

# Edit distance and comparison scores

January 10, 2007

## 1 Sequence distance metrics

The simplest form of sequence comparison is based on editing strings and counting the number of edits required to get from one sequence to the other. The formulation of string-edit distance  $d_e$  balances two different types of edits. The simplest is replacement of a single letter by another letter. To start with, we need a metric on the set of letters in the alphabet  $\Sigma$  for our sets of sequences. Let  $D_\Sigma$  be a metric on the alphabet  $\Sigma$ . Then  $D_\Sigma(\xi, \eta)$  measures the difference between the two letters  $\xi$  and  $\eta$  in  $\Sigma$ . Now we show how to extend this to a metric  $d_e$  on the set of all finite strings  $\Sigma^*$ .

If two strings  $x$  and  $y$  differ only in the  $k$ -th position, then we set  $d_e(x, y) = D_\Sigma(x_k, y_k)$ . In general, when there are multiple replacements, string edit distance is based on just summing the effects. However, string-edit distance also allows a different kind of change as well: insertion and deletion. For example, we can define  $x_{\hat{k}}$  to mean the string  $x$  with the  $k$ -th entry removed. It might be that  $x_{\hat{k}}$  agrees perfectly with the string  $y$ , and so we assign  $d(x, y) = \delta$  where  $\delta$  is the deletion penalty. Similarly, insertions of characters are allowed to determine edit distance. Clearly, if  $y = x_{\hat{k}}$ , then adding  $x_k$  to  $y$  at the  $k$ -th position yields  $x$ . Again, the effect of multiple insertions/deletions is additive, and this allows strings of different lengths to be compared.

The use of both replacements and insertion/deletions to determine edit distance means that an edit path from  $x$  to  $y$  is not unique. Edit distance is therefore defined by taking the minimum over all possible representations, as we define formally in (1.4). But this will not in general define a metric unless appropriate conditions on  $\delta$  and  $D_\Sigma$  are satisfied. These conditions can be defined by extending the alphabet  $\Sigma$  and metric  $D_\Sigma$  to include a “gap” as a character, say “-” (let  $\tilde{\Sigma}$  denote the extended alphabet), and by assigning a distance  $D_{\tilde{\Sigma}}(x, -)$  for each character  $x$  in the original alphabet.

*Theorem 9.4 of [1] says that  $d_e$  is a metric on strings of letters in  $\Sigma$  whenever  $D_{\tilde{\Sigma}}$  is a metric on the extended alphabet.*

### 1.1 Two-letter alphabets

The simplest non-trivial example is an alphabet with two letters, say  $x$  and  $y$ , when there is only one distance  $D_\Sigma(x, y)$  that is non-zero. The requirement that the triangle inequality hold for  $D_{\tilde{\Sigma}}$  reduces to three inequalities that can be expressed as

$$|D_{\tilde{\Sigma}}(x, -) - D_{\tilde{\Sigma}}(y, -)| \leq D_\Sigma(x, y) \leq D_{\tilde{\Sigma}}(x, -) + D_{\tilde{\Sigma}}(y, -). \quad (1.1)$$

Together with the condition that all distances be non-negative, we see that (1.1) characterizes completely the requirement for  $D_{\tilde{\Sigma}}$  to be a metric in the case of a two-letter alphabet  $\Sigma$ .

## 1.2 General alphabets

For a general alphabet  $\Sigma$ , if

$$\alpha \leq D_{\Sigma}(x, y) \leq 2\alpha \tag{1.2}$$

for all  $x \neq y$  (including  $-$ ) for some  $\alpha > 0$ , then  $D_{\Sigma}$  is a metric (that is, the triangle inequality holds). This is because

$$D_{\Sigma}(x, y) \leq 2\alpha \leq D_{\Sigma}(x, z) + D_{\Sigma}(z, y) \tag{1.3}$$

for any  $z \in \Sigma$ . One simple choice for a metric on letters is to choose  $D_{\Sigma}(x, y) = 1$  for all  $x \neq y$ , and then to take  $D_{\tilde{\Sigma}}(x, -) = 2$ ; the resulting  $D_{\tilde{\Sigma}}$  satisfies (1.2) for  $\tilde{\Sigma}$ . However, condition (1.2) is far from optimal as the example (1.1) shows.

The edit distance  $d_e$  is derived from the extended alphabet distance  $D_{\tilde{\Sigma}}$  as follows. We introduce the notion of *alignment*  $\mathcal{A}$  of sequences  $(x^*, y^*) = \mathcal{A}(x, y)$  where  $x^*$  has the letters of  $x$  in the same order but possibly with gaps  $-$  inserted, and similarly for  $y^*$ . We suppose that  $x^*$  and  $y^*$  have the same length even if  $x$  and  $y$  did not, which can always be achieved by adding gaps at one end or the other. Then

$$d_e(x, y) = \min_{\mathcal{A}} \sum_i D_{\tilde{\Sigma}}(x_i^*, y_i^*). \tag{1.4}$$

for any  $z \in \Sigma$ . The minimum is over all alignments  $\mathcal{A}$  and the sum extends over the length of the sequences. Fortunately, string-edit distance  $d_e$ , and even more complex metrics involving more complex gap penalties, can be computed efficiently by the dynamic programming algorithm [1].

The simple string-edit distance  $d_e$  described here is useful in many contexts. However, more complex metrics would be required in other applications.

## 1.3 Distance versus score

Typically biologists prefer to work with a “score” that is large when two sequences are close as opposed to a distance which is small in such a case. The dynamic programming algorithm can equivalently be used to minimize the distance or maximize a score. There is a formal correspondence that can be made between scores and distances, as follows [1].

## References

- [1] Michael Waterman. *Introduction to Computational Biology*. Chapman & Hall/CRC Press, 1995.