New Entropy-Based Measures of Gene Significance and Epistasis

Dong-Il Seo, Yong-Hyuk Kim, and Byung-Ro Moon

School of Computer Science & Engineering, Seoul National University
Sillim-dong, Kwanak-gu, Seoul, 151-742 Korea
{diseo,yhdfly,moon}@soar.snu.ac.kr
http://soar.snu.ac.kr/~{diseo,yhdfly,moon}/

Abstract. A new framework to formulate and quantify the epistasis of a problem is proposed. It is based on Shannon's information theory. With the framework, we suggest three epistasis-related measures: gene significance, gene epistasis, and problem epistasis. The measures are believed to be helpful to investigate both the individual epistasis of a gene group and the overall epistasis that a problem has. The experimental results on various well-known problems support it.

1 Introduction

In the context of genetic algorithms, the difficulty of an optimization problem is explained in various aspects. The aspects are categorized into *deception* [1], *multimodality* [2], *noise* [3], *epistasis* [4,5], and so on. Among them, the epistasis is observed in most GA-hard problems. In biology, we refer to the suppression of gene expression by one or more other genes as epistasis. But, in the community of evolutionary algorithms, the term has a wider meaning; it means the interaction between genes.

In addition to the concepts to explain the problem difficulty, various measures quantifying the difficulty have been proposed recently. The epistasis variance, suggested by Davidor [5], is a measure quantifying the epistasis of a problem. He interpreted the epistasis as the nonlinearity embedded in the fitness landscape of the problem. The measure was explained more formally by Reeves and Wright [6] from the viewpoint of experimental design. The measures are, however, somewhat "macroscopic," i.e., they concern the epistasis merely as a factor of GA-hardness of a problem. In fact, the epistasis of a problem consists of many individual epistases between small groups of genes. This idea already affected various branches of evolutionary algorithms such as probabilistic model-building genetic algorithms (PMBGAs) [7,8], also called estimation-of-distribution algorithms (EDAs), and topological linkage-based genetic algorithms (TLBGAs) [9]. The epistases are estimated algorithmically or heuristically in the algorithms.

In this paper, we propose new algorithm-independent "microscopic" measures of epistases. We suggest a new framework for the formulation and quantification of the epistases. The framework is based on Shannon's information theory [10,

E. Cantú-Paz et al. (Eds.): GECCO 2003, LNCS 2724, pp. 1345–1356, 2003.

11]. We propose three measures: *gene significance*, *gene epistasis*, and *problem epistasis*. They are helpful to investigate both the individual epistasis of a gene group and the overall epistasis that a problem has.

The rest of this paper is organized as follows. The basic concepts of Shannon's entropy are introduced in Sect. 2. We establish a probability model and define new epistasis measures in Sect. 3. We provide the results of experiments on various well-known problems in Sect. 4. Finally, the conclusions are given in Sect. 5.

2 Shannon's Entropy

Shannon's information theory [10,11] provides manners to quantify and formulate the properties of random variables. According to the theory, the amount of information contained in a message notifying an event is defined to be the number of digits being required to describe the event. That is, the amount of information contained in a message notifying an event of probability p is defined to be $\log \frac{1}{p}$. The log is to the base 2 and the value is measured in bits. The lower the probability of the event is, the larger amount of information the message contains. The average amount of information contained in events is the amount of uncertainty of the random variable on the events. Thus, the uncertainty of a random variable is defined as

$$H(X) = -\sum_{x \in \mathfrak{X}} p(x) \log p(x) \tag{1}$$

where $\mathfrak{X}$ and $p(x)$ are the alphabet and the probability mass function (pmf), respectively. The quantity is called the *entropy* of X. It means the average number of bits being required to describe a random variable. The convention $0 \log 0 = 0$ is used in the equation, which is easily justified by continuity since $x \log x \to 0$ as $x \to 0$. Entropy is always nonnegative. Similarly, the *joint entropy* of two random variables is defined as

$$H(X,Y) = -\sum_{x \in \mathfrak{X}} \sum_{y \in \mathfrak{Y}} p(x,y) \log p(x,y) \tag{2}$$

where $\mathfrak{X}$ and $\mathfrak{Y}$ are the alphabets of random variables X and Y, respectively, and $p(x,y)$ is the joint pmf of the random variables. The *conditional entropy* of X given Y is defined as

$$H(X|Y) = -\sum_{x \in \mathfrak{X}} \sum_{y \in \mathfrak{Y}} p(x,y) \log p(x|y). \tag{3}$$

It means the average uncertainty of X when the value of Y is known. The conditioning reduces entropy, i.e., $H(X|Y) \leq H(X)$. The average amount of uncertainty of X reduced by knowing the value of Y is the amount of information about X contained in Y. The quantity is called *mutual information* between X

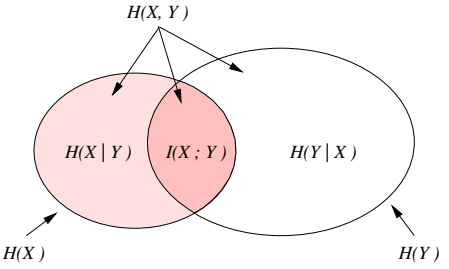


Fig. 1. Relationship between entropy and mutual information

and Y and is formally written as

$$I(X;Y) = H(X) - H(X|Y) \tag{4}$$

$$= \sum_{x\in\mathfrak{X}} \sum_{y\in\mathfrak{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{5}$$

Mutual information is symmetric and nonnegative. Two random variables are mutually independent if and only if the mutual information between them is zero. The Equation (4) can be rewritten as

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \tag{6}$$

It is deduced from the equation that the random variables are independent if and only if the joint entropy of them is equal to the summation of the two marginal entropies. The relationship between entropy and mutual information is illustrated in Fig. 1. Table 1 shows examples of joint random variables. Table 1(a) is an example of mutually independent random variables and Table 1(b) is an example of mutually dependent random variables.

Table 1. Example joint pmf's of two pairs of random variables whose alphabets $\mathfrak{X}$ and $\mathfrak{Y}$ are $\{0,1\}$. (a) Two mutually independent random variables. $H(X) = 1$, $H(Y) = 0.81$, $H(X,Y) = 1.81$, and $I(X;Y) = 0$. (b) Two mutually dependent random variables. $H(X) = 1$, $H(Y) = 0.81$, $H(X,Y) = 1.75$, and $I(X;Y) = 0.06$

(a)

X	$Y=0$	$Y=1$	
0	1/8	3/8	1/2
1	1/8	3/8	1/2
	1/4	3/4	

(b)

X	$Y=0$	$Y=1$	
0	1/16	7/16	1/2
1	3/16	5/16	1/2
	1/4	3/4	

3 Probability Model and Epistasis Measures

3.1 Probability Model

Assume that a problem is encoded into n genes, and let the fitness function of the problem be $f : \mathcal{U} \to \mathbb{R}$ where $\mathcal{U}$ is the set of all feasible[1] solutions, called *universe* of the problem. When we do random sampling on $\mathcal{U}$, the probability that a feasible solution $(x_1, x_2, \ldots, x_n)$ will be chosen, is $1/|\mathcal{U}|$. By the probability model, random variables for the genes and the fitness are defined. Let the random variable for gene i be X_i and the random variable for the fitness be Y, and the set of allele values of gene i and the set of all possible fitness values be A_i and F, respectively, then the probability mass function is defined as

$$p(x_1, x_2, \ldots, x_n, y) = \begin{cases} 1/|\mathcal{U}| & \text{if } (x_1, x_2, \ldots, x_n) \in \mathcal{U} \\ & \quad \text{and } y = f(x_1, x_2, \ldots, x_n) \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

for $x_i \in A_i$, $i \in \{1, 2, \ldots, n\}$ and $y \in F$. It is practical to use a set of sampled solutions in the Equation (7) instead of the universe $\mathcal{U}$ for large-sized problems because of the spatial or computational limitations. But, in the case, the size of the set must be not too small for getting results of low levels of distortion.

3.2 Epistasis Measures

Three epistasis-related measures are proposed in this section. They are based on the probability model described in Sect. 3.1. They quantify *gene significance*, *gene epistasis*, and *problem epistasis*, respectively.

The significance of a gene i is defined to be the amount of its contribution to the fitness. It could be understood as the amount of information contained in X_i about Y, i.e., $I(X_i; Y)$. Since the minimum and the maximum of the mutual information are 0 and $H(Y)$, respectively, a normalization could be done by dividing $I(X_i; Y)$ by $H(Y)$. As a result, the significance ξ_i of a gene i is defined as

$$\xi_i = \frac{I(X_i; Y)}{H(Y)}. \tag{8}$$

It ranges from 0 to 1; if the value is zero, the gene has no contribution to the fitness and if the value is one, the gene wholly determines the fitness value.

The epistasis (often referred to as interaction) between genes means the dependence of a gene's contribution to the fitness upon the value of other genes. The contribution of gene i and gene j to the fitness are quantified as $I(X_i; Y)$ and $I(X_j; Y)$, respectively. And the contribution of the gene pair (i, j) to the fitness is quantified as $I(X_i, X_j; Y)$. Therefore, the epistasis between the two genes could be written as $I(X_i, X_j; Y) - I(X_i; Y) - I(X_j; Y)$. A normalization

[1] A solution is feasible if the fitness function is defined on it.

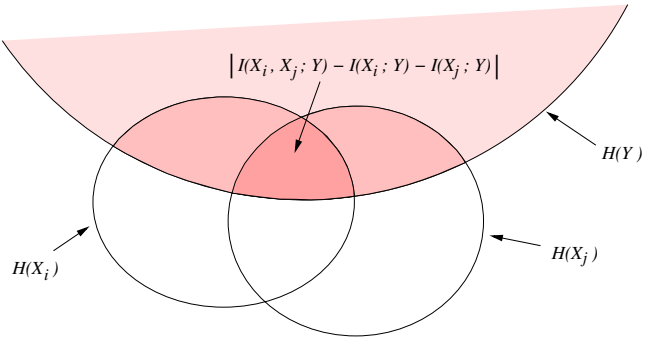


Fig. 2. An illustration of the pairwise epistasis

could be done by dividing the quantity by $I(X_i, X_j; Y)$. As a result, the gene epistasis ε_{ij} between gene i and gene j is defined as

$$\varepsilon_{ij} = \begin{cases} 1 - \dfrac{I(X_i;Y) + I(X_j;Y)}{I(X_i, X_j;Y)} & \text{if } I(X_i, X_j;Y) \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Since the minimum and the maximum of the fraction in the Equation (9) are 0 and 2, respectively, the epistasis ranges from -1 to 1. It has a positive value if $I(X_i, X_j; Y) > I(X_i; Y) + I(X_j; Y)$ and it has a negative value otherwise. The former case means that the genes interact constructively with each other, and the latter case means that they interact destructively with each other. We call the epistasis of the former case *positive* gene epistasis, and that of the latter case *negative* gene epistasis. If the two genes are mutually independent, the gene epistasis is zero. Figure 2 shows an illustration of the above definition.

The mean absolute of the gene epistases of all gene pairs could be used as a measure of the epistasis of a problem, i.e., the problem epistasis η is defined as

$$\eta = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j<i} |\varepsilon_{ij}|. \tag{10}$$

Since each ε_{ij} ranges from -1 to 1, η ranges from 0 to 1. The larger value η a problem has, the more epistatic the problem is.

3.3 Fitness Discretization

In general, the fitness function of a problem is defined on a continuous domain, while each gene has discrete allele values in many cases. So, the fitness value needs to be discretized to apply the measures described in Sect. 3.2. The most simple methods are *equal-width discretization* and *equal-frequency discretization* [12]. In the equal-width discretization, the whole range is divided into k intervals

of equal widths, while the whole range is divided into k intervals that include the same number of samples in the equal-frequency discretization. We use equal-frequency discretization with $k = 10$ in Sects. 4.3 and 4.4. Ten is the most widely used number of intervals.

3.4 An Example

Table 2 shows two example fitness functions. The function f_{pos} has four feasible solutions, while f_{neg} has only three feasible solutions. We can compute the values of the measures, proposed in Sect. 3.2, of the example functions as follows. First, we make a joint pmf table for each function as Tables 3a–b. Then, we apply the equations in Sect. 2 to each of the tables to compute the entropies and mutual informations. Finally, we use the Equations (8), (9), and (10) to compute the gene significance, gene epistasis, and problem epistasis, respectively. Table 4 shows the intermediate values and the resultant measure values. We can see that the two genes of f_{pos} have positive gene epistasis, while the two genes of f_{neg} have negative gene epistasis. The table shows that the gene 1 of f_{neg} is more significant than gene 2, and the function f_{neg} is more epistatic than f_{pos}.

Table 2. Two example functions f_{pos} and f_{neg}

x_1	x_2	f_{pos}	f_{neg}
0	0	0	0
0	1	0	undefined
1	0	0	1
1	1	1	1

Table 3. Joint pmf's of the example functions

(a) Joint pmf of f_{pos}.

X_1	X_2	$Y=0$	$Y=1$	
0	0	1/4	0	1/4
0	1	1/4	0	1/4
1	0	1/4	0	1/4
1	1	0	1/4	1/4
		3/4	1/4	

(b) Joint pmf of f_{neg}.

X_1	X_2	$Y=0$	$Y=1$	
0	0	1/3	0	1/3
0	1	0	0	0
1	0	0	1/3	1/3
1	1	0	1/3	1/3
		1/3	2/3	

Table 4. The calculation of the measures for the example functions f_{pos} and f_{neg}

	$H(Y)$	$I(X_1;Y)$	$I(X_2;Y)$	$I(X_1,X_2;Y)$	ξ_1	ξ_2	ε_{12}	η
f_{pos}	0.811	0.311	0.311	0.811	0.384	0.384	0.233	0.233
f_{neg}	0.918	0.918	0.252	0.918	1.000	0.274	−0.274	0.274

Table 5. Davidor's four example functions f_1, f_2, f_3, and f_4

x_1	x_2	x_3	f_1	f_2	f_3	f_4
0	0	0	0	0	0.0	7
0	0	1	1	0	0.5	5
0	1	0	2	0	1.0	5
0	1	1	3	0	1.5	0
1	0	0	4	0	2.0	3
1	0	1	5	0	2.5	0
1	1	0	6	0	3.0	0
1	1	1	7	28	17.5	8

Table 6. The gene significance ξ_i, gene epistasis ε_{ij}, and problem epistasis η of Davidor's four example functions f_1, f_2, f_3, and f_4

	ξ_1	ξ_2	ξ_3	ε_{12}	ε_{13}	ε_{23}	η	σ_ε^2
f_1	0.333	0.333	0.333	0.000	0.000	0.000	0.000	0
f_2	0.254	0.254	0.254	0.060	0.060	0.060	0.060	49
f_3	0.333	0.333	0.333	0.000	0.000	0.000	0.000	12.25
f_4	0.304	0.188	0.188	0.082	0.082	0.298	0.154	8.57

4 Experimental Results

4.1 Davidor's Examples

Davidor tested his epistasis variance on four example functions as shown in Table 5. They are a linear function (f_1), a delta function (f_2), a mixture of the linear function and delta function ($f_3 = \frac{f_1+f_2}{2}$), and a minimal deceptive function (f_4). Table 6 shows the gene significance ξ_i, gene epistasis ε_{ij}, and problem epistasis η of each example function. The epistasis variance σ_ε^2 in the final column was quoted from the Davidor's paper [5] for comparison. The results are somewhat different. The problem epistasis and the epistasis variance of f_1 are zero in common. But, the epistasis variance of f_3 is not zero, while the problem epistasis of the function is zero. At the same time, the most problem epistatic function among them is f_4, while the function with the largest epistasis variance is f_2. The difference comes mainly from the reasons in the following. The epistasis variance treats the fitness as a scalar quantity. But, the proposed measures treat the fitness as a categorical index, i.e., the proposed measures do not individually concern the magnitude of the fitness. The proposed measures only concern whether the fitness values of solutions are the same or not.

4.2 Royal Road Function

Royal Road function is a function proposed by Forrest and Mitchell [13] to investigate precisely and quantitatively how schema processing actually takes place during the typical evolution of a genetic algorithm. To do so, the function

was designed to have obvious building blocks and an optimal solution. Royal Road function is defined as

$$f(x_1, x_2, \ldots, x_n) = \sum_i c_i \delta_i(x_i, x_2, \ldots, x_n) \tag{11}$$

where c_i is a predefined coefficient corresponding to a schema s_i, and $\delta_i : \{0,1\}^n \rightarrow \{0,1\}$ is a function that returns 1 if the solution contains the schema s_i, and returns 0 otherwise. Generally, the coefficient c_i is defined as the order of schema s_i. Table 7 shows the two Royal Road functions used in our experiments. The function R_1 has four building blocks of order 2, while R_2 has the building blocks of R_1 and two more building blocks of order 4. Figures 3a–b shows the illustrations of the gene epistases of R_1 and R_2, respectively. We can see that the genes of the building blocks have relatively strong gene epistasis with each other. The figure shows that the gene epistases between the genes in order-2 building blocks are larger than those of order-4 building blocks. The problem epistasis η of R_1 and R_2 were 0.126 and 0.236, respectively. It means that R_2 has stronger problem epistasis than R_1.

4.3 *NK*-Landscape

The NK-landscape model is a model proposed by Kauffman [14] to define a family of fitness functions that have various dimensions of search space and degrees of epistasis (see also [15,16]). The functions are tuned by two parameters:

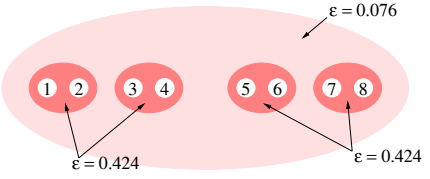
Table 7. Two Royal Road functions

(a) R_1

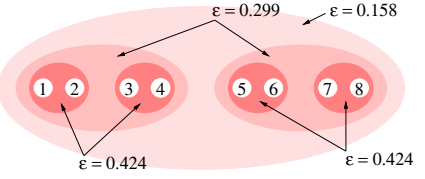
Schema	Coefficient
$s_1 = 1\,1 * * * * * *$	$c_1 = 2$
$s_2 = * * 1\,1 * * * *$	$c_2 = 2$
$s_3 = * * * * 1\,1 * *$	$c_3 = 2$
$s_4 = * * * * * * 1\,1$	$c_4 = 2$
$s_{opt} = 1\,1\,1\,1\,1\,1\,1\,1$	

(b) R_2

Schema	Coefficient
$s_1 = 1\,1 * * * * * *$	$c_1 = 2$
$s_2 = * * 1\,1 * * * *$	$c_2 = 2$
$s_3 = * * * * 1\,1 * *$	$c_3 = 2$
$s_4 = * * * * * * 1\,1$	$c_4 = 2$
$s_5 = 1\,1\,1\,1 * * * *$	$c_5 = 4$
$s_6 = * * * * 1\,1\,1\,1$	$c_6 = 4$
$s_{opt} = 1\,1\,1\,1\,1\,1\,1\,1$	



(a) R_1



(b) R_2

Fig. 3. Gene epistasis ε_{ij} of Royal Road functions

Table 8. Gene epistasis ε_{ij} of NK-landscape ($N = 12$)

	Co-Contribution Frequency												
K	0	1	2	3	4	5	6	7	8	9	10	11	12
2	0.072	0.206	0.249										
3	0.092	0.196	0.242	0.254									
4	0.112	0.186	0.228	0.263	0.262								
5	0.136	0.193	0.233	0.257	0.271	0.282	0.303						
6		0.231	0.249	0.269	0.289	0.305	0.307	0.310					
7			0.260	0.281	0.287	0.300	0.305	0.311	0.325				
8				0.295	0.305	0.309	0.311	0.317	0.321	0.327	0.317		
9						0.312	0.320	0.320	0.324	0.330	0.329	0.330	
10								0.315	0.337	0.328	0.334	0.338	0.339
11													0.334

N and K. The parameters N and K determines the dimension of the problem space and the degree of epistasis between the genes constituting a chromosome, respectively. The fitness f of a solution $(x_1, x_2, \ldots, x_N)$ is defined as

$$f(x_1, x_2, \ldots, x_N) = \frac{1}{N} \sum_{i=1}^{N} f_i(x_i, x_{j_{i1}}, x_{j_{i2}}, \ldots, x_{j_{iK}}) \tag{12}$$

where the fitness contribution f_i depends on the value of gene i and the values of K other genes $j_{i1}, j_{i2}, \ldots, j_{iK}$. The function $f_i : \{0,1\}^{K+1} \to \mathbb{R}$ assigns a random number, distributed uniformly between 0 and 1, to each of its 2^{K+1} inputs. The values for $j_{i1}, j_{i2}, \ldots, j_{iK}$ are chosen from $\{1, 2, \ldots, N\}$ at random.

The gene values $x_i, x_{j_{i1}}, x_{j_{i2}}, \ldots, x_{j_{iK}}$ contribute together to the fitness contribution f_i. We define the *co-contribution frequency* of gene i and j as the number of cases that the gene values x_i and x_j contribute together to the fitness f. Intuitively, we can say that two genes are strongly correlated if the co-contribution frequency of the genes is high.

The gene epistases of gene pairs of NK-landscapes were listed along with their co-contribution frequencies in Table 8. We discretized the fitness into 10 intervals by the equal-frequency discretization method in computing the measures. For each K, 200 independently generated functions were used for statistical stability. In the table, the (i, j) entry represents the average gene epistasis of the gene pairs of the NK-landscapes of $K = i$, that co-contribute j times. We can see that the gene epistasis increases as the co-contribution frequency increases for small K's, but it tends to converge for larger K's. Figure 4 shows the problem epistases of the NK-landscapes for various K's. We can see that the problem epistasis increases as the K increases. Both of the results support our intuitive predictions.

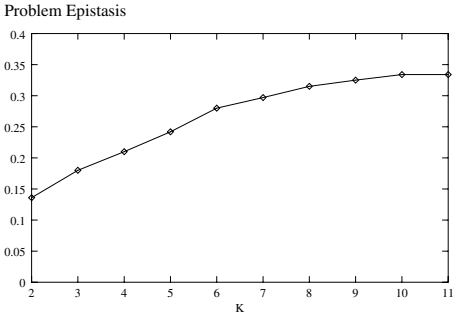


Fig. 4. Problem epistasis η of NK-landscapes ($N = 12$)

4.4 Traveling Salesman Problem

Given n cities, the traveling salesman problem (TSP) is the problem of finding a shortest Hamiltonian cycle visiting the cities. TSP is a well-known NP-hard problem [17]. It is one of the most popular optimization problems and has served as an initial proving ground for new problem solving techniques for decades.

We apply the locus-based encoding to the problem as in [18]; one gene is allocated for every city and the gene value represents the index of its next city in the Hamiltonian cycle. By the encoding, the fitness f of a solution $(x_1, x_2, \ldots, x_n)$ that represents a Hamiltonian cycle is written as

$$f(x_1, x_2, \ldots, x_n) = C_{max} - \sum_{i=1}^{n} d_{ix_i} \tag{13}$$

where d_{pq} is the distance from city p to city q and C_{max} is the cycle length of the worst solution. The subtraction in the equation forces the fitness to be nonnegative and the problem becomes a maximization problem. It is notable that the absolute value of C_{max} does not affect the epistasis measures when the equal-frequency discretization is used. We computed the problem epistasis η on TSP and compared it with a problem difficulty measure, *fitness distance correlation*.

The fitness distance correlation (FDC) is a measure of problem difficulty proposed by Jones and Forrest [19]. FDC is defined to be the correlation coefficient of the fitness and the distance to the nearest global optimum of sampled solutions. Thus, it ranges from -1 to 1. As the value approaches -1, a problem is believed to become easier.

When a genetic algorithm is hybridized with a local optimization algorithm, what the algorithm can see are only local optima. Thus, it is valuable to examine the space of local optima. For each problem instance, the solution set used for the computation of FDC and problem epistasis, was chosen as follows. First, we generate ten thousand solutions at random and apply a local optimization algorithm to them. Then, we discard the duplicated copies from the resultant solutions. We used 2-Opt [20] as the local optimization algorithm because it is

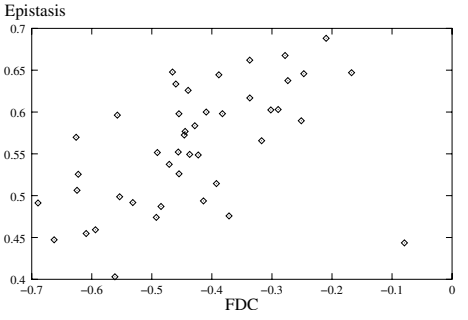


Fig. 5. Fitness-distance correlation (FDC) vs. problem epistasis of TSP

one of the most simple and basic heuristics. The fitness was discretized into 10 intervals by the equal-frequency discretization method as in the case of NK-landscape. As the distance measure, we used Hamming distance that is defined to be the number of genes with different values. Figure 5 shows the relationship between the problem epistasis and the FDC for 44 instances taken from TSPLIB [21]. They are all instances available whose numbers of cities lie inbetween 100 and 700. The figure shows that the two measures are strongly correlated. It means that the problem epistasis works well as the problem difficulty measure.

5 Conclusions

We provided a new framework to formulate and quantify the epistasis of a problem based on Shannon's entropy. With the framework, three measures were proposed: gene significance, gene epistasis, and problem epistasis. They are for choosing significant genes, detecting epistatic gene pairs, and quantifying the epistasis of a problem as a difficulty measure, respectively. They are different from Davidor's epistasis variance in the way of treating the fitness. They treat a fitness value as a categorical index, while the epistasis variance treats it as a scalar quantity. The experimental results on various well-known problems, such as Royal Road function, NK-landscape, and traveling salesman problem, support their usefulness and appropriateness. Future studies include extensions of the framework and applications of the measures.

Acknowledgments. This work was partly supported by Optus Inc. and Brain Korea 21 Project. The RIACT at Seoul National University provided research facilities for this study.

References

1. D. E. Goldberg. Simple genetic algorithms and the minimal deceptive problem. In *Genetic Algorithms and Simulated Annealing*, pages 74–88, 1987.

2. J. Horn and D.E. Goldberg. Genetic algorithm difficulty and the modality of fitness landscapes. In *Foundations of Genetic Algorithms 3*, pages 243–270. 1995.
3. H. Kargupta. Signal-to-noise, crosstalk, and long range problem difficulty in genetic algorithms. In *International Conference on Genetic Algorithms*, pages 193–200, 1995.
4. J. Holland. *Adaptation in Natural and Artificial Systems.* The University of Michigan Press, 1975.
5. Y. Davidor. Epistasis variance: Suitability of a representation to genetic algorithms. *Complex Systems*, 4:369–383, 1990.
6. C.R. Reeves and C.C. Wright. An experimental design perspective on genetic algorithms. In *Foundations of Genetic Algorithms 3*, pages 7–22. 1995.
7. P. Larrañaga and J.A. Lozano. *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation.* Kluwer Academic Publishers, 2002.
8. M. Pelikan, D.E. Goldberg, and F. Lobo. A survey of optimization by building and using probabilistic models. *Computational Optimization and Applications*, 21(1):5–20, 2002.
9. D.I. Seo and B.R. Moon. A survey on chromosomal structures and operators for exploiting topological linkages of genes. In *Genetic and Evolutionary Computation Conference*, 2003.
10. C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.
11. T. Cover and J. Thomas. *Elements of Information Theory.* John Wiley & Sons, 1991.
12. D. Chiu, A. Wong, and B. Cheung. Information discovery through hierarchical maximum entropy discretization and synthesis. In *Knowledge Discovery in Databases.* MIT Press, 1991.
13. S. Forrest and M. Mitchell. Relative building-block fitness and the building-block hypothesis. In *Foundations of Genetic Algorithms 2*, pages 109–126. 1993.
14. S.A. Kauffman. Adaptation on rugged fitness landscapes. In D. L. Stein, editor, *Lectures in the Sciences of Complexity*, pages 527–618. Addison-Wesley, 1989.
15. B. Manderick, M. de Weger, and P. Spiessens. The genetic algorithm and the structure of the fitness landscape. In *International Conference on Genetic Algorithms*, pages 143–157, 1991.
16. P. Merz and B. Freisleben. On the effectiveness of evolutionary search in high-dimensional NK-landscapes. In *IEEE Conference on Evolutionary Computation*, pages 741–745, 1998.
17. M.R. Garey and D.S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* Freeman, 1979.
18. T.N. Bui and B.R. Moon. A new genetic approach for the traveling salesman problem. In *IEEE Conference on Evolutionary Computation*, pages 7–12, 1994.
19. T. Jones and S. Forrest. Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In *International Conference on Genetic Algorithms*, pages 184–192, 1995.
20. G.A. Croes. A method for solving traveling salesman problems. *Operations Research*, 6:791–812, 1958.
21. TSPLIB. http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95/.