Similarity Measures

Stewart Adcock

Last updated: 15th May 2002

Abstract

GA-based applications, it is often desirable to have a measure of the similarity between two chromosomes or between two sets of chromosomes. Such cases include the detection of genetic convergence in a population, and the selction of genetically distant entities. This document details the background and definitions for the similarity measures provided with GAUL. The background information is fairly general and is hopefully useful for other applications. For example, molecular similarity is an important concept in computer-aided molecular design. If this is what interests you, read "chromosome" as "physico-chemical fingerprint".

1 Equality tests

In GAUL, two chromsomes are defined to be equal if, and only if, all alleles share exactly the same values. It is possible that two given chromosomes share exactly the same phenotype but are not necessarily equal (e.g. if the modulus of each allele value is critical to the fitness, rather than the absolute value of each allele).

2 Distance and similiarity measures

Several definitions of the distance between pairs of chromosomes exist. These may all be further subdivided into two variants. "Continuous" distance measures are appropriate for real-valued alleles, whilst "dichotomous" distance measures are appropriate for binary valued alleles.

Sometimes it is more convenient to define the distance in terms of a "similarity" coefficient. A similarity coefficient is simply a complementary measure to distance, that is, how close two chromsomes are in the genetic sapce. In the following text, the distance between two chromosomes or sets of chromosomes, A and B, are represented by D_{AB}, and the similarity (often just $1 - D_{AB}$) represented by S_{AB}. In all equations, N is the total number of alleles and A_i is the value of the ith allele of chromosome A.

Please be aware that the distance and similarity measured described below are often known by alternative names. For example, the Dice similarity coefficient from statistics is known as the Hodgkin index in the fields of computational and quantum chemistry. Since I am a computational chemist in the real world (i.e. outside of my apartment), I shall naturally refer to this measure as the Hodgkin similarity. If you are suspicious of my naming conventions, then please refer to these measures by whichever terms you feel most comfortable with...

3 Dichotomous similarity measures

A general definition of similarity for binary data is the Tversky coefficient [**?**]:

$$S_{AB}^{Tversky} = \frac{n}{\alpha\,(a-n) + \beta\,(b-n) + n} \tag{1}$$

alpha and *beta* are adjustable variables. a is the number of "1"s in chromosome A. b is the number of "1"s in chromosome B. n is the number of alleles which are "1" in both chromsomes (i.e. $A\&B$).

You may notice that this Tversky coefficient is asymmetric, that is $S_{AB}^{Tversky} \neq S_{BA}^{Tversky}$. Special cases exist for $\alpha = 1$, $\beta = 1$ (the Dice coefficient) and $\alpha = \frac{1}{2}$ and $\beta = \frac{1}{2}$ (the Jaccard or Tanimoto coefficient).

$$S_{AB}^{Tanimoto} = \frac{n}{a+b-n} \tag{2}$$

$$S_{AB}^{Dice} = \frac{2n}{a+b} \tag{3}$$

These three measures, $S_{AB}^{Tversky}$, $S_{AB}^{Tanimoto}$ and S_{AB}^{Dice} all yeild results in the range 0 to 1.

More common distance measures include $D_{AB}^{Euclidean}$, Euclidean distance, and $D_{AB}^{Hamming}$, Hamming, or Manhattan, or the city-block distance. These are defined as follows:

$$D_{AB}^{Euclidean} = (a+b-2n)^{\frac{1}{2}} \tag{4}$$

$$D_{AB}^{Hamming} = a + b - 2n \tag{5}$$

To convert these distance measures into more appropriate similarity corefficients, they may be normallised and subtracted from 1:

$$S_{AB}^{Euclidean} = 1.0 - \frac{(a + b - 2n)^{\frac{1}{2}}}{N} \tag{6}$$

$$S_{AB}^{Hamming} = 1.0 - \frac{a + b - 2n}{N} \tag{7}$$

An alternative similarity measure is the Cosine coefficient, S_{AB}^{Cosine}:

$$S_{AB}^{Cosine} = \frac{n}{(ab)^{\frac{1}{2}}} \tag{8}$$

4 Continuous similarity measures

All six of the dichotomous similarity measures defined above provide values in the range 0 to 1 which is very convenient. Unfortuneately the continuous measures do not. Normalisation of these coefficients is often not straight forward because the range of allele values may not be clear. You have been warned.

If A_i is the ith element of vector A, the Tanimoto, Dice and Cosine similarity measures are defined as:

$$S_{AB}^{Tanimoto} = \frac{\sum_{i=1}^{N} A_i B_i}{\sum_{i=1}^{N} (A_i)^2 + \sum_{i=1}^{N} (B_i)^2 - \sum_{i=1}^{N} A_i B_i} \tag{9}$$

$$S_{AB}^{Dice} = \frac{2 \sum_{i=1}^{N} A_i B_i}{\sum_{i=1}^{N} (A_i)^2 + \sum_{i=1}^{N} (B_i)^2} \tag{10}$$

$$S_{AB}^{Cosine} = \frac{\sum_{i=1}^{N} A_i B_i}{\left\{ \sum_{i=1}^{N} (A_i)^2 \sum_{i=1}^{N} (B_i)^2 \right\}^{\frac{1}{2}}} \tag{11}$$

These three measures have the ranges $-\frac{1}{3} \geqslant S_{AB}^{tanimoto} \geqslant +1$, $-1 \geqslant S_{AB}^{Dice} \geqslant +1$ and $-1 \geqslant S_{AB}^{Cosine} \geqslant +1$.

Two very common distance measures are the Euclidean and Hamming distances. These both have a value of 0 for identical vectors and may have an value upto inf for non-identical vectors.

5 Summary

Several measures for the evaluation of similarity between two arbitrary vectors have been described. They are ideal for use in, for example, cluster analysis. These have been implemented for bitstring vectors, integer vectors and double-precision floating-point vectors within the GAUL GA library. Standalone code may be furnished by Stewart upon request.

One problem with all of the above measures is that they may only be applied to assess the similarity of vectors with equal lengths. When comparing, say, two protein sequences then the vectors will probably have differing lengths. There are numerous approaches to overcome this difficulty. They are not discussed in this report.

References