Similarities and Distances in Biology

Elena and Michel DEZA

September 23, 2007

This conference is entitled ”Similarity and Dissimilarity in Computational Biology” in allusion to **XU Guangqi (1562-1633)** treatise on ”Similarity and Dissimilarity in Measurement” and to mark the 400th anniversary of the publication of his translation of Euclid's ”Elements of Geometry”.

The abstraction of measurment, in terms of mathematical notions distance, similarity, metric, etc. was originated by M.Fréchet (1906) and F.Hausdorff (1914). But the triangle inequality, crucial in it, should be already known to XU Guangqi since it appears in Euclid's ”Elements of Geometry”.

Given a set X, a **distance** (or **dissimilarity**) on it is a function $d: X \times X \to \mathbb{R}_{\geq 0}$ with all $d(x,x) = 0$ and $d(x,y){=}d(y,x)$ (**symmetry**).

A **similarity** is a symmetric function $s: X \times X \to \mathbb{R}_{\geq 0}$ such that $s(x,y) \leq s(x,x)$ holds for all $x, y \in X$ with equality if and only if $x = y$.

A **metric** is a symmetric function $d: X \times X \to \mathbb{R}_{\geq 0}$ with $d(x,y) = 0$ iff $x = y$ and **triangle inequality** $(d(x,y) \leq d(x,z) + d(z,y)$ if $x, y, z \in X)$.

Metric space (a set X with a metric d defined on it: (X, d)) started century ago as a special case of an infinite topological space.

However, from K.Menger (1928) and L.M.Blumenthal (1953), an explosion of interest in both, finite and infinite metric spaces, occurred.

By now, theories involving distances and similarities florished in many areas of Mathematics including Geometry, Probability, Coding/Graph Theory. Many mathematical theories, in the process of their generalization, settled down on the level of metric space. It is ongoing process in Riemannian Geometry, Real Analysis, Approximation Theory.

On the other hand, devising most suitable distances/similarities became an essential task in many applications incl. Pattern Recognition, Networks, Astronomy/Cosmology, and esp. Computational Biology, Image/Audio Analisys, Information Retrieval. But Biology still lags behind last two in using, besides distances themselves, powerful distance-related notions and paradigms: transforms, various numerical invariants, etc.

CONTENTS

BIRDVIEW OF METRIC SPACES

1. Metric repairs
2. Variations of triangle inequality
3. Transform metrics
4. Numeric invariants of metric spaces
5. Relevant notions: special subsets, mappings, convexity
6. Main classes of metric spaces

METRIC REPAIRS

Let X be a set. A function $d : X \times X \rightarrow \mathbb{R}_{\geq 0}$ with all $d(x,x) = 0$ is called a **quasi-distance** (or, in Topology, **prametric**) on X.

A quasi-distance is a **distance** if $d(x,y) = d(y,x)$ and

semi-metric (or, in Topology, **pseudo-metric**) if, moreover, it holds $d(x,y) \leq d(x,z) + d(z,y)$ (**triangle inequality**) for all $x, y, z \in X$.

A quasi-distance is a **quasi-metric** if $d(x,y) > 0$ for all $x \neq y$ and triangle inequality holds.

A **metric** is both, semi- and quasi-metric.

Main transforms used to obtain a **distance** $d \leq 1$ **from a similarity** s are: $d = \arccos s$, $d - \ln s$, $d = 1 - s$, $d = \frac{1-s}{s}$, $d = \sqrt{1-s}$, For a **distance** d, the function, defined by $D(x,x)=0$ and, for $x \neq y$ by $D_1(x,y) = d(x,y) + \max_{x,y,z \in X}(d(x,y) - d(x,z) - d(y,z))$ is a semi-metric. Also, $D_2(x,y) = d(x,y)^c$ is a semi-metric for sufficiently small $c \geq 0$. The function $D_3(x,y) = \inf \sum_i d(z_i, z_{i+1})$, where the infimum is taken over all sequences $x = z_0, \dots, z_{n+1} = y$, is also a semi-metric.

For a **semi-metric** d on X, define equivalence relation by $x \sim y$ if $d(x,y)=0$; let $[x]$ be the equivalence class containing x. Then $D([x],[y]) = d(x,y)$ is a metric on the set $\{[x] : x \in X\}$ of eqv. classes.

For a **quasi-metric** d, functions $\max\{d(x,y), d(y,x)\}$, $\min\{d(x,y), d(y,x)\}$ and $\frac{d(x,y)+d(y,x)}{2}$ are metrics.

For a **metric** d, the function $D(x,y) = \frac{d(x,y)}{1+d(x,y)} < 1$, is a 1-bounded metric.

1. $d(x,y) \le d(x,z) + d(z,y)$ (**triangle inequality**), i.e., a **metric**;

2. $d(x,y)d(u,z) \le d(x,u)d(y,z) + d(x,z)d(y,u)$, a **Ptolemaic metric**;

3. $d(x,y) + d(z,u) \le \max(d(x,z) + d(y,u), d(x,u) + d(y,z))$ (**4-point inequality**), a $\mathbb{R}_{>0}$**-edge-weighted tree metric** (it is 2, 5, 7);

4. $d(x,y) \le \max(d(x,z), d(z,y))$, an **ultrametric** (it is 3);

5. $d(x,y) + d(z,u) \le 2\delta + \max\{d(x,z) + d(y,u), d(x,u) + d(y,z)\}$ **for** $\delta \ge 0$, a δ**-hyperbolic metric**;

6. $d(x,y) \le d(x,z) + d(y,z) - d(x,z)d(z,y)$ (**equivalent to** $(1 - d(x,y)) \ge (1 - d(x,z))(1 - d(z,y)))$, a P**-metric**;

7. $\sum_{1 \le i < j \le n} b_i b_j d(x_i, x_j) \le 0$ **for** $b \in \mathbb{Z}^n$**,** $\sum_{i=1}^{n} b_i = 1$, **a hypermetric**;

8. $d(x,y) \le C(d(x,z) + d(z,y))$ **for a constant** $C \ge 1$, a **near-metric**;

9. $d(x,y) \le d(x,z) + d(z,y) - d(z,z)$ **for** $0 \le d(z,z) \le \inf_u d(z,u)$**, i.e., self-distances are small**, a **partial metric**.

A **2-metric** is function $d : X \times X \times X \to \mathbb{R}_{\geq 0}$ which is **totally symmetric** (i.e., $d(x_1, x_2, x_3)$ is unchanged by any permutation of arguments),

zero conditioned (i.e., $d(x_1, x_2, x_3) = 0$ iff $x_i = x_j$ for some $1 \leq i < j \leq 3$) and satisfy **tetrahedron inequality**

$$d(x_1, x_2, x_3) \leq d(x_4, x_2, x_3) + d(x_1, x_4, x_3) + d(x_1, x_2, x_4).$$

A m**-metric** (or m**-volume**) is defined by m**-simplex inequality**. The cases $m = 1, 2$ correspond to usual metric (length) and area, respectively.

A **proximity space** is a set X with a **proximity**, i.e., symmetric binary relation δ on the **power set** $P(X)$ (of all its subsets) with $A \delta A$ iff $A \neq \emptyset$

and $A\delta(B \cup C)$ if and only if $A\delta B$ or $A\delta C$ (**additivity**).

Every metric space (X, d) is a proximity space: define $A\delta B$ iff $d(A, B) = \inf_{x \in A, y \in B} d(x, y) = 0$.

Consider a set X and a map $cl : P(X) \to P(X)$ with $cl(\emptyset) = \emptyset$. The maps $cl(A)$ (for $A \subset X$), its dual $int(A) = X \backslash cl(X \backslash A)$ and $N : X \to P(X)$ with $N(x) = \{A \subset X : x \in int(A)\}$ are called **closure**, **interior** and **neighborhood** map, resp. A subset $A \subset X$ is **closed** if $A = cl(A)$ and **open** if $A = int(A)$. Consider the following possible properties of (X, cl):

1. $A \subseteq B$ implies $cl(A) \subseteq cl(B)$ (**isotony**);
2. $A \subseteq cl(A)$(**enlarging**);
3. $cl(A \cup B) = cl(A) \cup cl(B)$ (**linearity**, and, in fact, 3. implies 1.);
4. $cl(cl(A)) = cl(A)$ (**idempotency**).

The pair (X, cl) is called **extended topology** if 1. hold, **Brissaud space** (Brissaud, 1974) if 2. hold, **neighborhood space** (Hammer, 1964) if 1., 2. hold, **Smyth space** (Smyth, 1995) if 3. hold, **pretopology** (Čech, 1966) if 2., 3. hold, and **closure space** (Soltan, 1984) if 1., 2, 4. hold.

(X, cl) is usual topology, in closure terms, if 2., 3., 4. hold.

The **pseudo-Euclidean distance** of signature $(p, q = n - p)$ on $\mathbb{R}^n$ is

$$d_{pE}(x,y) = \sum_{i=1}^{p}(x_i - y_i)^2 - \sum_{i=p+1}^{n}(x_i - y_i)^2.$$

The pseudo-Euclidean space of signature $(p, q = n - p)$ is a real vector space equipped with a non-degenerate, indefinite, symmetric bilinear function $\langle \cdot, \cdot \rangle$. A basis $e_1, \dots, e_{p+q}$ is orthonormal if $\langle e_i, e_j \rangle = 0$ for $i \neq j$, $\langle e_i, e_i \rangle = +1$ for $1 \leq i \leq p$ and $\langle e_i, e_i \rangle = -1$ for $p + 1 \leq i \leq p + q$. Given an orthonormal basis, the inner product of two vectors x and y is $\langle x, y \rangle = \sum_{i=1}^{p} x_i y_i - \sum_{i=p+1}^{p+q} x_i y_i$.

The pseudo-Euclidean space can be seen as $\mathbb{R}^p \times i\mathbb{R}^q$, where $i = \sqrt{-1}$.

The ”norm” $\langle x, x \rangle$ of non-zero vector x can be positive, negative or zero; then x is called **space**, **time** or **light** vector, respectively.

The case $(p, q) = (1, 3)$ is used as space-time model of Special Relativity.

METRIC TRANSFORMS

A **transform metric** is a metric on a set X which is obtained as a function of a given metric (or metrics) on X. Examples obtained from a given metric d (or metrics d_1 and d_2) on X follow (here $t > 0$):

1. $td(x, y)$ (t-**scaled metric**, or **dilated metric**);
2. $\min\{t, d(x, y)\}$ (t-**truncated metric**, or **t-bounded metric**);
3. $\max\{t, d(x, y)\}$ for $x \neq y$ (t-**discrete metric**);
4. $d(x, y) + t$ for $x \neq y$ (t-**translated metric**);
5. $\frac{d(x,y)}{1+d(x,y)}$;
6. $\max\{d_1(x, y), d_2(x, y)\}$;
7. $\alpha d_1(x, y) + \beta d_2(x, y)$, where $\alpha, \beta > 0$ (so, **semi-metric cone** on X);
8. $d^z(x, y) = \frac{d(x,y)}{d(x,z)+d(y,a)+d(x,y)}$ where z is an fixed element of X (**biotope transform metric**).

Given a metric space (X, d) and a point $z \in X$, the **Farris transform metric** on $X \backslash \{z\}$ is defined by $D_z(x, x) = 0$ and, for $x \neq y$, by

$$D_z(x, y) = C - (x.y)_z,$$

where $C > 0$ is a constant and $(x.y)_z = \frac{1}{2}(d(x, z) + d(y, z) - d(x, y))$ is the **Gromov product**. It is a **metric** if and only if $C \geq C_0$ for some number $C_0 \in (\max_{x,y \in X \backslash \{z\}, x \neq y}(x.y)_z, \max_{x \in X \backslash \{z\}} d(x, z)]$.

Farris transform is an **ultrametric** if and only if d is a $\mathbb{R}_{>0}$**-edge-weighted tree metric**.

In Phylogenetics, where it was applied first, the term *Farris transform* is used for function $d(x, y) - d(x, z) - d(y, z)$.

- Given a metric space (X, d) and $0 < \alpha \leq 1$, the **power transform metric** (or **snowflake transform metric**) on X is $(d(x, y))^{\alpha}$.

  It is a **metric**, for any positive α if and only if d is an **ultrametric**.

- Given a metric space (X, d) and a point $z \in X$, the **involution transform metric** on $X \backslash \{z\}$ is

  $$d_z(x, y) = \frac{d(x, y)}{d(x, z)d(y, z)}.$$

  It is a **metric**, for any $z \in X$, if and only if d is a **Ptolemaic metric**.

- Given a metric space (X, d) and $\lambda > 0$, the **Schoenberg transform metric** on X is

  $$1 - e^{-\lambda d(x,y)}.$$

  The Schoenberg transform metrics are exactly ***P*-metrics**.

- An **induced metric** is a restriction of a metric (X, d) to $X^{'} \subset X$.

- Given metric spaces (X, d_X), (Y, d_Y) and injective mapping $g : X \to Y$, the **pullback metric** (of (Y, d_y) by g) on X is $d_Y(g(x), g(y))$.

- Given a metric space (X, d) and an equivalence relation $\sim$ on X, the **quotient semi-metric** on the set $\overline{X} = X/\sim$ of equivalence classes is $\overline{d}(\overline{x}, \overline{y})$=$\inf_{m\in\mathbb{N}} \sum_{i=1}^{m} d(x_i, y_i)$, where the infimum is over all sequences $x_1, y_1, \dots, x_m, y_m$ with $x_1 \in \overline{x}$, $y_m \in \overline{y}$ and $y_i \sim x_{i+1}$ if $1 \leq i =\leq m-1$

- Given $n \leq \infty$ metric spaces (X_1, d_1), (X_2, d_2), ..., (X_n, d_n), the **product metric** is any metric on their **Cartesian product** $X_1 \times X_2 \times \cdots \times X_n = \{x = (x_1, x_2, \dots, x_n) : x_1 \in X_1, \dots, x_n \in X_n\}$, defined as a function of $d_1, \dots, d_n$.

- Given a metric space (X, d) with any points $x, y \in X$ joined by a **rectifiable curve** (i.e., of finite length), the **intrinsic metric** $D(x, y)$ is the infimum of the lengths of rectifiable curves connecting x and y.

  (A (metric) **curve** γ is a continuous mapping $\gamma : I \to X$ from an interval I of $\mathbb{R}$ into X. The **length** $l(\gamma)$ of a curve $\gamma : [a, b] \to X$ is

  $$l(\gamma) = \sup\{ \sum_{1 \leq i \leq n} d(\gamma(t_i), \gamma(t_{i-1})) : n \in \mathbb{N}, a = t_0 < t_1 < \cdots < t_n = b\}).$$

- The **Riemannian metric** of a connected n-dim. smooth **manifold** M^n, is a collection of positive-definite symmetric bilinear forms $((g_{ij}))$ on the tangent spaces of M^n which varies smoothly from point to point.

  The length of a curve γ on M^n is $\int_\gamma \sqrt{\sum_{i,j} g_{ij} dx_i dx_j}$.

  The **Riemannian distance** (between two points of M^n) is intrinsic metric on M^n, i.e. the infimum of lengths of curves, connecting them.

NUMERIAL INVARIANTS OF METRIC SPACES

- For a metric space (X, d) and any $q > 0$, let $N_X(q)$ be the minimal number of sets with diameter $\leq q$ needed in order to cover X. The number $\mathbf{dim_{metr}} = \lim_{q \to 0} \frac{\ln N(q)}{\ln(1/q)}$ (if it exists) is called its **metric dimension** (or **packing dimension**, **box-counting dimension**).

- For any compact metric space (X, d), its **topological dimension** is $\mathbf{dim_{top}(X, d)}$=$\inf_{d'}(\dim_{Haus}(X, d'))$, where d' is any metric on X topologically equivalent to d and $\mathbf{dim_{Haus}}$ is **Hausdorff dimension**.

  Two metrics d_1, d_2 on a set X are **equivalent** if they define same **topology** on X (for any $x_0 \in X$, any open d_1- metric ball centered at x_0 contains an open d_2-metric ball centered at x_0 and conversely).

- For any $p, q > 0$, let $M_p^q(X) = \inf \sum_{i=1}^{+\infty} (diam A_i)^p$, where infimum is taken over all countable coverings $\{A_i\}$ of X with diameter of $A_i < q$.

  The **Hausdorff dimension** (or **fractal dimension**, **capacity dimension**) of X is $\mathbf{dim_{Haus}} = \inf\{p : \lim_{q \to 0} M_p^q(X) = 0\}$.

  It holds $dim_{top} \leq dim_{Haus} \leq dim_{metr}$. A **fractal** is a metric space for which $dim_{top} < dim_{Haus}$.

- The **Assouad-Nagata dimension** $\mathbf{dim_{AN}}$ of a metric space (X, d) is the smallest integer n for which there exist a constant $C > 0$ such that, for all $s > 0$, there exists a covering of X by its subsets of diameter at most Cs with no point of X belonging to more than $n + 1$ elements.

  d called a **doubling metric** if $dim_{AN} < \infty$. It holds $dim_{top} \leq dim_{AN}$.

- The **metric diameter** (or **diameter**, **width**) is $\sup_{x,y \in X} d(x,y)$.

  If (X,d) is A**-bounded** ($A = \sup_{x,y \in X} A < \infty$) and a**-discrete** ($a = \inf_{x,y \in X, x \neq y} d(x,y) > 0$), then its **metric spread** is $\frac{A}{a}$.

- The **metric radius** of metric space (X,d) is $\inf_{x \in X} \sup_{y \in X} d(x,y)$.

  Some authors call *radius* the half-diameter.

  The **packing radius** of $M \subset X$ is the largest r such that the open metric balls of radius r with centers at the elements of M are pairwise disjoint, i.e., $\inf_{x \in M} \inf_{y \in M \setminus \{x\}} d(x,y) > 2r$.

  The **covering radius** of $M \subset X$ is $\sup_{x \in X} \inf_{y \in M} d(x,y)$, i.e., the smallest number R such that the open metric balls of radius R with centers at the elements of M cover X. It is $d_{Haus}(X,M)$.

- A metric space (X,d) has the **order of congruence** n if every finite metric space which is not **isometrically embeddable** in (X,d) has a subspace with $\leq n$ points which is not isometrically embeddable in it.

- Given a compact connected metric space (X, d), there exists a unique number $r(X, d) > 0$, **rendez-vous number** (or **magic number**) such that for all $x_1, \dots, x_n \in X$ and any n, there exists an $x \in X$ with $\frac{1}{n}\sum_{i=1}^{n} d(x_i, x) = r(X, d)$.

- Given a set $D \subset \mathbb{R}_{>0}$, the ***D*-chromatic number** of (X, d) is the standard **chromatic number** of the ***D*-distance graph** of (X, d), i.e., the graph with the vertex-set X and the edge-set $\{xy : d(x, y) \in D\}$. Usually, (X, d) is an l_p**-space** and $D = \{1\}$ or $D = [1 - \epsilon, 1 + \epsilon]$.

- The **average distance** is the number $\frac{1}{|X|(|X|-1)}\sum_{x,y \in X} d(x, y)$.

  The **Wiener index** (used in Chemistry) is $\frac{1}{2}\sum_{x,y \in X} d(x, y)$.

- The ***p*-energy** is the number $\sum_{x,y \in X, x \neq y} \frac{1}{d^p(x,y)}$; usually, $p = 1, 2$.

  A **center of mass** is a point $x \in X$ minimizing $\sum_{y \in X} d^2(x, y)$.

RELEVANT NOTIONS: SUBSETS, MAPPINGS, CONVEXITY

- Given distinct points $x, y \in X$, the **midset** (or **bisector**) is the set $\{z \in X : d(x,z) = d(y,z)\}$ of **midpoints** z.

- $M \subset X$ is a **metric basis** of X if $d(x,z) = d(y,z)$ for all $z \in M$ implies $x = y$. The numbers $d(x,z), z \in M$, are the **metric coordinates** of x.

- Given a finite or countable semi-metric space $(X = \{x_1, \cdots, x_n\}, d)$, its **distance matrix** is the symmetric $n \times n$ matrix $((d_{ij}))$, where $d_{ij} = d(x_i, x_j)$ for any $1 \leq i, j \leq n$.

  The **semi-metric cone** is the set of all distance matrices on X.

- The **proximity** (or **underlying**) **graph of** metric space (X, d) is a graph with the vertex-set X and xy being an edge if no point $z \in X$ with $d(x,y) = d(x,z) + d(z,y)$ exists.

- The **point-set distance** $d(x, M)$ between $x \in X$ and $M \subset X$ is $\inf_{y \in M} d(x, y)$. The function $f_M(x) = d(x, M)$ is **distance map**. Distance maps are used in MRI (M being gray/white matter interface) as cortical maps, in Image Processing (M being image boundary), in Robot Motion (M being the set of obstacle points).

- A subset $M \subset X$ is **Chebyshev set** (or **gated set**) if for every $x \in X$, there is **unique** $z \in M$ with $d(x, z) = d(x, M)$.

- The **set-set distance** between two subsets $A, B \subset X$ is $\inf_{x \in A,} d(x, B)$=$\inf_{x \in A, y \in B} d(x, y)$. In Cluster Analysis, it is **single linkage**, while $\sup_{x \in A, y \in B} d(x, y)$ is **complete linkage**.

- The **Hausdorff metric** (on all compact subspaces of (X, d)) is $d_{Haus}(A, B)$=$\max\{d_{dHaus}(A, B), d_{dHaus}(B, A)\}$, where $d_{dHaus}(A, B)$=$\max_{x \in A} \min_{y \in B} d(x, y)$ is **directed Hausdorff distance**.

MAPPINGS FOR METRIC SPACES

- Given metric spaces (X, d_X) and (Y, d_Y), a function $f : X \to Y$ is an **isometric embedding** of X into Y if it is injective and $d_Y(f(x), f(y)) = d_X(x, y)$ holds for all $x, y \in X$.

  An **isometry** is a bijective isometric embedding.

- Two metric spaces (X, d_X) and (Y, d_Y) are **homeomorphic** if there exists a bijection $f : X \to Y$ with **continuous** f and f^{-1} , i.e., all points close to x map to points close to $g(x)$.

- Given metric spaces (X, d_X) and (Y, d_Y), a function $f : X \to Y$ is called a **short mapping** from X to Y if, for all $x, y \in X$, holds $d_Y(f(x), f(y)) \leq d_X(x, y)$. The **category of metric spaces** (Isbell), denoted by Met, has metric spaces as objects and **short mappings** as morphisms. In Met, the **isomorphisms** are **isometries**.

- Again, given metric spaces (X, d_X) and (Y, d_Y), a function $f : X \to Y$ is an **isometric embedding** of X into Y if it is injective and $d_Y(f(x), f(y)) = d_X(x, y)$ holds for all $x, y \in X$.

  An **isometry** is a bijective isometric embedding.

- A function $f : X \to Y$ is a **quasi-isometry** if there are numbers $C > 1$ and $c > 0$ such that $C^{-1}d_X(x, y) - c \leq d_Y(f(x), f(y)) \leq Cd(x, y) + c$, and for every point $y \in Y$ there is a point $x \in X$ with $d_Y(y, f(x)) \leq c$.

  A quasi-isometry with $C = 1$ is **coarse** (or **rough**) **isometry**.

- A metric space (X, d) is **homogeneous** if, for each two finite isometric subsets $Y = \{y_1, \dots, y_m\}$ and $Z = \{z_1, \dots, z_m\}$ of X, there exists a self-isometry (motion) of (X, d) mapping Y to Z.

- (X, d) is **symmetric** if for any $p \in X$ there is a **symmetry relative to** p, i.e., a **motion** (self-isometry) f_p of (X, d) such that $f_p(f_p(x)) = x$ for all $x \in X$ and p is an isolated fixed point of f_p.

CONVEXITY NOTIONS FOR METRIC SPACES

Given a metric space (X, d), a **metric curve** (or **curve**) γ in it is a continuous mapping $\gamma : I \to X$ from an interval I of $\mathbb{R}$ into X.

The **length** $l(\gamma)$ of a curve $\gamma : [a, b] \to X$ is defined by

$$l(\gamma) = \sup\{ \sum_{1 \le i \le n} d(\gamma(t_i), \gamma(t_{i-1})) : n \in \mathbb{N}, a = t_0 < t_1 < \cdots < t_n = b\}.$$

A **geodesic segment** (or **shortest path**) $[x, y]$ from x to y is

(the image of) an isometric embedding $\gamma : [a, b] \to X$ with $\gamma(a) = x$ and $\gamma(b) = y$.

- A metric space (X, d) is called **geodesic metric space** (or **convex**) if any two points are joined by a geodesic segment.

- (X, d) is **midpoint convex** (or **admitting midpoint map**) if, for any different points $x, y \in X$, there exists a third point $z \in X$, a **midpoint** $m(x, y)$, for which $d(x, y) = d(x, z) + d(z, y)$ and $d(x, z) = \frac{1}{2}d(x, y)$.

- (X, d) is **Busemann convex** (or **globally non-positively Busemann curved**) if it is midpoint convex and, for any three points $x, y, z \in X$ and midpoints $m(x, z)$ and $m(y, z)$, it holds

$$d(m(x, z), m(y, z)) \leq \frac{1}{2}d(x, y).$$

- **ball convex** if it is midpoint convex and for all $x, y, z \in X$ it holds

$$d(m(x, y), z) \leq \max\{d(x, z), d(y, z)\}.$$

- **distance convex** if it is midpoint convex and for all $x, y, z \in X$ holds

$$d(m(x, y), z) \leq \frac{1}{2}(d(x, z) + d(y, z)).$$

- **Menger convex** (or **M-convex**) if, for any different points $x, y \in X$, there exists a third point $z \in X$ for which $d(x,y) = d(x,z) + d(z,y)$.

- (X,d) is **metrically convex** if, for any different points $x, y \in X$ and any $\lambda \in (0,1)$, there exists a third point $z = z(x,y,\lambda) \in X$ for which $d(x,y) = d(x,z) + d(z,y)$ and $d(x,z) = \lambda d(x,y)$.

  (X,d) is **strictly metrically convex** if the point $z(x,y,\lambda)$ is unique for all $x, y \in X$ and $\lambda \in (0,1)$.

- (X,d) is **hyperconvex** (or **injective**) if it is metrically convex and its metric balls have the **infinite Helly property**, i.e., any family of mutually intersecting closed balls in X has non-empty intersection.

MAIN CLASSES OF METRICS

- Given a connected graph $G = (V, E)$, the **path metric** between two vertices is the number of edges of a shortest path connecting them.

- Given a finite set X and a finite set $\mathcal{O}$ of (unary) **editing operations** on X, the **editing metric** on X is the path metric of the graph with the vertex-set X and xy being an edge if y can be obtained from x by one of the operations from $\mathcal{O}$.

- On a **normed vector space** $(V, ||.||)$, the **norm metric** is $||x - y||$.
- The **l_p-metric**, $1 \leq p \leq \infty$, is $||x - y||_p$ norm metric on $\mathbb{R}^n$ (or on $\mathbb{C}^n$), where $||x||_p = (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$ for $p < \infty$ and $||x||_\infty = \max_{1 \leq i \leq n} |x_i|$.

  The **Euclidean metric** (or **Pythagorean distance**, **as-crow-flies distance**, **beeline distance**) is l_2-metric on $\mathbb{R}^n$.
- **Banach-Mazur distance** between n-dim. **normed spaces** V and W is $\ln \inf_T \{||T|| \cdot ||T^{-1}||$, where $T : V \to W$ is an isomorphism.
- **Lipschitz distance** between metric spaces (X, d_X) and (Y, d_Y) is $\inf_f \{||f||_{Lip} \cdot ||f^{-1}||_{Lip}\}$, where infimum is over all bijective functions $f : X \to Y$ and the **Lipschitz norm** is $||f||_{Lip} = \sup\{\frac{d_Y(f(x), f(y))}{d_X(x,y)} : x, y \in X, x \neq y\}$.

- Given a **measure space** $(\Omega, \mathcal{A}, \mu)$, the **symmetric difference** (or **measure**) **semi-metric** on the set $\mathcal{A}_\mu = \{A \in \mathcal{A} : \mu(A) < \infty\}$ is $\mu(A \triangle B)$ (where $A \triangle B = (A \cup B) \backslash (A \cap B)$ is the **symmetric difference** of the sets $A, B \in \mathcal{A}_\mu$) and 0 if $\mu(A \triangle B) = 0$.

  Identifying $A, B \in \mathcal{A}_\mu$ if $\mu(A \triangle B) = 0$, gives the **measure metric**.

  If $\mu(A) = |A|$, then $|A \triangle B| = 0$ iff $A = B$ and $|A \triangle B|$ is a metric.

- Given a **measure space** $(\Omega, \mathcal{A}, \mu)$, the **Steinhaus semi-metric** on the set $\mathcal{A}_\mu = \{A \in \mathcal{A} : \mu(A) < \infty\}$ is 0 if $\mu(A \cup B) = 0$ and

$$\frac{\mu(A \triangle B)}{\mu(A \cup B)} = 1 - \frac{\mu(A \cap B)}{\mu(A \cup B)}, \text{ otherwise.}$$

  The **biotope** (or **Tanimoto**) **metric** $\frac{|A \triangle B|}{|A \cup B|}$ is the case $\mu(A) = |A|$.

The **Hamming metric** on $\mathbb{R}^n$ is $d_H = |\{i : 1 \leq i \leq n, x_i \neq y_i\}|$.

On vertices of unit cube $\{0,1\}^n$ it is l_1**-metric** and squared l_2**-metric**. Eqv., for subsets $A, B \subset X$ with $|X| = n$, it is **measure metric** $|A \triangle B|$.

The **Bray-Curtis distance** on $\mathbb{R}^n$ is $\frac{\sum |x_i - y_i|}{\sum (x_i + y_i)}$.

The **Canberra distance** on $\mathbb{R}^n$ is $\sum \frac{|x_i - y_i|}{|x_i| + |y_i|}$.

The **Mahalonobis distance** (or **statistical distance**) on $\mathbb{R}^n$ is

$$\sqrt{(det A)^{\frac{1}{n}} (x-y) A^{-1} (x-y)^T},$$

where A is a positive-definite matrix.

The **Hellinger distance** on $\mathbb{R}^n_+$ is $\sqrt{2 \sum \left(\sqrt{\frac{x_i}{\overline{x}}} - \sqrt{\frac{y_i}{\overline{y}}} \right)^2}$.

Metrics on real plane $\mathbb{R}^2$

- Given a **norm** $||.||$ on $\mathbb{R}^2$, the **French Metro metric** on $\mathbb{R}^2$ is $||x-y||$ if $x = cy$ for some $c \in \mathbb{R}$ and $||x|| + ||y||$, otherwise.

  For Euclidean norm, it is called **Paris metric** (or **hedgehog metric**)

- Given a **norm** $||.||$ on $\mathbb{R}^2$ (in general, on $\mathbb{R}^n$), the **British Rail metric** (or **Post Office metric**, **caterpillar metric**, **shuttle metric**) is $||x|| + ||y||$ for $x \neq y$ (and it is equal to 0, otherwise).

- Let d be a metric on $\mathbb{R}^2$ (in general, on any metric space) and let f be a fixed point (a **flower-shop**) in the plane.

  The **flower-shop metric** (or **SNCF metric**) on $\mathbb{R}^2$ is $d(x, f) + d(f, y)$ for $x \neq y$ (and it is equal to 0, otherwise). If $d(x, y) = ||x - y||$ and $f = (0, 0)$, it is the **British rail metric**.

- The **lift metric** (or **raspberry picker metric** or **metric ”river”**) on $\mathbb{R}^2$ is $|x_1 - y_1|$ if $x_2 = y_2$ and $|x_1| + |x_2 - y_2| + |y_1|$ if $x_2 \neq y_2$.

- The **Central Park metric** on $\mathbb{R}^2$ is the length of a shortest l_1-path (**Manhattan path**) between two points $x, y \in \mathbb{R}^2$ at the presence of a given set of areas which are traversed by a shortest Euclidean path (for example, Central Park in Manhattan).

- Let $\mathcal{O} = \{O_1, \dots, O_m\}$ be a collection of pairwise disjoint polygons on the Euclidean plane representing a set of obstacles which are neither transparent, nor traversable. The **collision avoidance distance** (or **piano movers distance**) is a metric on $\mathbb{R}^2 \backslash \{\mathcal{O}\}$, defined as the length of the shortest path among all possible continuous paths, connecting x and y, that do not intersect obstacles $O_i \backslash \partial O_i$,

Metrics on digital plane $\mathbb{Z}^2$

A **computer image** is a subset of $\mathbb{Z}^n$ (**digital nD space**). Usually, n=2. The points of $\mathbb{Z}^2$ and $\mathbb{Z}^3$ are **pixels** and **voxels**, respectively.

A **digital metric** is any integer-valued metric on a digital nD space.

Main digital metrics are: the **l_1-, l_∞-metrics** and (rounded to nearest, upper or lower, integer) **l_2-metric**.

A list of **neighbors** of a pixel can be seen as a list of permitted **one-step moves** on $\mathbb{Z}^2$. Associate a positive weight to each type of such move. Many digital metrics are the minimum, over all admissible paths (sequences of permitted moves) of the sum of their weights.

- The **rook metric** is a metric on $\mathbb{Z}^2$, defined as the minimum number of moves a chess rook need to travel from x to $y \in \mathbb{Z}^2$. It is $\{0, 1, 2\}$-valued and coincides with the **Hamming metric** on $\mathbb{Z}^2$.

- The **grid metric** is the l_1**-metric** on $\mathbb{Z}^n$. It is the **path metric** of an infinite graph: two points of $\mathbb{Z}^n$ are adjacent if their l_1-distance is 1. For $n = 2$, this metric is restriction on $\mathbb{Z}^2$ of **Manhattan metric** and it called **4-metric** since each point has exactly 4 l_1-neighbors in $\mathbb{Z}^2$.

- The **lattice metric** is the l_∞**-metric** on $\mathbb{Z}^n$. It is the **path metric** of an infinite graph: two points of $\mathbb{Z}^n$ are adjacent if their l_∞-distance is 1. For $\mathbb{Z}^2$, the adjacency corresponds to the king move in chessboard terms, and this metric is called **chessboard metric** (or **king metric**, **8-metric** since each point has exactly 8 l_∞-neighbors in $\mathbb{Z}^2$).

- The **hexagonal metric** is a metric on $\mathbb{Z}^2$ with an **unit sphere** $S^1(x)$: $S^1(x) = S^1_{l_1}(x) \cup \{(x_1 \pm 1, x_2 - 1), (x_1 \pm 1, x_2 + 1)\}$ if x_2 is odd/even. Since $|S^1(x)| = 6$, the hexagonal metric is called also **6-metric**. The hexagonal metric is the **path metric** on the **hexagonal grid** of the plane. It approximates l_2-metric better than l_1- or l_∞**-metric**.

- The **knight metric** is a metric on $\mathbb{Z}^2$, defined as the minimum number of moves a chess knight would take to travel from x to $y \in \mathbb{Z}^2$.

- Let $p, q \in \mathbb{N}$ such that $p + q$ is odd, and $(p, q) = 1$.

  A (p, q)**-super-knight** (or (p, q)**-leaper**) is a (variant) chess piece a move of which consists of a leap p squares in one orthogonal direction followed by a 90 degree direction change, and q squares leap to the destination square. Chess-variant terms for an $(p, 1)$-leaper with p=$0, 1, 2, 3, 4$: **Wazir**, **Ferz**, usual **Knight**, **Camel**, **Giraffe** and for an $(p, 2)$-leaper with $p = 0, 1, 2, 3$: **Dabbaba**, **Knight**, **Alfil**, **Zebra**.

  A **super-knight metric** on $\mathbb{Z}^2$ is the minimum number of moves a (p, q)-super-knight would take to travel from x to $y \in \mathbb{Z}^2$.

  The **knight metric** is the $(1, 2)$-super-knight metric.

  The l_1**-metric** is $(0, 1)$-super-knight metric, i.e., the **Wazir metric**.

- Given $\alpha, \beta \geq 0$ with $\alpha \leq \beta < 2\alpha$, consider (α, β)**-weighted** l_∞**-grid**, i.e., pixel graph $(V = \mathbb{Z}^2, E)$ with $(xy) \in E$ if $|x - y|_\infty = 1$, and horizontal/vertical and diagonal edges having **weights** α and β, resp. Borgefors (α, β)**-chamfer metric** is the **weighted path metric** of this graph. The main cases are (α, β)=$(1, 0)$ (l_1**-metric**), $(3, 4)$, $(1, 1)$ (l_∞**-metric**), $(1, \sqrt{2})$ (**Montanari metric**), $(5, 7)$ (**Verwer metric**), $(2, 3)$ (**Hilditch-Rutovitz metric**).

- An (α, β, γ)**-chamfer metric** is the weighted path metric of voxel graph $(V = \mathbb{Z}^3, E)$ with $(xy) \in E$ if $|x - y|_\infty = 1$, and moves to 6 face, 12 edge, 8 corner neighbors having **weights** α, β, γ, respectively.

  The cases (α, β, γ)=$(1, 1, 1)$ (l_∞**-metric**), $(3, 4, 5)$, $(1, 2, 3)$ are the most used ones for digital $3D$ images.

DISTANCES IN BIOLOGY

1. DISTANCES FOR FREQUENCY, DNA/RNA, PROTEIN DATA
2. OTHER BIO DISTANCES (FOR GENOMES, REACTIONS ETC.)
3. DISTANCES ON TREES
4. BIOLOGICAL DISTANCE MODELS
5. VISUAL, AUDITORY AND HAPTIC SPACES
6. REAL-WORLD BIOLOGICAL DISTANCES
7. IMAGE DISTANCES
8. AUDIO DISTANCES

The distances are mainly used in **Biology** to pursue basic classification tasks, for instance, for reconstructing the evolutionary history of organisms in the form of **phylogenetic trees**.

In the classical approach those distances were based on the comparative morphology, physiology, mating studies, paleontology and immunodiffusion.

The progress of modern **Molecular Biology** allowed also to use nuclear- and/or amino-acid sequences to estimate distances between genes, proteins, genomes, organisms, species, etc..

DNA is a sequence of **nucleotides** (or **nuclei acids**) A, T, G and C, and it can be seen as a word over this alphabet of 4 letters.

In **RNA**, it is uracil U instead of T.

Two strands of DNA are held together (in the form of a **double helix**) by (weak hydrogen) bonds between corresponding nucleotides (necessarily, a **purine** A, G and a **pyrimidine** T, C) in the strands alignment. Those pairs are called **base pairs**.

A **mutation** is a substitution of a base pair.

DNA molecules occur (in the nuclei of eukaryote cells) in the form of long strings, called **chromosomes**.

A **gene** is a contiguous stretch of DNA, which encodes a protein or an RNA molecule. The location of a gene on its chromosome is **gene locus**. Different versions (states) of a gene are called its **alleles**.

A **proteins**, i.e., hormones, catalysts (enzymes), antibodies etc. are large molecules formed by **amino acids**. There are 20 amino acids; the three-dimensional shape of a protein is defined by the (linear) sequence of amino acids, i.e., by a word in this alphabet in 20 letters.

The **genetic code** is the correspondence, universal to (almost) all organisms, between some **codons** (ordered triples of nucleotides) and 20 amino acids. It express the **genotype** (information, contained in genes, i.e., in DNA) as the **phenotype** (proteins).

A **genome** is entire genetic constitution of a species or of a living organism.

IAM (for infinite-alleles model of evolution) assumes that an allele can change from any given state into any other given state.

It corresponds to primary role for **genetic drift** (i.e. random variation in gene frequencies from one generation to another); especially in small populations, over **natural selection** (stepwise mutations).

SMM (for step-wise mutation model of evolution) is more convenient for (recently, most popular) micro-satellite data. Micro-satellites are highly variable repeating short sequences of DNA; their mutation rate is 1 per 1000-10000 replication events, while it is 1/1000000 for allozymes. Micro-satellite data (for example, for DNA fingerprinting) consists of numbers of repeats of micro-satellites for each allele.

The term **taxonomic distance** is used for every distance between two **taxa**, i.e., entities or groups, which are arranged into an hierarchy (a tree indicating relationship).

Linnean taxonomic hierarchy is arranged in ascending series of ranks: Zoology (7 ranks: Kingdom, Phylum, Class, Order, Family Genus, Species) and Botany (12 ranks).

A **phenogram** is an hierarchy expressing **phenetic relationship**, i.e., unweighted overall similarity. A **cladogram** is a strictly genealogical (by ancestry) hierarchy in which no attempt is made to represent amount of genetic divergence between taxa.

A **phylogenetic tree** is an hierarchy representing a hypothesis of **phylogeny**, i.e., evolutionary relationships within and between taxonomic levels, especially the patterns of lines of descent.

Distances between any two taxa (points on phylogenetic tree) are:

Phenetic distance: a measure of the difference in phenotype.
Phylogenetic (or **cladistic**, **genealogical**) **distance**: the minimum number of edges, separating them in a phylogenetic tree.
Evolutionary (or **patristic**, general **genetic**) distance: a measure of genetic divergence estimating the **divergence time**, i.e., the time that has past since those populations existed as a single population.
General **immunological distance**: a measure of the strength of antigen-antibody reactions. Precise terms for immunological and genetic distances will be defined below.

The main way to estimate genetic distance between DNA, RNA or proteins is to compare their (nucleotide or amino acid) sequences. Main non-sequencing techniques are **immunology**, **annealing** (cf. **DNA hybridization metric**) and comparing images under **gel electrophoresis** (separation by an electric charge) and dye staining.

Proponents of **molecular clock hypothesis** estimate that 1 unit of **immunological albumin distance** between two taxa corresponds to $\approx 540,000$ years of their divergence time, and that 1 init of **Nei standard genetic distance** corresponds to $18-20$ million years.

Sarich and Watson,1967, estimated albumin immunological differences for pairs humans-monkeys, apes-monkeys, humans-apes as are 6%, 6%, 1%, resp. Since the hominoids/monkeys divergence time is 30 mya and their immunological distance is 6%, they deduced humans/apes divergence time as $\frac{1}{6}$ of it, i.e., 5 mya ago.

Zuckerkandl and Pauling, 1960, sequenced the hemoglobin amino acids of several species. The difference was roughly proportional to estimated geologucal time since these species had a common ancestor. The hemoglobins of 3 mammals (human, horse, mouse) originated ≈ 70 mya) differed pairwisely by ≈ 20 amino acids, while, for each of them and shark (originated ≈ 470 mya), they differed by ≈ 80 amino acids.

- An **antigen** is any molecule eliciting immune response. **Antibodies** are specific proteins that bind to the antigen.

  The **index of dissimilarity** $id(x,y)$ between two taxa x and y is the factor by which the **heterologous** (reacting with an antibody not induced by it) antigen concentration must be raised to get a reaction as strong as that to the **homologous** (reacting with its specific antibody) antigen. The **immunological distance** is $100(\log id(x,y)+\log id(x,y))$

  Earlier immunodiffusion procedure compared the amount of precipitate when heterologous bloods were added in similar amount as homologous ones, or compared highest dilution giving positive reaction.

  The name of applied antigen (target protein) can be used to specify immunological distance, say, albumin, transferrin, lysozyme distances.

  An antiserum **titer** is a measurement of concentration of antibodies found in a serum. Titers are expressed in their highest positive dilution.

DISTANCES FOR FREQUENCY, DNA, PROTEIN DATA

Those distances between populations measure evolutionary divergence by counting the number of allelic substitutions by loci.

A **population** is represented by a double-indexed vector $x = (x_{ij})$ with $\sum_{j=1}^{n} m_j$ components, where x_{ij} is the frequency of ith **allele** (the label for a state of a gene) at the jth gene locus m_j is the number of alleles at the jth locus and n is the number of considered loci.

$\sum$ denotes summation over all i and j. It holds $x_{ij} \geq 0$, $\sum_{i=1}^{m_j} x_{ij} = 1$.

- **Dps distance** is $-\ln \frac{\sum \min(x_{ij}, y_{ij})}{\sum_{j=1}^{n} m_j}$.
- **Prevosti-Ocana-Alonso distance** is $\frac{\sum |x_{ij} - y_{ij}|}{2n}$.
- **Roger metric** is $\frac{1}{\sqrt{2n}} \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{m_j} (x_{ij} - y_{ij})^2}$.
- **Cavalli-Sforza arc distance** is $\frac{2}{\pi} \arccos(\sum \sqrt{x_{ij} y_{ij}})$.

- **Nei-Tajima-Tateno D_A distance** is $1 - \frac{1}{n}\sum\sqrt{x_{ij}y_{ij}}$.
- **Nei minimum genetic distance** is $\frac{1}{2n}\sum(x_{ij} - y_{ij})^2$.
- **Nei standard genetic distance** is $-\ln I$, where I is Nei **normalized identity of genes** defined by $\frac{\langle x,y\rangle}{||x||_2 \cdot ||y||_2}$.

  Cf. **Bhattacharya distance** and **angular semi-metric**.
- **Sangvi χ^2 distance** is $\frac{2}{n}\sum\frac{(x_{ij}-y_{ij})^2}{x_{ij}+y_{ij}}$.
- **Latter F-statistics distance** is $\frac{\sum(x_{ij}-y_{ij})^2}{2(n-\sum x_{ij}y_{ij})}$.
- **Goldstein and al. distance** is $\frac{1}{n}\sum(ix_{ij} - iy_{ij})^2$ or $\frac{1}{n^2}(\sum(ix_{ij} - iy_{ij}))^2$.
- **Average square distance** is $\frac{1}{n}\sum_{k=1}^{n}(\sum_{1\le i<j\le m_j}(i-j)^2 x_{ik}y_{jk})$.

- The **kinship distance** $-\ln\langle x, y\rangle$ and **kinship coefficient** $\langle x, y\rangle$.

- **Reynolds-Weir-Cockerham distance** (or **co-ancestry distance**) $-\ln(1-\theta)$, where **co-ancestry coefficient** $\theta(x, y)$ of two individuals (or populations) is the probability that a randomly picked allele from one is **identical by descent** (i.e. corresponding genes are copies of the same ancestral gene) to a randomly picked allele in another. Two genes can be **identical by state** (having same allele label) but not by descent.

  $\theta(x, y)$ is the **inbreeding coefficient** F of their next generation.

Distances between DNA, RNA or protein sequences are usually measured in terms of substitutions, i.e. mutations, between them.

A **DNA sequence** is a string $x = (x_1, \dots, x_n)$ over the alphabet $\{A, C, G, T\}$ of nucleotides; $\sum$ denotes $\sum_{i=1}^{n}$.

- **No. of differences** is just the **Hamming distance** $\sum 1_{x_i \neq y_i}$.

  "Non-corrected"

- **p-distance** is $d_p(x, y) = \frac{\sum 1_{x_i \neq y_i}}{n}$.

- **Jukes-Cantor nucleotide distance** is $-\frac{3}{4} \ln(1 - \frac{4}{3} d_p)$.

- **Tajima-Nei distance** is $-b \ln\left(1 - \frac{d_p(x,y)}{b}\right)$, where

  $b = \frac{1}{2}\left(1 - \sum_{j=A,T,C,G}\left(\frac{1_{x_i=y_i=j}}{n}\right)^2 + \frac{1}{c}\sum\left(\frac{1_{x_i \neq y_i}}{n}\right)^2\right)$ and

  $c = \frac{1}{2}\sum_{i,k \in \{A,T,G,C\}, j \neq k} \frac{\left(\sum 1_{(x_i,y_i)=(j,k)}\right)^2}{(\sum 1_{x_i=y_i=j})(\sum 1_{x_i=y_i=k})}$.

- **Hybridization** is the process of combining, into a single molecule, complementary, single-stranded nucleic acids.

  Annealing is binding of two strands by interchange of all A, T, G, C by T, A, C, G, resp. (**Watson-Crick complementation**).
  Denaturation is the reverse process of separating two strands of the double stranded DNA/RNA molecule (heating breaks hydrogen bonds between bases). The rate of annealing of two strands (or t^0 at which denaturation occurs) measures similarity of their base sequences.

  Garson et al. hybridization metric between DNA cubes A and B is $\min_{x \in A, y \in B} \mathbf{H(x,y)}$, where, for DNA n-sequences x and y, $H(x,y)$ is $\min_{-n \le k \le n} \sum 1_{x_i \neq y^*_{i+k}}$. Here indexes $i+k$ are modulo n and y^* is the reversal of y followed by Watson-Crick complementation. A **DNA cube** is any maximal set of DNA n-sequences with all $H(x,y) = 0$.

A **protein sequence** is a sequence $x = (x_1, \ldots, x_n)$ over alphabet of 20 amino acids; $\sum$ denotes $\sum_{i=1}^{n}$.

Among notions of similarity/distance on the set of 20 amino acids (based on their hydrophilicity, polarity, charge, shape etc.), most important is 20 × 20 **Dayhoff PAM250** matrix expressing relative mutability of 20 amino acids.

- **PAM distance** (or **Dayhoff-Eck distance**) between two protein sequences is the minimal number of accepted (fixed) point mutations per 100 amino acids, needed to transform one protein into another.

  1 PAM is a **unit of evolution**: it corresponds to 1 point mutation per 100 amino acids. PAM values 80, 100, 200, 250 correspond to the distance (on %) 50, 60, 75, 92 between proteins.

- **No. of differences** is the **Hamming distance** $\sum 1_{x_i \neq y_i}$.
- **Amino p-distance** (or **uncorrected distance**) is $d_p(x,y) = \frac{\sum 1_{x_i \neq y_i}}{n}$.
- **Amino Poisson correction distance** is $-\ln(1-d_p)$.
- **Amino γ distance** (or **Poisson correction γ distance**) is $a((1-d_p)^{-1/a}-1)$, if mutation rate is γ-distributed with parameter a.

  For $a = 2.25$, it is **Dayhoff distance**.
- **Jukes-Cantor protein distance** is $-\frac{19}{20}\ln(1-\frac{20}{19}d_p)$.
- **Kimura protein distance** is $-\ln(1-d_p-\frac{d_p^2}{5})$.

OTHER BIOLOGICAL DISTANCES

- An **RNA sequence** (or **RNA primary structure**) is a string over the alphabet $\{A, C, G, U\}$ of nucleotides. Inside a cell, such string folds in $3D$ space (as **RNA tertiary structure**), because of pairing of nucleotide bases (usually, by bonds A–U, G–C and G–U).

  The **RNA secondary structure** is, roughly, the set of helices (or the list of paired bases) making up the RNA. This structure can be represented as planar graph and further, as rooted tree.

  An **RNA structural distance** between two RNA sequences is a distance between their secondary structures.

  Examples are: **tree edit distance** (and other distances on rooted trees), and the **base-pair distance**, i.e., the **symmetric difference metric** between secondary structures seen as sets of paired bases.

- Represent **RNA secondary structure** by a graph $(V = \{1, \ldots, n\}, E)$ such that, for $1 \leq i \leq n$, $(i, i+1) \notin E$ and $(i, j), (i, k) \in E$ imply $j = k$. Let $E = \{(i_1, j_1), \ldots, (i_k, j_k)\}$ and let (ij) denote the transposition of i, j. Then $\pi(G) = \prod_{t=1}^{k}(i_t j_t)$ is an **involution**.

  The **Reidys-Stadler-Rosello metrics** between $G = (V, E)$ and $G' = (V', E')$ are $(\ln 2)|E \Delta E'|$ and $|E \Delta E'| - 2T$, where T is the number of cyclic orbits of length greater than 2 induced by the action on V of the subgroup $\langle \pi(G), \pi(G') \rangle$ of the group Sym_n. The second metric is the number of transpositions needed to represent $\pi(G)\pi(G')$.

  Let $I_G = \langle x_i x_j : (x_i, x_j) \in E \rangle$ be the monomial ideal (in the ring of polynomials in variables $x_1, \ldots, x_n$ with coefficients $0, 1$) and $M(I_G)_m$ be the set of monomials of degree $\leq m$ belonging to I_G.

  For any $m \geq 3$, a Liabrés-Rosello **monomial metric** between G and G' is $|M(I_G)_{m-1} \Delta M(I_{G'})_{m-1}|$.

- The **fuzzy polynucleotide metric** (or **NTV-metric**) is the metric $\frac{\sum_{1 \leq i \leq 12} |x_i - y_i|}{\sum_{1 \leq i \leq 12} \max\{x_i, y_i\}}$ (Nieto, Torres and Valques-Trasande, 2003) on the 12-dimensional unit cube I^{12}.

  Coding letters U, C, A, G of RNA alpabet as $(1, 0, 0, 0,)$, $(0, 1, 0, 0)$, $(0, 0, 1, 0)$, $(0, 0, 0, 1)$, resp,, one can see 64 possible triplet codons of the genetic code as vertices of I^{12}. Then any point $x = (x_1, \dots, x_{12}) \in I^{12}$ can be seen as a **fuzzy polynucleotide codon** with x_i expressing the grade of membership of element i, $1 \leq i \leq 12$, in the **fuzzy set** x. 64 vertices of the cube are the **crisp sets**.

  Dress and Lokot: $\frac{\sum_{1 \leq i \leq n} |x_i - y_i|}{\sum_{1 \leq i \leq n} \max\{|x_i|, |y_i|\}}$ is a metric on whole $\mathbb{R}^n$.

  On $\mathbb{R}^n_{\geq 0}$ this metric is $1 - s(x, y)$, where $s(x, y) = \frac{\sum_{1 \leq i \leq n} \min\{x_i, y_i\}}{\sum_{1 \leq i \leq n} \max\{x_i, y_i\}}$ is the **Ruzicka similarity**.

- Given a connected graph $G = (V, E)$, the **path metric** between two vertices is the number of edges of a shortest path connecting them.

- Given a finite set $\mathcal{O}$ of (unary) **editing operations** on a finite set X, the **editing metric** on X is the path metric of the graph with the vertex-set X and xy being an edge if and only if y can be obtained from x by operations from $\mathcal{O}$.

  An **alphabet** is a set $\mathcal{A}$, $2 \leq |\mathcal{A}| \leq \infty$ of **characters**. A **string** is a sequence of characters over $\mathcal{A}$; $W(\mathcal{A})$ is the set of all finite strings.

  Main editing operations on strings are: **character replacement**, **character indel** (insertion or deletion of a character), **character swap** (interchange of adjacent characters) and **blok reversal**.

  On $2^n n!$ signed permutations, for example, **signed reversal** is a move from $x_1, \dots, x_n$ to $x_n^*, \dots, x_1^*$, where $x_i^* = -x_{n-i}$.

- The **Levenstein metric** (or **Hamming+Gap metric**, **shuffle Hamming distance**, **character edit metric**) is an editing metric on $W(\mathcal{A})$ with $\mathcal{O}$ consisting of only character replacements and indels.

  The Levenstein metric between strings $x = x_1 \dots x_m$ and $y = y_1 \dots y_n$ is equal to $\min\{\mathbf{d_H(x^*, y^*)}\}$, where x^*, y^* are strings of length k, $k \geq \max\{m, n\}$, over alphabet $\mathcal{A}^* = \mathcal{A} \cup \{*\}$, so that after deleting all new characters $*$, strings x^* and y^* shrink to x and y, respectively. Here, the **gap** is the new symbol $*$, and x^*, y^* are **shuffles** of strings x and y with strings consisting of only $*$.

- If $(\mathcal{A}, d)$ is a metric space, the **Needleman-Wunsch-Sellers metric** (or **Levenstein distance with costs**, **global alignment metric**) is an **editing distance with costs** on $W(\mathcal{A})$ obtained for $\mathcal{O}$ consisting of only indels, each of fixed cost $q > 0$, and character replacements, where the cost of replacement of i by j is $d(i, j)$. This metric is the minimal total cost of transforming x into y by those operations.

  The **Gotoh-Smith-Waterman distance** is a more specialized editing metric with costs. It discounts mismatching parts in the beginning and end of the strings x, y and has one indel cost for starting an **affine gap** (contiguous block of indels) and lower cost for extending a gap.

- The **genomes of unichromosomal** species or 1-chromosome organelles (as small viruses and mitochondria) are represented by the order of genes along chromosomes, i.e., as **permutations** (or **rankings**) of given set of n homologous genes.

  If the **directionality** of the genes is accounted for, a chromosome is described by a **signed permutation**.

  The **circular** genomes are represented by **circular (signed) permutations** $x = (x_1, \ldots, x_n)$, where $x_{n+1} = x_1$.

  Given a set of considered mutation moves, a **genomic distance** between two such genomes is the **editing metric** with editing operations being these moves, i.e., the minimal number of moves needed to transform one (signed) permutation into another.

In addition (usually, instead) of local mutations (as character indels or replacements in the DNA sequence), the **large rearrangement** (those happening on large portion of the chromosome) mutations are considered, and corresponding genomic editing metrics are called **genome rearrangement distances**. Such mutations being rarer, these distances estimate better true genomic evolutionary distance.

The main genome (chromosomal) rearrangements are:

for permutations, **inversions** (block reversals), **transpositions** (exchanges of two adjacent blocks), **inverted transposition** (inversion combined with transposition)

and, for signed permutations only, **signed reversals** (sign reversal combined with inversion).

Main genome rearrangement distances between two unichromosomal genomes are: **reversal metric** and **signed reversal metric**;

transposition distance: the minimal number of transpositions needed to transform (permutation representing) one into another;

ITT-distance: the minimal number of inversions, transpositions and inverted transpositions needed to transform one of them into another.

Given two circular signed permutations $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ (so, $x_{n+1} = x_1$ etc.), a **breakpoint** is a number i, $1 \le i \le n$, such that $y_{i+1} \ne x_{j(i)+1}$, where the number $j(i)$, $1 \le j(i) \le n$, is defined by $y_i = x_{j(i)}$.

The **breakpoint distance** (Watterson-Ewens-Hall-Morgan, 1982) between genomes (represented by x,y) is the number of breakpoints.

- **Multichromosomal genomes** can be seen as unordered collections of **systeny sets** of genes, where two genes are **systenic** if they appear in the same chromosome.

  The **syntenic distance** (Ferretti-Nadeau-Sankoff, 1996) between such genomes is the minimal number of mutation moves:

  translocations (exchanges of genes between two chromosomes), **fusions** (merging of two chromosomes into one), **fissions** (split of one chromosome into two) needed to transfer one genome into another.

  Above three mutation moves correspond to interchromosomal genome rearrangements, which are rarer than intrachromosomal ones; so, they give information about deeper evolutionary history.

Example of **distance function selection** for PR in neuronal network.

To gain information about functional connectivity of a neuronal network, one needs to classify neurons, in terms of their firing similarity; so, to select a distance function and a clustering algorithm. A classical example: simple and complex cells discrimination between in the primary visual cortex.

A human brain has $\approx 10^{11}$ of **neurons** (nerve cells). Neuronal response to a stimulus is a continuous time series. It can be reduced, by a threshold criterion, to much simpler discrete series of **spikes** (short electrical pulses),

A **spike train** is a sequence $x = (t_1, \dots, t_s)$ of s events (neuronal spikes, or hearth beats, etc.) listing absolute spike times or inter-spike time intervals.

"Good" **distances between spike trains** should minimize bias (due to predefining analysis parameters if any) and resulting clusters should well match the stimuli and reproduce some control clustering.

Main **distances between spike trains** $x = x_1, \dots, x_m$ and $y = y_1, \dots, y_n$:

1. $\frac{|n-m|}{\max\{m,n\}}$ (**spike count distance**); no bias by predefining analysis parameters, but the temporal structure of trains is missed.

2. $\sum_{1 \leq i \leq s}(x'_i - y'_i)^2$, where, say, $x' = x'_1, \dots, x'_s$ is the sequence of local firing rates of train $x = x_1, \dots, x_m$ partitioned in s time intervals of length T_{rate} (**firing rate distance**); bias due to predefinition of T_{rate}.

3. Let $\tau_{ij} = \frac{1}{2} \min\{x_{i+1} - x_i, x_i - x_{i-1}, y_{i+1} - y_i, y_i - y_{i-1}\}$ and $c(x|y)=$ $\sum_{i=1}^{m} \sum_{j=1}^{n} J_{ij}$, where $J_{ij} = 1, \frac{1}{2}, 0$ if $0 < x_i - y_i \leq \tau_{ij}$, $x_i{=}y_i$, else, resp. **Event sinchronization distance** (Quiroga et al., 2002) is $1 - \frac{c(x|y)+c(y|x)}{\sqrt{mn}}$. Two metrics (above and below) have no parameter presetting time scale.

4. Let $x_{isi}(t) = \min\{x_i : x_i > t\} - \max\{x_i : x_i < t\}$ for $x_1 < t < x_m$, and let $I(t) = \frac{x_{isi}(t)}{y_{isi}(t)-1}$ if $x_{isi}(t) \leq x_{isi}(t)$ and $I(t) = 1 - \frac{y_{isi}(t)}{x_{isi}(t)}$, otherwise. Kreuz et al., 2007, **ISI distances** are $\int_{t=0}^{T} dt|I(t)|$ and $\sum_{i=1}^{m} |I(t_i)|$.

5. **information distances** (**Kullback-Leibler distance** or **Bennet et al.**: Kolmogorov complexity $K(x|y)$ of train x given train y, i.e., the length of the shortest program to compute x if y is provided as an auxiliary input.

The **Kolmogorov complexity** (or **algotithmic entropy**) $K(x)$ of a binary string x is the length of a shortest binary program x^* (the ultimate compressed version of x) to compute x on an universal computer usung a **Turing-complete** language.

6. The **Lempel-Ziv distance** between two binary n-strings x and y is $\max\{\frac{LZ(x|y)}{LZ(x)}, \frac{LZ(y|x)}{LZ(y)}\}$, where $LZ(x) = \frac{|P(x)| \log |P(x)|}{n}$ approximates uncomputable **Kolmogorov complexity** $K(x)$, and $LZ(x|y) = \frac{|P(x) \backslash P(y)| \log |P(x) \backslash P(y)|}{n}$. Here $P(x)$ is the set of non-overlapping substrings into which x is parsed sequentially, so that new substring is not yet contained in the set of substrings generated so far. For example, such **Lempel-Ziv parsing** for $x = 001100101010011$ is $0|01|1|00|10|101|001|11$.

Bias in above 2. and 3. due to transforming the trains into bitstrings.

7. the minimal cost of transforming x into y by the following operations: insert a spike (cost 1), delete a spike (cost 1), shift a spike by time t (cost qt) (**Victor-Purpura distance**); bias due to presetting time scale q.

8. **van Rossum distance**, 2001, is $\sqrt{\int_0^\infty (f_t(x) - f_t(y))^2)dt}$, where x is convoluted with $h_t = \frac{1}{\tau}e^{-t/\tau}$ and $\tau \approx 12$ ms (best); $f_t(x) = \sum_0^m h(t - x_i)$.

Victor-Purpura distance ≈ van Rossum L_1-distance with $h_t{=}\frac{q}{2}$ if $0 \leq t < \frac{2}{q}$

9. **cross-correlation distances**, i.e., as $1 - \frac{\langle x,y \rangle}{||x||||y||}$, if components of x, y are seen as the samples of two zero-mean random variables:

$1 - \frac{\langle f(x),f(y) \rangle}{||f(x)||||f(y)||}$, where $f(x)$ is the train x filtered by convolution with a kernel function $f(\cdot)$ exponential in Haas-White, 2002, or Gaussian in Schreiber et al., 2003; bias due to predefinition of function $f(x)$.

10. **Aronov et al. distance** between two sets of labelled (by firing neuron) spike trains is the minimal cost of transforming one to the other by spike operations insert/delete, shift by time t, relabel with costs 1, qt, k, resp.

- The **genome distance** between two **loci on a chromosome** is the number of base pairs separating them on the chromosome.

- The **map distance** between two **loci on a genetic map** is the recombination frequency expressed as a percentage. It is measured in centimorgans cM, where 1 cM corresponds to their stat. corrected recombination frequency 1%. 1 cM corresponds to $\approx 10^6$ base pairs.

- The **marital distance** is one between birthplaces of spouses (zygotes).

- The **gerontologic distance** between individual of age x and y from a population with **survival fraction distributions** $S_1(t)$ and $S_2(t)$, respectively, is $|\ln \frac{S_2(y)}{S_1(x)}|$. Here a distribution $S(t)$ can be either empirical, or a parametric one based on modeling.

- The **ontogenetic depth** is the number of cell divisions, from fertilized egg to the adult metazoan capable of reproduction (viable gametes).

- **Telomeres**: repetitive DNA sequences ($(TTAGGG)_n$ in vertebrates) at both ends of each linear chromosome in the cell nucleus. They are long stretches of noncoding DNA protecting coding DNA.

  The number n of TTAGGG repeates is **telomere length**; it is ≈ 2000 in humans. Cell can divide if each of its telomeres has positive length; otherwise, it became **senescent** and die.

  Human telomeres are 3-20 kilobases in length; they lose ≈ 100 base pairs (16 repeats) at each mitosis (happening each 20-180 min). Mean leucocyte telomere length decreases with age by 9% per decade. There is correlation between telomere length and longetivity in humans, and between chronic emotional stress in women and telomere shortening.

  But telomere length can increase (by action of enzyme **telomerase** or transfer of repeats between daughter telomers); moreover, the cells of germline, unicellular eukaryotes and some cancer cells are immortal.

- The **metabolic distance** between two **enzymes** is the minimum number of metabolic steps separating them in the metabolic pathways.

- The **Gendron et al. distance** between two **base-base interactions** (represented by 4 × 4 **homogeneous transformation matrices** X and Y) is $\frac{[S(XY^{-1})+S(X^{-1}Y)]}{2}$, where $S(M)=\sqrt{l^2+(\theta/\alpha)^2}$ and l,θ, α: translation length, rotation angle, scaling translation/rotation factor.

- Let $\{s_1,\dots,s_n\}$ be the set of **stimuli** and let q_{ij} be the conditional probability that a subject will perceive stimulus s_j, when the stimulus s_i was shown; so, $q_{ij}\geq 0$ and $\sum_{j=1}^{n} q_{ij}=1$.

  The **Oliva et al. perception distance** between stimuli s_i and s_j is $\frac{1}{q_i+q_j}\sum_{k=1}^{n}|\frac{q_{ik}}{q_i}-\frac{q_{jk}}{q_j}|$, where q_i is the probability of presenting s_i.

- **Biotopes** here are represented as binary sequences $x = (x_1, \dots, x_n)$, where $x_i = 1$ means the presence of the species i. The **biotope distance** (or **Tanimoto distance**) is $\frac{|\{1 \le i \le n : x_i \ne y_i\}|}{|\{1 \le i \le n : x_i + y_i > 0\}|} = \frac{|A \triangle B|}{|A \cup B|}$.
- The **dispersal distance** is a **range distance** to which a species maintains or expand the distribution of a population. It refer, for example, to seed dispersal by pollination, to natal dispersal, to breeding dispersal, to migration dispersal, etc.
- Given a finite metric space (X, d) (usually, a Euclidean space) and selected, as typical by some criterion, vertex $x_0 \in X$, called **prototype** (or **centroid**), the **prototype distance** of $x \in X$ is $d(x, x_0)$.

  Usually, elements of X represent phenotypes or morphological traits. The average of $d(x, x_0)$ by $x \in X$ estimates corresponding **variability**.

DISTANCES ON TREES

Let T be a **rooted tree** (a tree with a fixed vertex **root**).

The **depth** of a vertex v, $depth(v)$, is the length of shortest path from v to the root. A vertex v is **parent** of a vertex u, $v = par(u)$ (and u is **child** of v) if they are adjacent and $depth(u) = depth(v) + 1$.

Two vertices are **siblings** if they have the same parent. **In-degree** of a vertex is the number of its children. $T(v)$ is the subtree of T, rooted at a node $v \in V(T)$. If $w \in V(T(v))$, then v is an **ancestor** of w, and w is a **descendant** of v; $nca(u, v)$ is the **nearest common ancestor** of the vertices u and v. T is **labeled tree** if a symbol from a fixed finite alphabet $\mathcal{A}$ is assigned to each node. T is **ordered tree** if a left-to-right order among siblings in T is given.

On the set $\mathbb{T}_{rlo}$ of all rooted labeled ordered trees there are three main **editing operations**:

1. **Relabel**: change the label of a vertex v;

2. **Deletion**: delete a non-root vertex v with parent $v^{'}$ so that children of v become the children of $v^{'}$; the children are inserted instead of v as a subsequence in the left-to-right order of the children of $v^{'}$;

3. **Insertion**: the complement of deletion (insert v as a child of $v^{'}$ making v the parent of a consecutive subsequence of the children of $v^{'}$.

For unordered trees the editing operations can be defined similarly, but insert and delete operations work on a subset instead of a subsequence.

We assume that there is a **cost function** defined on each editing operation, and the **cost** of a sequence of editing operations is the sum of costs of these operations.

- The **tree edit distance** on $\mathbb{T}_{rlo}$ is the minimum cost of a sequence of editing operations (relabels, insertions, and deletions) turning one tree into another. The edit tree distance can be defined in similar way on the set of all rooted labeled unordered trees.

- The **Selkow distance** (or **degree-1 edit distance**) on $\mathbb{T}_{rlo}$ is the minimum cost of a sequence of editing operations (relabels, insertions, and deletions) turning one tree into another if insertions and deletions are restricted to leaves of the trees. The root of T_1 must be mapped to the root of T_2, and if a node v is to be deleted (inserted), then subtree rooted at v, if any, is to be deleted (inserted).

- The **constrained edit distance** on $\mathbb{T}_{rlo}$ is the minimum cost of a sequence of editing operations (relabels, insertions, and deletions) turning one tree into another with the restriction that disjoint subtrees should be mapped to disjoint subtrees.

- The **alignment distance** on $\mathbb{T}_{rlo}$ is the minimum cost of an **alignment** of T_1 and T_2. It corresponds to a restricted edit distance, where all insertions must be performed before any deletions.

  Thus, one inserts **spaces** (vertices labeled with a **blank symbol** λ) into both trees so they become isomorphic when labels are ignored; the resulting trees are overlayed on top of each other giving the **alignment** $T_{\mathcal{A}}$ which is a tree, where each vertex is labeled by a pair of labels.

- The **unit cost edit distance** on $\mathbb{T}_{rlo}$ is the minimum number relabels, insertions and deletions turning one tree into another.

- The **splitting-merging distance** on $\mathbb{T}_{rlo}$ is the minimum number of vertex splittings and mergings needed to turn one tree into another.

- The **degree-2 distance** on the set $\mathbb{T}_l$ of all labeled trees is minimum cost of a sequence relabels, insertions and deletions turning one tree into another if any inserted/deleted vertex has ≤ 2 neighbors.

A **phylogenetic X-tree** is an unordered, unrooted tree with the labeled leaf set X and no vertices of degree two. Let $\mathbb{T}(X)$ denote the set of all such trees. If every interior vertex has degree three, the tree is called **binary** (or **fully resolved**).

A **cut** $A|B$ of X is a partition $X = A \cup B$. Removing an edge e from a tree $T \in \mathbb{T}(X)$ induces a cut of X called **cut associated with** e.

- The **Robinson-Foulds metric** on $\mathbb{T}(X)$ between $T_1, T_2 \in \mathbb{T}(X)$ is $\frac{1}{2}|\Sigma(T_1) \triangle \Sigma(T_2)| = \frac{1}{2}|\Sigma(T_1) - \Sigma(T_2)| + \frac{1}{2}|\Sigma(T_2) - \Sigma(T_1)|$, where $\Sigma(T)$ is the family of cuts of X associated with edges of T.

- The **crossover metric** on $\mathbb{T}(X)$ is the minimum number of **nearest neighbor interchanges** needed to get T_1 from T_2.

  A **nearest neighbor interchange** consists of swapping two subtrees that are adjacent to the same internal edge.

- The **subtree prune-regraft distance** on $\mathbb{T}(X)$ is the minimum number of **subtree prune-regraft transformations** needed to get T_1 from T_2. Such transformation proceeds in 3 steps: remove an edge uv of the tree, thereby dividing it into two subtrees T_u (containing u) and T_v (containing v); then subdivide an edge of T_v, giving a new vertex w; then connect u and w by an edge, and remove all vertices of degree 2.

- The **tree bisection-reconnection metric** (or **TBR-metric**) on $\mathbb{T}(X)$ is the minimum number of **tree bisection and reconnection** transformations needed to get T_1 from T_2.

  Such transformation proceeds in 3 steps: remove an edge uv of the tree, thereby dividing it into two subtrees T_u (containing u) and T_v (containing v); then subdivide an edge of T_v, giving a new vertex w, and an edge of T_u, giving a new vertex z; then connect w and z by an edge, and remove all vertices of degree 2.

- Let $\mathbb{T}_b(X)$ denote the set of all binary phylogenetic X-trees. The **quartet distance** between $T_1, T_2 \in \mathbb{T}_b(X)$ is the number mismatched **quartets** (from the total number $\binom{n}{4}$ possible quartets) for T_1 and T_2.

  This distance is based on the fact that given four leaves $\{1, 2, 3, 4\}$ of a tree, they can only be combined in a binary subtree in 3 different ways: $(12|34)$, $(13|24)$, or $(14|23)$: a notation $(12|34)$ refers to the binary tree with the leaf set $\{1, 2, 3, 4\}$ in which removing the inner edge yields the trees with the leave sets $\{1, 2\}$ and $\{3, 4\}$.

- The **triples distance** between $T_1, T_2 \in \mathbb{T}_b(X)$ is the number of triples (from the total number $\binom{n}{3}$ possible triples) that differ for T_1 and T_2.

- Given set $A = \{1, \ldots, 2k\}$, a **perfect matching** of A is its partition A into k pairs. A **rooted binary tree with n labeled leaves** has a root and $n-2$ internal vertices distinct from the root. It can be identified with a perfect matching on $2n-2$, different from the root, vertices.

  Let $\mathbb{T}_{br}(X)$ denote the set of all rooted binary phylogenetic X-trees with n the set X of labeled leaves. The **perfect matching distance** between $T_1, T_2 \in \mathbb{T}_{br}(X)$ is the minimum number of exchanges needed to bring the perfect matching of T_1 to the perfect matching of T_2.

- Let $\mathbb{T}_n(X)$ denote the set of all rooted ordered binary trees with n interior vertices. The **rotation distance** between $T_1, T_2 \in \mathbf{T}_n$ is the minimum number of **rotations** needed to get T_1 from T_2.

  Given interior edges uv, vv', vv'' and uw of a binary tree the **rotation** is replacing them by edges uv, uv'', vv' and vw.

  There is 1-1-correspondence between edge flipping operations in in triangulations of convex polygons with $n+2$ vertices and rotations in binary trees with n interior vertices.

- The **greatest agreement subtree distance** between **any two trees** is the minimum number of leaves removed to obtain a **common pruned tree**, i.e., an identical subtree that can be obtained from both trees by pruning leaves with the same label.

- An **attributed tree** is a triple (V, E, α), where $T = (V, E)$ is a tree and α is a function assigning an **attribute vector** $\alpha(v)$ to every vertex $v \in V$. Given two attributed trees (V_1, E_1, α) and (V_2, E_2, β), the set of all their **subtree isomorphisms**, i.e., all isomorphisms $f : H_1 \rightarrow H_2$, $H_1 \subset V_1$, $H_2 \subset V_2$, between their **induced subtrees**. Given a similarity s on the set of attributes, the **similarity** between isomorphic induced subtrees is $W_s(f) = \sum_{v \in H_1} s(\alpha(v), \beta(f(v)))$. Let ϕ be the isomorphism with maximal similarity $W_s(\phi) = W(\phi)$.

  The following **semi-metrics on attributed trees** are used: $\max\{|V_1|, |V_2|\} - W(\phi)$, $|V_1| + |V_2| - 2W(\phi)$, $1 - \frac{W(\phi)}{\max\{|V_1|,|V_2|\}}$, $1 - \frac{W(\phi)}{|V_1|+|V_2|-W(\phi)}$. They become metrics on equivalences classes of attributed trees: two attributed trees (V_1, E_1, α) and (V_2, E_2, β) are **equivalent** if there exists an isomorphism $g : V_1 \rightarrow V_2$ between the trees T_1 and T_2, such that, for any $v \in V_1$, we have $\alpha(v) = \beta(g(v))$. Then $|V_1| = |V_2| = W(g)$.

BIOLOGICAL DISTANCE MODELS

- **Long-distance dispersal** (or **LDD**) refer to the rare events of biological dispersal (especially, plants) on distances an order of magnitude greater than median **dispersal distance**.

  Together with **vicarience theory** (land bridges based on continental drift), LDD emerged in Biogeography as main factor of biodiversity and species migration patterns. LDD is shown to be more important for the regional survival of some plants than local (median-distance) dispersal.

  Transoceanic LDD by wind currents is a probable source of the strong floristic similarities among landmasses in southern hemisphere. Examples of other LDD vehicles are: rafting by water (corals can traverse 40000 km during their lifetime), migrating birds (great albatross can fly 300000 km), human trasport, extreme climatic events.

- The **isolation-by-distance** predicts that the genetic distance between populations increases exponentially with respect to their geographic distance. Emergence of regional differences (races) and new species is explained by restricted gene flow and adaptive variations. Isolation by distance was studied, for example, via surnames.

- The **Lasker distance** between two human populations x and y, characterized by surname frequency vectors (x_i) and (y_i), is the number $-\ln 2R_{x,y}$, where $R_{x,y} = \frac{1}{2}\sum_i x_i y_i$ is Lasker's **coefficient of relationship by isonymy**.

  Surname structure is related to inbreeding and (in patrilinear societies) to random genetic drift, mutation and migration. Surnames can be considered as alleles of one locus, and so, distributed as neutral mutations. An isonymy points to a common ancestry.

- **Surname distance model**

  In Collado et al. the preference transmission from parents to children was estimated by comparing, for 47 provinces of Spain, 47×47 distance matrices for **surname distance** with those of **consumption** and **cultural** distances. The distances were L_1-distances $\sum_i |x_i - y_i|$ between the frequency vectors (x_i), (y_i) of provinces x, y, where z_i is, for the province z, either the frequency of i-th surname, or the budget share of i-th good, or the population rate for i-th cultural issue (rate of weddings, newspaper readership etc.), respectively.

  Other distance matrices were for **geographical distance** (in km, between the capitals of two provinces), **income distance** $|m(x) - m(y)|$ where $m(z)$ is mean income in the province z, **climatic distance** $\sum_{1 \le i \le 12} |x_i - y_i|$ where z_i is the average temperature in the province z during i-th month, **migration distance** $\sum_{1 \le i \le 47} |x_i - y_i|$ where z_i is the percentage of people (living in the province z) born in i.

- The **distance model of altruism** (by Koella) suggests that altruists spread locally (i.e. with small **interaction distance** and **offspring dispersal distance**), while the egoists invest in increasing of those distances. The intermediate behaviors are not maintained, and evolution will lead to a stable bimodal spatial pattern.

- The **distance running model** is a model of antropogenesis (Bramble and Lieberman) explaining the transition (from australopithecines to non-animal genus Homo, about 2 million years ago) by adaptations to running long distances in the savanna. Endurance running could define the human body form, producing balanced head, low/wide shoulders, narrow chest, short forearms, large hip etc.

- The **probability-distance hypothesis** (in Psychophysics): the probability of discrimination between two stimuli is a (continuously increasing) function of some subjective quasi-metric between them.

VISUAL, AUDITORY AND HAPTIC SPACES

1. Selected vision distances
2. Size-distance phenomena
3. Distortion of sensual versus physical space
4. Distance cues

- **Selected vision (Ophthalmology) distances**

  Inter-ocular distance: the distance (≈ 6.35 cm) between the centers of the pupils of the two eyes when the visual axes are parallel.

  Near distance: the distance between the object plane and the **spectacle** (eyeglasses) plane.

  Vertex distance: the distance between corneal and spectacles planes.

  Infinite distance: the distance ≥ 6 m (rays entering the eye from an object at that distance appear as parallel as if comung from infinity).

  Resting point of vergence: the distance at which the eyes are set to **converge** (turn inward toward the nose) if there is no close object to converge on. It is ≈ 1.14 m if looking straight ahead, and ergonomists recommend it as eye-screen distance in sustained viewing.

  Default accommodation distance: the distance at which the eyes focus when there is nothing to focus on.

Examples of **size-distance phenomena** in visual perception follow.

Emmert's law: a retinal image is proportional in perceived size S of object to the perceived distance D of the surface it is projected upon. In fact, S doubles every time D is cut in half and vice versa. Emmert's law accounts for **constancy scaling** (that the size of an object is perceived to remain constant despite the changes in the retinal image).

The **size-distance centration** is size overestimation of objects located near the focus of attention and underestimation of it at the periphery.

The **size-distance invariance hypothesis**: the ratio of perceived ones size and distance is the tangent of the physical visual angle. So, the objects which appear closer should also appear smaller. But with **moon illusion** (not understood yet) appears **size-distance paradox**: despite of constancy of its visual angle (≈ 0.52 degree), the horizon moon (similarly, Sun) may appear to be about twice the diameter of the zenith moon (Sun).

- **Visual space** refers to a stable percept (internal representation) of the environment provided by vision, while **haptic space** (or **tactile space**) and **auditory space** refers to such representation provided by the senses of pressure perception and audition. The geometry of these spaces and eventual mappings between them are unknown.

  Main proposals for the visual space: a Riemannian space of constant negative curvature (Luneburg, 1947), a general Riemannian/Finsler space, or an affinely connected (so, not metric, in general) space.

  (An **affine connection** is a linear map sending two vector fields into a third one.) But expansion of perceived depth on near distances and its contraction on far distances indicate that the mapping between visual and physical space is not affine. There is evidence that visial space is almost affine.

  Observed **distorions** and **size-distance phenomena** should be incorporated in good model of visual space.

Main kinds of **distortion of vision and haptic spaces versus physical space** follow; first 3 were observed for auditory space also.

Horopter lines: perceived frontparallel (to observer) lines are physically parallel only at certain subject/task depending distance.

Parallel-alleys: perceived parallel (to the medial plane of the observer) lines are, actually, some hyperbolic curves.

Distance-alleys: lines with corresponding points perceived equidistant, are, actually, some hyperbolic curves. The parallel-alleys are lying inside of distance-alleys and, for visual space, their difference is small on the distances larger than 1.5 m.

Oblique effects: performance of certain tasks is worse when the orientation of stimuli is oblique than in horizontal or vertical case.

Equidistant circles: **egocentric distance** is direction-dependent (the points subject perceives equidistant lie on egg-like curves).

- In Psychology, **symbolic distance effect** is that the brain compares two concepts (or objects) with higher accuracy and faster reaction time if they differ more on the relevant dimension.
- The **subjective distance** (or **cognitive distance**) is a mental representation of actual distance molded by an individual's social, cultural and general life experiences.

  Cognitive distance errors occur either because information about two points is not coded/stored in the same branch of memory, or because of errors in retrieval of this information. For example, the length of a route with many turns and landmarks is usually overestimated.

- The **egocentric distance** is the perceived absolute distance from the self (observer or listener) to an object or a stimulus (say, sound source). Usually, visual egocentric distance underestimates actual physical distance to far objects, and overestimates it for near objects.

  Exocentric distance is perceived relative distance between objects.

- **Distance cues** are cues used to estimate **egocentric distance**.

  For a listener from a fixed location, main **auditory distance cues** are: **intensity** (in open space it decreases of 5 dB for each doubling of the distance; **direct-to-reverberant energy ratio** (in the presence of sound reflecting surfaces), **spectrum**,**binaural differences**.

Main **visual distance cues** include:

relative size, **relative brightness**, **light and shade**;

height in the visual field (in the case of flat surfaces lying below the level of the eye, the more distant parts appear higher);

motion perspective (stationary objects appear, if observer moves, to glide past);

interposition (one object partially occludes view of another);

binocular disparities, **convergence accommodation**;

aerial perspective, **distance hazing** (the objects in the distance became bluer, paler, decreased in contrast, more fuzzy).

REAL-WORLD BIOLOGICAL DISTANCES

1. Selected medical distances

2. Selected human and animal distances

3. Length magnitudes in Biology

- Example of **range distances** (emphasizing a maximum distance) in Biology: the **dispersal distance** refers to seed dispersal by pollination, to natal dispersal, to breeding dispersal, to migration dispersal, etc.

- Examples of **spacing distances** (emphasizing a minimum distance) in Biology: **nearest-neighbor distance** which an animal maintain, in directional movement of large groups from its neighbors, and **isolation distance**: a minimum one required (because of pollination) to be maintained between variations of the same species of crop for the purpose to keep seed pure (for example, 10 feet $\approx$ 3m for rice).

- **Selected medical distances**

  Inter-occlusal distance: in Dentistry, the distance between the occluding surfaces of the maxillary and mandibular teeth.

  Interproximal distance: spacing distance between adjacent teeth;

  Inter-pediculate distance: the distance between the vertebral pedicles as measured on the radiograph.

  Source-skin distance: the distance from the focal spot on the target of the x-ray tube to the skin of the subject.

  Inter-aural distance: the distance between the ears.

  Inter-ocular distance: the distance between the eyes.

  Anogenital distance: the length of the *perineum* (region between anus and genital area). For a male it is normally twice what it is for a female; so, it measures physical masculinity.

The **sedimentation distance** (or ESR, **erythrocyte sedimentation rate**): the distance red blood cells travel in one hour in a sample of blood as they settle to the bottom of a test tube. ESR indicates inflammation and increases in many diseases.

The **margin distance**, in Oncology: the tumor-free surgical margin (after formalin fixation) of tumor resection, done in order to prevent local recurrence. Is margin 8 mm enough, instead of present 2-3 cm?

Main distances used in Ultrasound Biomicroscopy (esp. for glaucoma treatment) are the **angle-opening distance** (anterior iris/corneal endothelium) and the **trabecular ciliary process distance** (from a particular point on **trabecular meshwork** to **ciliary process**).

Magnetic Resonance Imaging uses for **cortical maps** (outer layer regions of cerebral hemispheres representing sensory inputs or motor outputs) **MRI distance map** from gray/white matter interface and **cortical distance** of activation locuses of spatially adjacent stimuli.

- **Distances between people** (types of informal space, by Hall):

  intimate distance for embracing or whispering (15.2 – 45 cm),

  personal distance for conversations among friends (45 – 120 cm),

  social distance for conversations among acquaintances (1.2 – 3.6 m),

  public distance used for public speaking (over 3.6 m).

  For an average westerner, personal space is

  about 70 cm in front, 40 cm behind and 60 cm on either side.

- **Selected animal distances**

  The **individual distance**: the distance which an animal attempts to maintain between itself and other animals.

  The **group distance**: the distance which a group of animals attempts to maintain between it and other groups.

  The **nearest-neighbor distance**: about constant distance which an animal maintain, in directional movement of large groups (schools of fish or flocks of birds), from its immediate neighbors.

  The **distance-to-shore**: the distance to the coastline used, for example, to study clustering of whale strandings.

  The **escape distance**: the distance on which the animal reacts on the appearance of a predator or dominating animal of the same species. Such flight initiation distance is shorter than related **alert distance**.

The **reaction distance**: the distance on which the animal reacts to the appearance of prey; **catching distance**: the distance on which the predator can strike a prey.

The **communication distance** of animal vocalizations (incl. human speech): maximal distance on which the receiver still can get the signal.

Example of simple **distance estimation** (for prey recognition) by some animals: the velocity of the mantid's head movement is kept constant during peering, and so, the distance to the target is inversely proportional to the velocity of the retinal image.

A **distance pheromone** is a soluble (for example, in the urine) and/or evaporable substance emitted by an animal, as an olfactory chemosensory cue, in order to send a message (on alarm, sex, food trail, recognition, etc.) to other members of the same species. In contrast, a **contact pheromone** is such insoluble non-evaporable substance; it coats the animal's body and is a contact cue.

ORDERS OF LENGTH MAGNITUDE IN BIOLOGY (in meters)

$10^{-10} = 1$ **angström**: diameter of a typical atom, EM resolution limit;

$10^{-9} = 1$ **nanometer**: diameter of typical molecule;

2×10^{-9}: diameter of the DNA helix;

1.1×10^{-8}: diameter of prion (smallest self-replicating bio. entity);

2×10^{-8}: smallest nanobes - filament structures in rocks/sediments - (some see them as merely crystal growths since DNA still not found);

9×10^{-8}: HIV virus; in general, known viruses range from 2×10^{-8} (adeno-associated virus) to 4×10^{-7} (Mimivirus); there is a contoversy: consider them as living (and classify as 4th domain, Asytota) or not;

10^{-7}: size of chromosomes and maximum size of a particle that can fit through a surgical mask;

2×10^{-7}: limit of resolution of the light microscope;

$3.8 - 7.4 \times 10^{-7}$: wavelength of visible (to humans) light, violet/red;

4×10^{-7}: diameter of the smallest known archeaum;

$10^{-6} = 1$ **micrometer** (formerly, **micron**);

$10^{-6} - 10^{-5}$: diameter of a typical bacterium; in general, 1.5×1^{-7} is the diameter of smallest known (in non-dormant state) bacteria, **Micoplasma genitalium**, while for largest one, it is 7.5×10^{-4};

7×10^{-6}: diameter of the nucleus of a typical eukaryotic cell;

8×10^{-6}: mean width of human hair (range: $1.8 \times 10^{-6} - 18 \times 10^{-6}$);

$\approx 2 \times 10^{-4}$: the lower limit of the human eye to discern an object;

5×10^{-4}: diameter of a human ovum and typical Amoeba proteus;

5×10^{-3}: length of average red ant; in general, insects range from 1.7×10^{-4} (Megaphragma caribea) to 3.6×10^{-1} (Pharnacia kirbyi);

5.5, and 30.1: height of the tallest animal, the giraffe, and length of a blue whale, the largest animal;

115.3: height of the world's tallest tree, a sequoia Coast Redwood;

8 km: length of largest organism on Earth, sea grass plant **Posidonia oceanica** near Balear Islands, 100,000 years old;

43 hectares: area of Pando, a clonal colony of **Populus tremuloides** tree in U.S. state Utah, 80,000 years old;

$5 \times 10^4 = 50$ km: the maximal distance on which the light of a match can be seen; (at least 10 photons arrive on the retina during 0.1 s);

1.5×10^4–1.5×10^7: wavelength of audible sound (20 Hz - 20 kHz);

$2,000$ km: length of Great Barrier Reef, largest known superorganism;

But, perhaps, it is Gaia?

DICTIONARY OF DISTANCES

The concept of distance is one of the basic ones in human experience; this is the first book treating it in full generality.

Distance metrics have become an essential tool in many areas of Mathematics and its applications. However, a lot of information on distances is too scattered and hardly accessible for non-experts. In view of the growing need (especially in Information Retrieval with respect to Image, Audio, Internet and Biology) for an accessible interdisciplinary source on distances, we have expanded our private collection into this Dictionary.

It aims to be a thought-provoking archive: besides distances, many distance-related notions and paradigms are collected also, in ready-to-use fashion.

In a time when over-specialization and the use of terminology isolates researchers, this Dictionary tries to be “centripetal” and “ecumenical”, providing some access and altitude of vision but without taking the route of scientific vulgarization.

The Dictionary is divided into 28 chapters grouped into 7 Parts. Parts II, III and IV, V require some culture in, respectively, pure and applied Mathematics. Part VII can be read by a layman.

The chapters are thematic lists which can be read independently. When necessary, a chapter or a section starts with a short introduction. Each chapter consists of items ordered in a way that hints of connections between them. All item titles and key terms can be traced via the Index.

Many nice curiosities appear in this “Who is Who” of distances. Also distances having physical meaning show up; they range from 1.6 x 10^{-35} m (Planck length) to 7.4 x 10^{26} m (estimated size of the observable Universe).

The target audience consists of all researchers working on some measuring schemes, of students and the general public interested in science.

ELSEVIER books.elsevier.com

ISBN 0 444 52087 2

9 780444 520876

DEZA
DEZA

DICTIONARY OF DISTANCES

ELSEVIER

ELSEVIER

Dictionary of DISTANCES

Elena Deza and
Michel Marie Deza

I MATHEMATICS OF DISTANCES (Chapters 1–5)

1 *General definitions*

2 *Topological Spaces*

3 *Generalizations of Metric Spaces*

4 *Metric Transforms*

5 *Metrics on Normed Structures*

II GEOMETRY AND DISTANCES (Chapters 6–9)

6 *Distances in Geometry*

7 *Riemannian and Hermitian Metrics*

8 *Distances on Surfaces and Knots*

9 *Distances on Convex Bodies, Cones and Simplicial Complexes*

III DISTANCES IN CLASSICAL MATHEMATICS (Chapters 10–14)

10 *Distances in Algebra*

11 *Distances on Strings and Permutations*

12 *Distances on Numbers, Polynomials and Matrices*

13 *Distances in Functional Analysis*

14 *Distances in Probability Theory*

IV DISTANCES IN APPLIED MATHEMATICS (Chapters 15–18)

15 *Distances in Graph Theory*

16 *Distances in Coding Theory*

17 *Distances and Similarities in Data Analysis*

18 *Distances in Mathematical Engineering*

V COMPUTER-RELATED DISTANCES (Chapters 19–22)

IMAGE DISTANCES

Image Processing treat signals such as photographs, video, or tomographic output. In particular, **Computer Graphics** consists of image synthesis from some abstract models, while **Computer Vision** extracts some abstract information. From ≈ 2000: mainly digitally.

Computer graphics (and our brains) deals with **vector graphics** images, i.e., those represented geometrically by curves, polygons, etc. A **raster graphics image** (or **digital** image) is a representation of $2D$ image as a finite set of digital values, **pixels**, on square ($\mathbb{Z}^2$) grid.

Video and tomographic (MRI) images are $3D$ ($2D$ plus time).

A **digital binary** image corresponds to only two values 0,1 with 1 being interpreted as logical “true” and displayed as black. A **binary continuous** image is a compact subset of Euclidean space $\mathbb{E}^n$, $n{=}2,3$

The **gray-scale images** can be seen as point-weighted binary images. In general, a **fuzzy set** is a point-weighted set with weights (**degrees of membership**. **Histogram** of a a gray-scale image gives the frequency of brightness values in it.

Humans can differ between ≈ 350000 colors but only 30 gray-levels.

For **color images**, (RGB)-representation is most known, where space coordinates R, G, B indicate red, green and blue level.

Among other color models (spaces) are: (CMY) cube (Cyan, Magenta, Yellow colors), (HSL) cone (Hue-color type given as angle, Saturation in %, Luminosity in %), and (YUV), (YIQ) used in PAL, NTSC TV.

(RGB) converts into gray-level luminance by $0.299R + 0.587G + 0.114B$

A **color space** is a 3-parameter description of colors. Exactly 3 are needed because 3 kinds of receptors (cells on the retina) exist in the human eye: for short, middle, long wavelengths, i.e., blue, green, red.

The basic assumption of Colorimetry is that the perceptual color space admits a metric, the true **color distance**. It is expected to be locally Euclidean, i.e., a **Riemannian metric**. Another assumption: there is a continuous mapping from this metric to the one of light stimuli.

Probability-distance hypothesis: the probability with which one stimulus is discriminated from another is a (continuously increasing) function of some subjective quasi-metric between these stimuli.

Such **uniform color scale**, where equal distances in the color space correspond to equal differences in color, is not obtained yet and existing **color distances** are various approximations of it.

Images are often represented by **feature vectors**, including color histograms, color moments, textures, shape descriptors, etc.

Examples of feature (parameter) spaces are:

raw intensity (pixel values), **edges** (contours, boundaries, surfaces), **salient features** (corners, line intersections, points of high curvature), and **statistical features** (moment invariants, centroids). Typical video features are in terms of overlapping frames and motions.

Image Retrieval (similarity search) consists of (as for DNA/protein sequences, audio, text documents, etc.) finding images whose features values are similar between them, or to given query or in given range.

Distances are, for Image Retrieval, between feature vectors of a query and reference, and, for Image Processing, they are between approximated and “true” digital images (to evaluate algorithms).

There are two methods to compare images directly (without features): intensity-based (color and texture histograms) and geometry-based (shape representations as **medial axis**, **skeletons**).

Unprecise term **shape** is used for the extent (silhouette) of the object, for its local geometry or geometrical pattern (conspicuous geometric details, points, curves, etc.), or for that pattern modulo a similarity transformation group (translations, rotations, and scalings).

Unprecise term **texture** means all what is left after color and shape have been considered, or it is defined via structure and randomness.

The similarity between vector representations of images is measured usually by l_p**-**, **weighted editing**, **probabilistic** distances, etc.

The main distances used for compact subsets X and Y of $\mathbb{R}^n$ (usually, $n = 2, 3$) or their digital versions are: **Asplund**, **Shephard metrics**, $vol(X\Delta Y)$ and variations of the **Hausdorff distance**.

- For a given $3D$ color space and a list of n colors, let (c_{i1}, c_{i2}, c_{i3}) be the representation of the i-th color of the list in this space.

  For a color histogram $x = (x_1, \dots, x_n)$, its **average color** is the vector $(x_{(1)}, x_{(2)}, x_{(3)})$, where $x_{(j)} = \sum_{i=1}^{n} x_i c_{ij}$ (for example, the average red, blue and green values in (RGB)).

  The **average color distance** between two color histograms is the Euclidean distance of their average colors.

- Given an image (as a subset of $\mathbb{R}^2$), let p_i be the area percentage of it occupied by the color c_i. A **color component** of the image is (c_i, p_i).

  The **Ma-Deng-Manjunath distance** between color components (c_i, p_i) and (c_j, p_j) is $|p_i - p_j| \cdot d(c_i, c_j)$, where $d(c_i, c_j)$ is the distance between colors c_i and c_j in a given color space.

- Given two color histograms $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ (with x_i, y_i representing number of pixels in the bin i), their Swain-Ballard's **histogram intersection quasi-distance** is $1 - \frac{\sum_{i=1}^{n} \min\{x_i, y_i\}}{\sum_{i=1}^{n} x_i}$.

  For normalized histograms (total sum is 1) above quasi-distance is the usual l_1**-metric** $\sum_{i=1}^{n} |x_i - y_i|$. Their Rosenfeld-Kak's **normalized cross correlation** is a similarity $\frac{\sum_{i=1}^{n} x_i, y_i}{\sum_{i=1}^{n} x_i^2}$.

- Given two color histograms $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ (usually, $n = 256$ or $n = 64$) representing the color percentages of two images, their **histogram quadratic distance** is **Mahalonobis distance**, defined by $\sqrt{(x - y)^T A (x - y)}$, where $A = ((a_{ij}))$ is a symmetric positive-definite matrix, and weight a_{ij} is some, perceptually justified, similarity between colors i and j.

  For example, $a_{ij} = 1 - \frac{d_{ij}}{\max_{1 \le p, q \le n} d_{pq}}$, where d_{ij} is the Euclidean distance between 3-vectors representing i and j in some color space.

- Let $f(x)$ and $g(x)$ denote brightness values of two digital gray-scale images f and g at the pixel $x \in X$, where X is a raster of pixels. Any distance between point-weighted sets (X, f) and (X, g) can be applied as **gray-scale image distance** between f and g. The main used ones:

  RMS (**root mean-square error**) $\left(\frac{1}{|X|}\sum_{x\in X}(f(x)-g(x))^2\right)^{\frac{1}{2}}$;

  Signal-to-noise ratio $SNR(f,g) = \left(\frac{\sum_{x\in X} g(x)^2}{\sum_{x\in X}(f(x)-g(x))^2}\right)^{\frac{1}{2}}$;

  Pixel misclassification error rate $\frac{1}{|X|}|\{x \in X : f(x) \neq g(x)\}|$;

  Frequency RMS $\left(\frac{1}{|U|^2}\sum_{u\in U}(F(u)-G(u))^2\right)^{\frac{1}{2}}$, where F, G are the discrete Fourier transforms of f, g, and U is the frequency domain;

  Sobolev norm of order δ error $\left(\frac{1}{|U|^2}\sum_{u\in U}(1+|\eta_u|^2)^{\delta}(F(u)-G(u))^2\right)^{\frac{1}{2}}$, where $0 < \delta < 1$ is usually $\frac{1}{2}$), and η_u is the $2D$ frequency vector associated in U with position u.

- Given a number r, $0 \leq r < 1$, the **image compression L_p-metric** is the usual L_p**-metric** on $\mathbb{R}^{n^2}_{\geq 0}$ (the set of gray-scale images seen as $n \times n$ matrices) with p being a solution of the equation $r = \frac{p-1}{2p-1} \cdot e^{\frac{p}{2p-1}}$. So, $p = 1, 2, \infty$ for, respectively, $r = 0, r = \frac{1}{3}e^{\frac{2}{3}} \approx 0.65$, $r \geq \frac{\sqrt{e}}{2} \approx 0.82$. Here r estimates **informative** (i.e., filled with non-zeros) part of the image. It is a quality metric to select a lossy compression scheme.

- The **digital volume metric** (a digital analog of the **Nikodym metric**) on bounded subsets (images) of $\mathbb{Z}^n$) is $vol(A \triangle B)$, where $vol(A) = |A|$ (number of pixels in A), and $A \triangle B$ is the **symmetric difference** of sets A and B.

Consider two binary images, seen as non-empty subsets A and B of a finite metric space (say, a raster of pixels) (X, d).

- Their Baddeley's ***p*-th order mean Hausdorff distance** is $\left(\frac{1}{|X|}\sum_{x \in X} |d(x, A) - d(x, B)|^p\right)^{\frac{1}{p}}$, where $d(x, A) = \min_{y \in A} d(x, y)$. For $p = \infty$, it is proportional to usual Hausdorff metric.
- Their Dubuisson-Jain's **modified Hausdorff distance** is $\max\left\{\frac{1}{|A|}\sum_{x \in A} d(x, B), \frac{1}{|B|}\sum_{x \in B} d(x, A)\right\}$.
- If $|A| = |B| = m$, $\min_f \max_{x \in A} d(x, f(x))$, where f is any bijective mapping between A and B, is their **bottleneck distance**.

  Variations of above distance are: **minimum weight matching** $\min_f \sum_{x \in A} d(x, f(x))$, **uniform matching** $\min_f(\max_{x \in A} d(x, f(x))$-$\min_{x \in A} d(x, f(x)))$ and **minimum deviation matching** $\min_f(\max_{x \in A} d(x, f(x))$- $\frac{1}{|A|}\sum_{x \in A} d(x, f(x)))$.

Consider two two images, seen as non-empty compact subsets A and B of a metric space (X, d).

- Their **non-linear Hausdorff metric** (or **wave distance**) is the **Hausdorff distance** $d_{Haus}(A \cap B, (A \cup B)^*)$, where $(A \cup B)^*$ is the subset of $A \cup B$ which forms a closed contiguous region with $A \cap B$, and the distances between points are allowed to be measured only along paths wholly in $A \cup B$.

- Their **Hausdorff distance up to G**, for given group $(G, \cdot, id)$ acting on the Euclidean space $\mathbb{E}^n$, is $\min_{g \in G} d_{Haus}(A, g(B))$. Usually, G is the group of all isometries or all translations of $\mathbb{E}^n$.

- Their **hyperbolic Hausdorff distance** is the **Hausdorff metric** between $MAT(A)$ and $MAT(B\text{MAT(A)})$ of (X, d_{hyp}), where the **hyperbolic distance** $d_{hyp}(x,y)$ is $\max\{0, d_E(x',y') - (r_y - r_x)\}$ for elements $x = (x', r_x)$ and $y = (y', r_y)$ of X.

  Here $MAT(C)$ denotes, for any compact $C \subset \mathbb{R}^n$, its **Blum's medial axis transform**, i.e., the subset of $X = \mathbb{R}^n \times \mathbb{R}_{\geq 0}$ of all pairs $x = (x', r_x)$ of the centers x' and the radii r_x of the maximal inscribed (in C) l_2-balls, in terms of the Euclidean distance d_E in $\mathbb{R}^n$.

- Let (X, d) be a metric space, and let $M \subset X$. The **medial axis** of X is the set $MA(X) = \{x \in X : |\{m \in M : d(x, m) = d(x, M)\}| \geq 2\}$. $MA(X)$ consists of all points of boundaries of **Voronoi regions** (**zones of influence**) of points of M. The **skeleton** $Skel(X)$ of X is the set of the centers of all balls, in terms of the distance d which are inscribed in X and **maximal**, i.e., not belong to any other such ball. The **cut locus** of X is the closure $\overline{MA(X)}$ of the medial axis.

  In general, $MA(X) \subset Skel(X) \subset \overline{MA(X)}$.

  The **medial axis, skeleton, cut locus transforms** are those three point-weighted sets with $d(x, M)$ being the weight of $x \in X$.

  Usually, $X \subset \mathbb{E}^n$, and M is the boundary of X. For $2D$ binary images X, the skeleton is a curve, a single-pixel thin one, in digital case.

  The **exoskeleton** of X is the skeleton of the complement of X, i.e., of the background of the image for which X is the foreground.

- Given a metric space (X, d) $(X = \mathbb{Z}^2$ or $\mathbb{R}^2)$ and a binary image $M \subset X$, the **distance transform** (or **distance field**, **distance map**) is a function $f_M : X \to \mathbb{R}_{\geq 0}$, where $f_M(x) = d(x, M) = \inf_{u \in M} d(x, u)$. So, it can be seen as a gray-scale image where pixel gray-level is labeled by its distance to the nearest pixel of the background.

  The **Voronoi surface** of M is $\{(x, d(x, M)) : x \in X = \mathbb{R}^2\}$.

- Let see two digital images as binary $m \times n$ matrices $x = ((x_{ij}))$ and $y = ((y_{ij}))$, where a pixel x_{ij} is black or white if it is 1 or 0, resp.

  For each pixel x_{ij}, the **fringe distance map** to the nearest pixel of opposite color $D_{BW}(x_{ij})$ is the number of **fringes** expanded from (i, j) (where each fringe consists of pixels that are equi-distant of (i, j)) until the first fringe with a pixel of opposite color is reached. Then $\sum_{1 \leq i \leq m} \sum_{1 \leq j \leq n} |x_{ij} - y_{ij}|(D_{BW}(x_{ij}) + D_{BW}(y_{ij}))$ is **pixel distance**.

- In any metric space (X, d), the **point-set distance** $d(x, M)$ between $x \in X$ and $M \subset X$ is $\inf_{y \in M} d(x, y)$.

  The function $f_M(x) = d(x, M)$ is a (general) **distance map**.

- The **set-set distance** between two subsets $A, B \subset X$ is $\inf_{x \in A,} d(x, B)$. The **Hausdorff metric** is $\max\{d_{dHaus}(A, B), d_{dHaus}(B, A)\}$, where $d_{dHaus}(A, B)$=$\max_{x \in A} \min_{y \in B} d(x, y)$ (for compact subsets $A, B \subset X$).

- If the boundary $B(M)$ of the set M is defined, then

  the **signed distance function** g_M is defined as $-\inf_{u \in B(M)} d(x, u)$ for $x \in M$ and $\inf_{u \in B(M)} d(x, u)$, otherwise.

  If M is a (closed and orientable) manifold in $\mathbb{R}^n$, then g_M is the solution of the **eikonal equation** $|\nabla g| = 1$ for its **gradient** ∇.

- The shape can be represented by a parameterized simple plane curve. Let $X = X(x(t))$, $Y = Y(y(t))$ be two parameterized curves, where $x(t)$, $y(t)$ are continuous on $[0, 1]$ and $x(0) = y(0) = 0$, $x(1) = y(1) = 1$. The most used **parameterized curves distance** is the minimum, over all monotone increasing $x(t)$, $y(t)$, of $\max_t d_E(X(x(t)), Y(y(t)))$. It is Euclidean case of the **dogkeeper distance** which is, in turn, the **Fréchet metric** for the case of curves.

- Consider a digital representation of curves. Fix $r \geq 1$ and let $A = \{a_1, \dots, a_m\}$, $B = \{b_1, \dots, b_n\}$ be finite ordered sets of consecutive points on two closed curves. For any order-preserving correspondence f between all points of A and B, the **stretch** $s(a_i, b_j)$ **of** $(a_i, f(a_i) = b_j)$ is r if either $f(a_{i-1}) = b_j$ or $f(a_i) = b_{j-1}$, or zero, otherwise.

  The **elastic matching distance** is $\min_f \sum(s(a_i, b_j) + d(a_i, b_j))$, where $d(a_i, b_j)$ is the difference between the tangent angles of a_i and b_j. It is a **near-metric** for some r: all $d(x, y) \leq C(d(x, z) + d(z, y))$ for $C \geq 1$.

- For a plane polygon P, its **turning function** $T_P(s)$ is the angle between the counterclockwise tangent and the x-axis as the function of the arc length s. This function increases with each left hand turn and decreases with right hand turns.

  Given two polygons of equal perimeters, their **turning function distance** is the L_p**-metric** between their turning functions.

- For a plane graph $G = (V, E)$ and a **measuring function** f on its vertex-set V (say, the distance from $v \in V$ to the center of mass of V), the **size function** $S_G(x, y)$ on the points $(x, y) \in \mathbb{R}^2$ is the number of connected components of the restriction of G on vertices $\{v \in V : f(v) \leq y\}$ containing a point v' with $f(v') \leq x$.

  Given two plane graphs with vertex-sets belonging to a raster $R \subset \mathbb{Z}^2$, their Uras-Verri's **size function distance** is the normalized l_1-metric between their size functions over raster pixels.

- The **time series video distances** are objective wavelet-based spatial-temporal **video quality metrics**.

  A video stream x is processed into time series $x(t)$ (seen as a curve on coordinate plane) which then (piecewise linearly) approximated by a set of n contiguous line segments that can be defined by $n+1$ endpoints (x_i, x'_i), $0 \le i \le n$, on coordinate plane.

  Wolf-Pinson's distances between video streams x and y are:

  1. $Shape(x,y) = \sum_{i=0}^{n-1} |(x'_{i+1} - x'_i) - (y'_{i+1} - y'_i)|$;
  2. $\text{Offset}(x,y) = \sum_{i=0}^{n-1} |\frac{x'_{i+1}+x'_i}{2} - \frac{y'_{i+1}+y'_i}{2}|$.

AUDIO DISTANCES

Audio (speech, music, etc.) **Signal Processing** is the processing of analog (continuous) or, mainly, digital representation of the air pressure waveform of the sound. A **sound spectrogram** (or **sonogram**) is a visual 3D representation of an acoustic signal. It is obtained either by series of bandpass filters (an analog processing), or by application of the **short-time Fourier transform** to the electronic analog of an acoustic wave.

Three axes represent time, frequency and **intensity**. Often this 3D curve is reduced to 2D by indicating the intensity with, say, more thick lines.

Sound is called **tone** if it is periodic (the lowest **fundamental** frequency plus its multiples, **harmonics**) and **noise**, otherwise. The frequency is measured in **cps** (the number of complete cycles per second) or Hz (Herzs). The range of audible sound frequencies to humans is 20Hz–20kHz.

Decibel dB is the unit used to express relative strength of two signals. Audio signal's amplitude in dB is $20\log_{10}\frac{A(f)}{A(f')} = 10\log_{10}\frac{P(f)}{P(f')}$, where f' ia a reference signal selected to correspond 0 dB (threshold of human hearing). The threshold of pain is about $120-140$ dB.

Pitch and **loudness** are psycho-acoustic (auditory subjective) terms for frequency and amplitude.

Mel scale is a pitch scale, corresponding to the auditory sensation of tone height and based on **mel**, a unit of pitch. It is connected to frequency f Hz scale by $Mel(f) = 1127\ln(1+\frac{f}{700})$.

Bark scale is a scale of loudness scale: it range from 1 to 24 corresponding to the first 24 critical bands of hearing $(0, 100, \dots, 950, 12000, 15500$ Hz).

Those bands correspond to spatial regions of the basilar membrane of the inner ear, where oscillations produced by the sound activate the hair cells and neurons. $Bark(f) = 13\arctan(0.76f) + 3.5\arctan(\frac{f}{0.75})^2$ in f kHz scale.

Human **phonation** (speech, song, laughter) is controlled usually by **vocal tract** (the throat and mouth) shape. This shape, i.e., the cross-sectional profile of the tube from the closure in the **glottis** (the space between the vocal cords) to the opening (lips), is represented by the cross-sectional area function $Area(x)$, where x is the distance to glottis.

The vocal tract acts as a resonator during vowel phonation, because it is kept relatively open. Those resonances reinforce the source sound (ongoing flow of lung air) at particular **resonant frequencies** (or **formants**) of the vocal tract, producing peaks in the **spectrum** of the sound. Each **vowel** has two characteristic formants, depending of the vertical and horizontal position of the tongue in the mouth.

If the vocal tract is modeled as a sequence of concantenated tubes of constant cross-sectional area of equal length, then ratios $\frac{Area(x_{i+1})}{Area(x_i)}$ for consecutive tubes can be computed. The **log area ratio distance** between discrete spectra x and y of length n is $(\frac{1}{n}\sum_{i=1}^{n} 10(\log_{10}\frac{Area(x_i)}{Area(y_i)})^2)^{\frac{1}{2}}$.

The **spectrum** of a sound is the distribution of magnitude (dB) of the components of the wave. The **spectral envelope** is a smooth contour connecting spectral peaks. Estimation of the spectral envelopes is based on either LPC (linear predictive coding), or FTT (fast Fourier transform).

FTT maps time-domain functions into frequency-domain. The **cepstrum** of the signal $f(t)$ is $FT(\ln(FT(f(t) + 2\pi mi)))$, where m is the integer needed to unwrap the angle or imaginary part of the complex log function.

(The complex and real cepstrum use, respectively, complex and real log function. The real cepstrum uses only the magnitude of the original signal $f(t)$, while the complex cepstrum uses also phase of $f(t)$.)

FFT performs Fourier transform on the signal and sample the discrete transform output at the desired frequencies in mel scale.

Parameter-based distances used in recognition and processing of speech data are usually derived by LPC, modeling speech spectrum as a linear combination of the previous samples (as in autoregressive process).

Majority of **distortion measures between sonograms** are variations of **squared Euclidean distance** (including **Mahalanobis distance**) and probabilistic distances (f**-divergence of Csizar**, **Chernoff distance**, generalized **total variation metric**).

The distances for sound processing below are between vectors x and y representing two signals to compare. For **recognition**, they are a template reference and input signal, while for **noise reduction**, they are (as in Image Processing) original (reference) and distorted signal.

Often distances are calculated for small segments, between vectors representing short-time spectra, and then averaged.

- The **RMS log spectral distance** between discrete spectra $x = (x_i)$ and $y = (y_i)$ is Euclidean distance $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ln x_i - \ln y_i)^2}$.

  The **log area ratio distance** $LAR(x, y)$ between x and y is $(\frac{1}{n}\sum_{i=1}^{n} 10(\log_{10} \frac{Area(x_i)}{Area(y_i)})^2)^{\frac{1}{2}}$, where $Area(z_i)$ means cross-sectional area of the segment of the vocal tract tube corresponding to z_i.

- The **segmented signal-to-noise ratio** $SNR_{seg}(x, y)$ between signals $x = (x_i)$ and $y = (y_i)$ is $\frac{10}{m}\sum_{m=0}^{M-1}\left(\log_{10}\sum_{i=nm+1}^{nm+n}\frac{x_i^2}{(x_i-y_i)^2}\right)$, where n is the number of frames, and M is the number of segments.

  Usual **signal-to-noise ratio** $SNR(x, y)$ is $10\log_{10}\frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i-y_i)^2}$.

  Also used, to compare waveforms x and y in time-domain, their **Czekanovski-Dice distance** $\frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{2\min\{x_i,y_i\}}{x_i+y_i}\right)$.

- The **Klatt slope metric** between discrete spectra $x = (x_i)$ and $y = (y_i)$ with n channel filters is $(\sum_{i=1}^{n}((x_{i+1} - x_i) - (y_{i+1} - y_i))^2)^{\frac{1}{2}}$.

- The **Bark spectral distance** is a perceptual distance $BSD(x, y) = \sum_{i=1}^{n}(x_i - y_i)^2$, i.e., is the **squared Euclidean distance** between **Bark spectra** (x_i) and (y_i) of x and y, where i-th component corresponds to i-th auditory critical band in Bark scale.

- The **Itakura-Saito quasi-distance** (or **maximum likelihood distance**) $IS(x, y)$ between LPC-derived spectral envelopes $x = x(\omega)$ and $y = y(\omega)$ is $\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\ln\frac{x(w)}{y(w)} + \frac{y(w)}{x(w)} - 1\right)dw$.

  The **cosh distance** is defined by $IS(x, y) + IS(y, x)$.O

- The **log likelihood ratio quasi-distance** (or **Kullback-Leibler distance**) $KL(x, y)$ between LPC-derived spectral envelopes $x = x(\omega)$ and $y = y(\omega)$ is defined by $\frac{1}{2\pi}\int_{-\pi}^{\pi}x(w)\ln\frac{x(w)}{y(w)}dw$. The **Jeffrey divergence** $KL(x, y) + KL(y, x)$ is also used.

“Quefrency”, “cepstrum”: anagrams of “frequency”, “spectrum”, resp

- The **RMS log spectral distance** (or **root-mean-square distance**) $LSD(x,y)$ between discrete spectra $x=(x_i)$ and $y=(y_i)$ is Euclidean distance $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ln x_i - \ln y_i)^2}$. The square of it, via cepstrum representation $\ln x(\omega)=\sum_{j=-\infty}^{\infty} c_j e^{-ij\omega}$ is the **cepstral distance**.

- The **cepstral distance** (or **squared Euclidean cepstrum metric**) $CEP(x,y)$ between LPC-derived spectral envelopes $x=x(\omega)$ and $y=y(\omega)$ is $\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\ln x(w)-\ln y(w)\right)^2 dw = \sum_{j=-\infty}^{\infty}(c_j(x)-c_j(y))^2$, where $c_j(z)=\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{iwj}\ln|z(w)|dw$ is j-th cepstral (real) coefficient of z derived by Fourier transform or LPC.

  The **quefrency-weighted cepstral distance** (or **weighted slope distance**) between x and y is $\sum_{i=-\infty}^{\infty} i^2(c_i(x)-c_i(y))^2$.

  The **Martin cepstrum distance** between two AR (autoregressive) models is, in terms of their cepstrums, $(\sum_{i=0}^{\infty} i(c_i(x)-c_i(y))^2)^{\frac{1}{2}}$.

- A **phone** is a sound segment that possess distinct acoustic properties, the basis sound unit.

  (Cf. **phoneme**, i.e., a family of phones that speakers usually hear as a single sound; the number of phonemes range, among about 6000 spoken now languages, from 11 in Rotokas to 112 in !Xóõ (languages spoken by $\approx$ 4000 people in Papua New Guinea and Botswana, respectively.)

  Two main classes of **phone distance** between phones x and y are:

  Spectrogram-based distances: physical-acoustic distortion measures between the sound spectrograms of x and y;

  Feature-based phone distances: usually **Manhattan distance** $\sum_i |x_i - y_i|$ between vectors (x_i) and (y_i) representing phones x and y with respect to given inventory of phonetic features (for example, nasality, stricture, palatalization, rounding, sillability).

- The **Laver consonant distance** refers, for 22 consonantal phonemes of English, the improbability of confusing them, developed by Laver, 1994, from subjective auditory impressions.

  The smallest distance, 15%, is between [p] and [k], the largest one, 95%, is, for example, between [p] and [z]. Laver also proposed a quasi-distance based on the likehood that one consonant will be misheard as another by an automatic speech-recognition system.

- Liljencrans and Lindlom, 1972, developed a **vowel space** of 14 vowels. Each vowel, after a procedure maximizing contrast among them, is represented by a pair (x, y) of resonant frequencies of the vocal tract (1st and 2nd formants) in linear mel units with $350 \leq x \leq 850$ and $800 \leq y \leq 1700$). Higher x values correspond to lower vowels and higher y values to less rounded or farther front vowels. For example, [u], [a], [i] are represented by $(350, 800)$, $(850, 1150)$, $(350, 1700)$, resp.

- The **phonetic word distance** between two words x and y is the cost-based **editing metric** (for phone sustitutions and indels).

  A word is seen as a string of phones. Given a **phone distance** $r(u, v)$ on the International Phonetic Alphabet with additional phone 0 (the silence), the cost of substitution of phone u by v is $r(u, v)$, while $r(u, 0)$ is the cost of insertion or deletion of u.

- The **linguistic distance** (or **dialectology distance**) between language varieties X and Y is the mean, for fixed sample S of notions, **phonetic word distance** between **cognate** (i.e., having the same meaning) words s_X and s_Y, representing the same notion $s \in S$ in X and Y, respectively.

- **Stover's distance** between phrases with the same key word is the sum $\sum_{-n \leq i \leq +n} a_i x_i$, where $0 < a_i < 1$ and x_i is the proportion of non-mathched words between the phrases within a moving window.

- **Pitch** is a subjective correlate of the fundamental frequency linearly ordered collection of pitches (notes).

  A **pitch distance** (or **musical distance**) is the size of the section of the linearly-perceived pitch-continuum bounded by those two pitches, as modeled in a given scale. The pitch distance between two successive notes in a scale is called a **scale step**.

  In Western music now, the most used one is the **chromatic scale** (octave of 12 notes) of **equal temperament**, i.e., divided into 12 equal steps with the ratio, between any two adjacent frequencies, being $\sqrt[12]{2}$. The scale step here is a **semitone**, i.e., the distance between two adjacent keys (black and white) on a piano. The **distance between notes** whose frequencies are f_1 and f_2 is $12\log_2(\frac{f_1}{f_2})$ semitones.

  A MIDI (Musical Instrument Digital Interface) number of fundamental frequency f is defined by $p(f) = 69 + 12\log_2 \frac{f}{440}$. The distance between notes, in terms of MIDI numbers, is **natural metric** $|m(f_1) - m(f_2)|$.

- A rhythm timeline (music pattern) is represented, besides standard music notation, as binary vector, pitch vector, pitch difference vector, chronotonic histogram or, for example as:

  1. a **inter-onset interval vector** $t = (t_1, \dots, t_n)$ of n time intervals between consecutive onsets.

  2. a **rhythm difference vector** $r = (r_1, \dots, r_{n-1})$, where $r_i = \frac{t_{i+1}}{t_i}$.

  Examples of general **distances between rhythms** are Hamming distance, **swap metric**, **Earth Mover distance** between their given vector representations. The **Euclidean interval vector distance** is the Euclidean distance between two inter-onset interval vectors.

  Coyle-Shmulevich **interval-ratio distance** is $1 - n + \sum_{i=1}^{n-1} \frac{\max\{r_i, r_i'\}}{\min\{r_i, r_i'\}}$, where r and r' are rhythm difference vectors of two rhythms.