On the Similarity Metric and Distance Metric

Kaizhong Zhang

Department of Computer Science
The University of Western Ontario
London, Ontario, Canada, N6A 5B7

Joint work with Shihyen Chen and Bin Ma
Introduction

Similarity measure and distance measure are used in many applications.

It is often important to define a proper similarity or distance measure.

How do we define a good similarity measure or distance measure?

- What are the essential requirements for similarity or distance measure?

- What is the relationship between similarity measure and distance measure?

- How do we modify an existing similarity or distance measure for our application?
Similarity Measure and Distance Measure

What is a similarity measure and why do we need it?

- A large distance score means that the two objects being measured are different.

- A large similarity score means that the two objects being measured are similar.

So they are opposite and is this the only difference?
An example of similarity metric used in Bioinformatics.

**BLOSUM62 Substitution Matrix**

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>S</th>
<th>T</th>
<th>P</th>
<th>A</th>
<th>G</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>Q</th>
<th>H</th>
<th>R</th>
<th>K</th>
<th>M</th>
<th>I</th>
<th>L</th>
<th>V</th>
<th>F</th>
<th>Y</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>9</td>
<td>-1</td>
<td>-1</td>
<td>-3</td>
<td>0</td>
<td>-3</td>
<td>-3</td>
<td>-4</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>S</td>
<td>-1</td>
<td>4</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
</tr>
<tr>
<td>T</td>
<td>-1</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
</tr>
<tr>
<td>P</td>
<td>-3</td>
<td>-1</td>
<td>1</td>
<td>7</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-4</td>
<td>-3</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>4</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
</tr>
<tr>
<td>G</td>
<td>-3</td>
<td>0</td>
<td>1</td>
<td>-2</td>
<td>0</td>
<td>6</td>
<td>-2</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-4</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
</tr>
<tr>
<td>N</td>
<td>-3</td>
<td>1</td>
<td>0</td>
<td>-2</td>
<td>-2</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
</tr>
<tr>
<td>D</td>
<td>-3</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-3</td>
<td>-4</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
</tr>
<tr>
<td>E</td>
<td>-4</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
</tr>
<tr>
<td>Q</td>
<td>-3</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>H</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>R</td>
<td>-3</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>-2</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
</tr>
<tr>
<td>K</td>
<td>-3</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>2</td>
<td>5</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
</tr>
<tr>
<td>M</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>0</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>I</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-1</td>
<td>-4</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>-1</td>
<td>-3</td>
<td>-3</td>
</tr>
<tr>
<td>L</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-1</td>
<td>-4</td>
<td>-3</td>
<td>-4</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>V</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>0</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>-1</td>
<td>-1</td>
<td>-3</td>
</tr>
<tr>
<td>F</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-4</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-3</td>
<td>-3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>6</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Y</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>2</td>
<td>-2</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>3</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>W</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-4</td>
<td>-3</td>
<td>-2</td>
<td>-4</td>
<td>-4</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
When defining a similarity or distance measure, we ask:

- Is this measure what we really want? Have we considered all aspects? If not, then
  - What is missing?
  - How should we modify it?
To conveniently answer these questions:

- establish a set of general guidelines (minimum requirements)
- use these guidelines to ensure that we are on the right track
  - Meeting all these guidelines is not mandatory.
  - However, if a guideline is not met, there should be a good reason.
Definition: Distance Metric (well known)

A distance metric on a set $X$ is a nonnegative function $d(x, y)$ on the Cartesian product $X \times X$ satisfying the following properties. For all $x, y, z \in X$:

1. $d(x, y) \geq 0$

2. $d(x, y) = d(y, x)$

3. $d(x, z) \leq d(x, y) + d(y, z)$

4. $d(x, y) = 0$ if and only if $x = y$
Definition: Similarity Metric (our work)

Given a set $X$, a real-valued function $s(x, y)$ on the Cartesian product $X \times X$ is a similarity metric if, for all $x, y, z \in X$, it satisfies:

1. $s(x, x) \geq 0$

2. $s(x, y) = s(y, x)$

3. $s(x, x) \geq s(x, y)$

4. $s(x, y) + s(y, z) \leq s(x, z) + s(y, y)$

5. $s(x, x) = s(y, y) = s(x, y)$ if and only if $x = y$
Conditions 1, 2, and 3 are clear.

Conditions 4 and 5 need some explanation.

Under a distance metric, when do we know that two elements, $x$ and $y$, are the same?

$$d(x, y) = 0 \iff \forall z \; d(x, z) = d(y, z)$$

Under a similarity metric, when do we know that two elements, $x$ and $y$, are the same?

$$s(x, y) = s(x, x) = s(y, y) \iff \forall z \; s(x, z) = s(y, z)$$
Why $s(x, y) + s(y, z) \leq s(x, z) + s(y, y)$ (condition 4)?

Why not just $s(x, y) + s(y, z) \leq s(x, z)$

An example where the similarity between two persons is measured by the number of common friends they have.
Two alignments: which one is better/correct?

Alignment 1
V A V A V A V A V A V A V A V A V A V A V A V A V A V
V I V I V I V I V I V I V I V I V I V I V I V I V I V

Alignment 2
V A V A V A V A V A V A V A V A V A V A V A V A V A V
- V I V I V I V I V I V I V I V I V I V I V I V I V I V
Two alignments: which one is better/correct?

Alignment 1
V V A V V V A V V V A V V V A V V V A V V V A V V V A V V V A
V V I V V V I V V V I V V V I V V V I V V V I V V V I

Alignment 2
V V A V V V A V V V A V V V A V V V A V V V A V V V A V V V A V V V A
- V V I V V V I V V V I V V V I V V V I V V V I V V V I
This is an example using Blosum70.

In blosum70, \( s(A,V) = 0 \), \( s(V,I) = 3 \), \( s(A,I) = -2 \) and \( s(V,V) = 4 \) which violate condition 4.

However before rounding, \( s(A,V) = -0.1480 \), \( s(V,I) = 1.2795 \), \( s(A,I) = -0.7728 \), and \( s(V,V) = 2.0474 \) which satisfy condition 4.
Why condition 4?

Consider the similarity between two objects as their common properties, then it can be represented as set intersection.

\[ |X \cap Y| + |Y \cap Z| = |X \cap Y \cap Z| + |X \cap Y \cap \bar{Z}| + |X \cap Y \cap Z| + |\bar{X} \cap Y \cap Z| \]

Since \(|X \cap Y \cap Z| \leq |X \cap Z|\) and \(|X \cap Y \cap \bar{Z}| + |X \cap Y \cap Z| + |\bar{X} \cap Y \cap Z| \leq |Y \cap Y|\),

\[ |X \cap Y| + |Y \cap Z| \leq |X \cap Z| + |Y \cap Y| \]
Some examples satisfying our general definition of similarity metric:

- The set intersection \(|A \cap B|\)
- The mutual information \(I(X, Y)\) between variables \(X\) and \(Y\)
- The amino acid similarity based on BLOSUM-62 matrix
- The protein sequence similarity obtained by using the Smith-Waterman algorithm based on the BLOSUM-62 matrix
• Distance metric is a special case of similarity metric

\[ -d(x, y) \] is a similarity metric

\[ -s(x, y) \] is not a distance metric

• The distance and similarity metrics are often interconvertible

\[ d_s(x, y) = s(x, x) + s(y, y) - 2s(x, y) \]

\[ s_d(x, y) = \frac{d(x, o) + d(y, o) - d(x, y)}{2}, \text{ where } o \text{ is any fixed element} \]

• Similarity is more general than distance?

\[ d_{sd}(x, y) = d(x, y) \]

\[ s_{ds}(x, y) = s(x, y), \text{ if } \exists o \text{ s.t. } s(o, z) = 0 \text{ for any } z \]
Properties

• Let \( s_1(x, y), s_2(x, y) \) be similarity metrics, then

\[
s_1(x, y) + s_2(x, y)
\]

is a similarity metric.

• Let \( s_1(x, y), s_2(x, y) \) be nonnegative similarity metrics, then

\[
s_1(x, y) \times s_2(x, y)
\]

is a similarity metric.
Normalized Similarity/Distance Metric

- In many applications, one needs to convert a distance/similarity metric to a normalized one.

- Examples:

  Similarity metric: \( s(a, a) = 1 \) and \( s(g, t) = -1 \)

    \[
    x = aaa \quad y = aaa, \quad s(x, y) = 3
    \]

    \[
    v = agaagaagaaga \quad w = ataataataata, \quad s(v, w) = 4
    \]

  Distance metric: \( d(g, t) = 1 \)

    \[
    x = gg \quad y = tt, \quad d(x, y) = 2
    \]

    \[
    v = aaagaagaagaag \quad w = aaataaataaat, \quad d(v, w) = 3
    \]
• What is the definition of normalized similarity or distance metric?

• Given a similarity/distance metric, how do we derive a normalized metric?
  
  There are research work on normalizing the edit distance between sequences.

• We show general solutions.
Definition: Normalized Similarity and Distance Metric

- A distance metric \( d(x, y) \) is a normalized distance metric if
  \[
d(x, y) \leq 1
  \]

- A similarity metric \( s(x, y) \) is a normalized similarity metric if
  \[
  |s(x, y)| \leq 1
  \]
Examples of normalized similarity metric

• Some examples satisfying our definition of normalized similarity metric:
  – For set intersection $|A \cap B|$: $\frac{|A \cap B|}{|A \cup B|}, \frac{|A \cap B|}{\max\{|A|,|B|\}}$
  – For mutual information $I(X,Y)$: $\frac{I(X,Y)}{H(X,Y)}, \frac{I(X,Y)}{\max\{H(X),H(Y)\}}$

• Some examples of normalized distance metric:
  – For set intersection $|A \cap B|$: $\frac{|A-B|+|B-A|}{|A \cup B|}, \frac{|A-B|+|B-A|}{\max\{|A|,|B|\}}$
  – For mutual information $I(X,Y)$: $\frac{H(X|Y)+H(Y|X)}{H(X,Y)}, \frac{H(X|Y)+H(Y|X)}{\max\{H(X),H(Y)\}}$
Relationships

• If \( d(x, y) \) is a normalized distance metric, then

\[ s_d(x, y) = 1 - d(x, y) \]

is a normalized similarity metric.

• If \( s(x, y) \) is a nonnegative normalized similarity metric and \( s(x, x) = 1 \) for any \( x \), then

\[ d_s(x, y) = 1 - s(x, y) \]

is a normalized distance metric.
Let $d_1(x, y), d_2(x, y), \cdots d_n(x, y)$ be normalized distance metrics, then the following is a normalized distance metric.

$$1 - (1 - d_1(x, y)) \times (1 - d_2(x, y)) \times \cdots \times (1 - d_n(x, y))$$
Normalized Similarity Metric

Suppose $s(x, y)$ is a similarity metric:

- $\frac{s(x, y)}{s(x,x) + s(y,y) - s(x,y)}$

Suppose $s(x, y)$ is a nonnegative similarity metric:

- $\frac{s(x, y)}{\max\{s(x,x), s(y,y)\}}$

Normalized Distance Metric

Suppose $d(x, y)$ is a distance metric and $o \in X$:

- $\frac{2d(x, y)}{d(x,o) + d(y,o) + d(x,y)}$

- $\frac{d(x, y)}{2 \max\{d(x,o), d(y,o)\}} - \frac{\min\{d(x,o), d(y,o)\}}{2 \max\{d(x,o), d(y,o)\}} + \frac{1}{2}$
Similarity metric is more general?

- Normalized similarity metric formulae are more intuitive

- Normalized distance metric formulae can be derived from normalized similarity metric formulae
Given a distance metric \( d(x, y) \), \( s_d(x, y) = d(x, o) + d(y, o) - d(x, y) \) is a nonnegative similarity metric.

Therefore

\[
1 - \frac{s_d(x, y)}{\max\{s_d(x, x), s_d(y, y)\}}
\]

is a normalized distance metric.

Since \( s_d(x, x) = 2d(x, o) \) and \( s_d(y, y) = 2d(y, o) \), we have

\[
1 - \frac{s_d(x, y)}{\max\{s_d(x, x), s_d(y, y)\}} = 1 - \frac{d(x, o) + d(y, o) - d(x, y)}{2 \max\{d(x, o), d(y, o)\}}
\]

Finally

\[
\frac{d(x, y)}{2 \max\{d(x, o), d(y, o)\}} - \frac{\min\{d(x, o), d(y, o)\}}{2 \max\{d(x, o), d(y, o)\}} + \frac{1}{2}
\]

is a normalized distance metric.
Set similarity and distance metric

- $|A \cap B|$ is a similarity metric.

- $|A \cup B| - |A \cap B|$ is a distance metric.

- \[
\frac{|A \cap B|}{\max\{|A|,|B|\}}
\] is a normalized similarity metric.

- \[
\frac{\max\{|A-B|,|B-A|\}}{\max\{|A|,|B|\}}
\] is a normalized distance metric.

- \[
\frac{|A \cap B|}{|A \cup B|}
\] is a normalized similarity metric.

- \[
\frac{|A-B|+|B-A|}{|A \cup B|}
\] is a normalized distance metric.
Information similarity and distance metric

- $I(X, Y)$ is a similarity metric.
- $H(X|Y) + H(Y|X)$ is a distance metric.
- $\frac{I(X,Y)}{\max\{H(X), H(Y)\}}$ is a normalized similarity metric.
- $\frac{\max\{H(X|Y), H(Y|X)\}}{\max\{H(X), H(Y)\}}$ is a normalized distance metric.
- $\frac{I(X,Y)}{H(X,Y)}$ is a normalized similarity metric.
- $\frac{H(X|Y)+H(Y|X)}{H(X,Y)}$ is a normalized distance metric.
Sequence edit distance and similarity

• If the costs of insertion, deletion, and substitution is a distance metric, then the sequence edit distance \(d(s, t)\), between two sequences \(s\) and \(t\), is also a distance metric.

• Several normalized edit distances have been proposed and studied.

\[
\frac{d(s, t)}{|s| + |t|}, \quad \frac{d(s, t)}{\min\{|s|, |t|\}}, \quad \frac{d(s, t)}{\max\{|s|, |t|\}}
\]

\[n(s, t) = \min\left\{ \frac{p(s, t)}{|p|} \mid p \text{ is a path that change } s \text{ to } t \right\}
\]

• In fact, they are not distance metric.

• Our result: \[\frac{d(s, t)}{\max\{|s|, |t|\}} - \frac{\min\{|s|, |t|\}}{\max\{|s|, |t|\}} + 1\] is a distance metric.
General “Normalized” Similarity Metric

Suppose \( s(x,y) \) is a similarity metric:

\[
\frac{s(x,y)}{f(s(x,x)+s(y,y)-s(x,y))}
\]

where \( f(x) \) is concave over \([0, \infty)\), \( f(0) \geq 0 \), \( f(x) > 0 \) if \( x > 0 \), \( f(x) \geq f(y) \) if \( x \geq y \)

Suppose \( s(x,y) \) is a nonnegative similarity metric:

\[
\frac{s(x,y)}{g(\max\{s(x,x),s(y,y)\})}
\]

where \( g(x) \) is over \([0, \infty)\), \( g(0) \geq 0 \), \( g(x) > 0 \) if \( x > 0 \), \( g(x) \geq g(y) \) if \( x \geq y \)

Suppose \( s(x,y) \) is a nonnegative similarity metric:

\[
\frac{s(x,y)}{f(\max\{s(x,x),s(y,y)\})+\lambda(\min\{s(x,x),s(y,y)\}-s(x,y))}, \quad 0 \leq \lambda \leq 1
\]

where \( f(x) \) is concave over \([0, \infty)\), \( f(0) \geq 0 \), \( f(x) > 0 \) if \( x > 0 \), \( f(x) \geq f(y) \) if \( x \geq y \)
Suppose $d(x, y)$ is a distance metric:

- $\frac{d(x,y) - \min\{d(x,o),d(y,o)\}}{g(\max\{d(x,o),d(y,o)\})} + \frac{\min\{d(x,o),d(y,o)\}}{g(\min\{d(x,o),d(y,o)\})}$,

  where $g(x)$ is over $[0, \infty)$, $g(0) \geq 0$, $g(x) > 0$ if $x > 0$, $g(x) \geq g(y)$ if $x \geq y$

- $\frac{d(x,y) - \min\{d(x,o),d(y,o)\}}{f(\max\{d(x,o),d(y,o)\})} + \max\{d(x,o),d(y,o)\}$

- $\frac{d(x,y)}{f(d(x,o) + d(y,o) + d(x,y))}$

where $f(x)$ is concave over $[0, \infty)$, $f(0) \geq 0$, $f(x) > 0$ if $x > 0$, $f(x) \geq f(y)$ if $x \geq y$
Let $s(x, y)$ be a similarity metric, then for $1 \leq p$ $d_{L_p}(x, y)$ is a distance metric.

$$d_{L_p}(x, y) = \sqrt[p]{[s(x, x) - s(x, y)]^p + [s(y, y) - s(x, y)]^p} \quad 1 \leq p$$

Suppose $s(x, y)$ is a similarity metric:

- $\frac{s(x, y)}{s(x, x) + s(y, y) - s(x, y)} = \frac{s(x, y)}{d_{L_1}(x, y) + s(x, y)}$

Suppose $s(x, y)$ is a nonnegative similarity metric:

- $\frac{s(x, y)}{\max\{s(x, x), s(y, y)\}} = \frac{s(x, y)}{d_{L_\infty}(x, y) + s(x, y)}$

How about $d_{L_p}(x, y)$?

- $\frac{s(x, y)}{d_{L_p}(x, y) + s(x, y)}$?
New: “Normalized” Similarity and Distance Metric

Let $s(x, y)$ be a similarity metric, then for $1 \leq p$ $d_{L^p}(x, y)$ is a distance metric.

$$d_{L^p}(x, y) = \sqrt[p]{[s(x, x) - s(x, y)]^p + [s(y, y) - s(x, y)]^p} \quad 1 \leq p$$

<table>
<thead>
<tr>
<th>Formula</th>
<th>$L_p$</th>
<th>$s(x, y)$</th>
<th>$f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{s(x, y)}{f(d_{L^1}(x, y) + s(x, y))}$</td>
<td>$p = 1$</td>
<td>any</td>
<td>concave</td>
</tr>
<tr>
<td>$\frac{s(x, y)}{f(d_{L^p}(x, y) + s(x, y))}$</td>
<td>$1 &lt; p &lt; \infty$</td>
<td>$\geq 0$</td>
<td>concave</td>
</tr>
<tr>
<td>$\frac{s(x, y)}{f(d_{L^\infty}(x, y) + s(x, y))}$</td>
<td>$p = \infty$</td>
<td>$\geq 0$</td>
<td>any</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Formula</th>
<th>$L_p$</th>
<th>$s(x, y)$</th>
<th>$f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_{L^1}(x, y)$</td>
<td>$p = 1$</td>
<td>any</td>
<td>concave</td>
</tr>
<tr>
<td>$\frac{d_{L^1}(x, y)}{f(d_{L^1}(x, y) + s(x, y))}$</td>
<td>$p = 1$</td>
<td>any</td>
<td>concave</td>
</tr>
<tr>
<td>$d_{L^p}(x, y)$</td>
<td>$1 &lt; p &lt; \infty$</td>
<td>$\geq 0$</td>
<td>concave</td>
</tr>
<tr>
<td>$\frac{d_{L^\infty}(x, y)}{f(d_{L^\infty}(x, y) + s(x, y))}$</td>
<td>$p = \infty$</td>
<td>$\geq 0$</td>
<td>concave</td>
</tr>
</tbody>
</table>
“Normalized” set similarity and distance metric

<table>
<thead>
<tr>
<th>Similarity metric</th>
<th>Distance metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>(</td>
<td>A \cap B</td>
</tr>
<tr>
<td>(\frac{</td>
<td>A \cap B</td>
</tr>
<tr>
<td>(\frac{</td>
<td>A \cap B</td>
</tr>
<tr>
<td>(\frac{</td>
<td>A \cap B</td>
</tr>
<tr>
<td>(\frac{</td>
<td>A \cap B</td>
</tr>
<tr>
<td>(\frac{</td>
<td>A \cap B</td>
</tr>
<tr>
<td>(\frac{</td>
<td>A \cap B</td>
</tr>
</tbody>
</table>
**“Normalized” information similarity and distance metric**

<table>
<thead>
<tr>
<th>Similarity metric</th>
<th>Distance metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>( I(X, Y) )</td>
<td>( \sqrt{H(X</td>
</tr>
<tr>
<td>( \frac{I(X,Y)}{f(\sqrt{H(X</td>
<td>Y)^p + H(Y</td>
</tr>
<tr>
<td>( \frac{I(X,Y)}{f(H(X,Y))} )</td>
<td>( \frac{H(X</td>
</tr>
<tr>
<td>( \frac{I(X,Y)}{\max{H(X), H(Y)}} )</td>
<td>( \max{H(X</td>
</tr>
<tr>
<td>( \frac{I(X,Y)}{H(X,Y)+k} )</td>
<td>( \frac{H(X</td>
</tr>
<tr>
<td>( \frac{I(X,Y)}{\sqrt{H(X,Y)}} )</td>
<td>( \frac{H(X</td>
</tr>
<tr>
<td>( \log(\max{H(X), H(Y)})+1 )</td>
<td>( \log(\max{H(X), H(Y)})+1 )</td>
</tr>
<tr>
<td>( \log(\sqrt{H(X</td>
<td>Y)^2 + H(Y</td>
</tr>
</tbody>
</table>
Smith-Waterman algorithm:

- similarity is defined in terms of the additive score of an alignment
- does not reveal the consistency of score distribution within the alignment
Inconsistent Score Distribution of an Alignment

The highest-scoring segment: \((S_1[i, j], S_2[k, l])\) (score = 60 − 45 + 50 = 65)

However, \((S_1[i, j], S_2[k, l])\) contains a poor-scoring segment (score = −45)

A better choice: the leftmost segment
  − score is still good (60 vs. 65)
  − better consistency

Q: How to find this better one?
A: Compute normalized similarity.
Computation View: Mosaic Effect

(0, 0)^{t_1 t_2} \rightarrow t_n

s_1

s_2

s_m

(m, n)
New Formulae: Normalized Local Similarity

**Input:** two sequences \( S_1 \) and \( S_2 \)

**Output:** local similarity between \( S_1 \) and \( S_2 \)

\( \theta_{12}(i', i, j', j) \): the similarity of the subsequences \( S_1[i', i] \) and \( S_2[j', j] \)

1. \[
\max\left\{ \theta_{12}(i', i, j', j) \right\}, \text{ where } 1 \leq i' \leq i \leq m, \ 1 \leq j' \leq j \leq n, \ h \geq 0
\]

2. \[
\max\left\{ \frac{\theta_{12}(i', i, j', j)}{f(\theta_{11}(i', i, i') + \theta_{22}(j', j, j') - \theta_{12}(i', i, j', j))} \right\}, \text{ where } 1 \leq i' \leq i \leq m, \ 1 \leq j' \leq j \leq n, \ f(x) \text{ is concave over } [0, \infty), \ f(0) \geq 0, \ f(x) > 0 \text{ if } x > 0, \ f(x) \geq f(y) \text{ if } x \geq y
\]

3. \[
\max\left\{ \frac{\theta_{12}(i', i, j', j)}{h + \max\{\theta_{11}(i', i, i'), \theta_{22}(j', j, j')\}} \right\}, \text{ where } 1 \leq i' \leq i \leq m, \ 1 \leq j' \leq j \leq n, \ h \geq 0
\]

4. \[
\max\left\{ \frac{\theta_{12}(i', i, j', j)}{g(\max\{\theta_{11}(i', i, i'), \theta_{22}(j', j, j')\})} \right\}, \text{ where } 1 \leq i' \leq i \leq m, \ 1 \leq j' \leq j \leq n, \ g(0) \geq 0, \ g(x) > 0 \text{ if } x > 0, \ g(x) \geq g(y) \text{ if } x \geq y
\]

5. \[
\max\left\{ \frac{\theta_{12}(i', i, j', j)}{h + \sqrt[\nu]{[\theta_{11}(i', i, i') - \theta_{12}(i', i, j', j)]^\nu + [\theta_{22}(j', j, j') - \theta_{12}(i', i, j', j)]^\nu + \theta_{12}(i', i, j', j)}} \right\}, \text{ where } 1 \leq i' \leq i \leq m, \ 1 \leq j' \leq j \leq n, \ h \geq 0
\]
Algorithmic Techniques

- fractional programming
- dynamic programming
Formula 2 with Dynamic Programming

Input:  \( S_1[1,m], S_2[1,n], h > 0 \)

Output:  \( \max\{ \frac{\theta_{12}(i',i,j',j)}{f(\theta_{11}(i,i,i,i)+\theta_{22}(j,j,j,j)-\theta_{12}(i,i,j,j))} | 1 \leq i' \leq i \leq m, 1 \leq j' \leq j \leq n, f(x) \text{ is concave over } [0, \infty), f(0) \geq 0, f(x) > 0, \text{ if } x > 0, f(x) \geq f(y) \text{ if } x \geq y \} \)

1:  \( K \leftarrow \theta_{11}(1,m,1,m) + \theta_{22}(1,n,1,n) \)
2:  for \( k \leftarrow 1 \) to \( K \) do
3:      \( x \leftarrow 0 \)  \( \Gamma(0,0,k) \leftarrow 0 \)
4:      for \( i \leftarrow 1 \) to \( m \) do
5:          \( \Gamma(i,0,k) \leftarrow 0 \)
6:      end for
7:      for \( j \leftarrow 1 \) to \( n \) do
8:          \( \Gamma(0,j,k) \leftarrow 0 \)
9:      end for
10:     for \( i \leftarrow 1 \) to \( m \) do
11:         for \( j \leftarrow 1 \) to \( n \) do
12:             \( \Gamma(i,j,k) \leftarrow \max\{0, \Gamma(i',j',k') + \theta_{12}(i,i,j,j) \times (i-i') \times (j-j') + \theta_{11}(i,i,0,0) \times (i-i') \times (1-j+j') + \theta_{22}(0,0,j,j) \times (1-i+i') \times (j-j') | (i',j') \in \{(i-1,j-1),(i-1,j),(i,j-1)\}, k' = k - \theta_{11}(i,i,i,i) \times (i-i') - \theta_{22}(j,j,j,j) \times (j-j'), k' \geq 0 \} \)
13:             if \( \Gamma(i,j,k) > x \) then
14:                 \( x \leftarrow \Gamma(i,j,k) \)
15:             end if
16:         end for
17:     end for
18:     \( a[k] \leftarrow x \)
19: end for
20: \( \lambda \leftarrow \frac{a[1]}{f(1-a[1])} \)
21: for \( k \leftarrow 1 \) to \( K \) do
22:     if \( \frac{a[k]}{f(k-a[k])} > \lambda \) then
23:         \( \lambda \leftarrow \frac{a[k]}{f(k-a[k])} \)
24:     end if
25: end for
26: return \( \lambda \)
## Algorithmic Results

<table>
<thead>
<tr>
<th>Formula</th>
<th>Time Complexity</th>
<th>Space Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\mathcal{O}(m \times n \times \log(m + n))$</td>
<td>$\mathcal{O}(\min{m, n})$</td>
</tr>
<tr>
<td>2</td>
<td>$\mathcal{O}(m \times n \times (m + n))$</td>
<td>$\mathcal{O}(m \times n)$</td>
</tr>
</tbody>
</table>

1. $\max\left\{ \frac{\theta_{12}(i', i, j', j)}{h + \theta_{11}(i', i, i') + \theta_{22}(j', j, j', j) - \theta_{12}(i', i, j', j)} \right\}$, where $1 \leq i' \leq i \leq m$, $1 \leq j' \leq j \leq n$, $h \geq 0$

2. $\max\left\{ \frac{\theta_{12}(i', i, j', j)}{f(\theta_{11}(i', i, i') + \theta_{22}(j', j, j', j) - \theta_{12}(i', i, j', j))} \right\}$, where $1 \leq i' \leq i \leq m$, $1 \leq j' \leq j \leq n$, $f(x)$ is concave over $[0, \infty)$, $f(0) \geq 0$, $f(x) > 0$, if $x > 0$, $f(x) \geq f(y)$ if $x \geq y$
Summary

• Formalized some aspects regarding similarity metric

• Studied relationship between similarity metric and distance metric

• Established new formulae based on our formal definition of normalized similarity metric

• Established new formulae of normalized distance metric

• Constructed algorithms for computing sequence normalized local similarity based on our proposed formulae