A 1-parameter family of metrics connecting Jaccard distance to normalized information distance

Bjørn Kjos-Hanssen*, Saroj Niraula, Soowhan Yoon, Sabrina Hardisty, Guanhong Li, Jacqueline Millard

October 12, 2020

Abstract

Jiménez, Becerra, and Gelbukh (2013) defined a family of symmetric Tversky ratio models S parametrized by $0 \leq \alpha \leq 1$ and $\beta > 0$. Letting $D = 1 - S$ we have a semimetric which we show is a metric if and only if $0 \leq \alpha \leq \frac{1}{2}$ and $\beta \geq 1/(1-\alpha)$.

For $\beta = 1/(1-\alpha)$, the two endpoints $\alpha = 0, \frac{1}{2}$ correspond to the normalized information distance and Jaccard distance, respectively.

1 Introduction

Distance metrics are used in a wide variety of scientific contexts. In bioinformatics, M. Li, Badger, Chen, Kwong, and Kearney [LBC$^+$01] introduced an information-based sequence distance. In an information-theoretical setting, M. Li, Chen, X. Li, Ma and Vitányi [LCL$^+$04] rejected the distance of [LBC$^+$01] in favor of a *normalized information distance* (NID). The Encyclopedia of Distances [DD16] describes the NID on page 205 out of 583, as

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

where $K(x \mid y^*)$ is the Kolmogorov complexity of x given a shortest program y^* to compute y. It is equivalent to be given y itself in hard-coded form:

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

*This work was partially supported by grants from the Simons Foundation (#704836 to Bjørn Kjos-Hanssen) and Decision Research Corporation (University of Hawai‘i Foundation Account #129-4770-4).

Another formulation sometimes used is

$$\frac{K(x,y)-\min\{K(x),K(y)\}}{\max\{K(x),K(y)\}}.$$

The fact that the NID is in a sense a normalized metric is proved in [LCL+04]. Then in 2017, while studying malware detection, Raff and Nicholas [RN17] suggested Lempel–Ziv Jaccard distance (LZJD) as a practical alternative to NID. As we shall see, this is a metric. In a way this constitutes a full circle: the distance in [LBC+01] is itself essentially a Jaccard distance, and the LZJD is related to it as Lempel–Ziv complexity is to Kolmogorov complexity. In the present paper we aim to shed light on this back-and-forth by showing that the NID and Jaccard distances constitute the endpoints of a parametrized family of metrics.

For comparison, the Jaccard distance between two sets X and Y, and our analogue of the NID, are

$$\frac{|X\setminus Y|+|Y\setminus X|}{|X\cup Y|}=1-\frac{|X\cap Y|}{|X\cup Y|}, \qquad \text{and} \tag{1}$$

$$\frac{\max\{|X\setminus Y|,|Y\setminus X|\}}{\max\{|X|,|Y|\}}, \tag{2}$$

respectively. Our main result Theorem 12 shows which interpolations between these two are metrics.

The way we arrived at (2) as an analogue of NID is via Lempel–Ziv complexity. While there are several variants [LZ76, ZL77, ZL78], the LZ 1978 complexity [ZL78] of a sequence is the cardinality of a certain set, the dictionary. It will not be used in our paper except in conferring the spirit of Kolmogorov complexity onto set distances by suggesting the following notation.

Definition 1. *Let X and Y be finite sets.*

Definition 2. *Let* LZSet(A) *be the Lempel–Ziv dictionary for a sequence A. We define LZ–Jaccard distance* LZJD *by*

$$\mathrm{LZJD}(A,B)=1-\frac{|\mathrm{LZSet}(A)\cap\mathrm{LZSet}(B)|}{|\mathrm{LZSet}(A)\cup\mathrm{LZSet}(B)|}.$$

It is shown in [LBC+01, Theorem 1] that the triangle inequality holds for a function which they call an information-based sequence distance. Later papers give it the notation d_s in [LCL+04, Definition V.1], and call their normalized information distance d. Raff and Nicholas [RN17] introduced the LZJD and did not discuss the appearance of d_s in [LCL+04, Definition V.1], even though they do cite [LCL+04] (but not [LBC+01]).

Kraskov et al. [KSAG03, KSAG05] use D and D' for continuous analogues of d_s and d in [LCL+04] (which they cite). The *Encyclopedia* calls it the normalized information metric,

$$\frac{H(X\mid Y)+H(X\mid Y)}{H(X,Y)}=1-\frac{I(X;Y)}{H(X,Y)}$$

Reference	Jaccard notation	NID notation
[LBC+01]	d	
[LCL+04]	d_s	d
[KSAG05]	D	D'
[RN17]	LZJD	NCD

Table 1: Overview of notation used in the literature. (It seems that authors use simple names for their favored notions.)

or Rajski distance [Raj61].

This d_s was called d by [LBC+01] — see Table 1. Conversely, [LCL+04, near Definition V.1] mentions mutual information.

Remark 3. *Ridgway [Ged10] observed that the entropy-based distance D is essentially a Jaccard distance. No explanation was given, but we attempt one as follows. Suppose X_1, X_2, X_3, X_4 are iid Bernoulli($p = 1/2$) random variables, Y is the random vector (X_1, X_2, X_3) and Z is (X_2, X_3, X_4). Then Y and Z have two bits of mutual information $I(Y, Z) = 2$. They have an entropy $H(Y) = H(Z) = 3$ of three bits. Thus the relationship $H(Y, Z) = H(Y)+H(Z)-I(Y, Z)$ becomes a Venn diagram relationship $|\{X_1, X_2, X_3, X_4\}| = |\{X_1, X_2, X_3\}| + |\{X_2, X_3, X_4\}| - |\{X_2, X_3\}|$. The relationship to Jaccard distance may not have been well known, as it is not mentioned in [KSAG05, CV07, LBC+01, CV05].*

A more general setting is that of STRM (Symmetric Tversky Ratio Models), Definition 11. These are variants of the Tversky index (Definition 5) proposed in [JBG13].

Definition 4. *A* semimetric *on $\mathcal{X}$ is a function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that satisfies the first three axioms of a metric space, but not necessarily the triangle inequality: $d(x, y) \geq 0$, $d(x, y) = 0$ if and only if $x = y$, and $d(x, y) = d(y, x)$ for all $x, y \in \mathcal{X}$.*

Definition 5. *For sets X and Y the Tversky index with parameters $\alpha, \beta \geq 0$ is a number between 0 and 1 given by*

$$S(X, Y) = \frac{|X \cap Y|}{|X \cap Y| + \alpha|X \setminus Y| + \beta|Y \setminus X|}.$$

We also define the corresponding Tversky dissimilarity $d^T_{\alpha,\beta}$ by

$$d^T_{\alpha,\beta}(X, Y) = \begin{cases} 1 - S(X, Y) & \text{if } X \cup Y \neq \emptyset; \\ 0 & \text{if } X = Y = \emptyset. \end{cases}$$

To motivate Definition 4, we include the following lemma without proof.

Lemma 6. *Suppose d is a metric on a collection of nonempty sets $\mathcal{X}$, with $d(X, Y) \leq 2$ for all $X, Y \in \mathcal{X}$. Let $\hat{\mathcal{X}} = \mathcal{X} \cup \{\emptyset\}$ and define $\hat{d} : \hat{\mathcal{X}} \times \hat{\mathcal{X}} \to \mathbb{R}$ by stipulating that for $X, Y \in \mathcal{X}$,*

$$\hat{d}(X, Y) = d(X, Y); \quad d(X, \emptyset) = 1 = d(\emptyset, X); \quad d(\emptyset, \emptyset) = 0.$$

Then $\hat{d}$ is a metric on $\hat{\mathcal{X}}$.

Theorem 7 (Gragera and Suppakitpaisarn [GS16, GS18])**.** *The optimal constant ρ such that $d^T_{\alpha,\beta}(X,Y) \leq \rho(d^T_{\alpha,\beta}(X,Y) + d^T_{\alpha,\beta}(Y,Z))$ for all X, Y, Z is*

$$\frac{1}{2}\left(1 + \sqrt{\frac{1}{\alpha\beta}}\right).$$

Corollary 8. *$d^T_{\alpha,\beta}$ is a metric only if $\alpha = \beta \geq 1$.*

Proof. Clearly, $\alpha = \beta$ is necessary to ensure $d^T_{\alpha,\beta}(X,Y) = d^T_{\alpha,\beta}(Y,X)$. Moreover $\rho \leq 1$ is necessary, so Theorem 7 gives $\alpha\beta \geq 1$. □

Definition 9. *The Szymkiewicz—Simpson coefficient is defined by*

$$\mathrm{overlap}(X,Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)}$$

We may note that $\mathrm{overlap}(X,Y) = 1$ whenever $X \subseteq Y$ or $Y \subseteq X$, so that $1 - \mathrm{overlap}$ is not a metric.

Definition 10. *The Sørensen–Dice coefficient is defined by*

$$\frac{2|X \cap Y|}{|X| + |Y|}.$$

Definition 11 ([JBG13, Section 2])**.** *Let $\mathcal{X}$ be a collection of finite sets. We define $S : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as follows. For sets $X, Y \in \mathcal{X}$ we define $m(X,Y) = \min\{|X \setminus Y|, |Y \setminus X|\}$ and $M(X,Y) = \max\{|X \setminus Y|, |Y \setminus X|\}$. The symmetric TRM is defined by*

$$S(X,Y) = \frac{|X \cap Y| + \mathrm{bias}}{|X \cap Y| + \mathrm{bias} + \beta\left(\alpha m + (1-\alpha)M\right)}$$

The unbiased symmetric TRM is the case where $\mathrm{bias} = 0$*, which is the case we shall assume we are in for the rest of this paper. The Tversky semimetric $D'_{\alpha,\beta}$ is defined by $D'_{\alpha,\beta}(X,Y) = 1 - S(X,Y)$, or more precisely*

$$D'_{\alpha,\beta} = \begin{cases} \beta\frac{\alpha m + (1-\alpha)M}{|X \cap Y| + \beta(\alpha m + (1-\alpha)M)}, & \text{if } X \cup Y \neq \emptyset; \\ 0 & \text{if } X = Y = \emptyset. \end{cases}$$

Note that for $\alpha = 1/2$, $\beta = 1$, the STRM is equivalent to the Sørensen–Dice coefficient. Similarly, for $\alpha = 1/2$, $\beta = 2$, it is equivalent to Jaccard's coefficient.

Our main result is (see Figure 1):

Theorem 12. *Let $0 \leq \alpha \leq 1$ and $\beta > 0$. Then $D'_{\alpha,\beta}$ is a metric if and only if $0 \leq \alpha \leq 1/2$ and $\beta \geq 1/(1-\alpha)$.*

Theorem 12 gives the converse to the Gragera and Suppakitpaisarn inspired Corollary 8:

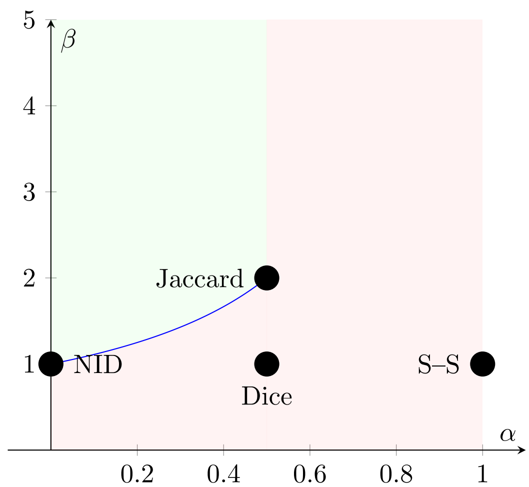

Figure 1: A Tversky semimetric $D'_{\alpha,\beta}$ is a metric if and only if (α, β) belongs to the green region. The parameter values corresponding to the Jaccard distance, the analogue of Normalized Information Distance (NID), the Sørensen–Dice semimetric, and the Szymkiewicz—Simpson semimetric are indicated.

Corollary 13. *The Tversky dissimilarity $d^T_{\alpha,\beta}$ is a metric iff $\alpha = \beta \geq 1$.*

Proof. Suppose the Tversky dissimilarity $d^T_{\alpha,\beta}$ is a semimetric. Let X, Y be sets with $|X \cap Y| = |X \setminus Y| = 1$ and $|Y \setminus X| = 0$. Then

$$1 - \frac{1}{1+\beta} = d^T_{\alpha,\beta}(Y, X) = d^T_{\alpha,\beta}(X, Y) = 1 - \frac{1}{1+\alpha},$$

hence $\alpha = \beta$. Let $\gamma = \alpha = \beta$.

Now, $d^T_{\gamma,\gamma} = D'_{\alpha_0,\beta_0}$ where $\alpha_0 = 1/2$ and $\beta_0 = 2\gamma$. Indeed, let $m = \min\{|X \setminus Y|, |Y \setminus X|\}$ and $M = \max\{|X \setminus Y|, |Y \setminus X|\}$. Since

$$D'_{\alpha_0,\beta_0} = \beta_0 \frac{\alpha_0 m + (1-\alpha_0)M}{|X \cap Y| + \beta_0 \left[\alpha_0 m + (1-\alpha_0)M\right]},$$

$$D'_{\frac{1}{2},2\gamma} = 2\gamma \frac{\frac{1}{2}m + (1-\frac{1}{2})M}{|X \cap Y| + 2\gamma \left[\frac{1}{2}m + (1-\frac{1}{2})M\right]}$$

$$= \gamma \frac{|X \setminus Y| + |Y \setminus X|}{|X \cap Y| + \gamma\left[|X \setminus Y| + |Y \setminus X|\right]} = 1 - \frac{|X \cap Y|}{|X \cap Y| + \gamma|X \setminus Y| + \gamma|Y \setminus X|} = d^T_{\gamma,\gamma}.$$

By Theorem 12, $d^T_{\gamma,\gamma}$ is a metric if and only if $\beta_0 \geq 1/(1-\alpha_0)$. This is equivalent to $2\gamma \geq 2$, i.e., $\gamma \geq 1$. □

The truth or falsity of Corollary 13 does not arise in Gragera and Suppakitpaisarn's work, as they require $\alpha, \beta \leq 1$ in their definition of Tversky index. We note that Tversky [Tve77] only required $\alpha, \beta \geq 0$.

2 Results

Lemma 14. *Let $u, v, w, \epsilon > 0$. Then*

$$\frac{1}{u} \leq \frac{1}{v} + \frac{1}{w} \implies \frac{1}{u+\epsilon} \leq \frac{1}{v+\epsilon} + \frac{1}{w+\epsilon}.$$

Proof. It is of course equivalent to show

$$vw \leq uw + uv \implies (v+\epsilon)(w+\epsilon) \leq (u+\epsilon)(w+\epsilon) + (u+\epsilon)(v+\epsilon),$$

which reduces to

$$(v+w)\epsilon \leq (u+w)\epsilon + (u+v)\epsilon + \epsilon^2,$$

which is clearly the case. □

Lemma 15. *Suppose $a(x,y) = a_{xy}$ and $b(x,y) = b_{xy}$ are functions. Suppose the function d given by $d(x,y) = a_{xy}/b_{xy}$ is a metric, and $\epsilon > 0$ is a real number. Let $\hat{d}(x,y) = \frac{a_{xy}}{b_{xy}+\epsilon a_{xy}}$. Then $\hat{d}$ is also a metric.*

Proof. The only nontrivial task is to verify the triangle inequality. Define further functions u, v, w by

$$u = b_{xy}/a_{xy}, \quad v = b_{xz}/a_{xz}, \quad w = b_{zy}/a_{zy}.$$

Since d is a metric we have

$$\frac{a_{xy}}{b_{xy}} \leq \frac{a_{xz}}{b_{xz}} + \frac{a_{zy}}{b_{zy}}$$

and hence $\frac{1}{u} \leq \frac{1}{v} + \frac{1}{w}$. We proceed by forward reasoning: we need the truth of the following equivalent conditions:

$$\begin{aligned} \frac{a_{xy}}{b_{xy}+\epsilon a_{xy}} &\leq \frac{a_{xz}}{b_{xz}+\epsilon a_{xz}} + \frac{a_{zy}}{b_{zy}+\epsilon a_{zy}}, \\ \frac{1}{u+\epsilon} &\leq \frac{1}{v+\epsilon} + \frac{1}{w+\epsilon}. \end{aligned}$$

By Lemma 14, we are done. □

Theorem 16. *For each α, the set of β for which $D'_{\alpha,\beta}$ is a metric is upward closed.*

Proof. Suppose D'_{α,β_0} is a metric and $\epsilon = \beta - \beta_0 \geq 0$. Let $a_{XY} := \alpha m(X,Y) + (1-\alpha)M(X,Y)$. Since

$$\begin{aligned} D'_{\alpha,\beta}(X,Y) &= \beta \frac{a_{XY}}{|X \cap Y| + \beta a_{XY}} \\ &= \beta \frac{a_{XY}}{|X \cap Y| + \beta_0 a_{XY} + \epsilon a_{XY}}, \end{aligned}$$

and since the upfront factor of β may be removed without loss of generality, the question reduces to Lemma 15. □

Some convenient notation to be used below includes

- $\overline{\alpha} = 1 - \alpha$,
- $\gamma := \beta\alpha \le 1$ with $\beta = 1/\overline{\alpha}$,
- $x_{\cap}y = |X \cap Y|$, $x = |X|$ etc.,
- x_y for $|X \setminus Y|$, $x_{zy} = |X \setminus (Z \cup Y)| = |(X \setminus Z) \setminus Y|$,
- $x_{000} = |\overline{X} \cap \overline{Y} \cap \overline{Z}|$, $x_{001} = |\overline{X} \cap \overline{Y} \cap Z|$, $x_{010} = |\overline{X} \cap Y \cap \overline{Z}|$, $x_{011} = |\overline{X} \cap Y \cap Z|$, $x_{100} = |X \cap \overline{Y} \cap \overline{Z}|$,$x_{101} = |X \cap \overline{Y} \cap Z|$, $x_{110} = |X \cap Y \cap \overline{Z}|$, $x_{111} = |X \cap Y \cap Z|$.

Theorem 17. *$\delta := \alpha m + \overline{\alpha} M$ satisfies the triangle inequality if and only if $\alpha \le 1/2$.*

Proof. We first show the *only if* direction. Let $X = \{0\}$, $Y = \{1\}$, $Z = \{0, 1\}$. Then

$$\begin{aligned} \alpha m(X,Y) + \overline{\alpha} M(X,Y) &= 1, \\ \alpha m(X,Z) + \overline{\alpha} M(X,Z) = \alpha m(Y,Z) + \overline{\alpha} M(Y,Z) &= 0 + \overline{\alpha}. \end{aligned}$$

The triangle inequality then is equivalent to $1 \le 2\overline{\alpha}$, i.e., $\alpha \le 1/2$.

Now let us show the *if* direction. The triangle inequality says

$$\begin{aligned} \alpha \min\{x_y, y_x\} + \overline{\alpha} \max\{y_x, x_y\} &\le \alpha \min\{x_z, z_x\} + \overline{\alpha} \max\{z_x, x_z\} \\ &+ \alpha \min\{z_y, y_z\} + \overline{\alpha} \max\{y_z, z_y\} \end{aligned}$$

By symmetry between x and y, we may assume that $y \le x$. Hence either $y \le z \le x$, $y \le x \le z$, or $z \le y \le x$. Thus our proof splits into three Cases, I, II, and III.

Case I: $y \le z \le x$: we must show that

$$\alpha y_x + \overline{\alpha} x_y \le \alpha z_x + \overline{\alpha} x_z + \alpha y_z + \overline{\alpha} z_y.$$

Since $y_x \le y_z + z_x$ and $x_y \le x_z + z_y$, this holds for all α.

Case II: $y \le x \le z$: We want to show that

$$\alpha y_x + \overline{\alpha} x_y \le \alpha x_z + \overline{\alpha} z_x + \alpha y_z + \overline{\alpha} z_y.$$

In terms of $\gamma = \alpha/\overline{\alpha}$ this says

$$\gamma y_x + x_y \le \gamma x_z + z_x + \gamma y_z + z_y,$$

or

$$0 \le (y_z + x_z - y_x)\gamma + z_x + z_y - x_y = C\gamma + D.$$

Since $y \le x \le z$, we have $y_x \le x_y$, $x_z \le z_x$, $y_z \le z_y$ and so

$$C = y_z + x_z - y_x \le z_x + z_y - x_y = D.$$

Subcase II.1: $C \geq 0$. Then $C\gamma + D \geq 2C \geq 0$, as desired.
Subcase II.2: $C < 0$. In order to show $C\gamma + D \geq 0$ for all $0 \leq \gamma \leq 1$ it suffices that $C + D \geq 0$, since then $C\gamma + D = D - |C|\gamma \geq D - |C| \geq 0$.

We have

$$\begin{aligned} C + D &= y_z + x_z - y_x + z_x + z_y - x_y \\ &= x_{110} + x_{001} + x_{110} + x_{001} \geq 0. \end{aligned}$$

Case III: $z \leq y \leq x$:

$$\alpha y_x + \overline{\alpha} x_y \leq \alpha z_x + \overline{\alpha} x_z + \alpha z_y + \overline{\alpha} y_z$$

Using the same analysis as in Case II, we need

$$0 \leq -y_x - x_y + z_x + x_z + z_y + y_z$$

which is the same equation as in Case II. □

Theorem 18. *The function $D'_{\alpha,\beta}$ is a metric only if $\beta \geq 1/(1-\alpha)$.*

Proof. Consider the same example as in Theorem 17. Ignoring the upfront factor of β, we have

$$D' = \frac{\delta}{|X \cap Y| + \beta\delta}.$$

In our example,

$$\begin{aligned} D'(X,Y) &= \frac{1}{0 + \beta \cdot 1} = \frac{1}{\beta}, \\ D'(X,Z) = D'(Y,Z) &= \frac{\overline{\alpha}}{1 + \beta \cdot \overline{\alpha}} = \frac{\overline{\alpha}}{1 + \beta\overline{\alpha}}. \end{aligned}$$

The triangle inequality is then equivalent to:

$$\frac{1}{\beta} \leq 2\frac{\overline{\alpha}}{1+\beta\overline{\alpha}} \iff \beta \geq \frac{1+\beta\overline{\alpha}}{2\overline{\alpha}} \iff \beta \geq 1/(1-\alpha).$$ □

Theorem 19. *The function $D'_{\alpha,\beta}$ is a metric on all finite power sets only if $\alpha \leq 1/2$.*

Proof. Suppose $\alpha > 1/2$. Let $Z_n = \{-(n-1), -(n-2), \ldots, 0\}$, a set of cardinality n disjoint from $\{1, 2\}$, and let $Y_n = Z_n \cup \{1\}$, $X_n = Z_n \cup \{2\}$. The triangle inequality says

$$\begin{aligned} \beta\frac{1}{n + \beta \cdot 1} = D'(X_n, Y_n) &\leq D'(X_n, Z_n) + D'(Z_n, Y_n) = 2\beta\frac{\overline{\alpha}}{n + \beta\overline{\alpha}} \\ n + \beta\overline{\alpha} &\leq 2\overline{\alpha}(n + \beta) \\ n(1 - 2\overline{\alpha}) &\leq \beta\overline{\alpha} \end{aligned}$$

Since $\alpha > 1/2$, we have $2\overline{\alpha} < 1$. Let $n > \frac{\beta\overline{\alpha}}{1-2\overline{\alpha}}$. Then the triangle inequality does not hold, so $D'_{\alpha,\beta}$ is not a metric on the power set of $\{-(n-1), -(n-2), \ldots, 0, 1, 2\}$. □

Proof of Theorem 12. We saw in Theorem 17 that δ is a metric for $0 \le \gamma \le 1$. (Recall that $\beta = 1/(1-\alpha)$, so that $\gamma = \alpha/\overline{\alpha}$.) In general if d is a metric and a is a function, we may hope that $d/(a+d)$ is a metric. We shall use the observation, mentioned by [Sra], that in order to show

$$\frac{d_{xy}}{a_{xy}+d_{xy}} \le \frac{d_{xz}}{a_{xz}+d_{xz}} + \frac{d_{yz}}{a_{yz}+d_{yz}},$$

it suffices to show the following pair of inequalities:

$$\frac{d_{xy}}{a_{xy}+d_{xy}} \le \frac{d_{xz}+d_{yz}}{a_{xy}+d_{xz}+d_{yz}} \tag{3}$$

$$\frac{d_{xz}+d_{yz}}{a_{xy}+d_{xz}+d_{yz}} \le \frac{d_{xz}}{a_{xz}+d_{xz}} + \frac{d_{yz}}{a_{yz}+d_{yz}} \tag{4}$$

Here (3) follows from d being a metric, i.e., $d_{xy} \le d_{xz} + d_{yz}$, since $c \ge 0 < a \le b \implies \frac{a}{a+c} \le \frac{b}{b+c}$.

Next, (4) would follow from $a_{xy} + d_{yz} \ge a_{xz}$ and $a_{xy} + d_{xz} \ge a_{yz}$. By symmetry between x and y and since $a_{xy} = a_{yx}$ in our case, it suffices to prove the first of these, $a_{xy} + d_{yz} \ge a_{xz}$. This is equivalent to

$$x{\cap}y + \gamma \min\{z_y, y_z) + \max\{z_y, y_z\} \ge x{\cap}z,$$

which holds for all $0 \le \gamma \le 1$ if and only if

$$x{\cap}y + \max\{z_y, y_z\} \ge x{\cap}z.$$

There are now two cases.
Case $z \ge y$: We have

$$x{\cap}y + z_y \ge x{\cap}z$$

since any element of $X \cap Z$ is either in Y or not.
Case $y \ge z$:

$$\begin{aligned} x{\cap}y + y_z &\ge x{\cap}z \\ x_{110} + x_{111} + x_{110} + x_{010} &\ge x_{101} + x_{111} \\ x_{110} + x_{110} + x_{010} &\ge x_{101} \end{aligned}$$

This is true since $z_y \ge x{\cap}z_y$. □

3 Application to phylogeny

The mutations of spike glycoproteins of coronaviruses are of great concern with the new SARS-CoV-2 virus causing the disease CoViD-19. We calculate several distance measures between peptide sequences for such proteins. The distance

$$Z_{2,\alpha}(x_0, x_1) = \alpha \min(|A_1|, |A_2|) + \overline{\alpha} \max(|A_1|, |A_2|)$$

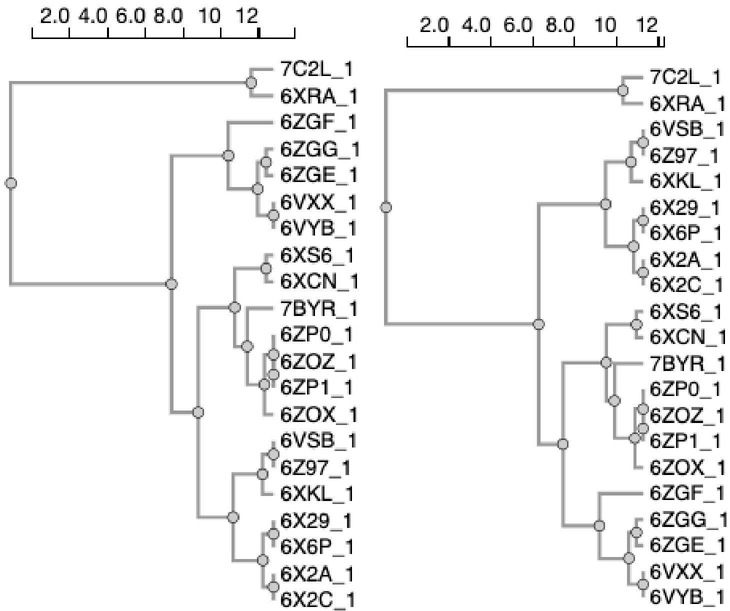

Figure 2: $\alpha = 0.21$ and 0.36.

where A_i is the set of subwords of length 2 in x_i but not in x_{1-i}, counts how many subwords of length 2 appear in one sequence and not the other. Our calculations are available in a URL format as follows:

`http://counter-automata.appspot.com/spike?metric=z2&alpha=0.36`

We used the Ward linkage criterion for producing Newick trees using the `hclust` package for the Go programming language. The calculated phylogenetic trees were based on the metric $Z_{2,\alpha}$.

We found one tree isomorphism class each for $0 \leq \alpha \leq 0.21$, $0.22 \leq \alpha \leq 0.36$, and $0.37 \leq \alpha \leq 0.5$, respectively (Figure 2, Figure 3). In Figure 3 we are also including the tree produced using the Levenshtein edit distance in place of $Z_{2,\alpha}$. We see that the various intervals for α can correspond to "better" or "worse" agreement with other distance measures. Thus, we propose that rather than focusing on $\alpha = 0$ and $\alpha = 1/2$ exclusively, future work may consider the whole interval $[0, 1/2]$.

4 Conclusion

Many researchers have considered metrics based on sums or maxima, but we have shown that these need not be considered in "isolation" in the sense that they form the endpoints of a family of metrics.

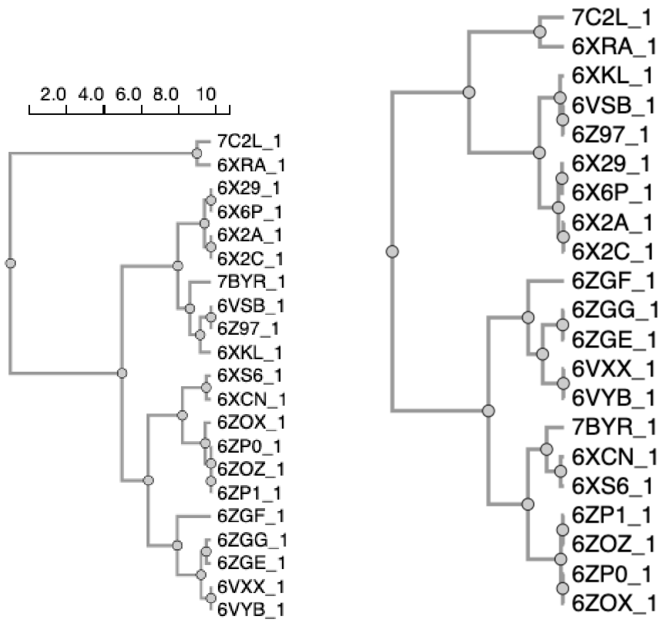

Figure 3: $\alpha = 0.5$ and edit distance.

More general set-theoretic metrics can be envisioned. The Steinhaus transform of δ with $\beta = 1/\overline{\alpha}$ is:

$$\begin{aligned} \delta'(X,Y) &= \frac{2\delta(X,Y)}{\delta(X,\emptyset)+\delta(Y,\emptyset)+\delta(X,Y)} \\ &= 2\frac{\gamma\min\{x_y,y_x)+\max\{x_y,y_x\}}{(x+y)+\gamma\min\{y_x,x_y\}+\max\{y_x,x_y\}} \end{aligned}$$

A question for future research is whether this Steinhaus transform is more or less useful than what Jiménez et al. [JBG13] considered. We can consider a general setting for potential metrics that contains both the Steinhaus transform of δ and the STRM metrics. In terms of $m(X,Y) = \min\{x_y, y_x\}$ and $M(X,Y) = \max\{x_y, y_x\}$, we can consider $\Delta_{\gamma,s} := \frac{\gamma m + M}{x\cap y + s(x\cup y) + (\gamma m + M)}$. When $s = 0$ this is our STRM metric. When $s = 1$ it is the Steinhaus transform, ignoring constant upfront factors.

References

[CV05] R. Cilibrasi and P. M. B. Vitanyi. Clustering by compression. *IEEE Transactions on Information Theory*, 51(4):1523–1545, 2005.

[CV07] Rudi L. Cilibrasi and Paul M. B. Vitanyi. The Google similarity distance. *IEEE Trans. on Knowl. and Data Eng.*, 19(3):370–383, March 2007.

[DD16] Michel Marie Deza and Elena Deza. *Encyclopedia of distances.* Springer, Berlin, fourth edition, 2016.

[Ged10] Ged Ridgway. Mutual information — Wikipedia, the Free Encyclopedia, Revision as of 14:55, 22 january 2010, 2010. [Online; accessed 14-May-2020].

[GS16] Alonso Gragera and Vorapong Suppakitpaisarn. Semimetric properties of Sørensen-Dice and Tversky indexes. In *WALCOM: algorithms and computation*, volume 9627 of *Lecture Notes in Comput. Sci.*, pages 339–350. Springer, [Cham], 2016.

[GS18] Alonso Gragera and Vorapong Suppakitpaisarn. Relaxed triangle inequality ratio of the Sørensen-Dice and Tversky indexes. *Theoret. Comput. Sci.*, 718:37–45, 2018.

[JBG13] Sergio Jiménez, Claudia Jeanneth Becerra, and Alexander F. Gelbukh. SOFTCARDINALITY-CORE: improving text overlap with distributional measures for semantic textual similarity. In Mona T. Diab, Timothy Baldwin, and Marco Baroni, editors, *Proceedings of the Second Joint Conference on Lexical and Computational Semantics, *SEM 2013, June 13-14, 2013, Atlanta, Georgia, USA*, pages 194–201. Association for Computational Linguistics, 2013.

[KSAG03] Alexander Kraskov, Harald Stögbauer, Ralph G. Andrzejak, and Peter Grassberger. Hierarchical clustering based on mutual information. *ArXiv*, q-bio.QM/0311039, 2003.

[KSAG05] A Kraskov, H Stögbauer, R. G Andrzejak, and P Grassberger. Hierarchical clustering using mutual information. *Europhysics Letters (EPL)*, 70(2):278–284, apr 2005.

[LBC+01] Ming Li, Jonathan H. Badger, Xin Chen, Sam Kwong, Paul E. Kearney, and Haoyong Zhang. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17 2:149–54, 2001.

[LCL+04] Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul M. B. Vitányi. The similarity metric. *IEEE Trans. Inform. Theory*, 50(12):3250–3264, 2004.

[LZ76] Abraham Lempel and Jacob Ziv. On the complexity of finite sequences. *IEEE Trans. Inform. Theory*, IT-22(1):75–81, 1976.

[Raj61] C. Rajski. Entropy and metric spaces. In *Information theory (Symposium, London, 1960)*, pages 41–45. Butterworths, Washington, D.C., 1961.

[RN17] Edward Raff and Charles K. Nicholas. An Alternative to NCD for Large Sequences, Lempel–Ziv Jaccard Distance. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

[Sra] Suvrit Sra. Is the Jaccard distance a distance? MathOverflow. URL:https://mathoverflow.net/q/210750 (version: 2015-07-03).

[Tve77] A. Tversky. Features of similarity. *Psychological Review*, 84(4):327––352, 1977.

[ZL77] Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Inform. Theory*, IT-23(3):337–343, 1977.

[ZL78] Jacob Ziv and Abraham Lempel. Compression of individual sequences via variable-rate coding. *IEEE Trans. Inform. Theory*, 24(5):530–536, 1978.