Distances in Biology

Michel DEZA (ENS, Paris) and Elena DEZA (SPU, Moscow)

Séminaire Brillouin, IRCAM, 31 mai 2012

The abstraction of measurment, in terms of distance, similarity, metric was originated by Maurice Fréchet (1906) and Felix Hausdorff (1914).

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

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

Devising most suitable distances/similarities became an essential task in many applications, especially, Computational Biology, Statistics, Ecology, Psychophysics, Medicine, Image/Audio Analisys, Information Retrieval.

But Biology still lags behind last two in using, besides distances themselves, powerful distance-related notions and paradigms: transforms, invariants, etc

General observations on **distance design** follow.

- Distance should be **invariant** with respect to small and irrelevant transformations of data.
- Distance can be **upgraded** to metric and **corrected** for bias.
- Distance can be between a **prototype** and **input** (or query) or between **true** and **distorted** (or approximated) data.
- Distance can be **abstract** $\mathbb{R}_{\geq 0}$-valued or **physical** (as length):

  between **spatial** or **temporal** points (the length of journey till, say, speed or concentration reach fixed value).
- Usually, **several** distances on the same data should be compared.
- Distance/similarity can be **implicite** as in Clustering.
- Instead of **fixing** distance/similarity function, it can be **learned**.
- The choice of good distance/similarity is rather an art.

- 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 $\approx$ 3 m for rice).

- Example of **relative distances** (emphasizing a reference point) in Biology: in the lek mating, females in estrous visit a congregation of displaying males, the lek, for fertilization, and mate preferentially with males of higher **lekking distance rank**, i.e., relative distance from male territory (the median of his positions) to the center of the lek.

CONTENTS

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.

In general, the **life** is not well-defined, say, for viruses. DNA could be only relatively recent attribute of life. Neither life can be anything that undergoes Darwinian evolution, since the unit is this evolution (gene, cell, organism, group, species, etc.) is not clear. So, Lineweaver, 2012, defined the life as a **far-from-equilibrium dissipative system**, i.e., exchanging energy and matter with its environment.

DNA is a sequence of **nucleotides** (or **nuclei acids**) A, T, G and C. It can be seen as a word over this alphabet of 4 letters (or 2 letters purine/pyramidine). In **RNA**, it is uracil U instead of T.

Two strands of DNA are held together and in the opposite orientation (forming a **double helix**) by (weak hydrogen) bonds of 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. It is **transversion** if the exchange is purine/pyramidine, and **transition**, otherwise.

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

A **gene** is a functionally complete segment of DNA encoding for a protein or 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 3D shape of a protein is defined by the (linear) sequence of amino acids, i.e., by a word in this alphabet of 20 letters. The **protein length** is the number of amino acids in the chain; average protein length is around 300.

Many of protein amino acids and over 70 extraterrestrial ones were identified in the Murchison meteorite which is older than the Earth: 4.95 versus 4.54 billion years.

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 organism.

The **human genome** is the set of 23 chromosomes consisting of 3.1 billion bp in $20,000$ genes. But the flea *Daphnia pulex* has $31,000$ genes, and the flower *Paris japonica* genome has $\approx$ 150 billion bp.

Human *gamete cell* (sperm or egg) contains only one set of 23 chromosomes. The (normal) males and females differ only in the 23rd pair: XY for males, and XX for female. But a protozoan *Tetrahymena* occurs in 7 *sexes* (different variants); it reproduces in $\binom{7}{2} = 21$ ways.

A *hologenome* is the collection of genomes in a *holobiont* (host plus all of its symbiont microbiota), a possible unit of selection in evolution. Human microbiota consists of $\approx 10^{14}$ (mainly, bacterial and fungal) cells of $\approx$ 500 species with 3 million distinct genes.

Besides genetic and *epigenetic* (i.e., not modifying the sequence) changes of DNA, **evolution** (heritable changes) can happen by ”protein mutations“ (prions), **synthetic XNA** (DNA with modified backbones) or culturally (via behavior and symbolic communication).

The macroscopic life exist on Earth during past 600 Ma (million years). Discounting viruses, ≈ 1.9 million extant species are known among estimated 8.74 million living today.

Roughly, 80% of species are parasites of others, parasites including. Global live biomass is ≈ 560 billion tons C (organically bound carbon). At least half of it is contributed by $\approx 5 \times 10^{30}$ prokaryotes. Humans, domesticated animals and crops contribute $100, 700$ and $2, 000$ million.

99% of species that have ever existed on Earth went extinct. Mean mammalian species' longevity is ≈ 1 Ma. Our subspecies is young (≈ 0.2 Ma) and growing continuously since 1400: $6 - 9\%$ of all humans that have been ever born are living today.

Gott, 2007, estimated: with a 95% chance, humans will last another $5, 100$—$7, 800, 000$ years. Neanderthals had $200, 000$—$350, 000$ years. Earth will support life for another $0.5 - 2.3$ billion years. Hawking: life is common, but intelligence is rare/dangerous; so, go to space.

IAM (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 (stepwise 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 $\frac{1}{1000000}$ in IAM-used enzymes. Micro-satellite data (for example, for DNA fingerprinting) consists of numbers of micro-satellite repeats for each allele.

Besides *vertical gene transfer* (reproduction within species), the evolution is affected by horizontal gene transfer, when an organism incorporates genetic material from another one without being its offspring) and *hybridization* (extra-species sexual reproduction).

In general, evolution, without design and purpose, increased size, complexity and diversity of life. There are indications that evolution has a direction: convergent evolution (say, cognition of primates and crows), increase of energy flow per gram per second, etc.

Macroevolution, dominated by biotic factors (competition, predation, etc.), as in the *Red Queen model*, shape ecosystems locally and over short time spans. But species diversity and larger-scale (geographic and temporal) patterns are driven largely by extrinsic abiotic factors (climate, landscape, food supply, oceanographic and tectonic events, etc.), as in the *Court Jester model*.

The organisms evolve rapidly by adaptation, but most changes cancel each other out and macroevolution (say, speciation) appears slow, non-linear (or chaotic) driven by single accidental events.

In *Black Queen model*, evolution pushes microbs lose essential functions when there is another species around to perform them.

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**: the strength of antigen-antibody reactions. Precise terms will be defined below.

The main way to estimate genetic distance between DNA, RNA or proteins is to compare their sequences.

Main non-sequencing techniques are **immunology**, **annealing** (pairing by hydrogen bonds to a complementary sequence, in hybridization) and comparing images under **gel electrophoresis** (separation by an electric charge) and dye staining.

For two taxa, the **temporal remoteness** of their most recent common ancestor (**divergence time**) is the time (or number of generations) that has passed since those populations existed as a single one.

Proponents of **molecular clock hypothesis** estimate that 1 unit of **immunological albumin distance** between two taxa corresponds to ≈ 0.54 Ma (million years) of their divergence time, and that 1 init of **Nei standard genetic distance** corresponds to $18 - 20$ Ma.

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

- 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 taxa x,y is the factor $\frac{r(x,x)}{r(x,y)}$ 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 $100(\log id(x,y)+\log id(y,x))$

  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.

- **Protein kinases** are enzimes which transmit signals and control cells. Many drugs against cancer, inflamation, etc. are kinase blockers. But their high cross-reactivity (binding to other proteins) lead to toxic side effects. Given a set $\{a_1, \dots, a_n\}$ of used drugs, the **affinity vector** of kinase x is -($\ln B_1, \dots, \ln B_n$), where B_i is the **binding constant** (concentration of binded kinase is measured at equilibrium) for reaction of x with drug a_i, and $B_i = 1$ if no interaction was observed. The **pharmacological distance** (Fabian et al., 2005) between kinases x and y is Euclidean distance $(\sum_{i=1}^{n} (\ln B_i(x) - \ln B_i(y))^2)^{\frac{1}{2}}$.

  The **secondary structure** of a protein is given by the hydrogen bonds between its *residues* (specific monomers). A **dehydron** is a such solvant-accessible bond. The **dehydron matrix** of kinase x with residue-set $\{R_1, \dots, R_m\}$ is $m \times m$ matrix $((D_{ij}))$, where D_{ij} is 1 if residues R_i and R_j are paired by a dehydron and 0, otherwise. The **packing distance** is Hamming distance $\sum_{1 \le i,j \le m} |D_{ij}(x) - D_{ij}(y)|$.

- **FRET** (fluorescence resonance energy transfer; Főrster, 1948) is a quantum mechanical property of a **fluorophore** (molecule component causing its fluorescence) resulting in direct non-radiative energy transfer between the electronic excited states of two dye molecules, the **donor** fluorophore and a suitable **acceptor** fluorophore, via dipole.

  In FRET microscopy, fluorescent proteins are used in living cells as non-invasive probes since they fuse genetically to proteins of interest.

  The efficiency of FRET transfer decays as the inverse 6th power of the physical **donor-acceptor distance**. The distance at which this energy transfer is 50% efficient, i.e., 50% of excited donors are deactivated by FRET, is called **Főrster distance** of those two fluorophores.

  Measurable FRET occur only if the distance is < 10 nm (a typical protein size), mutual orientation is favorable and the spectral overlap of the donor emission with acceptor absorption is sufficient.

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 distance** is $\frac{1}{\sqrt{2}n} \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{m_j} (x_{ij} - y_{ij})^2}$.
- **Cavalli-Sforza arc distance** is $\frac{2}{\pi} \arccos(\sum \sqrt{x_{ij} y_{ij}})$.

- **Nei-Tajima-Tateno 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 $\frac{2}{n}\sum\frac{(x_{ij}-y_{ij})^2}{x_{ij}+y_{ij}}$.
- **Goldstein and al. distance** is $\frac{1}{n}\sum(ix_{ij} - iy_{ij})^2$.
- **Average square distance** is $\frac{1}{n}\sum_{k=1}^{n}(\sum_{1\leq i<j\leq m_j}(i-j)^2 x_{ik}y_{jk})$.

- **Kinship distance** $-\ln\langle x, y\rangle$ and **kinship coefficient** $\langle x, y\rangle$.
- **Latter distance** is $-\ln(1-\frac{\sum(x_{ij}-y_{ij})^2}{2(n-\sum x_{ij}y_{ij})})$.
- **Reynolds-Weir-Cockerham distance** $-\ln(1-\theta)$, where θ is another estimation of their **co-ancestry coefficient**.

  This coefficient is the probability that a randomly picked allele from genetic pool of one population (or from one individual) 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.

  It is the **inbreeding coefficient** F of their next generation.

- **Hereditary trees** (or **family trees**, **pedigree graphs**) are used to represent ancestory relations. Every vertex (person) has in-degree ≤ 2. For any two vertices x, y, smallest **inbreeding loop** containing them is formed by concatenating **ansestral** and descending paths connecting them. It is used to identify genes associated with genetic diseases.

  Generally, for a **directed acyclyc graph**, Bender et al., 2001, defined:

  ancestral path distance: the length of shortest directed path through a common ancestor;

  LCA ancestral path distance: such length via the least ancestor.

  Those distances also measure semantic noun relatedness in WorldNet.

- Unrelated **ancestral distance** of an extant taxon (Hearn and Huber, 2006) is the time (or the number of speciation events, node depth) separating it from its most recent ancestor with ≥ 1 extant descendants having an independent character (trait).

Distances between nucleotide (DNA/RNA) or protein sequences are usually measured in terms of substitutions (mutations) between them. Protein-coding nucleotide sequences are called **codon sequences**.

A **DNA sequence** is a string $x = (x_1, \ldots, x_n)$ 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" No. of mutations

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

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

  "Jukes-Cantor correction"

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

- Given two DNA sequences $x=(x_1,\ldots,x_n)$ and $y=(y_1,\ldots,y_n)$, denote by J the determinant of the 4×4 matrix $((N_{ij}))$, where $N_{ij}=|\{1\leq t\leq n : x_t=i, y_t=j\}|$ and indices $i,j=1,2,3,4$ represent nucleotides A,T,C,G, respectively. Let $f_i(x)$ denote the frequency of i-th nucletiode in sequence x, and let $f(x)=f_1(x)f_2(x)f_3(x)f_4(x)$. **Lake paralinear distance** (defined if $J>0$) is $-\frac{1}{4}\ln\frac{J}{\sqrt{f(x)f(y)}}$.

- The **Eigen-McCaskill-Schuster distance** between DNA n-sequences x and y is $|\{1 \le i \le n : \{x_i, y_i\} \neq \{A, G\}, \{T, C\}\}|$ (No. of positions i with only one of x_i, y_i being a purine, i.e., No of transversions).

  The **Watson-Crick distance** between DNA n-sequences x and $y \neq x$ is $d_H(x, \overline{y}) = |\{1 \le i \le n : \{x_i, y_i\} \neq \{A, T\}, \{G, C\}\}|$, where $\overline{y} = (\overline{y}_1, \dots, \overline{y}_n)$ is the **Watson-Crick complement** of y, i.e., $\overline{y}_i = A, T, G, C$ if $y_i = T, A, C, G$, respectively.

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

  Garzon 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 $x = (x_1, \dots, x_n)$ 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 (in %) 50, 60, 75, 92 between proteins.

- The **genetic code distance** (Fitch and Margoliash, 1967) between amino acids x and y is the minimum number of nucleotides that must be changed to obtain x from y. It is $1, 2$ or 3: amino acids have 3 bases.

- The **Miyata-Miyazawa-Yasanaga distance** (1979) between amino acids x and y with polarities p_x, p_y and volumes v_x, v_y, respectively, is

$$\sqrt{(\frac{|p_x - p_y|}{\sigma_p})^2 + (\frac{|v_x - v_y|}{\sigma_v})^2},$$

  where σ_p and σ_v are standard deviations of $|p_x - p_y|$ and $|v_x - v_y|$.

- Dividing amino acids in **polar** (C, D, E, H, K, N, Q, R, S, T, W, Y) or not, the **polar distance** is 1 if x, y are in different groups and 0, else.

  Dividing amino acids in three groups - **positive** (H, K, R), **negative** (D, E) and **neutral**, (the rest) - the **charge distance** is 1 if x, y are in different groups and 0, else.

- **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 $(V=\{1,\dots,n\},E)$ such that, for $1\le i\le n$, $(i,i+1)\notin E$ and $(i,j),(i,k)\in E$ imply $j=k$. Let $E=\{(i_1,j_1),\dots,(i_k,j_k)\}$ and let (ij) denote the transposition of i,j. Then $\pi(G)=\prod_{t=1}^{k}(i_tj_t)$ 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 $I_G=\langle x_ix_j:(x_i,x_j)\in E\rangle$ be the monomial ideal (in the ring of polynomials in variables $x_1,\dots,x_n$ with coefficients $0,1$) and $M(I_G)_m$ be the set of monomials of degree $\le m$ belonging to I_G.

  For any $m\ge 3$, 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 $\frac{\sum_{1\le i\le 12}|x_i-y_i|}{\sum_{1\le i\le 12}\max\{x_i,y_i\}}$ (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, $1\le i\le 12$, in the **fuzzy set** x. 64 vertices of the cube are the **crisp sets**.

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

  On $\mathbb{R}^n_{\ge 0}$ this metric is $1-s(x,y)$, where $s(x,y)=\frac{\sum_{1\le i\le n}\min\{x_i,y_i\}}{\sum_{1\le i\le n}\max\{x_i,y_i\}}$ 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 $x = x_1 \ldots x_m$ and $y = y_1 \ldots y_n$ 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, $1 \leq i \leq n$, such that $y_{i+1} \neq x_{j(i)+1}$, where the number $j(i)$, $1 \leq j(i) \leq n$, 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 a 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** $x = x_1, \ldots, x_m$ and $y = y_1, \ldots, y_n$:

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, $x' = x'_1, \ldots, x'_s$ is the sequence of local firing rates of train $x = x_1, \ldots, x_m$ 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 $1 - \frac{c(x|y) + c(y|x)}{\sqrt{mn}}$.

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** $\sum_i x_i \ln \frac{x_i}{y_i}$ or **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** (**algotithmic entropy**) $K(x)$ of x is the length of a shortest program x^* (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 $0|01|1|00|10|101|001|11$.

7. **Victor-Purpura distance** is the minimal cost of transforming x into y by operations: insert, delete, shift a spike by time t with costs $1, 1, qt$.

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 $\approx$ van Rossum L_1-distance with $h_t{=}\frac{q}{2}$ if $0 \leq t < \frac{2}{q}$ and 0, otherwise. Those two metrics are the most commonly used ones.

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

- 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.

- **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 $\approx 2,000$ in humans. Cell divides if all its telomere lengths are > 0; otherwise, it became **senescent** and die, or it succeed to self-replicate and became cancer. The **Hayflick limit** is the maximal number of divisions beneath which a normal differentiated cell stop dividing; for humans it is about 52. Our telomeres have 3-20 kb; they lose ≈ 100 bp (16 repeats) at each mitosis (each 20-180 min). There is correlation between telomere length and longetivity in humans.

  But telomere length can increase (by action of enzyme **telomerase** or transfer of repeats between telomers). Some cancer cells, *Turritopsis nutricula* and hydra are **biologically immortal**: there is no **aging** (sustained increase in rate of mortality with age).

- An **action at a distance along a DNA**: many genes are regulated by distant (up to 1 Mp and, possibly, on other chromosome) short (30-200 bp) regions of DNA, **enhancers**. They increase probability of such gene **transcription** (RNA synthesis from DNA template) at large **genome distance** from transcription initiation site (the **promoter**).

  DNA super-coiling is folding of DNA double helix (twisting around the helical axis every 10.4 bp of sequence, forming circles and figures eight) because it has been bent, overwound or underwound. It put enchancer **geometrically** closer to the promoter. There is evidence that genomes are organized into **enhancer-promoter loops**.

  But the long-range enhancer function is not fully understood yet. **Looping model**: enhancer freely diffuses through the nucleoplasm to loop with and activate the target promoter; **protein-tracking model**: it is guided by tracking along DNA to reach/loop with the promoter.

- 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.

- 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**.

- **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 $\frac{|\{1 \le i \le n : x_i \neq y_i\}|}{|\{1 \le i \le n : x_i + y_i > 0\}|} = \frac{|A \triangle B|}{|A \cup B|}$.

- 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.

- Let species be distributed in **subpopulations over a landscape**: a textured mosaic of **patches** (homogeneous areas of land use as fields, lakes, forest) and linear **frontiers** (river shores, hedges, road sides). Individuals move across the landscape by frontiers until they reach another patch or exceed a maximum **dispersal distance**.

  The **ecological distance** between two subpopuations (patches) x and y is (Vuilleumier-Fontanillas, 2007) where $D(x,y)$ is the distance an individual covers to reach patch y from patch x, avereged over all sucsessful dispersers from x to y. If no such dispersers exist, $D(x,y)$ is defined as $\min_z(D(x,z)+D(z,x))$.

  The ecological distance in some heterogeneous landscapes depends more on the genetic than geographic (Euclidean) distance.

  The term **ecological distance** was used also to compare species composition of two samples; cf. **biotope distance**.

- **Migration distance**, in Biogeography, is the distance between regular breeding and non-breeding areas within annual large-scale return movement of birds, fish and insects.

  The longest such recorded distance is an average of $70,900$ km pole-to-pole traveled each year by the Arctic tern. For a mammal, such record is $9,800$ km Brazil-Madagaskar traveled by a whale.

- **Migration distance**, in Cattle/Human Reproduction, is the distance in mm traveled by the vanguard spermatozoon during displacement in vitro through a capillary tube filled with homologous cervical mucus or a gel mimicking it. Sperm swim $1-4$ mm per minute. It estimate, under different specifications of incubation (temperature, duration, etc.), the ability of spermatozoa to colonize the oviduct in vivo.

  This term is also used in any measurements of directional biomotility using controlled migration; say, determining the molecular weight of unknown protein via its migration distance through a gel.

- **Penetration distance** is a general term used in (especially, biological) measurements for the distance from the given surface to the point where the concentration of the penetrating substance (say, a drug) in the medium (say, a tissue) had dropped to the given level; for example:

  Penetration distance of a drug in brain is the distance from probe surface to the point where the concentration is $\approx \frac{1}{2}$ its far-field value.

  During penetration of a macromolecular drug into the tumor interstitium, **tumor interstitial penetration** is the distance that drug carrier moved away from the source at a vascular surface.

  Penetration distance of chemicals into wood is the distance between the point of application and the 5 mm cut section in which the contaminants concentration was at least 3 % of the total.

  Forest edge-effect penetration distance is the distance at which species abundance ceased to be different to forest interior abundance.

- **Capillary diffusion distance**

  One of diffusion processes is **osmosis**, i.e., the net movement of water through a permeable membrane to a region of lower solvent potential. In the respiratory system (the alveoli of mammalian lungs), oxygen O_2 diffuses into the blood and carbon dioxide CO_2 diffuses out.

  Capillary diffusion distance is a general term used in biological measurements for the distance, from the capillary blood through the tissues to the mitochondria, to the point where the concentration of oxygen had dropped to the given low level.

  This distance is measured as, say, the average distance from the capillary wall to the mitochondria, or the distance between the closest capillary endothelial cell to the epidermis, or in percentage terms.

  For example, it can be the the distance where a given percentage (95% for maximal, 50% for average) of the fiber area is served by a capillary.

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 above operations (and so, distances) are defined similarly, but insert/delete work on a subset instead of a subsequence.

Let there is a **cost function** defined on each operation, and the **cost** of a sequence of editing operations is the sum of their costs.

- 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. It is **unit cost edit distance** if each operation costs 1.

- The **Selkow distance** on $\mathbb{T}_{rlo}$ is the minimum cost of a sequence of 3 editing operations turning one tree into another if insertions and deletions are restricted to leaves of the trees.
- The **constrained edit distance** on $\mathbb{T}_{rlo}$ is the minimum cost of a sequence of 3 editing operations turning one tree into another so that disjoint subtrees are 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.
- 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 **greatest agreement subtree distance** between **any two trees** is min. number of leaves removed to obtain a **common pruned tree** obtained from both trees by pruning leaves with the same label.

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** (swappings two subtrees that are adjacent to the same internal edge) needed to get T_1 from T_2.

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

BIOLOGICAL DISTANCE MODELS

- **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 in humans was studied via surnames.

  Speciation by force of distance is a speciation despite gene flow between populations. It was observed in **ring species** (2 species connected by gene flow through a chain of intergrading populations).

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

  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** $\sum_{1 \leq i \leq 12} |x_i - y_i|$ where z_i is the average temperature in the province z during i-th month, **migration distance** $\sum_{1 \leq i \leq 47} |x_i - y_i|$ where z_i is the percentage of people (living in the province z) born in i.

- **Long-distance dispersal** (or **LDD**) refer to the rare events of biological dispersal (esp. plants) on distances an order of magnitude greater than median **dispersal distance**. Together with **vicarience theory** (dispersal via land bridges) based on continental drift), LDD emerged as main factor of biodiversity and species migration patterns.

  Examples: invasive species, law biodiversity of microbs, cancer metastases, human colonization of Madagascar $\approx 2,000$ years ago.

  LDD is more important for the regional survival of some plants than local (median-distance) dispersal. LDD by wind currents explains strong floristic similarities of landmasses in southern hemisphere.

  Examples of other LDD vehicles: rafting by water (corals can traverse $40,000$ km during their lifetime), extreme climatic events, human transport, migrating birds (snails travel hundreds km inside bird guts: $\leq 10\%$ of eaten snails survive ≤ 5 hours until being ejected in feces).

- 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.

- The **distance running model** of antropogenesis (by Bramble and Lieberman) explains 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 **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.

- **Distance grooming model of language**

  In primates, being groomed produces mildly narcotic effects, because it stimulate the production of the body's natural opiates.

  Language, according to Dunbar, 1993, evolved in archaic Homo sapiens as more distance/time efficient replacement of social grooming. Their brain size expanded from 900 cm^3 in Homo Erectus to $1,300$ cm^3, and they lived in large groups (over 120 individuals) requiring cohesion. Language allowed to produce the reinforcing, social-bonding effects of grooming (through opiate production) at a distance and to use more efficiently the time, available for social interaction.

  Language achieves this through information transfer, gossip and emotional means (say, laughter, facial expression). Many primate species make extensive use of contact calls as, say, long-distance *pant-hoot* call of chimpanzees. Dunbar interpret such calls as a **grooming-at-a-distance**, from which language evolved.

- **Distance coercion model** (by Okada-Bingham) explains all unique properties of humans (complex symbolic speech, cognitive virtuosity, manipulation-proof transmission of fitness-relevant information, etc.) as elements and effects of extensive kinship-independent conspecific social cooperation in spite of conflicts of interest.

  Such non-kin cooperation can arise only from the pursuit of individual self-interest by animals able **to project coercive threat remotely**. The individual advantage of cooperation is a by-product of ongoing coercive threat conjointly with other group members. So, individuals display public behaviors seen as beneficial to other coalition members.

  Humans are the only animal with an innate biological capacity to project coercive threat remotely: to kill adult conspecifics with thrown projectiles from a distance, $18 - 27$ m by throwing a spear and up to 91 m by a bow. The model posits that this capacity briefly preceded the emergence of brain expansion and social support.

Comparing with Neanderthals, evidence of a huge number of injuries suggests that their hunting involved dangerously close contact with large prey animals. Moreover, their tools are rarely found more than 50 km from the source, while early modern humans maintained social networks over distances of up to 200 km.

Throwing and language capacities enabled humans to survive rapid climatic and environmental changes, to spread and to become the dominant large-scale (i.e., excluding insects and smaller) species on the planet. Humans are most efficient enforcers of cooperation (even relying mainly on indirect cues): our cognitive abilities expanded the range of situations in which cooperation can be favored.

Historical increases in the scale of human social cooperation could be associated with prior acquisition of a new coercive technology, for instance, the bow and agricultural civilizations, gunpowder weaponry and the modern state.

- **Body size rules**

  Payne et al., 2008: the **maximum size** of Earth's organisms increased by 16 orders of magnitude over the last 3.5 billion years. 75% of the increase happened in two leaps (about $1,900$ and $600-400$ Ma ago: appearance of eukaryotic cells and multicellularity) due leaps in O_2.

  Smith et al., 2010: the **maximum size of mammals** increased (from $10-100$ g) near-exponentially after the extinction of dinosaurs 65 Ma ago. It is leveled off after 40 Ma ago and remained nearly constant.

  By mean body size (67 kg now and 50 kg in Stone Age) **humans are a small megafauna** (i.e., ≥ 44 kg) species. A rapid average decline of $\approx 20\%$ in size-related traits was observed in human-harvested species. One of main human effects on nature is the decline of the apex consumers (top predators and large plant eaters).

Island rule is a principle that on islands, small mammal species evolve to larger while larger ones evolve to smaller.

Damuth, 1993, suggested that in mammals there is an optimum body size ≈ 1 kg for energy acquisition, and so, island species should, in the absence of usual competitors and predators, evolve to that size.

Insular dwarfism: reduction in size of large mammals when their gene pool is limited to very small environment (for example, islands).

Smaller animals need fewer resources and so are more likely to get past the breakpoint when population decline allows food sources to replenish

Island gigantism: size of animals isolated on an island increases dramatically over generations due the removal of constraints.

Abyssal gigantism: deep-sea species are larger of shallow-water ones. An adaptation for scarcer food resources, greater pressure and lower t^0?

The Galileo's **square-cube law**: as an object increases in size (linear dimension l), its volume V (and mass) increases as l^3 while surface area SA increases as l^2; so, the **ratio** $\frac{SA}{V}$ decreases.

For materials, high $\frac{SA}{V}$ speed up chemical reactions and processes minimizing free energy. Higher $\frac{SA}{V}$ permit to smaller cells gather nutrients and reproduce very rapidly. Smaller animals in hot and dry climates better loose heat through the skin and cool the body.

But lower $\frac{SA}{V}$ (and so, larger size) improves temperature control in unfavorable environments since smaller proportion of body being exposed results in slower heat loss or gain.

Bergmann-Mayr's rule is a principle that within a species, the body size increases with colder climate.

Allen's rule is a principle that animals from colder climates usually have shorter limbs than the equivalent ones from warmer climates.

Cope's rule: tendency of body size to increase over geological time. Large size enhances reproductive success, ability to avoid predators and capture prey, and improves thermal efficiency. In large carnivores, bigger species dominate better smaller competitors.

Cope's rule can follow from Bergmann's rule: species and lineages evolve toward larger sizes during episodes of climatic cooling.

Large body size favours the individual but renders the clade more susceptible to extinction via, for example, dietary specialization.

An **allometric law** is a relation between the size of an organism and the size of any of its parts or attributes. Examples follow.

Rensch's rule: in groups of related species, sexual size dimorphism is more pronounced in larger species.

Proportionalities of metabolic rate to $M^{\frac{3}{4}}$ (**Kleber's law**) and of breathing time to $M^{\frac{1}{4}}$, for body mass M, are allometric **power-laws**.

VISUAL, AUDITORY AND HAPTIC SPACES

1. Selected vision distances and size-distance phenomena
2. Distortion of sensual versus physical space
3. Distance cues and geographic distance biases

- **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 ($\approx 0.52^{\circ}$), the horizon moon appear to be about twice the diameter of the zenith moon. It could be cognitive: zenith moon appear approaching.

- **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 almost affinely connected (so, not metric, in general) space. (An **affine connection** is a linear map sending two vector fields into a third one.) There is evidence that if visial space admits a metric d, then d is a **projective metric**, i.e., $d(x, y) + d(y, z) = d(x, z)$ for any perceptually collinear points x, y, z.

  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 related **magnitude effect** (Weber-Fechner law): performance decreases for larger $\min(x, y)$

- In Gestalt Psychology, the **law of proximity** is that spatial or temporal proximity of elements may induce the mind to perceive a collective or totality.

- 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**, and **binaural differences**. The closer sound object is louder, wider, has more bass, high-frequencies, transient detail, dynamic contrast, more direct sound level over (and greater time delay until) its reflected sound.

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** (depending on the angle of optical axes of the eyes), **accommodation** (the state of eyes focus);

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

Examples of **geographic distance biases** are:

Observers are faster to respond to locations preceded by locations that were either close in distance or in the same region.

Distances are overestimated when they are near to a reference point.

Often subjective distance is assymmetrical as the perspective varies with the reference object (a small village versus a big city).

Routes segmented by features look longer than unsegmented routes.

Structural features (turns, barriers) breaking pathway into separate vistas increase subjective distance; is it the sum of vista distances?

Chicago-Rome illusion: belief that that some European cities are located far to the south of their actual location; in fact, Chicago and Rome are at the same latitude (42°).

Miami-Lima illusion: belief that east cost US cities are to the east of west cost South America cities; in fact, Miami is 3° west of Lima.

REAL-WORLD BIOLOGICAL DISTANCES

1. Selected human and animal distances
2. Length magnitudes in biology and extent of biosphere
3. Selected medical distances

- **Distances between people** (types of informal space, by Hall, 1969):

  intimate distance for embracing, whispering, touch (15 – 45 cm),

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

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

  public distance for public speaking (over 3.6 m).

  The distance which is appropriate for a given social situation depends on culture, gender and personal preference.

  For an average westerner, personal space is about 70 cm in front, 40 cm behind and 60 cm on either side.

  In interaction between strangers, the interpersonal distance between women is smaller than between woman and man.

  Distancing behavior of people can be measured by the **stop distance**, when the subject stops an approach since feel uncomfortable.

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

The **alert distance**: the distance from the disturbance source (say, a predator or dominating co-specific) when the animal changes its behavior (say, turns towards) in response to an approaching treat.
The **catching distance**: on which the predator can strike a prey.

Example of **distance estimation** by animals: the velocity of the mantid's head is constant during peering; so, the distance to the target is inversely proportional to the velocity of the retinal image.

A **distance pheromone** is a soluble and/or evaporable substance emitted by an animal, in order to send a message (on alarm, sex, food trail, recognition, etc.) to other members of the same species.

Gaze following: great apes, ravens, canids follow another's gaze direction into distant space, moreover, geometrically behind obstacle.

All dead distance: maximum distance from the toxicant source within which no targeted insects are found alive after a fixed period.

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 1.7×10^{-8} (Porsine circovirus 2) to 4.4×10^{-7} (Mimivirus);

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: $1.8 \times 10^{-6} - 18 \times 10^{-6}$);

2×10^{-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 33.6: height of the tallest animal, the giraffe, and length of a blue whale, the largest animal;

Among extinct dinosaurs, 18 and 40: height of **Sauroposeidon** and length of **Amphicoelias**; **Gigantopithecus** was an ape ≈ 3 m tall;

3.5 and 2.72: height of Neolitic *Giant of Castelnau* and Robert Wadlow (1918-1940);

The lion's mane jellyfish can have tentacles of 37 m; **Colossal Squid** up to 14 m with nerve cell 12 m were recorded;

longest known animal, bootlace worm **Lineus longissimus**, can reach a length of 55 m; the parasitic tapeworm **Diphyllobothrium klebanovski** can reach 19.4 m in the internal organs of a human;

115.3: height of the world's tallest tree, a sequoia Coast Redwood (while the circumference of the thickest recorded tree is 35.9 m);

60 m: the record-sized stem of green algae, **Macrocystis pyrifera**; it and a bamboo can grow 46 and 90 cm a day;

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

8.9 km^2: area of a genetically constant mycelium of fungus **Armillalia ostoyae** living in U.S. state Oregon, 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, entire biosphere (“Gaia”) is an extended organism?

- **Extent of Earth's biosphere**

  The known range for active life: $[-20°C, 122°C]$, pressure $[0.05, 1300]$ bar and acidity/alkalinity $[1, 12]$ on the pH scale.

  Fungi and bacterial spores were found at an altitude 77 km (at -69°C and 10^{-5} bar) and viable yet non-culturable bacteria, at 20-70 km.

  Jones and Lineweaver, 2010, estimated the depth $5-20$ km of the 122°C isotherm and the altitude $10-15$ km (a *tropopause* boundary of the vertical movement of particles) to be the boundaries of active life.

  Nussinov and Lysenko, 1991, estimated -30 km and 100 km (*Kármán line*) to be the boundaries of general biosphere.

  There are permanent human habitations at mean annual temperatures of 34.4°C, -46°C and at an altitude of 5.5 km. Some frogs, turtles and snake *Thamnophis sirtalis* survive the winter by freezing solid. A brine shrimp *Artemia* tolerate salt amounts 50%.

Proponents of **panspermia hypothesis** (that life propagates via extremophile bacteria surviving in space) expect microbe density to be $10^{-6} - 10^{-4}$ cells/m^3 at altitude 500 km. Large amount is expected at altitude of ISS ($278 - 460$ km). $10^{14} - 10^{16}$ microorganisms per annum are ejected from Earth at survivable t°.

No ubiquitous ultrasmall bacteria (< 0.1 micron) were found in the stratosphere, but large ($5 - 100$ micron) *Bacillus* and fungal spores. Such viable, but non-culturable, microorganisms could be incoming from space. The solar system could be surrounded by an expanding biosphere of radius > 5 parsecs containing $10^{19} - 10^{21}$ microorganisms.

The life on Mars, if any, (but not on Titan) is expected to be of the same origin as on Earth. Interstellar panspermia, when the Sun passes a starforming cloud, and even intergalactic panspermia, when galaxies collide, are debated. But on a cosmic scale, even enthusiasts of panspermia see it as a local, “a few megaparsecs”, phenomenon.

- **Selected medical distances**

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

  Inter-proximal distance: spacing distance between adjacent teeth;

  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 (the average is about 6.5 cm for men and 5.5 cm for women).

  Inter-cornual distance: the distance between uterine horns (2-4 cm).
  C-V distance: the distance between clitoris and vagina (2.5-3 cm).

  AGD (anogenital distance: the length of the **perineum** (region between anus and genital area). For a male it is 5 cm in average (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 **stroke distance**: the distance a column of blood moves during each heart beat, from the aortic valve to a point on the aortic arch.

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? The **tumor diameter** is the greatest vertical diameter of any section; **tumor growth** is the geometric mean of its 3 perpendicular diameters.

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.

- In Face Recognition, are used sets of (vertical/horizontal) **cephalofacial dimensions**, i.e., distances between **fiducial** (used as a fixed standard of reference for measurement) facial points. The distances are normalized, say, with respect of **inter-pupillary distance** for horizontal ones.

  For example, the following 5 independent facial dimensions are derived by Fellous, 1997, for facial gender recognition:

  distance E between external eye corners,

  nostril-to-nostril width N,

  face wigth at cheek W

  and two vertical distances: eye-to-eyebrow distance B and

  distance L between eye midpoint and horizontal line of mouth.

  In above terms, "femaleness" relies on large E, B and small N, W, L.

There is no object so large ... that at great distance from the eye it does not appear smaller than a smaller object near. (Leonardo da Vinci)

Where the telescope ends, the microscope begins. Which of the two has the grander view? (Victor Hugo)

Telescopes and microscopes are designed to get us closer to the object of our studies. Thats all well and good. But its as well to remember that insight can also come from taking a step back. (New Scientist, 31 March 2012)

The closer the look one takes at a world, the greater distance from which it looks back. (Karl Kraus)

Nature uses only longest threads to weave her patterns. (Richard Feynman)

And so man, as existing transcendence abounding in and surpassing toward possibilities, is a creature of distance. Only through the primordial distances he establishes toward all being in his transcendence does a true nearness to things flourish in him. (Martin Heidegger)