ON ORDER EQUIVALENCES BETWEEN DISTANCE AND SIMILARITY MEASURES ON SEQUENCES AND TREES

Martin Emms and Hector-Hugo Franco-Penya
School of Computer Science and Statistics, Trinity College, Dublin, Ireland

Keywords: Similarity, Distance, Tree, Sequence.

Abstract: Both 'distance' and 'similarity' measures have been proposed for the comparison of sequences and for the comparison of trees, based on scoring mappings, and the paper concerns the equivalence or otherwise of these. These measures are usually parameterised by an atomic 'cost' table, defining label-dependent values for swaps, deletions and insertions. We look at the question of whether orderings induced by a 'distance' measure, with some cost-table, can be dualized by a 'similarity' measure, with some other cost-table, and vice-versa. Three kinds of orderings are considered: alignment-orderings, for fixed source S and target T, neighbour-orderings, where for a fixed S, varying candidate neighbours T_i are ranked, and pair-orderings, where for varying S_i, and varying T_j, the pairings $\langle S_i, T_j \rangle$ are ranked. We show that (1) alignment-orderings by distance can be dualized by similarity, and vice-versa; (2) neigbour-ordering and pair-ordering by distance can be dualized by similarity; (3) neighbour-ordering and pair-ordering by similarity can sometimes *not* be dualized by distance. A consequence of this is that there are categorisation and hierarchical clustering outcomes which can be achieved via similarity but not via distance.

1 TREE DISTANCE AND SIMILARITY

In many pattern-recognition scenarios the data either takes the form of, or can be encoded as, sequences or trees. Accordingly, there has been much work on the definition, implementation and deployment of measures for the comparison of sequences and for the comparison of trees.

These measures are sometimes described as 'distances' and sometimes as 'similarities'. We are concerned in what follows in first distinguishing between these, and then with the question whether orderings induced by a 'distance' measure can be dualized by a 'similarity' measure, and vice-versa. To some extent this can be seen as applying the same kind of analysis to sequence and tree comparison measures as has been applied to set and vector comparison measures (Batagelj and Bren, 1995; Omhover et al., 2005; Lesot and Rifqi, 2010).

From statements such as the following

> *To compare RNA structures, we need a score system, or alternatively a distance, which measures the similarity (or the difference) between the structures. These two versions of the problem score and distance are equivalent.*
>
> (Herrbach et al., 2006)

which are not uncommon in the literature (Alves et al., 2002; Kondrak, 2003; Bose and van der Aalst, 2009), it would be easy to gain the impression that similarity and distance (on sequences and trees) are straightforwardly interchangeable notions. In section 1.1 several distinct kinds of equivalence are defined. Sections 2, 3.1 and 3.2 then show that while some kinds of equivalence hold, others do not.

To begin we need to clarify what we will mean by 'distance' and 'similarity' on sequences and trees. Because sequences can be encoded as vertical trees it suffices to give definitions for trees. Tai first proposed a tree-distance measure (Tai, 1979). Where S and T are ordered, labelled trees, a *Tai* mapping $\alpha : S \mapsto T$ is a *partial*, *1-to-1* function from the nodes of S into the nodes of T, which respects *left-to-right order* and *ancestry*[1]. For the purpose of assigning a score to such a mapping it is convenient to identify three sets:

$\mathcal{M}$	the $(i,j) \in \alpha$: the 'matches' and 'swaps'
$\mathcal{D}$	the $i \in S$ s.t. $\forall j \in T, (i,j) \notin \alpha$: the 'deletions'
$\mathcal{I}$	the $j \in T$ s.t. $\forall i \in S, (i,j) \notin \alpha$: the 'insertions'

Thus $\mathcal{M}$ just is the mapping, as a set of node pairs, and

[1] So if (i,j) and (i',j') are in the mapping then (T1) $left(i,i')$ iff $left(j,j')$ and (T2) $anc(i,i')$ iff $anc(j,j')$.

Emms M. and Franco-Penya H. (2012).
ON ORDER EQUIVALENCES BETWEEN DISTANCE AND SIMILARITY MEASURES ON SEQUENCES AND TREES.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 15-24
DOI: 10.5220/0003712500150024

$\mathcal{D}$ and $\mathcal{I}$ just the remaining nodes of S and T which are not 'touched' by the mapping. Let $(.)^{\gamma}$ give the label of a node and let C^{Δ} be a 'cost' table, indexed by $\{\lambda\} \cup \Sigma$, where Σ is the alphabet of labels, which assigns 'costs' to $\mathcal{M}$, $\mathcal{D}$ and $\mathcal{I}$ according to[2]:

for $(i,j) \in \mathcal{M}$	cost is $C^{\Delta}(i^{\gamma}, j^{\gamma})$
for $i \in \mathcal{D}$	cost is $C^{\Delta}(i^{\gamma}, \lambda)$
for $j \in \mathcal{I}$	cost is $C^{\Delta}(\lambda, j^{\gamma})$

Where $\alpha : S \mapsto T$ is any mapping from S to T, define $\Delta(\alpha : S \mapsto T)$ by

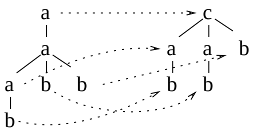
Definition 1. ('Distance' Scoring of an Alignment).

$$\Delta(\alpha : S \mapsto T) = \sum_{(i,j)\in\mathcal{M}} C^{\Delta}(i^{\gamma}, j^{\gamma}) + \sum_{i\in\mathcal{D}} C^{\Delta}(i^{\gamma}, \lambda) + \sum_{j\in\mathcal{I}} C^{\Delta}(\lambda, j^{\gamma})$$

From this costing of alignments, a 'distance' score on tree pairs is defined by minimization:

Definition 2. ('Distance' Scoring of a Tree Pair). The Tree- or Tai-distance $\Delta(S,T)$ between two trees S and T is the *minimum* value of $\Delta(\alpha : S \mapsto T)$ over possible *Tai*-mappings from S to T, relative to a chosen cost table C^{Δ}.

There is an illustration of the definitions in Figure 1



With $C^{\Delta}(x,\lambda) = C^{\Delta}(\lambda,x) = 1$, $C^{\Delta}(x,x) = 0$, $C^{\Delta}(x,y) = 1$ *for* $x \neq y$, *the alignment has score* $\Delta(\alpha) = 3$ *and this is minimal for the given* C^{Δ}

Figure 1: An illustration of tree distance.

$\Delta(S,T)$ can be computed by the algorithm of (Zhang and Shasha, 1989). Sequences can be encoded as vertical trees, and on this domain of trees the tree distance coincides with a well known comparison measure on sequences, the (alphabet-weighted) string edit distance (Wagner and Fischer, 1974; Gusfield, 1997).

We have formulated the definition[3] in terms of costs applied to mappings which respect tree-ordering properties. In contrast to this declarative perspective, there is procedural definition via the notion of an *edit-script* of atomic operations transforming S to T in a succession of stages. For both sequences and trees the mapping-based and script-based notions coincide (Wagner and Fischer, 1974; Tai, 1979; Kuboyama, 2007) and so we omit further details of the definition via edit-scripts.

While the correctness of the Tai 'distance' algorithm (Zhang and Shasha, 1989) – ie. that it truly finds the *minimal* value of $\Delta(\alpha : S \mapsto T)$ given cost-table C^{Δ}– does not require the cost-table C^{Δ} to satisfy any particular properties, some settings of C^{Δ} clearly make little sense. The combination of deletion/insertion cost-entries which are *negative* – $C^{\Delta}(x,\lambda) < 0, C^{\Delta}(\lambda,y) < 0$ – with swap/match cost entries which are *not negative* gives the counter-intuitive effect that a supertree of S is 'closer' – in the sense of having a lower Δ score – to S than S itself[4]. This is a rationale for the following non-negativity assumption

$$\forall x,y \in \Sigma(C^{\Delta}(x,y) \geq 0, C^{\Delta}(x,\lambda) \geq 0, C^{\Delta}(\lambda,y) \geq 0) \tag{1}$$

which is a pretty universal assumption, and from which it follows that $\Delta(S,T) \geq 0$, giving a minimum consistency with the every day notion of 'distance'. In what follows we will confine attention to 'distance' Δ based on a table C^{Δ} which satisfies at least (1).

When the cost-table $C^{\Delta}(x,y)$ is constrained more strictly than this to satisfy all the conditions of a *distance-metric*, then it is well known that $\Delta(S,T)$ will also be a distance-metric. Whether such further restriction is desirable is moot: in so-called stochastic variants (Ristad and Yianilos, 1998; Bernard et al., 2008; Emms, 2010), in which the entries in C^{Δ} are interpreted as negated logs of probabilities, these additional distance-metric assumptions are not fulfilled. In this article we shall only assume the cost-table C^{Δ} satisfies the non-negativity requiremnt of (1).

Turning now to 'similarity', rather than approach the problem of comparison by *minimizing* accumulated costs assigned to an alignment, a widely followed alternative, especially for sequence comparison, has been to *maximize* a score assigned to an alignment, with swaps/matches rewarded, and deletions/insertions punished.

Let C^{Θ} be a 'similarity' table, again indexed by $\{\lambda\} \cup \Sigma$, where Σ is the alphabet of labels, and where $\alpha : S \mapsto T$ is any mapping from S to T, and then let $\Theta(\alpha : S \mapsto T)$ be defined by

Definition 3. ('Similarity' Scoring of an Alignment).

$$\Theta(\alpha : S \mapsto T) = \sum_{(i,j)\in\mathcal{M}} C^{\Theta}(i^{\gamma}, j^{\gamma}) - \sum_{i\in\mathcal{D}} C^{\Theta}(i^{\gamma}, \lambda) - \sum_{j\in\mathcal{I}} C^{\Theta}(\lambda, j^{\gamma})$$

From this costing of alignments, a 'similarity' score on tree pairs is defined by maximisation:

[2] Note in this general setting even a pairing of two nodes with identical labels can in principal make a non-zero cost contribution.

[3] The literature contains quite a number of inequivalent notins, all referred to as 'tree distance'; in this article Definition 2 will be understood to define the term.

[4] Or a subtree.

Definition 4. ('Similarity' Scoring of a Tree Pair). The Tree- or Tai-similarity $\Theta(S,T)$ between two trees S and T is the *maximum* value of $\Theta(\alpha : S \mapsto T)$ over possible Tai-mappings from S to T, relative to a chosen cost table C^{Θ}

Applied to the same example as shown in Figure 1, with $C^{\Theta}(x,\lambda) = C^{\Theta}(\lambda,x) = 0$, $C^{\Theta}(x,x) = 2$, $C^{\Delta}(x,y) = 0$ for $x \neq y$, the shown alignment has score $\Theta(\alpha) = 9$, which is maximal for the given C^{Θ}.

$\Theta(S,T)$ can be computed via a simple modification of the algorithm of (Zhang and Shasha, 1989). Again on the domain of vertical trees this coincides with a well known approach to sequence comparison, the (alphabet-weighted) string similarity (Smith and Waterman, 1981; Gusfield, 1997).

As with Δ, while the correctness of the algorithm for Θ is not dependent on any assumptions about the cost-table C^{Θ}, some settings of C^{Θ} make little sense. Given the formulation in (3), which *subtracts* the contribution from deletions and insertions, a setting where deletion/insertion cost entries are negative – $C^{\Theta}(x,\lambda) < 0$, $C^{\Theta}(\lambda,x) < 0$ – gives the counter-intuitive effect that a supertree of S would be more 'similar' – in the sense of higher Θ score – to S than S itself. This gives a rationale for the nearly universal assumption of non-negative deletion/insertions entries in C^{Θ}:

$$\forall x,y \in \Sigma(C^{\Theta}(x,\lambda) \geq 0, C^{\Theta}(\lambda,y) \geq 0) \qquad (2)$$

In what follows we will confine attention always to 'similarity' Θ based on a table C^{Θ} satisfying (2)[5]. For the C^{Θ}-entries which are not deletions or insertions, it is quite common in biological sequence comparison to have both positive and negative entries. In contrast to the notion of a distance-metric, the notion of a set of axioms for a similarity Θ is less well established. (Chen et al., 2009) have recently made a proposal concerning this (see section 5).

To reiterate, for the purposes of this discussion a tree 'distance' measure will imply a cost-table C^{Δ}, satisfying (1), used in accordance to definitions 1 and 2 to score alignments and tree pairs. A tree 'similarity' measure measure will imply a cost-table C^{Θ}, satisfying (2), used in accordance to definitions 3 and 4 to score alignments and tree pairs. This is sufficient to distinguish the 'distance' approach from the 'similarity' approach in an intuitive way without commiting to any further axioms.

1.1 Order-equivalence Notions between Tai Distance and Similarity

Given a 'distance' Δ scoring of alignments, it can be set to work to induce orderings of at least three different kinds entities

Alignment Ordering. Given fixed S, and fixed T, rank the possible *alignments* $\alpha : S \mapsto T$ by $\Delta(\alpha : S \mapsto T)$

Neighbour Ordering. Given fixed S, and varying candidate neighbours T_i, rank the *neighbours* T_i by $\Delta(S,T_i)$ – typically used in k-NN classification.

Pair Ordering. Given varying S_i, and varying T_j, rank the *pairings* $\langle S_i,T_j \rangle$ by $\Delta(S_i,T_j)$ – typically used in hierarchical clustering.

Similarly a 'similarity' Θ scoring of alignments induces orderings of the above kinds of entities. Comparing these orderings motivates the following definition

Definition 5. (A-,N- and P-dual). When the alignment orderings induced by a choice of C^{Δ}(used in accordance with (1)) and by a choice C^{Θ} (used in accordance with (3)) are the *reverse* of each other, we will say that C^{Θ} is a **A-dual** of C^{Δ}. Similarly we will say we have an **N-dual** when neighbour ordering is reversed, and a **P-dual** where pair-ordering is reversed.

For example, the following are A-duals in this sense (proven in section 2):

Example 1.

$$\Delta \text{ with} \begin{cases} C^{\Delta}(x,\lambda) = 1 \\ C^{\Delta}(x,x) = 0 \\ C^{\Delta}(x,y) = 1 \end{cases} \quad \Theta \text{ with} \begin{cases} C^{\Theta}(x,\lambda) = 0 \\ C^{\Theta}(x,x) = 2 \\ C^{\Theta}(x,y) = 1 \end{cases}$$

Example 2.

$$\Delta \text{ with} \begin{cases} C^{\Delta}(x,\lambda) = 0.5 \\ C^{\Delta}(x,x) = 0 \\ C^{\Delta}(x,y) = 0.5 \end{cases} \quad \Theta \text{ with} \begin{cases} C^{\Theta}(x,\lambda) = 0 \\ C^{\Theta}(x,x) = 1 \\ C^{\Theta}(x,y) = 0.5 \end{cases}$$

A natural question that presents itself then is whether for *every* choice of C^{Δ}, there is a choice of C^{Θ} which is a A-dual, N-dual or P-dual, and vice-versa. More precisely there are the following

Order-relating Conjectures.

$$\textbf{A-duality} \begin{cases} (i) & \forall C^{\Delta} \exists C^{\Theta}(C^{\Delta} \text{ and } C^{\Theta} \text{ are A-duals}) \\ (ii) & \forall C^{\Theta} \exists C^{\Delta}(C^{\Delta} \text{ and } C^{\Theta} \text{ are A-duals}) \end{cases}$$

$$\textbf{N-duality} \begin{cases} (i) & \forall C^{\Delta} \exists C^{\Theta}(\ C^{\Delta} \text{ and } C^{\Theta} \text{ are N-duals}) \\ (ii) & \forall C^{\Theta} \exists C^{\Delta}(\ C^{\Delta} \text{ and } C^{\Theta} \text{ are N-duals}) \end{cases}$$

$$\textbf{P-duality} \begin{cases} (i) & \forall C^{\Delta} \exists C^{\Theta}(\ C^{\Delta} \text{ and } C^{\Theta} \text{ are P-duals}) \\ (ii) & \forall C^{\Theta} \exists C^{\Delta}(\ C^{\Delta} \text{ and } C^{\Theta} \text{ are P-duals}) \end{cases}$$

Arguably these notions go to the heart of the question whether there is really anything that can be accomplished using an alignment 'distance' score,

[5] While Definition 3 formulates Θ with deletion/insertion contributions subtracted, as is often done (Smith and Waterman, 1981; Stojmirovic and Yu, 2009), an alternative formulation has these treated additively (Gusfield, 1997). With the additive formulation, the same consideration suggests making deletion/insertions non-positive.

which cannot by accomplised via an alignment 'similarity' score, and vice-versa. For example, if it turns out that all these order conjectures hold, then any alignment outcome, any categorisation outcome via k-NN and any hierarchical clustering outcome, achieved by a particular distance can be replicated by a similarity, and vice-versa, making the choice merely a matter of personal taste. On the other hand, if these duality conjectures do not hold, then there is substantive difference, with the outcomes achievable by distances and similarities being distinct.

For a number of similarity and distance measures based on sets and vectors, notions analogous to N-dual and P-dual have been considered (Batagelj and Bren, 1995; Omhover et al., 2005; Lesot and Rifqi, 2010), motivated similarly by the question whether anything which can be accomplished with one or other such measure can be replicated by another such measure. It is for example shown there that a particular Dice measure will rank retrieval results inevitably the same as a particular Jaccard measure. In the case of alignment-based measures on sequences and trees, as far as we are aware, these notions seem not have been systematically considered and the following sections endeavour to fill that gap.

2 ALIGNMENT-DUALITY

The following lemma will be useful for considering the A-duality conjectures above:

Lemma 1. *For any C^{Δ}, and some choice δ such that $0 \leq \delta/2 \leq min(C^{\Delta}(\cdot,\lambda), C^{\Delta}(\lambda,\cdot))$ let C^{Θ} be defined according to (i) below. For any C^{Θ}, and choice δ such that $0 \leq \delta \geq max(C^{\Theta}(\cdot,\cdot))$ let C^{Δ} be defined according to (ii) below.*

$$\text{(i)} \begin{cases} C^{\Theta}(x,\lambda) = C^{\Delta}(x,\lambda) - \delta/2 \\ C^{\Theta}(\lambda,y) = C^{\Delta}(\lambda,y) - \delta/2 \\ C^{\Theta}(x,y) = \delta - C^{\Delta}(x,y) \end{cases}$$

$$\text{(ii)} \begin{cases} C^{\Delta}(x,\lambda) = C^{\Theta}(x,\lambda) + \delta/2 \\ C^{\Delta}(\lambda,y) = C^{\Theta}(\lambda,y) + \delta/2 \\ C^{\Delta}(x,y) = \delta - C^{\Theta}(x,y) \end{cases}$$

then in either case, for any $\alpha : S \mapsto T$

$$\Delta(\alpha) + \Theta(\alpha) = \delta/2 \times (\sum_{s \in S}(1) + \sum_{t \in T}(1)) \qquad (3)$$

Proof of Lemma 1. *If defining C^{Θ} from $C^{\Delta}$by (i), by the choice of δ we have the non-negativity of $C^{\Theta}(x,\lambda)$ and $C^{\Theta}(\lambda,y)$. If defining C^{Δ} from $C^{\Theta}$by (ii), by the choice of δ, we have the non-negativity of all entries in C^{Δ}.*

Whether defining C^{Θ} from $C^{\Delta}$by (i), or C^{Δ} from $C^{\Theta}$by (ii), it is straightforward to show

$$\Delta(\alpha) + \Theta(\alpha) = \delta/2 \times (2|\mathcal{M}| + |\mathcal{D}| + |I|)$$

But then (3) follows since

$$2|\mathcal{M}| + |\mathcal{D}| + |I| = \sum_{s \in \mathcal{S}}(1) + \sum_{t \in \mathcal{T}}(1)$$

□

Theorem 2. *A-duality (i) and (ii) hold*

Proof of Theorem 2. *A-duality (i): define C^{Θ} according to (i) in Lemma 1. Given the constant summation property of (3), the ordering on alignments by Δ must be the reverse of the ordering by Θ.*

A-duality (ii): similarly define C^{Δ} according to (ii) in Lemma 1 □

Example 1 Revisited. *The C^{Θ} of Example 1 can be seen as derived from the C^{Δ} with $\delta = 2$. Table below shows outcomes for other choices of δ*

	C^{Δ}	$C^{\Theta}(\delta = 2)$	$C^{\Theta}(\delta = 1)$	$C^{\Theta}(\delta = 0)$
(x,λ)	*1*	*0*	*0.5*	*1*
(x,x)	*0*	*2*	*1*	*0*
(x,y)	*1*	*1*	*0*	*-1*

As a corollary one can obtain the following concerning how one similarity table can be 'shifted' to an equivalent one, and similarly for distance tables.

Corollary 3. ('Shifting'). for any $C^{\Theta}{}_1$, an alignment equivalent $C^{\Theta}{}_2$ can be derived by the conversion (a) below, and for any $C^{\Delta}{}_1$, an alignment equivalent $C^{\Delta}{}_2$ can be derived by the conversion (b)

$$\text{(a)} \begin{cases} C^{\Theta}{}_2(x,\lambda) = C^{\Theta}{}_1(x,\lambda) - \kappa/2 \\ C^{\Theta}{}_2(\lambda,y) = C^{\Theta}{}_1(\lambda,y) - \kappa/2 \\ C^{\Theta}{}_2(x,y) = C^{\Theta}{}_1(x,y) + \kappa \end{cases}$$

$$\text{(b)} \begin{cases} C^{\Delta}{}_2(x,\lambda) = C^{\Delta}{}_1(x,\lambda) + \kappa/2 \\ C^{\Delta}{}_2(\lambda,y) = C^{\Delta}{}_1(\lambda,y) + \kappa/2 \\ C^{\Delta}{}_2(x,y) = C^{\Delta}{}_1 + \kappa \end{cases}$$

Proof of Corollorary 3. *(a) is the composition of (ii), for some δ_1, with (i), for some δ_2, giving $\kappa = \delta_2 - \delta_1$. (b) is the composition (i), for some δ_1, with (ii), for some δ_2, giving $\kappa = \delta_2 - \delta_1$* □

Example 1 Revisited Again. *The three A-dualizing similarities $C^{\Theta}(\delta = 2)$, $C^{\Theta}(\delta = 1)$ and $C^{\Theta}(\delta = 0)$ derived from the unit-cost distance table using varying δ in the (i) conversion of Lemma 1 can be seen as related to each other by the (a) 'shifting' conversion of Lemma 3, with $\kappa = -1$ each time.*

The property of alignment dualizability between distance and similarity (and vice-versa) expressed above in Lemma 1 and Theorem 2 was essentially first proven for the case of sequence comparison by (Smith and Waterman, 1981). On the basis of this perhaps it is tempting to consider the case closed and treat 'distance' and 'similarity' as interchangeable. However, as noted in Section 1.1, there is more than one kind of ordering that one might wish to be sure of replicating in switching between distance and similarity, with N-duality coming to the fore in the context of k-NN classification, and P-duality coming to the fore in the context of hierarchical clustering. Section 3.1 considers the N-duality (i) and P-duality (i) order conjectures, and Section 3.2 considers the N-duality (ii) and P-duality (ii) conjectures.

3 NEIGHBOUR AND PAIR ORDERING

3.1 Distance to Similarity

Having seen that A-duals can always be created in both directions, attention shifts to N-duals and P-duals.

The case of using $\delta = 0$ in the (i) conversion of Lemma 1 from C^{Δ} to C^{Θ} gives non-positive values for all non-deletion, non-insertion entries in C^{Θ}, and is an especially trivial case of dualizing a distance setting C^{Δ}, with the effect that $\Theta(S,T) = -1 \times \Delta(S,T)$. Because of this, this distance-to-similarity conversion not only makes A-duals, but also N-duals and P-duals.

Theorem 4. *N-duality (i) and P-duality(i) hold*

Proof of Theorem 4. *By choosing $\delta = 0$ in the (i) conversion of Lemma 1 from C^{Δ} to C^{Θ}, we have $\Theta(S,T) = -1 \times \Delta(S,T)$, and hence $\Theta(S_1,T_1) \leq \Theta(S_2,T_2) \Leftrightarrow \Delta(S_1,T_1) \geq \Delta(S_2,T_2)$* □

This distance-to-similarity by negation is well known. On the other hand, concerning similarity-to-distance, in the (ii) conversion of Lemma 1 from C^{Θ} to C^{Δ}, you can only choose $\delta = 0$ if all $C^{\Theta}(x,y) \leq 0$, and clearly there are many natural settings of C^{Θ} where that is not true.

3.2 Similarity to Distance

The remaining order-equivalence conjectures of section 1.1 are *N-duality(ii)* and *P-duality(ii)*, concerning the similarity-to-distance direction. Of the remaining conjectures, *P-duality(ii)* is stronger than *N-duality(ii)*. We can fairly easily show *P-duality(ii)* does not hold

Theorem 5. *P-duality (ii) does not hold, that is, there are C^{Θ} such that there is no C^{Δ} such that C^{Θ} and C^{Δ} are P-duals.*

Proof of Theorem 5. *It is clearly possible for C^{Θ} to be such that there is no maximum value for $\Theta(S,T)$. For example for table below:*

	C^{Θ}
(a,a)	1
(a,λ)	1

its clear we have $\Theta(a,a) = 1$, $\Theta(aa,aa) = 2$ and in general $\Theta(a^n,a^n) = n$. Let C^{Θ} be any table defining a similarity with no maximum. On the other hand, for each C^{Δ} there will be minimum value of $\Delta(S,T)$. Suppose some C^{Δ} is a P-dual to C^{Θ}. For any n let $[\Theta]_n$ (resp. $[\Delta]_n$) be the set of pairs with similarity (resp. distance) n. If C^{Δ} is a P-dual to C^{Θ}, there is some bijection between the set of similarity classes $\{[\Theta]_s\}$ and the set of distances classes of $\{[\Delta]_d\}$. Some similarity class $[\Theta]_{s_1}$ of Θ must correspond to the minimum distance class $[\Delta]_{d_0}$. Let $[\Theta]_{s_2}$ be a higher Θ class than $[\Theta]_{s_1}$. It must correspond to some Δ class $[\Delta]_{d_1}$ distinct from $[\Delta]_{d_0}$, and since $[\Delta]_{d_0}$ is the distance-minimum, this must be a higher distance class. Then for $(S_0,T_0) \in [\Delta]_{d_0}$, and $(S_1,T_1) \in [\Delta]_{d_1}$ you have $\Delta(S_0,T_0) < \Delta(S_1,T_1)$, but also $\Theta(S_0,T_0) < \Theta(S_1,T_1)$. So the supposed dual C^{Δ} does not reverse the pair-ordering of C^{Θ}. □

Of the order-relating conjectures of section 1.1 the only remaining one is *N-duality(ii)* – that is the question whether every neighbour-ordering via some C^{Θ} can be replicated by a neighbour ordering via some C^{Δ}. We can show that there are neighbour-orderings by a Tai-similarity which cannot be dualized by any Tai-distance whose deletion and insertion costs are symmetric.

Theorem 6. *There is C^{Θ} such that there is no C^{Δ} with $C^{\Delta}(x,\lambda) = C^{\Delta}(\lambda,x)$ such that C^{Θ} and C^{Δ} are N-duals*

Proof of Theorem 6. *Let $S = aa$, and the set of neighbours be $\{a, aaa\}$.*
Let $C^{\Theta}(a,a) = x > 0$, and $C^{\Theta}(a,\lambda) = C^{\Theta}(\lambda,a) = y > 0$.

For (aa,aaa), the alignments with 2,1, and 0 a-matches haves scores, $2x - y$, $x - 3y$ and $-5y$, respectively, so the alignments maximising Θ are those with two a-matches, and $\Theta(aa,aaa) = 2x - y$.

For (aa,a), the alignments with 1 and 0 a-matches have scores $x - y$ and $-3y$, respectively, so the alignments maximising Θ have one a-match, and $\Theta(aa,a) = x - y$.

Consider what is required for the Θ-decreasing neigbour ordering to be: $[aaa,\ a]$,

$$\begin{array}{rl} & \Theta(aa,aaa) > \Theta(aa,a) \\ \Leftrightarrow & 2x-y > x-y \\ \Leftrightarrow & x > 0 \end{array}$$

So there is a Θ*-decreasing neighbour-ordering* $[aaa,\ a]$.

Let $C^{\Delta}(a,a) = x'$, *and* $C^{\Delta}(a,\lambda) = C^{\Delta}(\lambda,a) = y'$. *Note this assumes symmetric insertion and deletion costs.*

For (aa,aaa), *the alignments with 2,1, and 0 a-matches haves scores,* $2x'+y'$, $x'+3y'$ *and* $5y'$, *respectively. We distinguish two cases (i)* $2y' < x'$ *and (ii)* $2y' \geq x'$.

For case (i), $x' = 2y'+\varepsilon$, *for some no-zero* $\varepsilon > 0$, *and the 2,1,and 0 a-matches scores become* $5y'+2\varepsilon$, $5y'+\varepsilon$ *and* $5y'$, *respectively, so taking the minimum,* $\Delta(aa,aaa) = 5y'$.

For case (ii), $y' = x'/2+\kappa$, *for some* $\kappa \geq 0$, *and the 2,1,and 0 a-matches scores become* $2.5x'+\kappa$, $2.5x'+3\kappa$ *and* $2.5x'+5\kappa$, *respectively, and 2-match case is amongst the minimal cases, so* $\Delta(aaa,aa) = 2.5x'+\kappa$.

For (aa,a), *the alignments with 1 and 0 a-matches haves scores,* $x'+y'$ *and* $3y'$ *respectively. We again distinguish between cases (i)* $2y' < x'$ *and (ii)* $2y' \geq x'$.

For case (i), the 1 and 0 a-matches scores become $3y'+\varepsilon$ *and* $3y'$ *respectively, so taking the minimum,* $\Delta(aa,a) = 3y'$.

For case (ii), the 1 and 0 a-match scores become $1.5x'+\kappa$ *and* $1.5x'+3\kappa$ *respectively, and the 1-match case is amongst the minimal cases, so* $\Delta(aa,a) = 1.5x'+\kappa$.

Summarising the Δ *possibilities*

$$\begin{array}{lll} & \Delta(aa,aaa) & \Delta(aa,a) \\ (i)2y' < x' & 5y' & 3y' \\ (ii)2y' \geq x' & 2.5x'+\kappa & 1.5x'+\kappa \end{array}$$

So in neither case (i) nor case (ii) is it possible to achieve a Δ*-ascending neighbour ordering* $[aaa,a]$, *which was the* Θ*-descending neighbour ordering which was achieved with the assumed* C^{Θ}. □

Remark. If we drop the requirement that the N-dualizing C^{Δ} have $C^{\Delta}(x,\lambda) = C^{\Delta}(\lambda,x)$, then the argument does not go through. The Θ-descending neighbour ordering $[aaa,a]$ can be replicated by a Δ-ascending neighbour ordering with $C^{\Delta}(a,\lambda) > C^{\Delta}(\lambda,a)$. For most applications of alignment-based ’distances’, such an asymmetric setting of deletion and insertion costs would be considered unnatural.

4 EMPIRICAL INVESTIGATION

(Lesot and Rifqi, 2010) consider distance and similarity measures often used in information retrieval. These are defined over finite vectors, whose features are either binary or real-valued. They basically consider the neighbour orderings produced by different measures. Besides demonstrating absolute equivalence between some measures, between other measures they empirically determine *equivalence degrees*, between 0 and 1, based on the Kendall-tau statistic for comparing orderings (Kendall, 1945). While their work concerned comparison measures on vectors, it is a natural to consider an analogous empirical quantified comparison of distance and similarity orderings on trees and sequences. Some preliminary findings of such a study are given below.

The (i) conversion of Lemma 1 converts distance settings to A-dual similarity settings and one thing to consider is the degree to which the derived similarities are also N-duals of the distance. Table 1 gives some distance and similarity settings: the first column gives the unit-cost settings for Δ and the columns to the right give different similarity settings C^{Θ} derivable by the (i) conversion of Lemma 1 as δ is varied through various values.

Table 1: Unit-cost distance setting and several A-dual similarity settings.

		dual C^{Θ} for varying δ						
	C^{Δ}	2	1.5	1	0.5	0.2	0.1	0
(x,λ)	1	0	0.25	0.5	0.75	0.9	0.95	1
(x,x)	0	2	1.5	1	0.5	0.2	0.1	0
(x,y)	1	1	0.5	0	-0.5	-0.8	-0.9	-1

An experiment was done to quantify how close the similarities defined by the varying C^{Θ}tables come to being N-duals for the distance. Using a set of 1334 trees[6], repeatedly a tree S was chosen, and neighbour files $N_{\Delta}(S)$ and $N_{\Theta}(S)$ were computed, with $N_{\Delta}(S)$ the ordering of the remaining trees by ascending Δ, and $N_{\Theta}(S)$ the ordering by descending Θ. $N_{\Delta}(S)$ and $N_{\Theta}(S)$ were then compared by the *kendall-tau* measure τ (see the Appendix for the definition). For each δ the average of this τ comparison between the distance and similarity neighbour files is shown in Figure 2.

The bottom-left corner, for $\delta = 0$ is the special case of Lemma 1 which amounts to the well-known trivial distance-to-similarity conversion, $\Theta(S,T) = -1 \times \Delta(S,T)$, noted in section 3.1. In this case the distance and similarity neighbour files are identical.

[6] See the Appendix for further details of this data set.

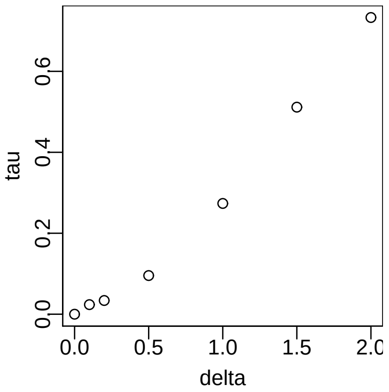


Figure 2: Average Kendall-tau comparison on neighbours using distance and derived similarities. Distance setting is first column of Table 1. Similarity settings are further columns of Table 1 defined by varying δ.

As the graph clearly shows, as δ increases, the neighbour files exhibit progressively greater difference in ordering, until at $\delta = 2$ the τ score is 0.73, which corresponds to a tendency more towards order reversal than to replication. This experiment shows that although each of these similarity settings is an A-dual of the simple distance setting, they are not at all equivalent to each other as far as neighbour ordering is concerned.

The (ii) conversion of Lemma 1 converts similarity settings to A-dual distance settings. Table 2 gives a similarity setting and then several distance settings derivable by the (ii) conversion as δ is varied through various values[7]

Table 2: A similarity setting and several A-dual distance settings.

		dual C^{Δ} for varying δ						
	C^{Θ}	1	1.5	2	2.5	3	3.5	4
(x,λ)	0.5	1	1.25	1.5	1.75	2	2.25	2.5
(x,x)	1	0	0.5	1	1.5	2	2.5	3
(x,y)	0	1	1.5	2	2.5	3	3.5	4

Figure 3 plots the average τ comparison between the similarity and distance neighbour files, as δ is varied to give different distances. Again this experiment shows that although each of the distance settings is an A-dual of the similarity setting, they are not equivalent to each other as far as neighbour ordering is concerned.

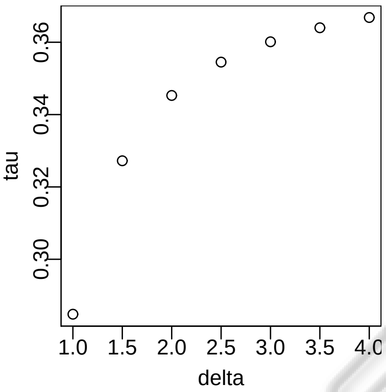


Figure 3: Average Kendall-tau comparison on neighbours using a similarity and derived distances. Similarity setting is first column of Table 2. Distance settings are further columns of Table 2 defined by varying δ.

Theorem 5 concerned the non-replicability by distance of pair-orderings by similarity. To illustrate this, consider a set of strings $\{a^5, a^4, a^3, a^2, a^1\}$. A table of pair-wise similarities of these was made with $C^{\Theta}(a,a) = 1, C^{\Theta}(a,\lambda) = 1$, and used to generate a single-link clustering, shown as the the uppermost dendrogram in Figure 4.

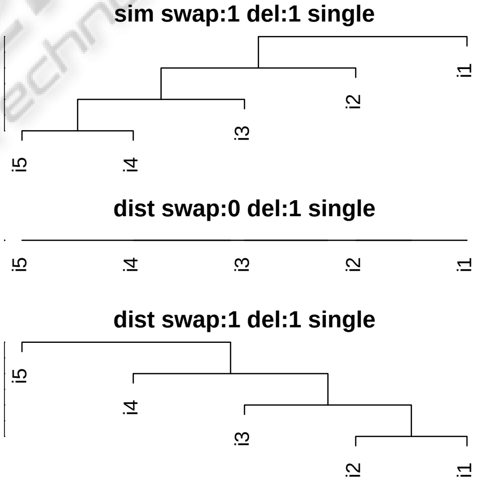


Figure 4: Similarity and distance clusterings. The instance labels $i5 \ldots i1$ represesent $a^5 \ldots a^1$.

No single-link clustering based on distance replicates this similarity clustering. The middle dendogram in Figure 4 is the result with $C^{\Delta}(a,a) = 0, C^{\Delta}(a,\lambda) = 1$, with all five shown on the same level because $\Delta(a^m, a^{m+1}) = 1$. The lowest dendogram in Figure 4 shows a result with $C^{\Delta}(a,a) = 1, C^{\Delta}(a,\lambda) = 1$. The same structure was found holding $C^{\Delta}(a,a) = 1$, and allowing the deletion/insertion cost to vary between 0.5 and 5.5 (which are $\geq C^{\Delta}(a,a)$) and between

[7] The nodes in these experiments have multi-part labels. Whilst the first experiment treated these simply as identical or not, for this second experiment, the base-line similarity node label are compared via $C^{\Theta}(x,y) = 1 - ham(x,y)$, $ham(x,y)$ is the standard hamming distance. The table thus shows the extreme values of $C^{\Theta}(x,y)$ and $C^{\Delta}(x,y)$.

0.4 and 0.1 (which are $< C^{\Delta}(a,a)$)

5 DISCUSSION AND COMPARISONS

In view of the outcomes noted in sections 2, 3.1 and 3.2 concerning the various ordering conjectures we can say that

- Any hierarchical clustering outcome achieved via Δ can be replicated via Θ, but *not* vice-versa.
- Any categorisation outcome using nearest-neighbours achieved via Δ can be replicated via Θ, but *not* vice-versa.

and in this sense 'similarity' and 'distance' comparison measures on sequences and trees are *not* interchangeable.

As far as we are aware this aspect of the choice between a similarity-based versus a distance-based comparison measure on sequences or trees has not been noted before.

There are a number of papers concerning conversion from a similarity-based sequence comparison measure to a distance-based comparison measure, and particularly one satisfying distance-metric axioms (Spiro and Macura, 2004; Stojmirovic and Yu, 2009). An aim of these papers is to find techniques for accelerating so-called range similarity queries, which are requests to find all neighbours within a similarity threshold $N_{\leq\theta}(S) = \{T : \Theta(S,T) \geq \theta\}$. To discuss these papers it will be as well to note the distance-metric axioms

Definition 6. (Distance Metric). A binary relation Δ is a distance-metric if it satisfies
$D1.\Delta(S,T) = \Delta(T,S)$
$D2.\Delta(S,T) \geq 0$
$D3.\Delta(S,V) \leq \Delta(S,T) + \Delta(T,V)$
$D4.\Delta(S,T) = 0$ iff $S = T$

It is a *pseudo-metric* if D4. is dropped. It is a *quasi-metric* if D1. is droppped

For a distance-metric on sequences there is a way to use the triangle-inequality to accelerate solution of a *distance* range query, $N_{\leq\delta}(S) = \{T : \Delta(S,T) \leq \delta\}$. Suppose S is a query, and T_1 is a training-set point known to be far from S, and that another training-set point T_2 is known to be close to T_1. Intuitively S is also going to be far from T_2. More specifically, if Δ is a distance-metric, an instance of the triangle-inequality will be:

$$\Delta(S,T_1) \leq \Delta(T_1,T_2) + \Delta(T_2,S) \qquad (4)$$

via which $\Delta(T_2,S)$ is bounded below by $\Delta(S,T_1) - \Delta(T_1,T_2)$. So if T_1 has already been excluded from a distance neigbhourhood, T_2 can be also immediately excluded if $\Delta(S,T_1) - \Delta(T_1,T_2)$ exceeds the threshold.

Most biological sequence comparison is done with similarity not distance and the concern of (Spiro and Macura, 2004) is to find a corresponding means of accelerating similarity range queries. In terms of the notations used here, they essentially propose the following conversion from similarity to distance cost-table

$$\begin{array}{ll}\forall x,y \in \Sigma & (C^{\Delta}(x,y) = C^{\Theta}(x,x) + C^{\Theta}(y,y) - 2C^{\Theta}(x,y)) \\ \forall x \in \Sigma & (C^{\Delta}(x,\lambda) = C^{\Theta}(x,\lambda)) \\ \forall y \in \Sigma & (C^{\Delta}(\lambda,x) = C^{\Theta}(\lambda,x))\end{array}$$

and they prove that, under some conditions imposed on C^{Θ}, the corresponding Δ will satisfy all the conditions of a distance-metric, in particular satisfying the triangle-inequality. and that the relation between Θ and Δ is then

$$\Delta(X,Y) = \Theta(X,X) + \Theta(Y,Y) - 2\Theta(X,Y) \qquad (5)$$

Substitution of (5) into the triangle-inequality and some re-arrangement gives that $\Theta(T_2,S)$ is bounded *above* by $\Theta(S,T_1) + \Theta(T_2,T_2) - \Theta(T_1,T_2)$, giving a means for rapid exclusion of T_2 from a similarity neighbourhood.

Beside the fact that equation (5) relating Θ and Δ holds only under particular assumptions concerning C^{Θ}, more importantly the obtained relationship in (5) is not sought in the context of deriving a P-dual or N-dual distance Δ from a given similarity Θ, and in fact (5) does not do this. Thus while Spiro et al do provide a conversion from a similarity to a distance, it addresses concerns somewhat orthogonal to those of this paper.

(Stojmirovic and Yu, 2009) is a paper with similar concerns to (Spiro and Macura, 2004). In terms of the notations used here, they propose the following conversion from similarity to distance cost-table:

$$\begin{array}{ll}\forall x,y \in \Sigma & (C^{\Delta}(x,y) = C^{\Theta}(x,x) - C^{\Theta}(x,y)) \\ \forall x \in \Sigma & (C^{\Delta}(x,\lambda) = C^{\Theta}(x,x) + C^{\Theta}(x,\lambda)) \\ \forall y \in \Sigma & (C^{\Delta}(\lambda,x) = C^{\Theta}(\lambda,x))\end{array}$$

and prove, under some assumptions concerning C^{Θ}, that the then derived 'distance' is a *quasi-metric* and that the relationship between Δ and Θ is then:

$$\Delta(S,T) = \Theta(S,S) - \Theta(S,T) \qquad (6)$$

Though not a distance-metric – it is *asymmetric* – it does satisfy the triangle-inequality $\Delta(X,Z) \leq \Delta(X,Y) + \Delta(Y,Z)$, and substituting (6) into the triangle-inquality and re-arranging again gives an upper bound which might be used to accelerate a similarity range query: $\Theta(S,T_2) \leq \Theta(S,T_1) + \Theta(T_2,T_2) - \Theta(T_2,T_1)$.

Though again this similarity to distance conversion is not sought in the context of finding P- or N-duals, Stojmirov et al's equation in (6) *does* make the derived distance an N-dual of the similarity. This is not, however, inconsistent with the example in section 3.2 of a similarity with no N-dualizing distance. Stojmirov et al's conversion generates *asymmetric* insertion and deletion entries in the distance cost-table C^{Δ}, whereas the proof in section 3.2 concerned the impossibily of a N-dualizing distance with *symmetric* insertion and deletion entries.

Our findings on the various order-relating conjectures concern notions with specific, though widely used, definitions (Defs.1, 2, 3 and 4). There are other closely related notions, and the corresponding questions concerning these have not been addressed. One variant is *stochastic*: in a stochastic similarity, probabilities are assigned to aspects of a mapping and *multiplied*. We conjecture that these will be A-, N- and P-dualisable to distance. This is because, under a logarithmic mapping, it seems such stochastic variants can be exactly simulated by a similarity as we have defined it. In the resulting table, all $C^{\Theta}(x,y) \leq 0$, allowing the (ii) conversion of Lemma 1 to define a C^{Δ} choosing $\delta = 0$. There are also *normalised* variants, which we have not considered. Throwing the net very much more widely, (Chen et al., 2009) study relationships between distance and similarity measures, in a very general setting, not restricted to measures based on sequence or tree alignment. Parallel to the well-known axioms of a distance-measure, they propose a set of similarity axioms, and they define conversions from similarity to distance and in the other direction, showing that the derived score satisfies the relevant axioms if the score that is input to the conversion does. Their work, however, does not address the question whether the conversions give N- or P-duals, that is whether they preserve relevant orderings.

Concerning directions for further work, the empirical investigation in section 4 was quite preliminary. For the Kendall-tau comparison of distance and similarity neighbourhoods, we looked at just one particular baseline distance and one particular baseline similarity, and compared only to A-duals as given by Lemma 1, so clearly there are other possibilities one could consider here. One is Spiro and Macura's relation in (5). The Appendix notes some further A-dualizing conversions, from distannce to similarity and from similarity to distance, which might be considered. It is also the case that we applied the Kendall-tau comparison to *full* rankings, and it would be of interest to look also at *top-k* ranking, as has been done for vector- and set-based measures (Lesot and Rifqi, 2010).

ACKNOWLEDGEMENTS

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College Dublin.

REFERENCES

Alves, C. E. R., Cáceres, E. N., and Dehne, F. (2002). Parallel dynamic programming for solving the string editing problem on a cgm/bsp. In *Proceedings of the fourteenth annual ACM symposium on Parallel algorithms and architectures*, SPAA '02, pages 275–281. ACM.

Batagelj, V. and Bren, M. (1995). Comparing resemblance measures. *Journal of Classification*, 12(1):73–90.

Bernard, M., Boyer, L., Habrard, A., and Sebban, M. (2008). Learning probabilistic models of tree edit distance. *Pattern Recogn.*, 41(8):2611–2629.

Bose, R. P. J. C. and van der Aalst, W. M. P. (2009). Context aware trace clustering: Towards improving process mining results. In *SAIM International Conference on Data Mining*, SDM, pages 401–412.

Chen, S., Ma, B., and Zhang, K. (2009). On the similarity metric and the distance metric. *Theoretical Computer Science*, 410(24-25):2365 – 2376.

Emms, M. (2010). Trainable tree distance and an application to question categorisation. In *KONVENS 2010*.

Emms, M. and Franco-Penya, H. (2011). Dataset used in Kendall-Tau experiments www.scss.tcd.ie/Martin.Emms/SimVsDistData.

Gusfield, D. (1997). *Algorithms on strings, trees, and sequences*. Cambridge Univ. Press.

Haji, J., Ciaramita, M., Johansson, R., Kawahara, D., Meyers, A., Nivre, J., Surdeanu, M., Xue, N., and Zhang, Y. (2009). The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009)*.

Herrbach, C., Denise, A., Dulucq, S., and Touzet, H. (2006). Alignment of rna secondary structures using a full set of operations. Technical Report 145, LRI.

Kendall, M. G. (1945). The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

Kondrak, G. (2003). Phonetic alignment and similarity. *Computers and the Humanities*, 37.

Kuboyama, T. (2007). *Matching and Learning in Trees*. PhD thesis, Graduate School of Engineering, University of Tokyo.

Lesot, M.-J. and Rifqi, M. (2010). Order-based equivalence degrees for similarity and distance measures. In *Proceedings of the Computational intelligence* *for knowledge-based systems design, and 13th international conference on Information processing and management of uncertainty*, IPMU'10, pages 19–28, Berlin, Heidelberg. Springer-Verlag.

Omhover, J.-F., Rifqi, M., and Detyniecki, M. (2005). Ranking invariance based on similarity measures in document retrieval. In *Adaptive Multimedia Retrieval*, pages 55–64.

Ristad, E. S. and Yianilos, P. N. (1998). Learning string edit distance. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 20(5):522–532.

Smith, T. F. and Waterman, M. S. (1981). Comparison of biosequences. *Advances in Applied Mathematics*, 2(4):482 – 489.

Spiro, P. A. and Macura, N. (2004). A local alignment metric for accelerating biosequence database search. *Journal of Computational Biology*, 11(1):61–82.

Stojmirovic, A. and Yu, Y.-K. (2009). Geometric aspects of biological sequence comparison. *Journal of Computational Biology*, 16:579–610.

Tai, K.-C. (1979). The tree-to-tree correction problem. *Journal of the ACM (JACM)*, 26(3):433.

Wagner, R. A. and Fischer, M. J. (1974). The string-to-string correction problem. *Journal of the Association for Computing Machinery*, 21(1):168–173.

Zhang, K. and Shasha, D. (1989). Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal of Computing*, 18:1245–1262.

APPENDIX

Proof of Alignment Sum Property from Lemma 1. In the proof of Lemma 1 it was claimed with C^{Δ} and C^{Θ} related according to the (i) or (ii) conversions that for any alignment α, $\Delta(\alpha) + \Theta(\alpha) = \delta/2 \times (2|\mathcal{M}| + |\mathcal{D}| + |I|)$. This is proven as follows.

If defining C^{Θ} from $C^{\Delta}$by (i), for $\Theta(\alpha)$ we have:

$$\begin{aligned}
& \sum_{(i,j)\in\mathcal{M}} [\delta - C^{\Delta}(i,j)] - \sum_{i\in\mathcal{D}} [C^{\Delta}(i,\lambda) - \delta/2] \\
& \qquad\qquad - \sum_{j\in I} [C^{\Delta}(\lambda,j) - \delta/2) \\
= \quad & \delta(|\mathcal{M}| + \frac{|\mathcal{D}|}{2} + \frac{|I|}{2}) \\
& - \sum_{(i,j)\in\mathcal{M}} [C^{\Delta}(i,j)] - \sum_{i\in\mathcal{D}} [C^{\Delta}(i,\lambda)] - \sum_{j\in I} [C^{\Delta}(\lambda,j)] \\
= \quad & \frac{\delta}{2}(2|\mathcal{M}| + |\mathcal{D}| + |I|) - \Delta(\alpha)
\end{aligned}$$

If defining C^{Δ} from $C^{\Theta}$by (ii), for $\Delta(\alpha)$ we have

$$\begin{aligned}
& \sum_{(i,j)\in\mathcal{M}} [\delta - C^{\Theta}(i,j)] + \sum_{i\in\mathcal{D}} [C^{\Theta}(i,\lambda) + \delta/2] \\
& \qquad\qquad + \sum_{j\in I} [C^{\Delta}(\lambda,j) + \delta/2) \\
= \quad & \delta(|\mathcal{M}| + \frac{|\mathcal{D}|}{2} + \frac{|I|}{2}) \\
& - \sum_{(i,j)\in\mathcal{M}} [C^{\Theta}(i,j)] + \sum_{i\in\mathcal{D}} [C^{\Theta}(i,\lambda)] + \sum_{j\in I} [C^{\Delta}(\lambda,j)] \\
= \quad & \frac{\delta}{2}(2|\mathcal{M}| + |\mathcal{D}| + |I|) - \Theta(\alpha)
\end{aligned}$$

Hence in either case the claim holds. □

Definition of Kendall-Tau (with Ties). Let N^1 and N^2 be two assignments of ranks to the same set of objects, U (with the possibility of ties). Where $\mathcal{P}$ is the set of all two-element sets of distinct objects from U, define a penalty function p on any $\{T_i, T_j\} \in \mathcal{P}$, such that (i) $p(\{T_i, T_j\}) = 1$ if the order in N^1 is the reverse of the order in N^2, (ii) $p(\{T_i, T_j\}) = 0.5$ if there is a tie in N^1 but not in N^2 or vice-versa and (iii) $p(\{T_i, T_j\}) = 0$ otherwise. The Kendall-Tau distance (with ties) between N^1 and N^2, $\tau(N^1, N^2)$, is $\sum_{\{T_i,T_j\}\in\mathcal{P}} [p(\{T_i, T_j\})] \times \frac{2}{m\times(m-1)}$

Details of the Data Set for Kendall-Tau Experiments. Section 4 reports experiments quantifying the difference between neighbour files computed by distance and similarity, when the two are related by the conversion in Lemma 1. The experiments used a set of 1334 trees, taking each tree in turn and ranking all the remaining trees. The trees represent syntax structures and originate in a data-set which was used in a shared-task on identifying inter-node semantic dependencies (Haji et al., 2009). See (Emms and Franco-Penya, 2011) for download information concerning this data.

Further A-dualizing Conversions. Concerning A-duals, there are besides the conversions given in Lemma 1, others which also generate A-duals.

Lemma 7. *For any C^{Δ}, for any k, let C^{Θ} be defined according to (iii) below.*

$$\textit{(iii)} \begin{cases} C^{\Theta}(x,\lambda) = kC^{\Delta}(x,\lambda) \\ C^{\Theta}(\lambda,y) = kC^{\Delta}(\lambda,y) \\ C^{\Theta}(x,y) = (1-k)(C^{\Delta}(x,\lambda) + C^{\Delta}(\lambda,y)) - C^{\Delta}(x,y) \end{cases}$$

Then for any $\alpha : S \mapsto T$

$$\Delta(\alpha) + \Theta(\alpha) = (1-k) \times (\sum_{s\in S}(C^{\Delta}(s,\lambda)) + \sum_{t\in T}(C^{\Delta}(\lambda,t)))$$

Lemma 8. *For any C^{Θ}, for any k, let C^{Δ} be defined according to (iv) below.*

$$\textit{(iv)} \begin{cases} C^{\Delta}(x,\lambda) = C^{\Theta}(x,\lambda) + kC^{\Theta}(x,x) \\ C^{\Delta}(\lambda,y) = C^{\Theta}(\lambda,y) + kC^{\Theta}(y,y) \\ C^{\Delta}(x,y) = k(C^{\Theta}(x,x) + C^{\Theta}(y,y)) - C^{\Theta}(x,y) \end{cases}$$

Then for any $\alpha : S \mapsto T$,

$$\Delta(\alpha) + \Theta(\alpha) = k \times (\sum_{s\in S}(C^{\Theta}(s,s)) + \sum_{t\in T}(C^{\Theta}(t,t)))$$

The proofs of these follow a similar pattern to that of Lemma 1 and are omitted. In a similar fashion both these conversions will give A-duals.