The Underlying Similarity of Diversity Measures Used in Evolutionary Computation

Mark Wineberg and Franz Oppacher

Computing and Information Science
University of Guelph
Guelph, Canada
wineberg@cis.uoguelph.ca

School of Computer Science
Carleton University
Ottawa, Canada
oppacher@scs.carleton.ca

In this paper we compare and analyze the various diversity measures used in the Evolutionary Computation field. While each measure looks quite different from the others in form, we surprisingly found that the same basic method underlies all of them: the distance between all possible pairs of chromosomes/organisms in the population. This is true even of the Shannon entropy of gene frequencies. We then associate the different varieties of EC diversity measures to different diversity measures used in Biology. Finally we give an $O(n)$ implementation for each of the diversity measures (where n is the population size), despite their basis in an $O(n^2)$ number of comparisons.

1 Introduction

In recent years there has been a growing interest in genetic diversity in the Evolutionary Computation field [1][2]. Diversity maintenance procedures are beginning to be emphasized, especially in the areas of multi-objective optimization [3], and dynamic environments in evolutionary systems[4][5].

There are many different diversity measures that can be found in the literature. The standard diversity measure is the sum of the Hamming distances between all possible pairs of chromosomes. Another popular measure is the use of the Shannon Entropy from Information Theory on gene frequencies. Even variance and standard deviation can be viewed as a diversity measure, especially when using real valued genes. This naturally leads to the question: which of these varied diversity measures is best?

Surprisingly, there appear to be a deep similarity between all of the above different measures. In this paper we analyze all of the above diversity measures and expose this similarity. We also present efficient procedures for computing these diversity measures. For example the time complexity for the "all possible pairs" diversity

measure would naively be thought to be $O(n^2)$ because there are $O(n^2)$ chromosome pairings in a population of size n. The algorithm we present computes the value in $O(n)$ time.

2 All-Possible-Pairs Diversity

The simplest definition of diversity comes from the answer to the question "how different is everybody from everybody else?" If every chromosome is identical, there is no difference between any two chromosomes and hence there is no diversity in the population. If each chromosome is completely different from one another, then those differences add, and the population should be maximally diverse. So the diversity of a population can be seen as the difference between all possible pairs of chromosomes within that population.

While the above definition makes intuitive sense, there is one aspect not covered: what do we mean by different. If a pair of chromosomes is only different by one locus, it only seems reasonable that this pair should not add as much to the diversity of the population as a pair of chromosomes with every locus different. Consequently the difference between chromosomes can be seen as the Hamming distance or chromosome distance, where the Hamming distance is the sum of all loci where the two chromosomes have differing genes. Hence the population diversity becomes the sum of the Hamming distances between all possible pairs of chromosomes; see [6]. In cluster theory this is called the *statistic scatter*, see [7].

Now, since the Hamming distance is symmetric, and is equal to 0 if the strings are the same, only the lower triangle in a chromosome-pairs table need be considered when computing the diversity. Consequently the all-possible-pairs diversity can be formalized as

$$\mathrm{Div}(P) = \sum_{i=1}^{n} \sum_{j=1}^{i-1} \mathrm{hd}(c_i, c_j) \qquad \textbf{(1)}$$

where P is the population and chromosome $c_i \in P$ and n is the population size and $\mathrm{hd}(c_i,c_j)$ is the Hamming distance between two chromosomes.

3 The Reformulation of the All-Possible-Pairs Diversity: A Linear Time Algorithm

A problem with formula (1) is its time complexity. Since the Hamming distance between any two pairs takes $O(l)$ time and there are n^2 possible pairs (actually $\frac{1}{2}n(n-1)$ pairs when symmetry is taken into account), then the time complexity

when using (1) is $O\left(l\,n^2\right)$. Since the time complexity of the GA is $O(l\cdot n)$ calculating the diversity every generation would be expensive.

Fortunately a reformulation of definition (1) can be converted into an $O(l\cdot n)$ algorithm to compute the all-possible-pairs diversity. This will be developed directly.

It seems unlikely such an efficient algorithm would not already be present somewhere in the literature. Unfortunately, most mathematical textbooks that deal with sets of binary strings seem to be interested in the maximum of the distances between all possible pairs and not the average. Unlike the sum or average, the time complexity of the maximum Hamming Distance between all possible pairs cannot be reduced from $O\left(l\,n^2\right)$ to $O(l\cdot n)$ time, so the issue is ignored. While the maximum distance between all possible pairs of chromosomes can also be considered as a diversity measure on a GA population, it is not a particularly good one since it is too sensitive to outlier chromosomes, and does not give fine grain information about changes in the population.

Gene Counts and the Gene Frequencies

Before we can give the reformulation, we shall introduce two terms that will be extensively used throughout the paper: the **gene count** across a population, and the **gene frequency** of a population.

The gene count $\mathrm{c}_k(a)$ is the count across the population of all genes at locus k that equals the symbol a. This means that

$$\mathrm{c}_k(a) = \sum_{i=1}^{n} d_{i,k}(a) \qquad (2)$$

where $d_{i,k}(a)$ is a Kronecker δ that becomes 1 when the gene at locus k in chromosome i equals the symbol α, and otherwise is 0. The array of the gene counts of each locus will be called the **gene count matrix**[1].

The gene frequency $\mathrm{f}_k(a)$ is the ratio of the gene count to the size of the population. In other words,

$$\mathrm{f}_k(a) = \frac{\mathrm{c}_k(a)}{n} \qquad (3)$$

The array of the gene frequencies of each locus will be called the **gene frequency matrix**[2].

[1] For non-evolutionary systems, such as those used for cluster analysis or symbolic machine learning, the gene count matrix could be called the *symbol count matrix*.

[2] Just as with the gene count matrix, the gene frequency matrix could be called the *symbol frequency matrix* when dealing with non-evolutionary systems.

The Reformulation

With the notation in place we can present the alternate form of writing the all-possible-pairs diversity:

Theorem 1a: The all-possible-pairs diversity can be rewritten as

$$\mathrm{Div}(P) = \frac{n^2}{2l} \sum_{k=1}^{l} \sum_{\forall a \in A} \mathrm{f}_k(a)\ (1 - \mathrm{f}_k(a)) \qquad (4)$$

Proof: Let us first examine a chromosome ch_i that at locus k has gene a. When computing all of the possible comparison pairs, the value 0 is obtained from comparisons of ch_i with any other chromosomes that also have the same gene a at locus k. There are $n\,\mathrm{f}_k(a)$ chromosomes with gene a at locus k (including ch_i). Consequently there are $n - n\,\mathrm{f}_k(a)$ comparisons with chromosomes that do not have gene a at locus k, and hence will return the value 1. So the component of the distance attributable to ch_i is $n - n\,\mathrm{f}_k(a)$. Since there are $n\,\mathrm{f}(a)$ chromosomes that have the same gene at locus k, the total distance contributed by chromosomes with gene a at locus k is $n\,\mathrm{f}_k(a)(n - n\,\mathrm{f}_k(a))$, which simplifies to $n^2\,\mathrm{f}_k(a)(1 - \mathrm{f}_k(a))$. Summing over all a will give us double the comparison count (since we are adding to the count both $\mathrm{hd}_k(ch_i, ch_j)$ and $\mathrm{hd}_k(ch_j, ch_i)$). So the true comparison count is $\frac{n^2}{2} \sum_{\forall a \in A} \mathrm{f}_k(a)(1 - \mathrm{f}_k(a))$. Averaging over all loci gives us the result we want. ■

Theorem 1b: The normalized all-possible-pairs diversity can be rewritten as

$$\overline{\mathrm{Div}(P)} = \begin{cases} \dfrac{a}{l \cdot \left((a-1) - \dfrac{r(a-r)}{n^2} \right)} \displaystyle\sum_{k=1}^{l} \sum_{\forall a \in A} \mathrm{f}_k(a) \cdot (1 - \mathrm{f}_k(a)) & a < n \\ \dfrac{n}{l \cdot (n-1)} \displaystyle\sum_{k=1}^{l} \sum_{\forall a \in A} \mathrm{f}_k(a) \cdot (1 - \mathrm{f}_k(a)) & a \geq n \end{cases}$$

where $r = n \cdot \mathrm{mod}\, a$.

Proof: To normalize this formula we must first find the maximum diversity value that can be obtained under any make-up of the population. Once this is found, the normalized diversity is just the regular diversity divided by this value.

Case 1) The alphabet size is strictly less than the size of the population ($a < n$)
In this case the diversity becomes maximal when all gene frequencies are as equal as possible[3], i.e. when $\mathrm{f}_k(a) = \frac{1}{a}$. Substituting this into the diversity equation (4) when modified slightly to handle the case when a does not evenly divide into n produces

$$\max(\mathrm{Div}(P)) = \frac{n^2}{2a}\left((a-1) - \frac{r(a-r)}{n^2}\right) \qquad (5)$$

which gives us the result we want when divided into equation (4).

Case 2) When the alphabet size is greater then or equal to the size of the population ($a \geq n$) each frequency becomes $\frac{1}{n}$. This is because there can only be the n different symbols in the population when maximally diverse. Also since $n \mid n$, $r = 0$. So, by setting $a = n$ and $r = 0$ in equation (5) and simplifying, then dividing this maximum into equation (4), the result required is obtained. ■

In most cases $a < n$ and $a \mid n$, so r is 0 and the normalized all-possible-pairs diversity can be written as

$$\overline{\mathrm{Div}(P)} = \frac{a}{l\,(a-1)} \sum_{k=1}^{l} \sum_{\forall a \in A} \mathrm{f}_k(a)\ (1 - \mathrm{f}_k(a)) \qquad (6)$$

Since, in the majority of GA implementations a binary alphabet is used with an even population size (because crossover children fill the new population in pairs), the above equation becomes

$$\overline{\mathrm{Div}(P)} = \frac{2}{l} \sum_{k=1}^{l} \sum_{\forall a \in A} \mathrm{f}_k(a)\ (1 - \mathrm{f}_k(a)) \qquad (7)$$

An $O(l \cdot n)$ All-Possible-Pairs Diversity Algorithm

The normalized all-possible-pairs diversity measure (6) can be translated into an algorithm that computes this diversity in $O(l \cdot n)$ time.

To keep its time complexity down, each alphabet symbol is replaced in the chromosome with its corresponding index in the alphabet set. For example, if the alphabet is $A = \{a, t, c, g\}$, then the corresponding indices are a=1, t=2, c=3 and g=4, and the chromosome *aatccgctatag* becomes 112334321214. This is done to allow constant time indexing into the gene frequency array, based on the gene values.

---

[3] This can be proven easily by solving for each of the frequencies from the set of equations produced by $\nabla \mathrm{Div}(P) = 0$ after taking into account the constraint $\sum_{\forall a} \mathrm{f}_k(a) = 1$.

The calculation is done in two parts. First the gene frequencies are found in 'findGeneFrequency' then the diversity is computed in 'APP_Diversity':

```
function findGeneFrequencies(p, lgth, a)

args:    population p – a population of chromosome
         int lgth – the length of the chromosome
         int a – the size of the alphabet

vars:    float[lgth,a] geneFreq; int[lgth,a] geneCount
         int[lgth] chr; int gene, i, j, k

code:    geneFreq := makeArray(float, lgth , a)
         geneCount := makeArray(int, lgth , a)
         initAllValues(geneCount, 0)
         for each chr in p
           for k := 1 to lgth
             gene := chr[k]
             geneCount[k,gene] := geneCount[k,gene] + 1
         for i := 1 to lgth
           for j := 1 to a
             geneFreq[i,j] := geneCount[i,j] / size(p)
         return geneFreq

function APP_Diversity(p, lgth, a)

args:    same as in findGeneFrequencies( )

vars:    float[lgth,a] geneFreq
         int max; float diversity = 0

code:    geneFreq := findGeneFrequencies(p, lngth, a)
         max := lgth * (a – 1) / a      assumes a | lgth;
         if a doesn't divide lgth, use max from Theorem1b
         for k := 1 to lgth
           for α := 1 to a
             diversity := diversity +
                       geneFreq[k,α] * (1-geneFreq[k,α])
         return diversity / max
```

Finding the gene frequencies costs $O(l \cdot n)$ time while the actual diversity is calculated in $O(l \cdot a)$ time. Since the alphabet size is considered to be a constant of the system (this is definitely true with the standard binary alphabet), the overall time complexity is $O(l \cdot n) + O(l) = O(l \cdot n)$. This is optimal time for this problem (each chromosome must at least be looked at once to affect the diversity value), so the all-possible-pairs diversity algorithm can be computed in $\Theta(l \cdot n)$ time. This is much faster than the $O(l \cdot n^2)$ time that the original naïve all-possible-pairs algorithm would take.

4 Diversity as Informational Entropy

In information theory, entropy is defined as [8]

$$H(X) = \sum_{x \in A} p(x) \log \frac{1}{p(x)} \tag{8}$$

where X is a discrete random variable with values taken from alphabet A and a probability mass function $p(x) = \Pr\{X = x\}$, $x \in A$. Equating the population at a locus with X and the gene frequencies at that locus $f_k(a)$ with the probabilities $p(x)$, the entropic diversity at a locus can be written as:

$$\mathrm{Div}_k(P) = \sum_{a \in A} \mathrm{f}_k(a) \log \frac{1}{\mathrm{f}_k(a)} \tag{9}$$

Averaging over all loci gives the actual entropic diversity of the population:

$$\mathrm{Div}(P) = \frac{1}{l} \sum_{k=1}^{l} \sum_{a \in A} \mathrm{f}_k(a) \log \frac{1}{\mathrm{f}_k(a)} \tag{10}$$

To distinguish between the two different types of diversity, the all-possible-pairs diversity shall be symbolized by $\mathrm{Div}_{\mathrm{X}}(P)$ and the entropic diversity by $\mathrm{Div}_{\mathrm{H}}(P)$ (the X represents the complete Cartesian cross of the all possible pairs, and the H represents the entropy).

The entropic diversity is closely tied to the all-possible-pairs diversity in behavior. Preliminary experiments done by the authors show that the correlation between the two is very close, although not identical to 1. If we compare the definition (4) of all-possible-pairs diversity with the entropic diversity definition above, we see that aside from the constants in front, the two forms are remarkably similar. The only difference is the use of $\log \frac{1}{\mathrm{f}_k(a)}$ term in the entropic diversity instead of the $(1 - \mathrm{f}_k(a))$ term as used in the all-possible-pairs diversity. Furthermore both diversities can be seen as just the expected value of those terms. However, if we perform a Taylor expansion of $\log \frac{1}{\mathrm{f}_k(a)}$ around $a = 1$, we get

$$\log \frac{1}{\mathrm{f}_k(a)} = (1 - f_k(a)) + \frac{(1 - f_k(a))^2}{2} + \frac{(1 - f_k(a))^3}{3} + \cdots + \frac{(1 - f_k(a))^i}{i} \tag{11}$$

Notice that the first term in the Taylor series is the same as the one used in the all-possible-pairs diversity definition. Also notice that the other terms are all less than 1 and rapidly approach 0 and consequently the early terms will dominate. So we can now see that $\mathrm{Div}_{\mathrm{X}}(P)$ is just the first term in the Taylor expansion about 1 of $\mathrm{Div}_{\mathrm{H}}(P)$, which accounts for the similarity in their behavior.

Now $(1-\mathrm{f}_k(a))$ can be thought of as the "probability" that, if selected at random, the gene at this location won't be α. The other terms then can be seen as the probability under random selection that gene α won't be selected after i selections. Therefore, the diversity can be regarded as the expectation that the selected gene will be "some other gene". Since in each generation the GA selects from the population multiple times, and since the Taylor series above rapidly converges, the entropic diversity is used as a more accurate measure of diversity.

Finally, it is well known that the maximum of the Shannon entropy occurs under a uniform probability distribution. This, as we would expect, is the same as we found with the all-possible-pairs diversity. Consequently, the normalized form for the entropic diversity in the usual case when $a < n$ and $n \mid a$ is

$$\overline{\mathrm{Div}(P)} = \frac{1}{l \log a} \sum_{k=1}^{l} \sum_{a \in A} \mathrm{f}_k(a) \log \frac{1}{\mathrm{f}_k(a)} \qquad (12)$$

(notice that if we choose the base of a for the logarithm, the original definition of entropic diversity (10) is already normalized). The corresponding formulae for the cases when $a < n$ but $n \nmid a$, and when $a \geq n$) are analogous to those developed for the all-possible-pairs diversity.

5 The Diversity Measures as Used in Biology and EC

Both of the all-possible-pairs formulations as well as the entropy diversity measure have been used in various biological fields of study such as genetics and ecology.

This diversity definition is actually used in the field of molecular genetics, although modified slightly to reflect the fact that, in practice, one only has a sampling of DNA sequences from a population. The modified formula is proportional to the all-possible-pairs definition of diversity,

$$\mathrm{Div}(P) = \frac{2}{l \cdot n \cdot (n-1)} \sum_{i=1}^{p} \sum_{j=1}^{i-1} \mathrm{hd}(c_i, c_j) \qquad (13)$$

and is called *nucleotide diversity*. For more details, see *Molecular Evolution* [9] pp. 237-238.

The reformulation of the all-possible-pairs diversity also has a biological interpretation; this time it is not in the recent area of molecular genetics, but in the older field of population genetics. Here the diversity is used to measure the variation of alleles in a population and is known as either the *gene diversity* or the *expected heterozygocity*. It is normally only defined for a single locus and is usually given in the form $1 - \sum_{\forall a \in A} f_a^2$, where A is the set of alleles at that locus. Remember, an allele at a locus may comprise an entire sequence, or even many sequences of nucleotides and so is working at a much higher level than the nucleotide diversity, even though the underlying mathematics is identical. For details, see [8] pp. 52-53.

This version of the (normalized) all-possible-pairs diversity has appeared in the GA literature [10] with reference made to the biological definition of heterozygocity, although the formula given had been modified to deal with binary values only. In the paper, no attempts were made to connect this diversity definition with the standard all-possible-pairs formulation.

The diversity measure based on the Shannon entropy also is commonly used in biology in the field of ecology, where it is used to compute the diversity of species, see [11] pp.7-8. While less common, entropic diversity has also been used for the genetic diversity of populations in the EC field [12].

6 Diversity in Populations with Real Valued Genomes

Until now we have concentrated our attention on populations with binary and symbolic gene valued chromosomes. There we found that many measures that are naively thought to be distinct are in fact connected by the underlying concept of an all-possible-pairs comparison. It would be reasonable to expect that the situation would be very different when considering real valued genes. However, this is not the case.

At first one might think that the all-possible-pairs diversity for a locus k would be

$$\frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n} D(x_{i,k}, x_{j,k}) \tag{14}$$

where D is the Euclidean distance between chromosomes. However, a simple summation of the distances is equivalent to taking the L_1-norm of the all-possible-pairs 'vector'. Furthermore, since D is based on the L_2-norm, it only makes sense to match the method of combination of all of the possible pairs with that used for the distance itself. There is nothing special about the averaging done by the L_1-norm. Rather the L_2-norm can be thought of as an 'average' that emphasizes the effects of larger differences. Consequently, using the L_2-norm, the all-possible-pairs diversity at a locus k becomes

$$Dv_k^2(P) = \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n} D^2(x_{i,k}, x_{j,k}) \tag{15}$$

To obtain the actual all-possible-pairs diversity by combining all the diversities at the various loci, we again use the L_2-norm producing

$$Dv^2(P) = \frac{1}{2}\sum_{k=1}^{l}\sum_{j=1}^{n}\sum_{i=1}^{n} D^2(x_{i,k}, x_{j,k}) = \frac{1}{2}\sum_{k=1}^{l}\sum_{j=1}^{n}\sum_{i=1}^{n} (x_{i,k} - x_{j,k})^2 \tag{16}$$

Reformulating the Real Valued All-Possible-Pairs Diversity for a Linear Time Algorithm

The above formula, coded as is, would produce an $O(l \cdot n^2)$ algorithm. However, just as with systems that use symbolic genes, the formula can be rearranged to produce a program that has $O(l \cdot n)$ time complexity. For the symbolic gene systems, this was accomplished by introducing the frequency of a gene at a locus. While this cannot be used directly for systems using numeric genes, there is an analogous measure: the average gene value at a locus. This can be used to produce a reformulation of the all-possible-pairs diversity:

Theorem 2: Let $\overline{x_k} = \frac{1}{n}\sum_{i=1}^{n} x_{i,k}$ and $\overline{x_k^2} = \frac{1}{n}\sum_{i=1}^{n} x_{i,k}^2$.

Then $Dv^2(P) = n^2 \sum_{k=1}^{l} \left(\overline{x_k^2} - (\overline{x_k})^2\right)$

Proof.

$$Dv^2(P) = \frac{1}{2}\sum_{k=1}^{l}\sum_{j=1}^{n}\sum_{i=1}^{n}(x_{i,k} - x_{j,k})^2 \qquad \text{from (15)}$$

$$= \frac{1}{2}\sum_{k=1}^{l}\sum_{j=1}^{n}\sum_{i=1}^{n} x_{i,k}^2 + \frac{1}{2}\sum_{k=1}^{l}\sum_{j=1}^{n}\sum_{i=1}^{n} x_{j,k}^2 - \sum_{k=1}^{l}\sum_{j=1}^{n}\sum_{i=1}^{n} x_{i,k}x_{j,k}$$

$$= \frac{n}{2}\sum_{k=1}^{l}\sum_{i=1}^{n} x_{i,k}^2 + \frac{n}{2}\sum_{k=1}^{l}\sum_{j=1}^{n} x_{j,k}^2 - \sum_{k=1}^{l}\left(\sum_{i=1}^{n} x_{i,k}\right)\left(\sum_{j=1}^{n} x_{j,k}\right)$$

$$= \frac{n^2}{2}\sum_{k=1}^{l}\overline{x_k^2} + \frac{n^2}{2}\sum_{k=1}^{l}\overline{x_k^2} - n^2\sum_{k=1}^{l}(\overline{x_k})^2$$

$$= n^2\sum_{k=1}^{l}\left(\overline{x_k^2} - (\overline{x_k})^2\right)$$

Looking at the diversity equation in the theorem, we see that the all-possible-pairs diversity is dependent directly on population size and implicitly on the square of the chromosome length. Since intuitively the diversity of a population should not increase by simply duplicating the exact population or by exactly duplicating genes, we will define the true diversity as

$$Div(P) = \frac{1}{l}\sqrt{\sum_{k=1}^{l}\left(\overline{x_k^2} - (\overline{x_k})^2\right)} \qquad (17)$$

which is simply the all-possible-pairs diversity divided by the population size and chromosome length.

We will now turn to the time complexity of an algorithm that implements the above definition of diversity. Since the average gene value, $\overline{x_k}$, and the average of

the gene value squared, $\overline{x_k^2}$, can be computed in $O(n)$ time for a single locus, and since there are l loci, all of the average gene values and gene values squared can be computed in $O(l \cdot n)$ time. Furthermore, once the average gene values are obtained, the diversity squared can be computed in $O(l)$ time. Consequently the total time complexity of an algorithm to find the squared diversity is $O(l \cdot n)$.

All-Possible-Pairs Diversity and Statistical Variance

The common statistical measure for scatter around the mean is the variance (the square root of which is the standard deviation). The definition of the variance is the sum of the squared difference between each value and the overall mean: $s^2(Y) = E[(Y-m)^2]$, where $m = E(Y)$ is the expected value of the random variable Y also called the mean. If the probability of each value in Y is unknown then if Y is tested by taking n samples, then $E(Y) = \frac{1}{n}\sum_{i=1}^{n} Y_i$, which is called the sample mean and is frequently written as $\overline{Y}$. The sample variance would therefore be

$$Var(Y) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \overline{(Y^2)} - \overline{Y}^2 \qquad (\mathbf{18})$$

This is a well-known result and can be found in any statistical textbook.

Now compare the above sample variance with (17) the linear time representation of the all-possible-pairs diversity formula. What we have is the square root variance in the population of each gene averaged across all loci. In other words the all-possible-pairs diversity corresponds to the 'average' standard deviation of each gene in a chromosome.

7 Conclusion

Many diversity measures exist in both the biological and EC literature: gene diversity or expected heterozygocity, nucleotide diversity, entropy and variance and standard deviation. In this paper we have shown that all are restatements or slight variants of the basic sum of the distances between all possible pairs of the elements in a system. Consequently, experiments need not be done to distinguish between the various measures, trying to find which of them provides a better measure of diversity, as they are really all the same measure.

By recognizing how all of the diversity measures have been formed, different diversity measures of this family can easily be created: merely change the distance measure used between the pairs. In the diversity measured examined in this paper, the Hamming Distance was used for binary and symbolic chromosomes while the Euclidean Distance was used for chromosomes with real valued genes.

However, we have pointed out that care must be taken to match the way the distances are to be combined. If the distance is formed by the combination of differences of component parts, the method used to combine the distances should match the method used to combine the parts. For example, the Hamming distance uses the L1-norm to combine the differences between genes and so the L1-norm was used to combine the distances between chromosomes; the L2-norm is used to combine the differences between genes in the Euclidean distances used for real valued genomes and so the L2-norm was used to combine the corresponding distances between pairs of chromosomes with real valued genes.

Finally we have shown that care must be taken when implementing all-possible-pair style diversity measures, as it is frequently easy to take the $O(n^2)$ comparisons and manipulate them so that it takes only $O(n)$ time to compute. We have given the associated $O(n)$ algorithms for all the diversity measures studied in this paper.

Reference

1. Hutter M.: Fitness Uniform Selection to Preserve Genetic Diversity. In: Fogel D. B. (ed.): CEC'02. IEEE Press (2002) 783-788
2. Motoki, T.: Calculating the Expected Loss of Diversity of Selection Schemes. In Evolutionary Computation 10-4. MIT Press, Cambridge MA (2002) 397-423
3. Ang, K. H., Chong, G. and Li, Y.: Preliminary Statement on the Current Progress of Multi-Objective Evolutionary Algorithm Performance Measurement. In: Fogel D. B. (ed.): CEC'02. IEEE Press (2002) 1139-1144
4. Oppacher F. and M. Wineberg (1999). The Shifting Balance Genetic Algorithm: Improving the GA in a Dynamic Environment. In Banzhaf W. et. al. (Eds.): GECCO'99. Morgan Kaufmann, San Francisco (1999) 504-510
5. Garrett, S. M., Walker, J. H.: Genetic Algorithms: Combining Evolutionary and 'Non'-Evolutionary Methods in Tracking Dynamic Global Optima. In Langdon W. B. et. al. (Eds.): GECCO-2002. Morgan Kaufmann, San Francisco (2002) 359-366
6. Louis, S. J. and Rawlins, G. J. E.: Syntactic Analysis of Convergence in Genetic Algorithms. In Whitley, L. D. (ed.); Foundations of Genetic Algorithms 2. Morgan Kaufmann, San Mateo, California. (1993) 141-151
7. Duran, B. S. and Odell, P. L.: Cluster Analysis: A Survey. Springer-Verlag, Berlin (1974)
8. Cover, T. M. and Thomas J. A.: Elements of Information Theory. John Wiley & Sons, New York (1991)
9. Li, W.-H.: Molecular Evolution. Sinauer Associates, Sunderland MA (1997)
10. Collins, R. J., and Jefferson, D. R.: Selection in Massively Parallel Genetic Algorithms. In: Belew, R. K., and Booker L. B. (eds,.): 4th ICGA. Morgan Kaufmann, San Mateo, California (1991) 249-256
11. Pielou, E. C.: Ecological Diversity. John Wiley & Sons, New York (1975)
12. Rosca, J.: Entropy-driven adaptive representation. In: J. Rosca (ed.): Proceedings of the Workshop on Genetic Programming: From Theory to Real-World Applications. Tahoe City, California (1995) 23-32