Hamming-like distances for ill-defined strings in linguistic classification

Luca Bortolussi, Andrea Sgarro
DMI, University of Trieste, 34100 Trieste Italy
`luca @ dmi.units.it`, `sgarro @ units.it`

Abstract. Ill-defined strings often occur in soft sciences, e.g. in linguistics or in biology. In this paper we consider ℓ-length strings which have in each position one of the three symbols 0 or *false*, 1 or *true*, $\flat$ or *irrelevant*. We tackle some generalisations of the usual Hamming distance between binary crisp strings which were recently used in computational linguistics. We comment on their metric properties, since these should guide the selection of the clustering algorithm to be used for language classification. The concluding section is devoted to future work, and the string approach, as currently pursued, is compared to alternative approaches.

****

Keywords: string distances, Hamming distance, fuzzy distances, computational linguistics, linguistic classification, incomplete knowledge

****

1 Ill-defined strings

Strings in linguistics or in biology, and more generally in *soft* sciences, are often *ill-defined*, and so the *crisp* tools of traditional mathematics fall short of the task. Ill-defined strings were recently used in papers devoted to the classification of languages following syntactic criteria, rather than lexical (LGG05), and *ad-hoc* distances were used in order to cluster such languages. More specifically, certain syntactic features are chosen, ℓ say, and to each language an ℓ-length string is associated, which has a 1 or a 0 in position i according whether feature i is present or, respectively, absent; in moot situations, however, a third symbol $\flat$ is written down.
Basically, distances as in (LGG05) fit into a family of generalised Hamming distances, which are parametrised by a parameter $w \in [0,1]$. For $w = 1/2$ one re-finds the so-called *fuzzy* Hamming distance, introduced independently by several authors into the fuzzy set literature: actually, the fuzzy Hamming

distance had already been used by the linguist Ž. Muljačić for a classification of Romance languages back in 1967, a remarkable fact, indeed ((Mul67), cf. also (Sga77) or (SB00)). The analogy to fuzzy distances, however, should not be over-emphasised: in our case, the meaning of $\flat$ is *"it doesn't really make sense"*, while in the fuzzy case the feature does make sense in itself, but it is not clear to what degree it is present (it is true); it is not surprising, therefore, that the generalised Hamming distance as pursued below and the fuzzy Hamming distance boil down to the same thing *only* when $w = 1/2$. We shall come back to this point in section 4, but below the term "fuzzy" will not be used for our distance, since it would be misleading: one is dealing here with yet another facet of representing and managing incomplete knowledge.
In this paper we investigate the metric properties of the family of distances d_w; in particular, we check whether the triangle inequality is or is not violated: bad news are in store for $w < 1/2$; cf. section 2 (by the way the fuzzy Hamming distance proper is *always* triangular, cf. (Sga77)). The variant preferred by (LGG05), called below δ, corresponds to $w = 0$, but it is normalised w.r. to "sound" (crisp) components only, and is definitely unruly; cf. Section 3. All this should be made good use of before choosing the clustering algorithm to be selected: that is why in Section 2 we check the triangular property with such fastidiousness. Specifically, in (LGG05) the authors use agglomerative hierarchical clustering algorithms, which work by merging at each step the two closest clusters. The delicate point here is the definition of the distance between two clusters, given the pairwise distances among their members. The method chosen in (LGG05), known as UPGMA (SS73), defines the distance between two groups as the average of the pairwise distances of their members. It is precisely the averaging that is a questionable operation when the distance is not a (pseudo)metric, as underlined by several authors, cf. (BS93). Actually, other aggregation mechanisms may be used instead of the average, see the final section 6.
Section 4 tackles dependencies between features; cf. also Section 6 on future work. Section 5 has a more matter-of-fact nature, and focuses on actual data as processed in (LGG05): it will turn out that the triple Greek - Romanian - Norwegian is especially worrying from a metric point of view. The concluding section 6 envisages future work; the string approach, as currently pursued, is compared to alternative approaches based on Q-matrices and tree distances, such as to better account for logical dependencies between the ℓ features, without having to resort to ANOVA-type techniques of feature reduction, whose adequacy here is dubious.

2 Metric properties of the generalised Hamming distance

If x and y are two binary string of the same length ℓ, their Hamming distance $d_H(x,y)$ is the number of positions where they are different. As well-known, the Hamming distance is a genuine metric distance; in particular, it verifies the *triangle inequality* $d_H(x,z) + d_H(z,y) \geq d_H(x,y)$ for any three binary strings x, y and z. We add a third symbol, called the *blank* symbol $\flat$, to obtain the ternary alphabet $\mathcal{A} = \{0, 1, \flat\}$. For any x, we denote by $b(x)$ the number of its blanks, i.e. the number of occurrences of the "odd" symbol $\flat$ in x. We say that a string is *ill-defined* whenever it has at least one blank, $b(x) \geq 1$. Instead, if $b(x) = 0$ we say that x is *crisp*; $\ell - b(x)$ is the number of positions where x is crisp. If $b = b(x,y)$ is the number of positions where *at least* one of the two strings has a blank, one has $b(x,y) \geq b(x,x) = b(x) \geq 0$; $\ell - d - b$ is the number of crisp coincidences.

Let w be a *weight* belonging to the real interval $[0,1]$; we extend the usual Hamming distance between crisp binary strings to the following *extended distance* $d_w(x,y) = d_w(x_1x_2 \ldots x_\ell, y_1y_2 \ldots y_\ell)$ between ill-defined ternary strings:

Definition 1 *Set* $d_w(x,y) = \sum_i d_w(x_i, y_i)$ *with the* $d_w(x_i, y_i)$*'s as in the following matrix:*

$d_w(x_i, y_i)$	0	1	$\flat$
0	0	1	w
1	1	0	w
$\flat$	w	w	w

The *balanced* option $w = 1/2$ corresponds (at least formally) to the fuzzy Hamming distance as covered in the literature when the logic involved is ternary, and the degree of truth of the statement ”in position i the symbols x_i and y_i are distinct" belongs to $\{0, 1/2, 1\}$. We find it convenient to deal separately with the limit cases $w = 0$ (blanks always yield crisp equalities) and $w = 1$ (blanks always yield crisp inequalities), whatever the symbol they are matched with. Beside the balanced option $w = 1/2$, the options $w = 0$ as basically used in (LGG05) and $w = 1$ appear to be the most appealing for applications; the general case of any w will be soon found by convexity; cf. below. Clearly, if $w > w'$ one has $d_w(x,y) \geq d_{w'}(x,y)$ with equality iff x, y are both crisp, when everything boils down to the usual (binary) Hamming distance. We start with $w = 0$.

Formal properties of $d_0(x,y)$:

- $d_0(x,y)$ is non-negative and symmetric: $d_0(x,y) = d_0(y,x) \geq 0$;

- $d_0(x,y) \geq d_0(x,x) = 0$;
- $d_0(x,y) = 0$ iff x and y coincide in those positions where there are no blanks;
- $d_0(x,y) \leq \ell$;
- $d_0(x,y) = \ell$ iff x and y are both crisp and they differ in each position;

This is straightforwardly proved (iff means if and only if). *The triangle inequality may fail to hold*; we deepen this point in theorem 1, after giving some notation.
Fix a position i, $0 \leq i \leq \ell$; the following might occur:

α) there is no blank, and $x_i = y_i \neq z_i$;

β) all three symbols occur, and z_i is crisp

ξ) all three symbols occur, and z_i is a blank

Whatever ℓ, let $\alpha = \alpha(x,z,y)$, $\beta = \beta(x,z,y)$ and $\xi = \xi(x,z,y)$ denote (also) the number of positions of type α, β, ξ, respectively (the slight notational ambiguity turns out to be convenient).

Theorem 1

1. $d_0(x,z) + d_0(z,y) - d_0(x,y) = 2\alpha + \beta - \xi$
2. $d_0(x,z) + d_0(z,y) - d_0(x,y) \geq -b(z)$
3. *For given z, the lower bound holds with equality iff x_i and y_i are crisp and different in those positions i where z has a blank, while in those positions i where z_i is crisp either x_i and y_i are both blank, or there is at least an equality involving z_i.*

Proof. Assume for the moment $\ell = 1$; then d can be either 0 or 1. The triangle inequality is *not* verified with equality, in the case $d_0(x,z) + d_0(z,y) = 2$, $d_0(x,y) = 0$, which corresponds to α, in the case $d_0(x,z) + d_0(z,y) = 1$, $d_0(x,y) = 0$ which corresponds to β (the case $d_0(x,z) + d_0(z,y) = 2$, $d_0(x,y) = 1$ cannot hold), and in the case $d_0(x,z) + d_0(z,y) = 0$, $d_0(x,y) = 1$, which corresponds to ξ. Given the additive nature of of our distance, this is enough to prove the equality. As for the lower bound, clearly $\Delta \geq -\xi \geq b(z)$. As for equality in the bound, one has to choose x, y so that case ξ always holds where z_i is a blank, while cases α and β never hold where it is crisp. ■

The triangle inequality *fails* whenever $2\alpha + \beta < \xi$; the lower bound $-b(z)$ can be met with equality even if x, y are constrained to be crisp. Observe

that the triangle inequality certainly holds for all x, y iff the "intermediate" string z is crisp.
Let us turn to the limit case $w = 1$; while in the case of d_0 a blank contributed for 0, now it contributes for 1, whatever symbol it is matched with.

Formal properties of $d_1(x, y)$:

- d_1 is non-negative and symmetric, $d_1(x, y) = d_1(y, x) \geq 0$;
- $d_1(x, y) \geq d_1(x, x) = b(x) \geq 0$;
- $d_1(x, y) = 0$ iff x and y are crisp and coincident;
- $d_1(x, y) \leq \ell$
- $d_1(x, y) = \ell$ iff x and y are differ in each position where they are both crisp;
- the triangle inequality holds: $d_1(x, z) + d_1(z, y) \geq d_1(x, y)$ (cf. Theorem 2)

Below let $\sigma = \sigma(x, y, z)$ and $\eta = \eta(x, y, z)$ be the number of positions where:

σ) x_i, y_i are crisp and coincide while z_i is a blank,

η) there are two or even three blanks, respectively, while $\alpha = \alpha(x, y, z)$, $\beta = \beta(x, y, z)$ and $\xi = \xi(x, y, z)$ are as above.

Observe that $[\alpha + \sigma]$ gives the number of positions where x, y are crisp and coincident, z is different whether crisp or blank, while $[\beta + \xi]$ gives the number of positions where all the three distinct symbols occur in x, y, z.

Theorem 2

1. $d_1(x, z) + d_1(z, y) - d_1(x, y) = 2\,[\alpha + \sigma] + [\beta + \xi] + \eta \; \geq 0.$
2. *The triangle inequality holds with equality iff in each position there is at least a crisp coincidence involving z; in particular, to have equality z must be crisp.*

Proof. The triangle inequality holds with equality iff $\alpha = \sigma = \beta = \xi = \eta = 0$; one can resort to a direct check, as in theorem 1. ■

Now, $d_w(x, y)$ is a weighted average of d_0 and d_1:

$$d_w(x, y) \; = \; (1 - w)\, d_0\big(x, y\big) \; + \; w\, d_1\big(x, y\big) \; , \; 0 \leq w \leq 1$$

In this case a blank contributes for w whatever symbol it is matched with, inclusive of itself. For $w = 1/2$, which corresponds to the usual arithmetic average, one re-obtains the "old" fuzzy Hamming distance $d_F = d_{1/2}$. For $w \geq 1/2$, $d_w(x, y)$ verifies the triangle inequality; the following statements are soon derived from what precedes.

Formal properties of $d_w(x,y)$, $0 < w < 1$:

- d_w is non-negative and symmetric, $d_w(x,y) = d_w(y,x) \geq 0$;
- $d_w(x,y) = 0$ iff x and y are crisp and coincident;
- $d_w(x,y) \geq d_w(x,x) = w\,b(x) \geq 0$;
- $d_w(x,y) \leq \ell$
- $d_w(x,y) = \ell$ iff x and y are both crisp and they differ in each position;

The triangle inequality: $d_w(x,z) + d_w(z,y) \geq d_w(x,y)$ holds only for $w \geq 1/2$, as now shown:

Theorem 3

1. $d_w(x,z) + d_w(z,y) - d_w(x,y) = 2\alpha + \beta + (2w-1)\xi + 2w\sigma + w\eta$
2. *If* $w < 1/2$ *one has* $d_w(x,z) + d_w(z,y) - d_w(x,y) \geq -(1-2w)b(z)$,
3. *Equality holds iff x_i and y_i are crisp and different in those positions i where z has a blank, while there is at least a coincidence with z_i in those positions i where z_i is crisp.*
4. *If* $w \geq 1/2$ *one has* $d_w(x,z) + d_w(z,y) - d_w(x,y) \geq 0$.
5. *If $w = 1/2$, equality holds iff in each position i either there is a crisp coincidence involving z_i, or x_i and y_i are crisp and different, while z_i is a blank. If $w > 1/2$, equality holds, as in the case of d_1, iff in each position i there is a crisp coincidence involving z_i.*

Proof. It will be enough to discuss conditions for equality, the rest being an obvious corollary of theorems 1 and 2. If $w < 1/2$, to have equality in the lower bound a necessary condition is clearly to have equality in the lower bound of theorem 1; however, one must further ensure $\sigma = \eta = 0$. Now, if z_i is a blank, the fact that ξ holds true implies that σ and η are both false, and nothing new is required; if z_i is crisp, unlike in theorem 1, x_i and y_i cannot be both blank, because this would imply $\eta > 0$. If $w > 1/2$ nothing changes with respect to theorem 2; instead, for $w = 1/2$, one can afford to have $\xi > 0 = \sigma = \eta$. ■

Remark 1. In the case of d_w with $w \geq 1/2$, the axioms of a pseudometric distance are violated only because one can have a positive value for the "self-distance" $d_w(x,x)$. This drawback is mild indeed, because a pseudometric can soon be obtained from d_w by "forcing" such self-distances to be zero; in practise any string x is forced to belong to all of its neighbourhoods.

Remark 2. We shall say that *z is between x and y* when in each position either $x_i \leq z_i \leq y_i$ or $y_i \leq z_i \leq x_i$ w.r. to the (unusual) ordering $0 < \flat < 1$. It is well-known that the triangle inequality for crisp Hamming distances is met with equality iff z is between x and y; instead, this condition is only *necessary* in the case of d_w with $w \geq 1/2$. For *sufficiency* one must add the requirement that z should be crisp in the case of d_w with $w > 1/2$; in the case of $d_{1/2}$, one must add the requirement that x, y should be crisp in those positions where z has a blank. Incidentally, (cf theorems 2 and 3) if the triangle inequality is met with equality for d_w with $w > 1/2$, then *a fortiori* it is met with equality for $d_{1/2}$.

3 The variable-normalisation approach

Sometimes, rather than using the *absolute* Hamming distance, one prefers to count the *percentage* of crisp differences: similarly, in (LGG05) the authors used a normalised variant of d_0, by counting the percentage of crisp differences over those positions where both strings are "sound" (are crisp):

$$\delta(x,y) = \frac{d_0(x,y)}{\ell - b(x,y)}$$

This distance is undefined whenever $b(x,y) = \ell$, and so in the sequel we will tacitly assume that the two strings in $\delta(x,y)$ have at least one crisp position in common, a condition easily met in practise. To have a distance which compares fairly to those in the previous sections, one may de-normalise multiplying by ℓ, and so obtain $\Delta(x,y) = \ell\,\delta(x,y)$. A serious objection is that the contribution to δ or Δ of a crisp difference in position i between x and y, or between x and z, is not the same unless $b(x,y)$ and $b(x,z)$ are equal. E.g. take $x = \underline{0}0$, $y = \underline{0}1$, $z = \underline{\flat}1$, where underbars denote runs of the same symbol of length $\ell - 1$. The two couples x, y and x, z crisply differ in just one position, but $\delta(x,y) = 1/\ell$, while $\delta(x,z)$ is equal to 1, i.e. to its upper bound.

Of course the triangle property is easily violated: just take z with blanks in a sub-string where x and y are crisp and different. E.g. take $x = 00...0$, $z = 0\flat...\flat$, $y = 01...1$. One has $\delta(x,z) = 0$, $\delta(z,y) = 0$, $\delta(x,y) = \frac{\ell-1}{\ell} \approx 1$. Even if mathematically rather weak, δ did perform quite well in (LGG05), an indication that one is dealing with stable linguistic structures which emerge even under moot classification criteria.

4 Metric properties on subspaces

While in the linguistic classifications of Muljačić (Mul67) the fuzzy truth value 1/2 was used in positions i where feature i was only weakly or partially

present, the situation is quite different in our case: here it happens that there are strong logical dependencies between the ℓ features, henceforth called *structural* dependencies, and one feature, j say, might "exist" only when another feature i has the binary value 1, else it doesn't make sense to talk about j. Such structural dependencies are accounted for through blanks: if feature number j is meaningless when feature number i has the value 0 rather than 1, say, one puts a blank in position j (the crisp value of feature i implies whether feature j does or does not make sense). Of course, one might have "deeper" dependencies: there might be a further feature u which makes sense only if feature j is equal to 1. An implication is that certain ternary strings are "structurally inadmissible", say string 010 or string $0\flat1$, assuming $i = 1, j = 2, u = \ell = 3$ Now, the basic piece of information derived from section 2 is that the triangular properties falls for $w < 1/2$: since not all ternary strings are structurally admissible, the hope is the the triangular property might be recovered on relevant subspaces even if $w < 1/2$. The bad news given by theorem 4 is that, even if one has only logical dependencies of length at most one (a bold assumption thinking of strings as used in (LGG05)), the triangular property is soon lost.
Let us go back to d_w; we now discuss a subspace of strings, which is defined so as to cope with structural implication of the type just hinted at, even if limited to depth one. Let $\mathcal{P} \subset \{1, \ldots, n\}$ be a (non-void) subset of positions, called henceforth *strong*, while the remaining positions are called *weak*; blanks may occur *only* in weak positions. Further, to each strong positions i we associate a set of weak positions (possibly empty) $\phi(i)$, such as to have: the value of the bit in position i determines whether the entries in positions $\phi(i)$ are all blank or all crisp (each weak position is obtained by exactly *one* strong position). Below, without real restriction, we shall assume that the strong bit which implies blanks is always 0, while bit 1 in position i implies that the corresponding $\phi(i)$ weak positions are all crisp. For fixed ϕ let us consider the subspace $\mathcal{S}$ made up of all the strings which verify the corresponding constraint. One has:

Theorem 4 *Take $w < 1/6$; on the specified subspace $\mathcal{S}$, the distance d_w is triangular iff $|\phi(i)| \leq 2$ for all strong positions i. If $w < 1/2$ there is an integer k^* such that the triangular inequality falls as soon as $|\phi(i)| \geq k^*$ for at least one strong position i.*

Proof. For $\ell \geq 4$ and $|\phi(1) = k \doteq \ell - 1|$, the triangle inequality is violated by the three strings $x = 1\underline{0}$, $z = 0\underline{\flat}$, $y = 1\underline{1}$ when $w < \frac{k-2}{2k}$; the latter number is equal to 1/6 for $k = 3$ and tends to 1/2 as k goes to infinity (an underbar denotes a k-length run of the same symbol). This argument is enough to prove that the triangular inequality falls in the situations specified by the theorem. Now, set $\alpha = 0\flat$, $\beta = 10$, $\gamma = 11$ for $\ell = 2$, and $a = 0\flat\flat$, $b = 100$, $c = 101$, $d = 110$, $e = 111$ for $\ell = 3$; one verifies by a check that the triangle inequality always holds. For $\phi(i) \leq 2$, suppress weak positions and write in the corresponding strong positions α, β, γ or a, b, c, d, e, as the case may be. In practise, we code ℓ-strings over $\{0, 1, \flat\}$ to shorter strings of length $\ell - \sum |\phi(i)|$ over $\{0, 1, \alpha, \beta, \gamma, a, b, c, d, e\}$, so as to keep (additive) distances (the sum is extended to all strong positions i). ■

Remark 3. On $\mathcal{S}$ the metric implication $d_w(x, y) = 0 \Rightarrow x = y$ holds true even for $w = 0$ by just assuming that there are weak positions, so that $\mathcal{S}$ is *strictly* included in the space of all ternary strings of length ℓ. Actually, if $d_0(x, y) = 0$ and in position $j \in \phi(i)$ there is a blank matched with a crisp bit, this gives no contribution to the distance, but then in position i the two strings must have distinct crisp bits, and so cannot coincide.

5 Triangles on real data: empirical results

To resume: for $w < \frac{1}{2}$ the triangular property falls even on limited subspaces, and so the use of standard clustering algorithms is at risk. A further possibility remains open: that our real data verify (always or at least mostly) the triangular property. We have checked the 24 natural languages classified in (LGG05), which give rise to $\binom{34}{3} = 2024$ distinct triangles; each of these might violate the triangle inequality in up to three ways, according to which languages is chosen as the "intermediate" one. We have counted triangles which are faulty in the sense that the triangle inequality is violated in at least one way.

Let us begin by taking the variable-normalisation distance δ, or equivalently Δ. The number of faulty triangles turns out to be 78, i.e. $\approx 3.85\,\%$. If one takes instead the unnormalised distance d_0 the number of faulty triangles slightly decreases: one finds 73 of them, which corresponds to $\approx 3.61\,\%$.

We have made a further check. As w increases one approaches a situation when the triangle inequality does hold. Actually, there is monotonicity, as soon proved: if $d_w(x, z) + d_w(z, y) \geq d_w(x, y)$, and $v > w$, then $d_v(x, z) + d_v(z, y) \geq d_v(x, y)$. We have checked $w > 0$ to see at which threshold faulty triangles disappear: it turns out that for $w = 0.16$ there are no faulty triangles left over. For $w = 0.15$ a single faulty triangle survives, i.e. modern Greek, Romanian, Norwegian when one takes Romanian as the intermediate language. In a way, it is "quicker" to move from Greek to Norwegian heading off to Romanian, rather that taking the unconstrained direct way.

6 Future work: strings vs. trees

Robust linguistic structures should emerge under different (but reasonable) classification criteria. Therefore it is interesting to take advantage of the fact that we have a family of distances d_w, $w \in [0, 1]$, rather than a single distance d_0, to see what changes as w spans its range. One might introduce the parameter w also in the case of variable-normalisations; in this case, however, one should presumably think of a weighted normalisation of the type

$$\delta_w(x, y) \;=\; \frac{d_w(x, y)}{\ell - (1 - w)\, b(x, y)} \;=\; \frac{d_w(x, y)}{[\ell - b(x, y)] + wb(x, y)}$$

which would have δ and d_1/ℓ at its extremes. Also, as sometimes one does (RLKB03), one might think of forcing triangularity by "optimally twisting" faulty triangles, a way-out which is tempting, since after all their percentage is comparatively low.

Interesting insights on the data may also emerge by using different strategies for calculating inter-cluster distances. For instance, the distance between two groups of languages can be computed by taking the minimum of the pairwise distances among their elements (*single linkage clustering*), or the maximum (*complete linkage clustering*). Other strategies can use the median or the distance between representatives of the clusters, like the mean or the median element (i.e. the elements minimising respectively the mean or the median distance within each group). In general, we expect that methods relying on "average-like" aggregators should work better, as they tend to smooth out the effect of "borderline" languages, which may unnaturally inflate or deflate the distance between two clusters. For a thorough review of clustering methods, see e.g. (JMF99).

In the current approaches taken by linguists, features, be they lexical as in (Mul67), or syntactic as in (LGG05), are always arranged in strings, only ternary rather than binary. Dependencies are a problem, and standard statistical methods for feature reduction appear to be of dubious usefulness, since it is not clear why the choice of features should by guided by statistics, let alone by multi-variate statistics for normal samples. A promising step forward is turning to Q-matrices (i.e., to infinitesimal generator matrices of Continuous Time Markov Chains (Nor97)), which have already been used with success in biological contexts (Fel04; SvH03). Up to this point the string approach has been taken for granted, but one may think of a bold change. In Section 4 features have been implicitly set at different depths to take care of structural dependencies, and so one might turn to trees rather than strings: this would make depth differences quite explicit. In (Bil05) distances between trees are reviewed, which might be used in our case.

Fabio Rossi (1943-2005) We express our gratitude and our friendship to Fabio Rossi, whose catching enthusiasm for mathematical research and mathematical teaching will remain with us forever. In his last years Fabio turned to bio-mathematics and bio-informatics and contributed to the launching of the International Summer School BCI on Biology, Computation and Information, which has now reached its 4th edition.

References

[Bil05] P. Bille. A survey on tree edit distance and related problems. *Theoretical Computer Science*, 337:217–239, 2005.

[BS93] J. Boberg and T. Salakoski. General formulation and evaluation of agglomerative clustering methods with metric and non-metric distances. *Pattern Recognition*, 26(9):1395–1406, 1993.

[Fel04] J. Felsenstein. *Inferring Phylogenies.* Sinauer, 2004.

[JMF99] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: a review. *ACM computing surveys*, 31(3), 1999.

[LGG05] G. Longobardi, C. Gianollo, and C. Guardiano. Syntactic measuring of language relatedness. In A. Kidwai et al., editor, *Proceedings of V Asian GLOW*, 2005.

[Mul67] Ž. Muljačić. Die Klassifikation der romanischen Sprachen. *Roman. Jahrbuch*, 1967.

[Nor97] J. R. Norris. *Markov Chains.* Cambridge University Press, 1997.

[RLKB03] V. Roth, J. Laub, M. Kawanabe, and J.M. Buhmann. Optimal clustering preserving embedding of nonmetric proximity data. *IEEE Trans. on pattern analysis and machine intelligence*, 2003.

[SB00] A. Sgarro and M. Borelli. *Finite versus infinite. Mathematical contributions to an eternal dilemma*, chapter A possibilistic distance for sequences of equal and unequal length, pages 27–38. Lecture Notes in Discrete Mathematics and Theoretical Computer Science. Springer-Verlag, Singapore, 2000.

[Sga77] A. Sgarro. A fuzzy Hamming distance. *Bulletin Math. de la Soc. Sci. Math. de la R.S. de Roumanie*, 21:137–144, 1977.

[SS73] P.H. Sneat and R.R. Sokal. *Numerical Taxonomy.* Freeman, London, UK, 1973.

[SvH03] K. Strimmer and A. von Haeseler. *The Phylogenetic Handbook*, chapter Nucleotide Substitution Models, pages 72–87. Cambridge University Press, 2003.