The Influence of Mutation on Protein-Ligand Docking Optimization: a Locality Analysis

Jorge Tavares, Alexandru-Adrian Tantar,
Nouredine Melab, and El-Ghazali Talbi

INRIA Lille - Nord Europe Research Centre
Parc Scientifique de la Haute Borne
59650 Villeneuve d'Ascq, France
jorge.tavares@inria.fr

Abstract. Evolutionary approaches to protein-ligand docking typically use a real-value encoding and mutation operators based on Gaussian and Cauchy distributions. The choice of mutation is important for an efficient algorithm for this problem. We investigate the effect of mutation operators by locality analysis. High locality means that small variations in the genotype imply small variations in the phenotype. Results show that Gaussian-based operators have stronger locality than Cauchy-based ones, especially if an annealing scheme is used to control the variance.

1 Introduction

Protein-ligand docking is an energy minimization search problem with the aim to find the best ligand conformation and orientation relative to the active site of a target protein [1]. The docking problem can be very difficult since the relative orientation and conformations of the two molecules must be considered. Typically, the receptor (usually a protein) is fixed in a three-dimensional coordinate system. By contrast, the ligand can be repositioned and rotated. In case that both receptor and ligand are allowed to be flexible, the problem difficulty increases. As such, the problem is classified, by increasing complexity, into the ensuing categories: rigid-structure docking (both molecules are rigid); rigid protein and flexible ligand; flexible protein and rigid ligand; and, both molecules are flexible. With both molecules flexible, usually the active site of the protein and the ligand, the problem becomes harder. In fact, a higher degree of flexibility implies a considerable increase of the search space size.

For the past years, numerous protein-ligand docking methods have been proposed using different techniques, e.g., incremental construction algorithms, stochastic algorithms and molecular dynamics. For more detailed descriptions, we refer the reader to several review studies [2, 3]. Evolutionary and swarm algorithms have recently become one of the dominant search techniques for docking methods and proved to be very successful [3, 4]. Although several applications exist, no comprehensive set of studies could be found *to understand why* these algorithms and their components are successful. To the best of our knowledge,

the only attempt was made in [5] where several parameters (e.g., population size) and some genetic operators are empirically investigated. When designing an evolutionary approach for this problem, to make it efficient is important to understand its components behavior and effects.

Locality is an important requisite to ensure the efficiency of search and it has been widely studied by the evolutionary computation community [6–8]. In general terms, this property indicates that small variations in the genotype space, usually originated by mutation, imply small variations in the phenotype space [6]. A locally strong search algorithm is able to efficiently explore the neighborhood of the current solutions. When this condition is not satisfied, the exploration performed by the algorithm is inefficient and, in a worse case scenario, tends to resemble random search.

The goal of this paper is to perform an empirical locality analysis on the evolutionary algorithm model [9, 3] that is usually adopted for protein-ligand docking optimization. Locality measures for the analysis are adopted from the framework proposed by [8] and extended by [10] to deal with real-valued encodings. One distance measure suitable for the selected representation is applied. Mutation is the most frequent operator considered in locality studies. The present study concentrates on the questions: do Gaussian and Cauchy mutation operators have a different effect on phenotypes? Which type of operator is more suitable for evolutionary approaches to protein-ligand? We expect to answer these questions by investigating the impact of the operators on locality. In spite of that, our main research focus is the study of representation properties and the effects of variation operators. The presented work is the first step of a wider study that includes analysis on locality, heritability and heuristic bias.

Results allow us to gain some insights about the degree of locality induced by different mutation operators. The search space is highly multimodal and its shape is influenced by the size, shape and topology of the ligand and the active site being docked [5]. As a consequence of this, even small modifications performed by genetic operators in the structure of an individual lead to large phenotypic changes. An evolutionary algorithm operating on its own is unable to deal with these difficulties. Thus, it is important to know how locality relates to mutation operators commonly used in evolutionary algorithms for molecular docking. Furthermore, understanding the role played by each algorithm's component may provide useful insights for future applications of evolutionary algorithms to this problem.

The rest of the paper is structured as follows. Section 2 contains an overview of the evolutionary algorithm's components used in our experimentation. In section 3 we present the locality analysis and respective discussion. Finally, section 4 contains the main conclusions.

2 Evolutionary Algorithms and Protein-Ligand Docking

Evolutionary algorithms applied to molecular docking can be found since 1993 [11]. A comprehensive review of these efforts, including an outline state-of-the

art applications, can be found in [12, 3]. One of the most important works is the evolutionary algorithm proposed in [9], commonly referred to as *AutoDock*. This approach is a conformational search method which uses an approximate physical model to evaluate possible protein-ligand conformations. It incorporates flexibility by allowing the ligand to change its conformation during the docking simulation. In addition, pairwise interactions between atoms are pre-calculated, considerably speeding up the docking simulation. To search the space of possible protein-ligand conformations, the approach uses an evolutionary algorithm with a local search method. When this method is applied, the genotype of the individuals is replaced with the new best solution found. This process is usually referred to as Lamarckian evolution.

In our analysis, we adopt an experimental model which uses the main components from [9], because *AutoDock* serves as a basis for the large majority of evolutionary-inspired approaches (e.g.,[3, 4]).

2.1 Encoding

During the docking process the protein remains rigid whilst the ligand is flexible. In this case, an individual represents only the ligand. The encoding is an indirect representation. A genotype of a candidate solution is encoded by a vector of real-valued numbers which represent the ligand's translation, orientation and torsion angles [9]. Cartesian coordinates represent the translation, three variables in the vector, whereas four variables defining a quaternion represent the orientation. A quaternion can be considered to be a vector (x, y, z) which specifies an axis of rotation with an angle θ of rotation for this axis. For each flexible torsion angle one variable is used. The phenotype of a candidate solution is composed of the atomic coordinates that represent the three-dimensional structure of the ligand. The atomic structure is built from the translation and orientation coordinates in the ligand crystal structure with the application of the torsion angles.

2.2 Evaluation

To evaluate each individual an energy evaluation function is used. The fitness for each candidate solution is given by the sum of the intermolecular interaction energy between the ligand and the protein, and the intramolecular energy that arises from the ligand itself [9]. An empirical free energy potential composed of five terms is used. The first three terms are pairwise interatomic potentials that account for weal long-range attractive forces and short-range electrostatic repulsive forces. The fourth term measures the unfavorable entropy of a ligand binding due to the restriction of conformational degrees of freedom. The fifth and last term uses a desolvation measure. Further details of the energy terms and how the potential is derived can be found in [9].

2.3 Genetic Operators

Common crossover and mutation operators are applied on the population. In *AutoDock* a standard two-point crossover is used. Cut points only occur between

related genes, i.e., separating translational values, orientation values and rotation torsion angles into separate blocks. This is done to avoid disruption of useful parts of the solution [9]. Since the encoding is a real-valued vector, mutation is performed by using evolutionary strategies based operators. The genetic operator acts in the following way: when undergoing mutation, the new value for a gene x' is obtained from the old value x by adding a random real number sampled from a distribution $U(0,1)$:

$$x' = x + \sigma \times U(0,1) \tag{1}$$

The common distribution used for $U(0,1)$ is the standard Gaussian distribution, $N(0,1)$. In spite of that, the *AutoDock* approach replaces the Gaussian distribution with a Cauchy distribution:

$$C(x,\alpha,\beta) = \frac{\beta}{\pi\beta^2 + (x-\alpha)^2} \tag{2}$$

where $\alpha \leq 0, \beta > 0, -\infty < x < +\infty$ (α and β are parameters that control the mean and spread of the distribution). The Cauchy distribution has a bias toward small variations. However, unlike the Gaussian distribution, it has thick tails which allows larger variations more frequently. Some evolutionary approaches to molecular docking use both distributions for mutation operators, e.g., [5].

One important aspect is the value for the parameter σ. If it is set too low, exploitation overcomes exploration and if set too high *vice versa*. The value can be fixed or self-adapted (e.g., if an evolutionary strategy approach is used). In [5], annealing schemes to control σ as a function of time, i.e., the number of generations are proposed. Results show that the following scheme presents good results, scaled with 0.1:

$$\sigma(t) = \frac{1}{\sqrt{1+t}} \tag{3}$$

We also include in the analysis the simple uniform mutation operator. It works in the following way: when applied to a gene, it assigns a new random value according to the gene bounds, sampled from a standard uniform distribution. This operator serves as a comparison baseline.

3 Locality Analysis

We selected several instances from the *AutoDock* test suite to perform the locality analysis. Due to space limitations, we will only present results obtained with the HIV-1 protease/XK 263 protein-ligand complex. It has 10 rotatable bounds with 8 torsional degrees of freedom and is one of the largest complexes in the suite. Results obtained with other instances (e.g., β-Trypsin/benzamidine) follow the same pattern. The parameter σ is set to a value of 0.1 which previous studies in computational chemistry problems have shown to be a good value [10].

3.1 Related Work

Several techniques have been proposed to estimate and study the behavior of evolutionary algorithms and their components. Some of these methods adopt measures that are, to some extent, similar to the locality property. We highlight the most relevant ones.

The concept of fitness landscapes, originally proposed by [13], establishes a connection between candidate solutions and their fitness. Moreover, [14] proposed fitness distance correlation as a way to determine the relation between fitness and distance to the optimum. If fitness values increase as the distance to the optimum decreases, then search is expected to be easy. An alternative way to analyze the fitness landscape is to determine its ruggedness. In [15], it is proposed the adoption of autocorrelation functions to measure the correlation of all points in the search space at a given distance. In [6], conditions for strong causality are studied. A search process is said to be locally strong causal if small variations in the genotype space imply small variations in the phenotype space. In this case, variations in genotypes are caused by mutation.

3.2 Definitions

Investigations with an evolutionary framework usually means considering two spaces: the genotype space Φ_g and the phenotype space Φ_p. Genetic operators are applied on Φ_g while the fitness function, f, is applied to solutions from the phenotype space: $f : \Phi_p \rightarrow \Re$. To establish the similarity between two individuals from Φ_p a phenotypic distance has to be defined. This measure captures the semantic difference between two solutions and is directly related to the problem being solved. The phenotypic distance can be determined with a structural distance measure. To evaluate a final ligand conformation we compare it with the experimental structures using the standard Cartesian root-mean-square deviation (RMSD):

$$RMSD_{lig} = \sqrt{\frac{\sum_{i=1}^{n} dx_i^2 + dy_i^2 + dz_i^2}{n}} \tag{4}$$

where n is the number of atoms in the comparison and dx_i^2, dy_i^2 and dz_i^2 are the deviations between the crystallographic structure and the corresponding coordinates from the predicted structure lig on Cartesian coordinate i. RMSD values below or near 1.5Å can be considered to be a success criterion. Thus, lower values mean that the observed and the predicted structures are similar. Therefore, our structural distance measure determines the difference between RMSD values of two phenotypes:

$$d_{struct}(A, B) = |RMSD_A - RMSD_B| \tag{5}$$

We adopt the innovation measure proposed by Raidl and Gottlieb [8] to study the effect of mutation on locality. To predict the effect of applying this operator we use the distance between individuals in a mutation step. Let X

be a solution and X^m the result of applying m mutation steps to X, then the Mutation Innovation (MI) is given by:

$$MI = dist(X, X^m) \tag{6}$$

MI illustrates how much innovation the mutation operator introduces, i.e., it aims to determine how much this operator modifies the semantic properties of an individual. Locality is directly related to this measure. The application of a locally strong operator implies a small modification in the phenotype of an individual. The distance between the two solutions is small. On the other hand, operators with weak locality allow large jumps on the search space. To evaluate MI, 1000 random individuals are generated. Afterwards, a sequence of mutation steps is applied to each one of them and the distance between the original individual and the new solution is measured. In our experimentation, we start by applying a single mutation step. Later, we repeat the experiment with k successive mutation steps, with $k \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512\}$.

3.3 Experimentation and Discussion

Table 1 shows the characteristic values for MI with a single mutation ($k = 1$). $P(MI = 0)$ represent the percentage of cases for which $MI = 0$. $E(MI|MI > 0)$ and $\sigma(MI|MI > 0)$ show the mean value and the standard deviation of MI, for $MI > 0$. They act as estimations for the expected values. $Max(MI)$ gives the maximum value for MI.

Table 1. Characteristic values for the Mutation Innovation MI with $k = 1$.

	Uniform	Gaussian 0.1	Cauchy 0.1	Gaussian AS	Cauchy AS
$P(MI = 0)[\%]$	0.30	7.30	4.70	9.10	7.40
$E(MI\|MI > 0)$	1.28	0.04	0.15	0.03	0.11
$\sigma(MI\|MI > 0)$	1.71	0.11	0.52	0.08	0.39
$Max(MI)$	7.49	1.19	5.94	0.75	5.54

We start by considering the case where mutation does not affect the phenotype, $MI = 0$ (occurring with probability $P(MI = 0)$). A large value of $P(MI = 0)$ indicates that mutation does not make often moves in the search space. In alternative, it may also be an evidence of redundancy or strong heuristic bias since many elements could map to the same phenotype. Table 1 shows that this is not the case. The probability of $MI = 0$ is low for every operator. Uniform mutation displays the lowest value (0.30) compared to Gaussian and Cauchy mutation. Since this operator replaces a complete gene in opposition to performing a small modification, this modification is enough to produce a new phenotype. For Gaussian and Cauchy operators the final result in behavior is similar. The modifications operated by these distributions will produce different phenotypes although the probability of generating a number that is small

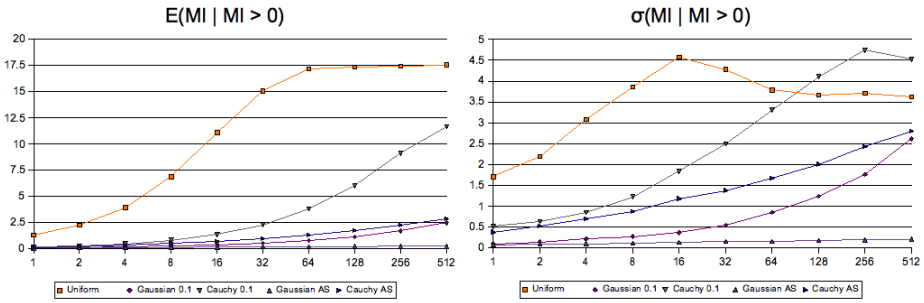

Fig. 1. $E(MI|MI > 0)$ and $\sigma(MI|MI > 0)$ over the number of mutations.

enough to induce the same individual is slightly larger. The small difference between Cauchy and Gaussian mutation is explained by the thick tails of the Cauchy distribution. These allow larger variations more frequently than with Gaussian distribution and as such, lower its $P(MI = 0)$.

Moving on to $E(MI|MI > 0)$, $\sigma(MI|MI > 0)$ and $Max(MI)$, in general, small values indicate high locality. A single mutation changes the phenotype only a little and thus, should be aspired [8]. Although lower values are good signs for a *good* locality, it should be noted that larger values for the standard deviation and for Max of MI may not necessarily be a bad indication. In our case, both distributions show low values for the locality measures. However, Cauchy mutation operators present larger values. For example, $E(MI|MI > 0)$ displays 0.15 and 0.11 in comparison to 0.04 and 0.03. The same pattern is observed for the remaining measures. To establish if these differences are statistically significant, we performed the Wilcoxon rank sum test with significance value $\alpha = 0.01$. We found significant differences between Gaussian and Cauchy mutation operators, with fixed and annealing schemes. Differences between operators with the same distribution were not found (e.g., Gaussian with fixed variance and Gaussian with annealing scheme).

A Gaussian operator displays better locality properties but, how does the distribution of mutation innovation changes when considering $k > 1$ mutations? We will now consider the case for $k \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512\}$. Figure 1 plots the empirically obtained mean values $E(MI|MI > 0)$ and $\sigma(MI|MI > 0)$ over the number of mutations k. Cauchy mutation without the annealing scheme shows higher mean and standard deviations than Gaussian operators and the Cauchy operator with the annealing scheme. For values of k larger than 32, the difference between this operator and the others increases considerably. This indicates weak locality with respect to the Cauchy operators. However, uniform mutation displays much higher values. When looking at the $E(MI|MI > 0)$ values, uniform mutation starts to express much larger values from $k = 1$, only stabilizing around $k = 64$. Nevertheless, the difference is very high showing the low locality properties induced by this operator. The combination of a Gaussian distribution and the annealing scheme displays the best behavior: for all the

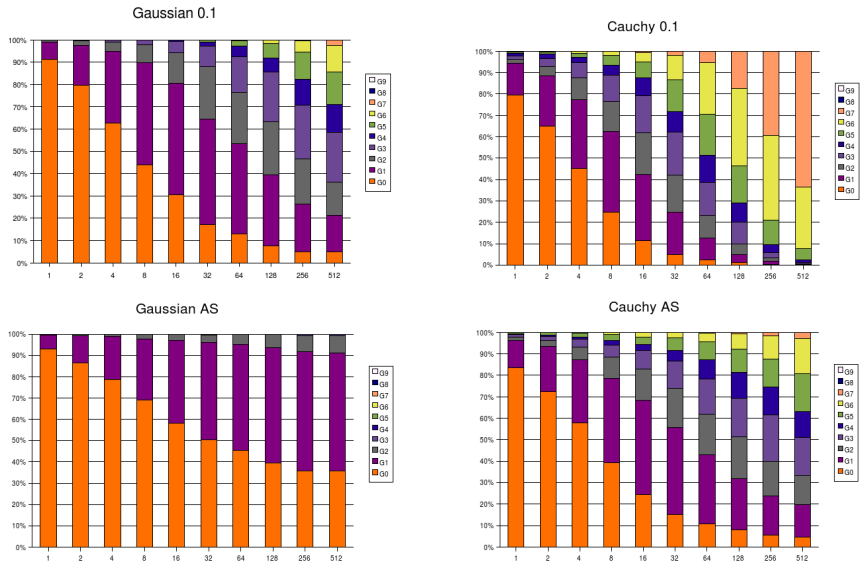

Fig. 2. Distribution of structural distances for $k \geq 1$ mutations.

mutation steps the mean and standard values remain low. This suggests a strong locality effect for this operator. The same is also true for the Gaussian operator with fixed variance and the Cauchy operator with annealing scheme operators. Nevertheless, for larger values of k, these two operators start to display a small $E(MI|MI > 0)$ increase.

Regarding $\sigma(MI|MI > 0)$ values, the pattern is similar but some remarks must be made. The most stable operator is Gaussian with annealing scheme whereas the most unstable are uniform mutation and Cauchy with fixed variance. The Cauchy operator starts with low standard deviation values but there is a shift of phase near $k = \{16, 32\}$. From this point on, the standard deviation values rise. For $k = 128$ the values are larger than uniform mutation. This is consistent with the mean values since by this time, uniform mutation has stabilized, although the distance between the mutated individuals and the originals is very large. At this point there is no semantic relation between the individuals. The Cauchy operator follows the same behavior. Here, the loss of semantic relationship occurs later in the process.

Grouping the distances between the original solution and the successive mutants allow us to observe the different types of changes operated by mutation for $E(MI|MI > 0)$. Given a structural distance d_{struct} between two phenotypes, the set G_i to which d_{struct} is assigned is determined the following way: $\{G0 : 0 \leq d_{struct} < 0.1; G1 : 0.1 \leq d_{struct} < 0.5; G2 : 0.5 \leq d_{struct} < 1; G3 : 1 \leq d_{struct} < 2; G4 : 2 \leq d_{struct} < 3; G5 : 3 \leq d_{struct} < 5; G6 : 5 \leq d_{struct} < 10; G7 : 10 \leq d_{struct} < 25; G8 : 25 \leq d_{struct} < 50; G9 : 50 \leq d_{struct}\}$. The specific values that were selected to determine intervals are arbitrary. The relevant information

is the distribution of the structural distances through the sets. Low order sets (i.e., small variations) suggest that locality is strong.

The charts in figure 2 show the distribution of structural distances for 1000 individuals, for all operators variants, with each column representing a mutation step. Important differences can be observed. Gaussian operators exhibit high locality: with the annealing scheme, $\geq 50\%$ of the distances belong to the first group until $k = 32$. From $k = 32$ to $k = 512$, the percentage of distances in the first groups stabilizes around 35%. Moreover, the majority of the remaining distances are within the lower three groups. This pattern is not observed elsewhere. This shows that this operator preserves the semantic properties of the individuals subject to mutation well. The Gaussian operator with fixed variance and the Cauchy operators demonstrate similar distributions. The main difference is given by Cauchy with fixed variance. The loss of the semantic properties can clearly be seen from the last four columns (representing the mutation steps for large k). Here the amount of individuals belonging to the last groups is considerable. This supports our previous plot analysis of $\sigma(MI|MI > 0)$.

4 Conclusions

Most evolutionary algorithms applied to this problem use one of these distribution operators (or variants based on them) but mostly Cauchy-based. However, no studies were performed to conclude about its efficiency and performance with the exception of [5]. The Gaussian operator with the annealing scheme is reported to attain the best optimization results. Nevertheless, an investigation on *why* the operator is able to achieve these results is not provided. Since Cauchy-based operators are commonly used in evolutionary approaches to molecular docking, it is important to understand their behavior and related operators.

We investigated the degree of locality induced by different mutation operators when applied to protein-ligand docking optimization. Results confirm that high locality is important and explain the behavior of different mutation operators. As such, the useful outcome from this work is twofold: 1) it explains in terms of locality the operators under investigation; 2) it provides hints on how future mutation operators can be developed. Is important for an operator to induce strong locality to obtain good optimization results. This result is sustained by the study described in [5] and experimentation performed by us (not shown due to space constraints). Gaussian mutation provides locally strong operators and this is especially true when used in conjunction with an annealing scheme. This is an indication that more fine-tuning of the conformations is allowed. On the other hand, Cauchy-based operators show a lesser degree of locality. The operator with the annealing scheme shows a locality similar to Gaussian mutation with fixed variance. Thus, these operators can provide a more exploratory role. In fact, the higher locality shown by Gaussian mutation with the annealing scheme could prove to be excessive, and therefore, difficulties to overcome traps in the search space could arise. Although results from optimization runs show this operator obtaining the best results [5], there are other algorithm's components which also

have a direct influence on the search process. As such, it is necessary to extend our research to other components, e.g., crossover, and perform additional optimization runs. Finally, in this work, local search methods were not considered. These techniques will be the focus of a future publication since the impact of local search is an important aspect of an evolutionary algorithm. As future research, we will extend this study to heritability and heuristic bias properties, to study the effects of representation and operators on this problem.

References

1. Neumaier, A.: Molecular modeling of proteins and mathematical prediction of protein structure. SIAM Review **39** (1997) 407–460
2. Morris, G.M., Olson, A.J., Goodsell, D.S.: Protein-ligand docking. In Clark, D.E., ed.: Evolutionary Algorithms in Molecular Design. Wiley-VCH (2000) 31–48
3. Thomsen, R.: Protein-ligand docking with evolutionary algorithms. In Fogel, G.B., Corne, D.W., Pan, Y., eds.: Computational Intelligence in Bioinformatics. Wiley-IEEE Press (2008) 169–195
4. Korb, O., Stützle, T., Exner, T.: An ant colony optimization approach to flexible protein-ligand docking. Swarm Intelligence **1** (2007) 115–134
5. Thomsen, R.: Flexible ligand docking using evolutionary algorithms: investigating the effects of variation operators and local search hybrids. Biosystems **72** (2003) 57–73
6. Sendhoff, B., Kreutz, M., Seelen, W.V.: A condition for the genotype-phenotype mapping: Casualty. In: 7th Int. Conf. on Genetic Algorithms. (1997) 73–80
7. Rothlauf, F.: On the locality of representations. In: Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2003). (2003) 1608–1609
8. Raidl, G.R., Gottlieb, J.: Empirical analysis of locality heriability and heuristic bias in evolutionary algorithms: A case study for the multidimensional knapsack problem. Evolutionary Computation Journal **13** (2005) 441–475
9. Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K., Olson, A.J.: Automated docking using a lamarckian genetic algorithm and and empirical binding free energy function. Journal of Computational Chemistry **19** (1998) 1639–1662
10. Pereira, F.B., Marques, J., Leitão, T., Tavares, J.: Analysis of locality in hybrid evolutionary cluster optimization. In: Proceedings of the 2006 IEEE Congress on Evolutionary Computation, Vancouver, Canada, IEEE Press (2006) 8049–8056
11. Dixon, J.S.: Flexible docking of ligands to receptor sites using genetic algorithms. In: Proc. of the 9th European Symposium on Structure-Activity Relationships, Leiden, The Netherlands, ESCOM Science Publishers (1993) 412–413
12. Moitessier, N., Englebienne, P., Lee, D., Lawandi, J., Corbeil, C.: Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. British Journal of Pharmacology **153** (2007) 1–20
13. Wright, S.: The roles of mutation, inbreeding, crossbreeding and selection in evolution. In: Proceedings of the VI International Conference on Genetics, Vol. 1. (1932) 356–366
14. Jones, T.: Evolutionary Algorithms, Fitness Landscapes and Search. PhD thesis, University of New Mexico, Albuquerque, New Mexico (1995)
15. Weinberger, E.D.: Correlated and uncorrelated fitness landscapes and how to tell the difference. Biological Cybernetics **63** (1990) 325–336