments [33], and phylogeny of chain letters in [5]. (We plan a further test on artificially generated data, where we know the right answer beforehand.)

Related Work: Preliminary applications of the current approach were tentatively reported to the biological community [21]—using an initial and partially improper distance. Subsequent work in the linguistics community, [2, 3], used compression-based methods to infer a language tree from different-language text corpora, as well as attributed authorship on basis of text corpora. Their methods approximate a certain

*Ming Li is with the Computer Science Department of the University of California, Santa Barbara, and with BioInformatics Solutions Inc., Waterloo, Canada. He is partially supported by NSF-ITR grant 0085801 and NSERC, and is on leave from University of Waterloo. Email: mli@cs.ucsb.edu. Xin Chen is with the Computer Science Department of the University of California, Santa Barbara. Email: chxin@cs.ucsb.edu. Xin Li and Bin Ma are with Computer Science Department, University of Western Ontario, London, ON, Canada N6A 5B7, and partially supported by NSERC grant RGP0238748. Email: { xinli,bma@csd.uwo.ca}. P. Vitányi was partially supported by the EU fifth framework project QAIP, IST–1999–11234, the NoE QUIPROCONE IST–1999–29064, the ESF QiT Programmme, and the EU Fourth Framework BRA NeuroCOLT II Working Group EP 27150. He is with the national Center of Mathematics and Computer Science (CWI) and the University of Amsterdam. Address: CWI, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands. Email `Paul.Vitanyi@cwi.nl`. `http://www.cwi.nl/~paulv`

type of empirical relative entropy—entropy a la Shannon is always zero since they deal with individual objects only—even though the authors refer to the information distance in [24], the actual method is essentially ad-hoc without supporting theory, and based on "distances" violating metric requirements like symmetry and triangle inequality. Together with our coauthors, we have studied various forms of information distance in [4] and [21] in the past. The information distance studied in [23, 24, 4], and subsequently investigated in [22, 16, 27, 29, 36], is "universal" and has other nice properties. This distance essentially says that the distance between two objects is the length of the shortest program (or amount of energy) that is needed to transform the two objects into each other. But this distance is not proper to measure evolutionary sequence distance. For example, *H. influenza* and *E. coli* are two closely related sister species. The former has about 1,856,000 base pairs and the latter has about 4,772,000 base pairs. However, using the information distance of [4], one would easily classify *H. influenza* with a short (of comparable length) but irrelevant species simply because of length, instead of with *E. coli.* The problem here is that the information distance of [4] does not deal with the fact that species may lose genes (by deletion) or gain genes (by duplication or insertion from external sources), relatively easily. Instead, deletion and insertion cost energy (proportional to the Kolmogorov complexity of deleted or inserted sequence) in the information distance of [4]. The paper [35] defined transformation distance between two species, and [15] defined compression distance. Both of these measures are essentially $K(x|y)$. Other than being asymmetric, they also suffer similar problems as the information distance of [4] as show in the above example.

This Work: Here we develop a new general tool for phylogeny analysis and other applications: a general mathematical theory of similarity based on a special type of normalized metrics, and construct a universal normalized metric (the normalized information distance). This normalized information distance is the first *universal* effective similarity measure, and is an objective recursively invariant notion by the Church-Turing thesis. To evidence the significance of this theory, we apply it computationally in widely diverse real-world areas. (In the application, of course, we cannot compute the theoretical normalized information distance precisely, we have to do with approximations we can achieve.) In order to demonstrate universal applicability in an experimental setting, we use the wide apart areas of (i) bio-molecular evolution studies, and (ii) natural language evolution. In area (i): In recent years, as the complete genomes of various species become available, it has become possible to do whole genome phylogeny (this overcomes the problem that different genes may give different trees). However, traditional phylogenetic methods on individual genes depended on multiple alignment of the related proteins and on the model of evolution of individual amino acids. Neither of these is applicable to the genome level. In the absence of such models, a method which can compute the shared information between two sequences is useful because biological sequences encode information, and the occurrence of evolutionary events (such as insertions, deletions, point mutations, rearrangements, and inversions) separating two sequences sharing a common ancestor will result in the loss of their shared information. Our theoretical approach is used experimentally to create a fully automated and reasonably accurate software tool based on such a distance to compare two genomes. We demonstrate that a whole mitochondrial genome phylogeny of the Eutherians which confirms [8], can be reconstructed automatically from *unaligned* complete mitochondrial genomes by use of our software implementing (an approximation of) our theory. These experimental confirmations of the effacity of our comprehensive approach contrasts with recent more specialized approaches such as [38] that have (and perhaps can) only be tested on small numbers of genes. They have not been experimentally tried on whole mitochondrial genomes that are, apparently, already numerically out of computational range. In area (ii) we fully automatically construct the language tree of 52 primarily Indo-European languages from translations of the "Universal Declaration of Human Rights"—leading to a grouping of language families largely consistent with current linguistic viewpoints. The main technical concepts in this work, distance metric, Kolmogorov complexity [24], information distance as in [4], are summarized in the—optional—Appendix.

2 Normalized Distance

In defining a class of acceptable metrics we want to exclude unrealistic distance metrics like $f(x,y) = \frac{1}{2}$ for *every* pair x, y, by restricting the number of objects within a given distance of an object. As in [4] we only consider distances $D(x,y)$ satisfying the density condition

$$\sum_{y:y\neq x} 2^{-D(x,y)} \leq 1. \tag{1}$$

However, to express a measure of dissimilarity between two objects, not every such distance metric will do. As remarked above, large objects that differ by a tiny part are intuitively closer than tiny objects that differ by the same amount. Therefore, we normalize the distance metric: If $D(x,y)$ is a distance satisfying the density condition above, then a proper *normalized distance* version is

$$d_D(x,y) = \frac{D(x,y)}{\max\{K(x),K(y)\}}.$$

Since we know from [4] that $D(x,y) \leq \max\{K(x \mid y), K(y \mid x)\}$ up to a logarithmic additive term it follows that $d_D(x,y)$ takes values in $[0,1]$ up to a vanishing $O(\log \max\{K(x \mid y), K(y \mid x)\}/\max\{K(x),K(y)\})$ additive term. If $D(x,y)$ satisfies the requirement (1), then we have

$$\sum_{y:y \neq x} 2^{-d_D(x,y)\max\{K(x),K(y)\}} \leq 1. \tag{2}$$

This leads to the following density property:

Lemma 2.1 *For every x, we have*

$$\{y : d_D(x,y) \leq d, \max\{K(x),K(y)\} \leq k\} \leq 2^{dk}. \tag{3}$$

Proof. Suppose the contrary. Then, starting from (2) we obtain the contradiction:

$$\begin{aligned} 1 &\geq \sum_{y:y \neq x} 2^{-d_D(x,y)\max\{K(x),K(y)\}} \\ &> \sum_{y:y \neq x \& d_D(x,y) \leq d \& \max\{K(x),K(y)\} \leq k} 2^{-d_D(x,y)\max\{K(x),K(y)\}} > 2^{dk}2^{-dk} = 1. \end{aligned}$$

•

This motivates the following:

Definition 2.2 *A* normalized distance *or* similarity distance *is a metric $m(x,y)$ that takes values in* $[0,1]$ *(up to an additive term that vanishes with growing* $\max\{x,y\}$*) for all x,y, and satisfies the density constraint (3).*

Clearly, every normalized distance measures a similarity type, and every similarity among a very wide class of types is measured by some normalized distance.

3 Normalized Information Distance

Clearly, information distance itself is not a proper evolutionary distance measure. Consider three species: *E. coli*, *H. influenza*, and some arbitrary bacteria X of similar length as *H. influenza*, but not related. Information distance d would have $d(X, H.influenza) < d(E.coli, H.influenza)$, simply because of the length factor. It would put two long and complex sequences that differ only by a tiny fraction of the total information as dissimilar as two short sequences that differ by the same absolute amount and are completely random with respect to one another.

Since the information distance is the number of bits in a two-way program to transform one sequence into the other both ways, it seems proper to consider the number of bits of the transformation program per bit of the more complex quantity involved. From the results in [4], summarized in the Appendix, it follows that that the new quantity gives the number of bits of the shortest two-way conversion program per bit of information in the target string containing the most information. In [21] we considered as first attempt (x^* denotes a shortest program for x; the conditional information contained in x^* is equivalent to that in $(x, K(x))$. See Appendix):

Definition 3.1 *Given two sequences x and y, define the function $d_s(x,y)$ by*

$$d_s(x,y) = \frac{K(x \mid y^*) + K(y \mid x^*)}{K(x,y)}. \tag{4}$$

Writing it differently, using (6) below,

$$d_s(x,y) = 1 - \frac{K(x) - K(x \mid y^*)}{K(x,y)},$$

where $K(x) - K(x \mid y^*)$ is mutual information. This distance satisfies the triangle inequality, up to a small error term, and universality (below), but only within a factor 2. Mathematically more precise and satisfying is the distance:

Definition 3.2 *Given two sequences x and y, define the function $d(x,y)$ by*

$$d(x,y) = \frac{\max\{K(x \mid y^*), K(y \mid x^*)\}}{\max\{K(x), K(y)\}}. \tag{5}$$

Remark 3.3 Several natural alternatives for the denominator turn out to be wrong:

(a) One approach to normalization in the d definition would be to divide by the length. Then, firstly we do not know which of the two length involved to divide by, possibly the sum or maximum, but furthermore the triangle inequality and the universality (domination) properties are not satisfied.

(b) In the d definition divide by $K(x,y)$. Then one has $d(x,y) = \frac{1}{2}$ whenever x and y are random (have maximal Kolmogorov complexity) relative to one another. This is improper.

(c) In the d_s definition dividing by length does not satisfy the triangle inequality. ◊

It is a priori not clear whether $d(x,y)$ is a distance metric. Below we show that the new quantity is, in fact, a normalized distance metric. It has a natural meaning as follows: The *information in x about y* is defined as $I(x:y) = K(x) - K(x \mid y^*)$. A deep result of Gács [12] shows that, up to additive constant terms $I(x:y) = I(y:x)$, that is

$$K(x) + K(y \mid x^*) = K(y) + K(x \mid y^*). \tag{6}$$

This allows us to give a very natural interpretation to $d(x,y)$: If $K(y) \geq K(x)$ then we can rewrite

$$d(x,y) = \frac{K(y) - I(x:y)}{K(y)}.$$

That is, the ratio $d(x,y)$ between x and y is the number of bits of information that is not shared between the two strings per bit of information that could be maximally shared between the two strings.

It is clear that $d(x,y)$ is symmetrical and satisfies the identity axiom:

$$d(x,x) = 0 + O(1/\max\{K(x), K(y)\}).$$

To show that it is a distance metric it remains to prove the triangle inequality.

Lemma 3.4 *$d(x,y)$ satisfies the triangle inequality, that is, $d(x,y) \leq d(x,z) + d(z,y)$ up to an additive error term of $O(1/\max\{K(x), K(y), K(z)\})$.*

Proof. **Case 1:** Suppose $K(z) \leq \max\{K(x), K(y)\}$. In [13], the following "directed triangle inequality" was proved: For all x, y, z, up to an additive constant term,

$$K(x \mid y^*) \leq K(x, z \mid y^*) \leq K(x \mid z^*) + K(z \mid y^*). \tag{7}$$

Then, dividing both sides of the triangle inequality by $\max\{K(x), K(y), K(z)\}$,

$$\begin{aligned} &\frac{\max\{K(x \mid y^*), K(y \mid x^*)\}}{\max\{K(x), K(y), K(z)\}} \\ &\leq \frac{\max\{K(x \mid z^*), K(z \mid x^*)\} + \max\{K(z \mid y^*), K(y \mid z^*)\}}{\max\{K(x), K(y), K(z)\}}, \end{aligned}$$

up to an additive term $O(1/\max\{K(x), K(y), K(z)\})$. Because of the assumption of this case, we can eliminate $K(z)$ in the denominator on the left-hand side without changing the left-hand side. Eliminating

$K(y)$ in the denominator of the first term of the right-hand side, and $K(x)$ in the denominator of the second term on the right-hand side can only increase the right-hand side (again, because of the assumption).

Case 2: Suppose $K(z) = \max\{K(x), K(y), K(z)\}$. Further assume that $K(x) \geq K(y)$ (the remaining case is symmetrical). Then, using the symmetry of information to determine the maxima, we also find $K(z \mid x^*) \geq K(x \mid z^*)$ and $K(z \mid y^*) \geq K(y \mid z^*)$. Then the maxima in the terms of the equation $d(x,y) \leq d(x,z) + d(y,z)$ are determined, and our proof obligation reduces to:

$$\frac{K(x \mid y^*)}{K(x)} \leq \frac{K(z \mid x^*)}{K(z)} + \frac{K(z \mid y^*)}{K(z)}, \tag{8}$$

up to an additive term $O(1/K(z))$. To prove (8) we proceed as follows:

Applying the triangle inequality (7) and dividing both sides by $K(x)$, we have

$$\frac{K(x \mid y^*)}{K(x)} \leq \frac{K(x \mid z^*) + K(z \mid y^*) + O(1)}{K(x)}.$$

The left-hand side is ≤ 1.

Case 2.1: The right-hand side is ≤ 1. Setting $K(z) = K(x) + \Delta$, and first adding Δ to both the nominator and the denominator of the first term in the right-hand side, and subsequently using (6) to obtain $K(z \mid x^*) + \Delta = K(x \mid z^*) + O(1)$,

$$\begin{aligned}\frac{K(x \mid y^*)}{K(x)} &\leq \frac{K(z \mid x^*) + K(z \mid y^*) + \Delta + O(1)}{K(x) + \Delta} \\ &= \frac{K(x \mid z^*) + K(z \mid y^*) + O(1)}{K(z)},\end{aligned}$$

which was what we had to prove.

Case 2.2: The right-hand side is ≥ 1. We proceed like in Case 1, and add Δ to both nominator and denominator. Although now the right-hand side decreases, it must still be ≥ 1. This proves Case 2.2. •

Clearly, $d(x,y)$ takes values in the range $[0, 1+O(1/\max\{K(x), K(y)\})]$. To show that it is a normalized distance, it is left to prove the normalization condition:

Lemma 3.5 *The function $d(x,y)$ satisfies the normalization condition (3).*

Proof. Assume that $K(y) \geq K(x)$ (the other case is symmetrical). Then, by (6) we also have $K(y \mid x^*) \geq K(x \mid y^*)$ up to an additive constant term. and rewriting $d(x,y) = K(y \mid x^*)/K(y) \leq d$ and $K(y) \leq k$, we obtain up to an additive constant term

$$K(y \mid x^*) \leq dK(y) \leq dk$$

That is, there are at most 2^{dk} binary programs to obtain a y from x^* satisfying the conditions, and hence at most that many such y. •

Since we have shown that $d(x,y)$ is a distance metric, takes values in $[0,1]$, (up to vanishing additive error terms) and satisfies the normalization condition, it follows:

Theorem 3.6 *The function $d(x,y)$ is a normalized information distance.*

4 Universality

We now show that $d(x,y)$ is universal in the sense that it incorporates every remotely computable type of dissimilarity: if two objects are dissimilar in normalized information in some computable sense, then they are less dissimilar in the $d(x,y)$ sense. We prove this by demonstrating that $d(x,y)$ minorizes every other normalized distance in a wide class—so wide that it will capture everything that can be remotely of interest. It is easy to show that there does not exist a normalized distance that minorizes every other normalized distance, so the result below is in a sense the best we can hope for.

Definition 4.1 *A real-valued function $f(x,y)$ is* upper semicomputable *if there exists a rational-valued function recursive $g(x,y,t)$ such that (i) $g(x,y,t+1) \leq g(x,y,t)$, and (ii) $\lim_{t\to\infty} g(x,y,t) = f(x,y)$. It is* lower semicomputable *if $-f(x,y)$ is upper semicomputable, and it is* computable *if it is both upper- and lower semicomputable.*

For example, it is easy to see that the functions $K(x)$ and $K(y \mid x^*)$ are upper semicomputable. But the function $d(x,y)$, being a ratio between two maxima of pairs of upper-semicomputable functions, is not itself upper-semicomputable. In fact, $d(x,y)$ has only a weaker computability property. Call a function $f(x,y)$ *computable in the limit* if there exists a rational-valued recursive function $g(x,y,t)$ such that $\lim_{t\to\infty} g(x,y,t) = f(x,y)$. Then $d(x,y)$ is in this class. It can be shown (in the full version) that this is precisely the set of functions that are Turing-reducable to the halting set.

The set of functions Turing-reducible to the halting set contains no function that multiplicatively dominates every function in the class. $d(x,y)$ doesn't dominate all normalized distances in this class. It may be meaningful to ask whether it dominates every normalized distance that is a ratio of two upper semicomputable functions. However, for us it suffices that $d(x,y)$ multiplicatively dominates every normalized distance in the very wide class of upper semicomputable normalized distances; this class contains every normalized distance we can be remotely be interested in.

Theorem 4.2 *The normalized information distance $d(x,y)$ dominates every upper-semicomputable normalized distance $f(x,y)$ up to a vanishing additive term: $d(x,y) \leq f(x,y) - O(\log k/k)$, where $k = \max\{K(x), K(y)\}$.*

Proof. Fix a normalized distance $f(x,y)$ and assume $f(x,y) = d$. By the normalization condition we have that, given x, the number of y, such that $f(x,y) \leq d$ and $\max\{K(x), K(y)\} \leq k$, is upper bounded by 2^{dk}. Hence, for fixed x^* and k we can recursively enumerate the y for which $f(x,y) \leq d$, and every y can be described by its index of length $\leq dk$ in this enumeration. Since the Kolmogorov complexity is the length of the shortest effective description, the binary length of the index must at least be as large as the Kolmogorov complexity, which yields $K(y \mid x^*, k) = K(y \mid x, K(x), k) \leq dk$. That is, since we can provide both k and $K(x)$ in $O(\log k)$ bits, $K(y \mid x^*) \leq dk - O(\log k)$. Given x and y, assume that $K(y) \geq K(x)$ (the other case is treated symmetricaly). Then, by (6), we also have $K(y \mid x^*) \geq K(x \mid y^*)$, and $d(x,y) = K(y \mid x^*)/K(y)$. Substitution gives:

$$d(x,y) = \frac{K(y \mid x^*)}{K(y)} \leq \frac{dk - O(\log k)}{k} = f(x,y) - O(\frac{\log k}{k}).$$

$\bullet$

The above results show that the normalized information distance is a universal similarity distance that incorporates all similarity aspects from a wide class of effecive similarity distances.

5 Application to Whole Mitochondrial Genome Phylogeny

Nothing is more ideal than DNA sequences to test our theory. We will use whole mitochondrial DNA genomes of 20 mammals and the problem of Eutherian orders to make a comprehensive examination of our measures. The problem we consider is this: It has been debated in biology which two of the three main groups of placental mammals are more closely related: Primates, Ferungulates, and Rodents. This is because the maximum likelihood method gives (Ferungulates, (Primates, Rodents)) grouping for half of the proteins in mitochondial genome, and (Rodents, (Ferungulates, Primates)) for the other half [8]. In [8], Cao *et al.* aligned 12 concatenated mitochondrial proteins taken from the following species: rat (*Rattus norvegicus*), house mouse (*Mus musculus*), grey seal (*Halichoerus grypus*), harbor seal (*Phoca vitulina*), cat (*Felis catus*), white rhino (*Ceratotherium simum*), horse (*Equus caballus*), finback whale (*Balaenoptera physalus*), blue whale (*Balaenoptera musculus*), cow (*Bos taurus*), gibbon (*Hylobates lar*), gorilla (*Gorilla gorilla*), human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), pygmy chimpanzee (*Pan paniscus*), orangutan (*Pongo pygmaeus*), Sumatran orangutan (*Pongo pygmaeus abelii*), using opossum (*Didelphis virginiana*), wallaroo (*Macropus robustus*) and platypus (*Ornithorhynchus anatinus*) as the outgroup, and built the maximum likelihood tree. The resulting phylogeny supports the currently accepted grouping (Rodents, (Primates, Ferungulates)).

5.1 Alternative Approaches:

Before applying our theory to further confirm this hypothesis, we first examine the alternative approaches, in addition to that of [8]. The mitochondrial genomes of the above 20 species were obtained from GenBank. Each is about 18k bases.

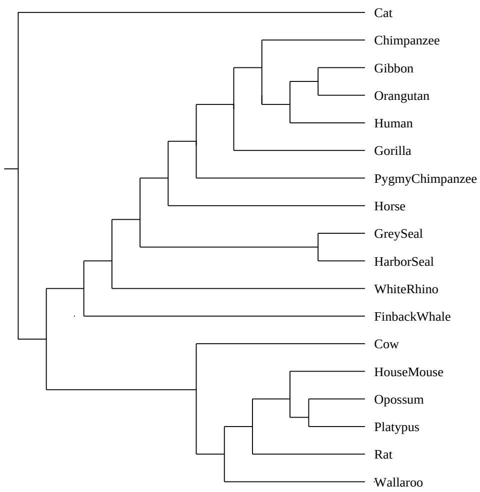

Figure 1: The evolutionary tree built from complete mammalian mtDNA sequences using frequency of k-gons.

k-gon Statistic: In the early years, researchers experimented on using G+C contents or slightly more general k-gons (or Shannon block entropy) to classify DNA sequences (in particular S. Wildman at Stanford). This approach uses the statistics of length k substrings in a genome and the phylogeny is constructed accordingly. To re-examine this approach, we performed simple experiments: Consider all length k blocks in each mtDNA, for $k = 1, 2, \ldots, 10$. There are $l = 4^{11} - 1$ different sequences (some may not appear in an mtDNA). We computed their number of occurrences in each mtDNA, obtaining a vector of aength l for each mtDNA. For two such vectors (representing two mtDNAs) p, q, their distance is computed as $d(p,q) = \sqrt{(p-q)^T(p-q)}$. Using neighbor joining [31], the resulting tree is the one given in Figure 1. Using hypercleaning method [7], we obtain equally absurd results. Similar experiments were repeated for size k blocks alone (for $k = 10, 9, 8, 7, 6$), without much improvement.

Gene Order: In [6] the authors propose to use the order of genes to infer the evolutionary history. This approach does not work for closely related species such as our example where all genes are in the same order in the mitochondrial genomes in all 20 species.

Gene Content: Gene content method, proposed in [11, 32], compares number of genes two species share divided by total number of genes. While this approach does not work here due to the fact that all 20 mammalian mitochondrial genomes share exactly the same genes, notice the similarity of gene content formula and our general formula.

Rearrangement Distance: Reversal and rearrangement distances in [18, 17, 28] compare genomes using other partial genome information such as number of reversals or translocations. These operations also do not appear in our mammalian mitochondrial genomes, hence the method again is not proper for our application.

Transformation Distance or Compression Distance: The transformation distance proposed in [35] and compression distance proposed in [15] are essentially defined as $K(x|y)$ which is asymmetric, and so, is not a distance. The measure $K(x|y)$ produces a wrong tree with one of the marsupials mixed up with ferungulates (the tree is not shown here).

5.2 The Present Approach

We have shown that d (and up to a factor 2 also d_s) is universal mong a wide class of computable normalized information measures. However the generality of d and d_s comes at the price of noncomputability. Kolmogorov complexity is not effectively computable and not even approximable [24]. Nonetheless, we can try to approximate the spirit of d and d_s at various levels of precision, as follows.

k-gons According to the New Measures: We have shown that using k-gon statistics alone does not work well. However, let us now combine the k-gon approach with the new measures. Consider the length-k substrings of the DNA sequence as words of the sequence. We denote the number of distinct, possibly overlapping, k-length words in a sequence x by $N(x)$. With k large enough, at least $\log_a(n)$, where a is the cardinality of the alphabet and n the length of x, we use $N(x)$ as a rough approximation

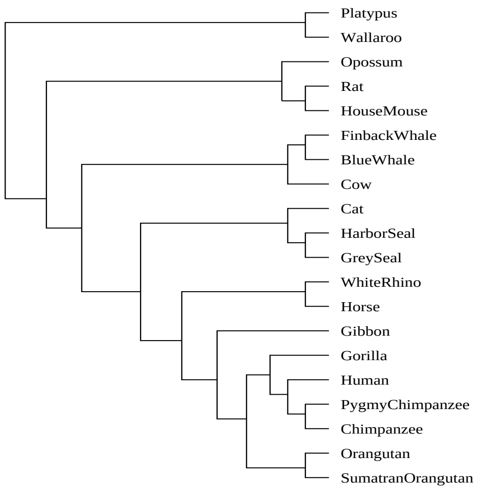

Figure 2: The evolutionary tree built from complete mammalian mtDNA sequences using block size $k = 7$ and d'.

to $K(x)$. We justify this by the pragmatic observation that, because the genomes evolve by duplications, rearrangements and mutations, [30], it can be argued that it is appropriate to use $N(x)$ to very roughly estimate for $K(x)$ in case x is a genome." Define $N(x|y)$ as $N(xy) - N(y)$. Given two sequences x and y, following the definition of d, (5), the distance between x and y can be defined as

$$d'(x, y) = \frac{\max\{N(x|y), N(y|x)\}}{\max\{N(x), N(y)\}}. \tag{9}$$

Similarly, following d_s, (4) we can also define another distance using $N(x)$,

$$d^*(x, y) = \frac{N(x|y) + N(y|x)}{N(xy)}. \tag{10}$$

Using d' and d^*, we computed the distance matrixes for the 20 mammal mitochondrial DNAs. Then we used hyperCleaning [7] to construct the phylogenies for the 20 mammals. Using either of d' and d^*, we were able to construct the tree correctly when $8 \leq k \leq 13$, as in Figure 3. A tree constructed with d' for $k = 7$ is given in Figure 2. We note that the opossum and a few other species are misplaced. The tree constructed with d^* for $k = 7$ is very similar, but it correctly positioned the opossum.

Spaced k-gons According to the New Measures In methods for doing DNA homology search, a pair of identical words, each from a DNA sequence, is called a "hit". Hits have been used as "seeds" to generate a longer match between the two sequences. We note that $N(x|y)$ is the number of distinct words that are in x and not in y, the more hits the two sequences have, the smaller the $N(x|y)$ and $N(y|x)$ are. Therefore, the previous two distances can also be interpreted as a function of the number of hits, each of which indicates some mutual information of the two sequences.

As noticed by the authors of [26], though it is difficult to get the first hit (of consecutive k letters) in a region, it only requires one more base match to get a second hit overlapping the existing one. This makes it inaccurate to attribute the same amount of information to each of the hits. For this reason, we also tried to use the "spaced model" introduced in [26] to compute our distances. A length-L, weight-k spaced model is a 0-1 string of length L and having k 1s. We overlap such a model with the DNA sequence at each of the positions in the DNA sequence, and take out the k bases covered by the 1s to form a length-k word. The number of those distinct words is then used to define the distances d' and d^* in Formula (9) and (10).

We applied the new defined distances to the 20 mammal data. The performance is slightly bettern than the performance of the distances defined in (9) and (10). The modified d' and d^* can correctly construct the mammal tree when $7 \leq k \leq 13$ and $6 \leq k \leq 13$, respectively.

Compression: To achieve the best approximation of Kolmogorov complexity, and hence most confidence in the approximation of d_s and d, we use a new version of the *GenCompress* program, [9], which achieves the currently best compression ratios for benchmark DNA sequences. *GenCompress* finds approximate matches (hence edit distance becomes a special case), approximate reverse complements, among other

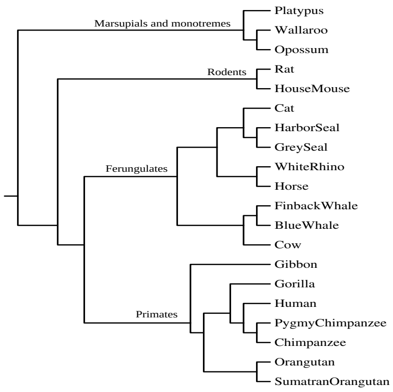

Figure 3: The evolutionary tree built from complete mammalian mtDNA sequences.

things, with arithmetic encoding when necessary. Online service of *GenCompress* is at UCSB Bioinformatics Lab website: http://cytosine.cs.ucsb.edu:8080/. We computed $d(x, y)$ between each pair of mtDNA x and y, using *GenCompress* to heuristically approximate $K(x|y)$, $K(x)$, and $K(x, y)$, and constructed a tree (Figure 3) using the neighbor joining [31] program in the MOLPHY package [1]. The tree is identical to the maximum likelihood tree of Cao, *et al.* [8]. For comparison, we used the hypercleaning program [7] and obtained the same result. The phylogeny in Figure 3 re-confirms the hypothesis of (Rodents, (Primates, Ferungulates)). Using the d_s measure gives the same result.

To further assure our results, we have extracted only the coding regions from the mtDNAs of the above species, and performed the same computation. This resulted in the same tree.

Evaluation: This new method for whole genome comparison and phylogeny does not require gene identification nor any human intervention, in fact, it is totally automatic. It is mathematically well-founded being based on general information theoretic concepts. It works when there are no agreed upon evolutionary models, as further demonstrated by the successful construction of a chain letter phylogeny [5] and when individual gene trees do not agree (*e.g.* Cao et al., [8]) as is the case for genomes. Next step would be to apply this method to larger genomes such as cpDNA and bacteria genomes.

6 The Language Tree

Normalized information distance is a totally general universal tool, not restricted to a particular application area. We show that it can also be used to successfully classify natural languages. Let us borrow from biology the "nature" (acquired by genetic mixture) versus "nurture" (acquired in the life of the individual) terminology. Any language tree built by only analyzing contemporary natural text corpora is partially corrupted by historical "nurture" contaminations. While according to Darwinism the genomes only change by inheritance (nature), languages acquire their characteristic by descent but also by interaction (nurture). Thus, while English is a Germanic Anglo-Saxon language, by the Norman conquest of 1066 it has absorbed about 25% French-Latin components. Similarly, Hungarian, often considered a Finn-Ugric language which consensus currently happens to be open to debate in the linguistic community, is known to have absorbed many Turkish and Slavic components. Thus, an automatic construction of a language tree based on contemporary text corpora, exhibits current linguistic relations (based on both nature and nurture) which do not necessarily coincide completely with the historic language family tree (based on nature). Having voiced these caveats, the result of our automatic experiment in language tree reconstruction is rather encouraging.

Our method improves the results of [2], using a linguistic corpus of "The Universal Declaration of Human Rights" [14] in 52 languages. The previous effort [2] used an asymmetric measure based on relative entropy, and the full matrix of the pair-wise measures between all 52 languages, to build a language classification tree. This resulted in some inconsistensies, such as English being isolated between Romance and Celtic languages, Romani-balkan and Albanian being isolated, and Hungarian (possibly a Finn-Ugric

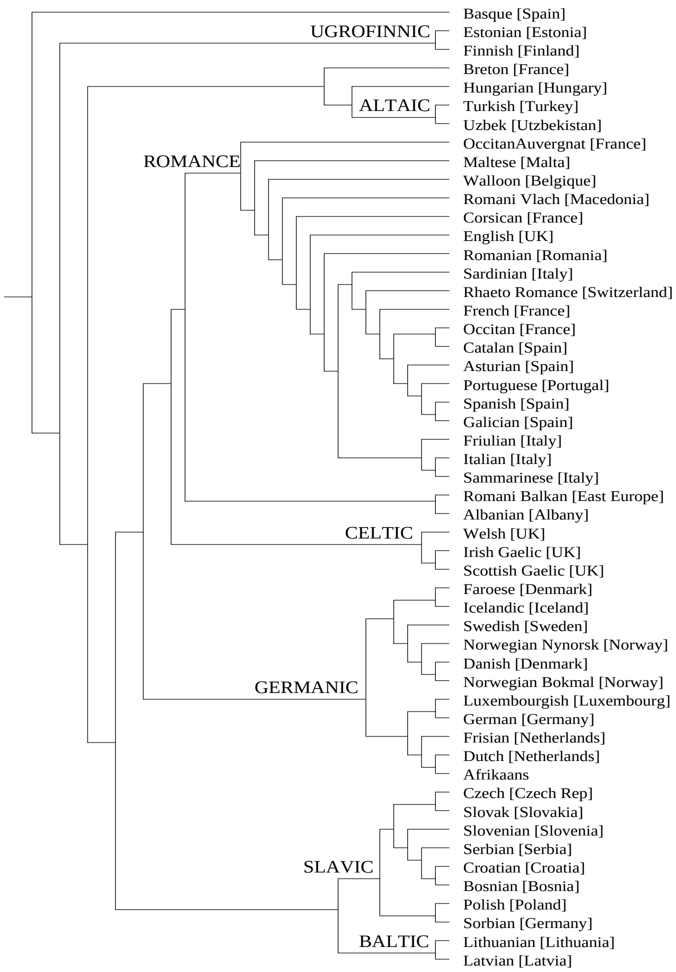

Figure 4: The language classification tree, built from "The Universal Declaration of Human Rights" [14] using approximated normalized information distance, d_s-version (9), and neighbor joining. This is a rooted tree, using Basque [Spain] as outgroup. The branch lengths are not proportional to the actual distances in the distance matrix.

language) being grouped with Turkish and Uzbek. The (rooted) trees resulting from our experiments (using Basque as out-group) seem more correct. We use Basque as outgroup since linguists regard it as a very old (3000 BC) language unconnected to any other language. In order to test the potential of normalized information distance d in classifying natural languages, a similar experiment was performed. First, transform each UNICODE character in the language text into an ASCII character by removing its vowel flag if necessary. Secondly, a LZ-type algorithm *gzip* is appropriate to compress these language sequences of sizes not exceeding the length of sliding window *gzip* use (32 kilobytes), and compression results can be used to approximate their Kolmogorov complexity. Instead of a complicated and less rigorous method to approximate relative entropy used in [2], we simply zip-compress several sequences that involve calculating the normalized information metric (both (9) and (5)) based on each pair of language sequences. In the last step, we applied the d_s-metric (9) with the neighbor-joining package to obtain Figure 4. Even better worked applying the d-metric (5) with the Fitch-Margoliash method [10] in the package PHYLIP [1]); the resulting language classification tree is given in Figure 5. We note that all the main linguistic groups can be successfully recognized, which includes Romance, Celtic, Germanic, Ugro-Finnic, Slavic, Baltic, Altaic as labeled in the figure.

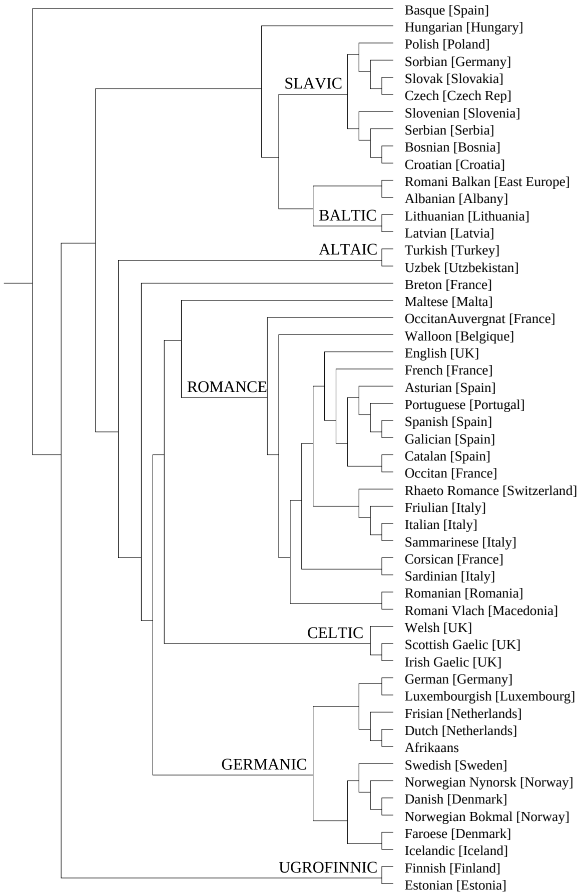

Figure 5: The language classification tree, built from "The Universal Declaration of Human Rights" [14] using approximated normalized information distance, d-version (5), and the Fitch-Margoliash method. This is a rooted tree using Basque [Spain] as outgroup. The branch lengths are not proportional to the actual distances in the distance matrix.

7 Conclusion

We developed a mathematical theory of similarity distances and shown that there is a universal similarity distance: the normalized information distance. This distance uncovers all effective similarities, and therefore estimates an evolutionary or relation-wise distance on strings. It has been shown to be applicable to whole genomes, but as well to chain letters ([5], not included here), to test tudents source code for plagiarism ([33], not included here), and to built a large language family tree from text corpora. It is perhaps useful to point out that the results reported in the figures were obtained at the very first runs and have not been selected by appropriateness from several trials. From the theory point-of-view we have obtained a general mathematical theory forming a solid framework spawning practical tools applicable in many fields. Based on the noncomputable notion of Kolmogorov complexity, the normalized information distance can only be approximated in an *ad hoc* manner, that is, without speed of convergence guarantees. Even so, the fundamental rightness of the approach is evidenced by the remarkable success (agreement with known phylogeny in Biology) of the evolutionary trees obtained and the building of language trees. From the applied side of genomics our work gives the first fully automatic generation of whole genome mitochondrial phylogeny; in computational linguistics it presents a fully automatic way to build language trees and determine language families.

References

[1] J. Adachi and M. Hasegawa. MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. *Comput. Sci. Monogr.*, *Inst. Stat. Math.*, 28. 1-150.

[2] D. Benedetto, E. Caglioti, and V. Loreto, Language trees and zipping, *Phys. Review Lett.*, 88:4(2002) 048702.

[3] Ph. Ball, Algorithm makes tongue tree, *Nature*, 22 Januari, 2002.

[4] C.H. Bennett, P. Gács, M. Li, P.M.B. Vitányi, and W. Zurek, Information Distance, *IEEE Transactions on Information Theory*, 44:4(1998), 1407–1423.

[5] C.H. Bennett, M. Li, and B. Ma, Linking chain letters. *Scientific American*, To appear.

[6] J.I. Boore and W.M. Brown, Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. *Curr. Opin. Genet. Dev.* 8(1998), 668-674.

[7] D. Bryant, V. Berry, P. Kearney, M. Li, T. Jiang, T. Wareham and H. Zhang. A practical algorithm for recovering the best supported edges of an evolutionary tree. *SODA'2000*.

[8] Y. Cao, A. Janke, P. J. Waddell, M. Westerman, O. Takenaka, S. Murata, N. Okada, S. Pääbo, M. Hasegawa, Conflict among individual mitochondrial proteins in resolving the phylogeny of Eutherian orders, *J. Mol. Evol.*, 47(1998), 307-322.

[9] X. Chen, S. Kwong, M. Li A compression algorithm for DNA Sequences, *IEEE-EMB Special Issue on Bioinformatics*, 20:4(2001), 61-66.

[10] W.M. Fitch and E. Margoliash, Construction of phylogenetic trees, *Science*, 155(1967), 279–284.

[11] S.T. Fitz-Gibbon and C.H. House, Whole genome-based phylogenetic analysis of free-living macroorganisms. *Nucleic Acids Res.* 27(1999), 4218-4222.

[12] P. Gács, On the symmetry of algorithmic information, *Soviet Math. Dokl.*, 15 (1974) 1477–1480. Correction: ibid., 15 (1974) 1480.

[13] P. Gács, J. Tromp, P. Vitányi, Algorithmic Statistics, *IEEE Transactions on Information Theory*, 47:6(2001), 2443–2463.

[14] United Nations General Assembly resolution 217 A (III) of 10 December 1948: Universal Declaration of Human Rights, http://www.un.org/Overview/rights.html

[15] S. Grumbach and F. Tahi, A new challenge for compression algorithms: genetic sequences, *J. Info. Process. Manage.*, 30(1994), 875-866.

[16] D. Hammer, A.E. Romashchenko, A.Kh. Shen', N.K. Verashchagin, Inequalities for Shannon entropies and Kolmogorov complexities, *Proc. 12th IEEE Conf. Computational Complexity*, pp. 13-23, 1997.

[17] S. Hannenhalli and P. Pevzner. Transforming cabbage into turnip. *STOC'95*. pp. 178-189.

[18] J. Kececioglu and D. Sankoff. Exact and approximation algorithms for the inversion distance. *Algorithmica*, 13(1995), 180-210.

[19] A.N. Kolmogorov, Three approaches to the quantitative definition of information, *Problems Inform. Transmission* 1:1 (1965) 1–7.

[20] E.V. Koonin, The emerging paradigm and open problems in comparative genomics, *Bioinformatics*, 15(1999), 265-266.

[21] M. Li, J.H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang, An information-based sequence distance and its application to whole mitochondrial genome phylogeny, *Bioinformatics*, 17:2(2001), 149–154.

[22] M. Li and P.M.B. Vitányi, Reversibility and adiabatic computation: trading time and space for energy, *Proc. Royal Society of London, Series A*, 452(1996), 769-789.

[23] M. Li and P.M.B. Vitányi, Reversibility and adiabatic computation: trading time and space for energy, *Proc. Royal Society of London, Series A*, 452(1996), 769-789.

[24] M. Li and P.M.B. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications*, Springer-Verlag, New York, 2nd Edition, 1997.

[25] M. Li, J. Tromp, and L. Zhang, On the neirest neighbor interchange distance between evolutionary trees, *J. Theor. Biol.*, 182:463–467, 1996.

[26] B. Ma, J. Tromp, and M. Li, PatternHunter: faster and more sensitive homology search, *Bioinformatics*, To appear.

[27] A.A. Muchnik and N.K. Vereshchagin, Logical operations and Kolmogorov complexity, *Proc. 16th IEEE Conf. Computational Complexity*, 2001.

[28] J.H. Nadeau and D. Sankoff. Counting on comparative maps. *Trends Genet.* 14(1998), 495-501.

[29] A. Romashchenko, A. Shen, and N. Vereshchagin, Combinatorial interpretation of Kolmogorov complexity, *Proc. 15th IEEE Conf. Computational Complexity*, 2000, 131–137.

[30] D. Sankoff, Mechanisms of genome evolution: models and inference, *Bull. International Statistical Institute* 47:3(1999), 461–475.

[31] N. Saitou and M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4(1987), 406-425.

[32] B. Snel, P. Bork, and M.A. Huynen, Genome phylogeny based on gene content. *Nature Genet.* 21(1999), 108-110.

[33] Shared Information Distance or Software Integrity Detection, Computer Science, University of California, Santa Barbara, http://dna.cs.ucsb.edu/SID/

[34] D. Sleator, R. Tarjan, W. Thurston, Short encodings of evolving structures, *SIAM J. Discrete Math.*, 5:428–450, 1992.

[35] J-S. Varre and J-P. Delahaye and É Rivals, The transformation distance: a dissimilarity measure based on movements of segments, *German Conf. Bioinformatics*, Koel, Germany, 1998.

[36] N.K. Vereshchagin and M. Vyugin Independent minimum length programs to translate between given strings, *Proc. 15th IEEE Conf. Computational Complexity*, 2000, 138–145.

[37] P.M.B. Vitányi, Meaningful information, Los Alamos National Laboratories Archives, http://xxx.lanl.gov/abs/cs.CC/0111053

[38] L.-S. Wang, T. Warnow, Estimating true distances between genomes, *Proc. 33rd ACM Symp. Theory Comput.*, 2001, 637–646.

[39] J.C. Wooley, Trends in computational biology: a summary based on a RECOMB plenary lecture, 1999, *J. Comput. Biol.*, 6(1999), 459-474.

[40] J. Ziv and N. Merhav, A measure of relative entropy between individual sequences with application to universal classification, *IEEE Transactions on Information Theory*, IT-39(4): 1270–1279, 1993.

A Appendix: Summary of Background

Distance Metric: Without loss of generality, a distance only needs to operate on sequences of 0's and 1's since any sequence can be represented by a binary sequence. Formally, a *distance* function D with nonnegative real values, defined on the Cartesian product $X \times X$ of a set X is called a *metric* on X if for every $x, y, z \in X$:

- $D(x, y) = 0$ iff $x = y$ (the identity axiom);
- $D(x, y) + D(y, z) \geq D(x, z)$ (the triangle inequality);
- $D(x, y) = D(y, x)$ (the symmetry axiom).

A set X provided with a metric is called a *metric space*. For example, every set X has the trivial *discrete metric* $D(x, y) = 0$ if $x = y$ and $D(x, y) = 1$ otherwise. All distances in this paper are defined on the set $X = \{0, 1\}^*$ and satisfy the metric conditions sometimes up to an additive vanishing constant term.

In our search for the proper definition of the distance between two, not necessarily equal length, binary strings, a natural choice is the length of the shortest program that can transform either string into the other one—both ways. This distance is known as *information distance*, which is one of the main concepts in this work and which we will discuss in detail below. However, such a distance measures an absolute distance, and we are more interested in a relative one. For example, if two strings of length 10^6 have distance 1000, then we are inclined to think that those strings are relatively more similar than two strings of length 1000 that have that distance and consequently are 100% different. Therefore, we want to normalize the information distance into a normalized distance. We do this by dividing it by the greater of the two lengths of the shortest programs that compute the strings concerned from scratch.

Kolmogorov Complexity: We write *string* to mean a finite binary string. Other finite objects can be encoded into strings in natural ways. The set of strings is denoted by $\{0, 1\}^*$. The *Kolmogorov complexity*, or *algorithmic entropy*, $K(x)$ of a string x is the length of a shortest binary program to compute x on a universal computer (such as a universal Turing machine). Intuitively, $K(x)$ represents the minimal amount of information required to generate x by any effective process, [19]. The conditional Kolmogorov complexity $K(x \mid y)$ of x relative to y is defined similarly as the length of a shortest program to compute x if y is furnished as an auxiliary input to the computation. The functions $K(\cdot)$ and $K(\cdot|\cdot)$, though defined in terms of a particular machine model, are machine-independent up to an additive constant and acquire an asymptotically universal and absolute character through Church's thesis, from the ability of universal machines to simulate one another and execute any effective process. The Kolmogorov complexity of a string can be viewed as an absolute and objective quantification of the amount of information in it. This leads to a theory of *absolute* information *contents* of *individual* objects in contrast to classical information theory which deals with *average* information *to communicate* objects produced by a *random source*. Since the former theory is much more precise, it is surprising that analogons of theorems in classical information theory hold for Kolmogorov complexity, be it in somewhat weaker form. See [24] for formal definitions of Kolmogorov complexity and its successful applications in physics, mathematics, and computer science.

Let x^* denotes the shortest program for x (if there are more than one of them then x^* is the first one in a standard enumeration). For technical reasons, to make our statements precise up to an additive constant term, we use $K(x \mid y^*)$ as the conditional Kolmogorov complexity (or conditional algorithmic entropy) of x given y^*, defined as the length of the shortest program causing a standard universal computer to output x on input y^*, and $K(x)$ is defined as $K(x \mid \epsilon)$, where ϵ is the empty string. Incidentally, $x, K(x)$ carries the same information as x^*, since either one can be computed from the other.

Information Distance: Here our goal is to study the question of an "absolute information distance metric" between individual objects. This should be contrasted with an information metric (entropy metric) such as $H(X|Y) + H(Y|X)$ between stochastic sources X and Y. Non-absolute approaches to information distance between individual objects have been studied in a statistical setting,see for example [40] for a notion of empirical information divergence (relative entropy) between two individual sequences. Other approaches include various types of edit-distances between pairs of strings: the minimal number of edit operations from a fixed set required to transform one string in the other string. Similar distances are defined on trees or other data structures. The huge literature on this ranges from pattern matching and cognition to search strategies on internet and computational biology. As an example we mention nearest neighbor interchange distance between evolutionary trees in computational biology, [34, 25]. *A priori* it is not immediate what is the most appropriate *universal* symmetric informational distance between two strings, that is, the minimal quantity of information sufficient to translate between x and y, generating either string effectively from the other. We summarize the approach and relevant details of [4]:

Algorithmic Information Distance: Define the information distance as the length of a shortest binary program that computes x from y as well as computing y from x. Being shortest, such a program should take advantage of any redundancy between the information required to go from x to y and the information required to go from y to x. The program functions in a catalytic capacity in the sense that it is required to transform the input into the output, but itself remains present and unchanged throughout the computation. We would like to know to what extent the information required to compute y from x can be made to overlap with that required to compute x from y. In some simple cases, *complete* overlap can be achieved, so that the same minimal program suffices to compute x from y as to compute y from x. For example if x and y are independent random binary strings of the same length n (up to additive contants $K(x \mid y) = K(y \mid x) = n$), then their bitwise exclusive-or $x \oplus y$ serves as a minimal program for both computations. Similarly, if $x = uv$ and $y = vw$ where u, v, and w are independent random strings of the same length, then $u \oplus w$ plus a way to distinguish x from y is a minimal program to compute either string from the other.

Maximal Correlation: Now suppose that more information is required for one of these computations than for the other, say,

$$K(y \mid x) > K(x \mid y).$$

Then the minimal programs cannot be made identical because they must be of different sizes. In some cases it is easy to see that the overlap can still be made complete, in the sense that the larger program (for y given x) can be made to contain all the information in the shorter program, as well as some additional information. This is so when x and y are independent random strings of unequal length, for example u and vw above. Then $u \oplus v$ serves as a minimal program for u from vw, and $(u \oplus v)w$ serves as one for vw from u.

A principal result of [4] shows that, up to an additive logarithmic error term, the information required to translate between two strings can be represented in this maximally overlapping way in every case. Namely, let

$$k_1 = K(x \mid y),\ k_2 = K(y \mid x),$$
$$l = k_2 - k_1$$

where we assume $k_1 \leq k_2$. Then there is a string q of length $k_1 + K(k_1, k_2)$ and a string d of length l such that q serves as the minimal program both to compute from xd to y and from y to xd. The term $K(k_1, k_2)$ has magnitude $O(\log k_2)$. This means that the information to pass from x to y can always be maximally correlated with the information to get from y to x. It is therefore never the case that a large amount of information is required to get from x to y and a large *but independent* amount of information is required to get from y to x. This demonstrates that

$$\max\{K(y \mid x), K(x \mid y)\}$$

equals the length of a shortest program $p := qd$ to compute x from y and y from x, up to a logarithmic additive term. (It is very important here that the time of computation is completely ignored: this is why this result does not contradict the idea of one-way functions.)

The process of going from x to y may be broken into two stages. First, add the string d; second, use the difference program q between xd and y. In the reverse direction, first use q to go from y to xd; second, erase d. Thus the computation from x to y needs both q and d, that is, the program $p = qd$, while the computation from y to x needs only q as program.

Meaningful Information: The information contained by an individual finite object (like a finite binary string) is objectively measured by its Kolmogorov complexity—the length of the shortest binary program that computes the object. Such a shortest program contains no redundancy: every bit is information; but is it meaningful information? If we flip a fair coin to obtain a finite binary string, then with overwhelming probability that string constitutes its own shortest program. However, also with overwhelming probability all the bits in the string are meaningless information, random noise. On the other hand, let an object x be a DNA string containing random junk. Then x can be described by the binary string pd, where p is the description of the meaningful DNA part, and d the description of the junk material: we can divide the information in x into meaningful information p and accidental information d. The main task for statistical inference and learning theory is to distil the meaningful information present in the data. The question arises whether it is possible to separate meaningful information from accidental information, and if so, how. In [13] this theory is developed as "Kolmogorov sufficient statistic." The generalization in [37] with p is a total recursive program with $p(d) = x$, $l(p) + l(d) \approx K(x)$, and $l(p)$ is minimal, is perhaps most appropriate here. Here, the program p embodies the meaningful information, and $l(p)$ is the number of meaningful bits thereof. In [37] it is shown that there is an algorithm that approximates p from $x, K(x)$ in the sense that it computes a sequence $p_1, p_2, \ldots p_f = p$ where f is bounded by a universal constant independent of x. Denoting the meaningful information associated with x by p_x, we can formulate the distances d_s and d in terms of the meaningful information, rather than the total (including the accidental information). The resulting distances will satisfy the same properties as the ones defined above, and will also be computable in the limit. Thus, the same theory applies. To eliminate random information in the normalized information distances above, it would perhaps be better to replace the objects, like x and y, by the "meaningful" information in x and y.