A Concept for Real-Valued Multi-objective Landscape Analysis Characterizing Two Biochemical Optimization Problems

Susanne Rosenthal and Markus Borschbach(✉)

Chair of Optimized Systems, Faculty of Computer Science,
University of Applied Sciences, FHDW,
Hauptstr. 2, 51465 Bergisch Gladbach, Germany
{Susanne.Rosenthal,Markus.Borschbach}@fhdw.de

Abstract. Landscape analysis is an established method to provide an insight into the characteristic properties of an optimization problem with the aim of designing a suitable evolutionary algorithm for a given problem. However, these conventional landscape structures require sophisticated notions for multi-objective optimization problems. This work presents a real-valued multi-objective landscape analysis concept that allows the investigation of multi-objective molecular optimization problems. Sophisticated definitions for ruggedness, correlation and plateaus on multi-objective real-valued landscapes are introduced and indicators are proposed for this purpose. This landscape concept is realized on a generic three- and four-dimensional biochemical minimization problem and the results of this analysis are discussed regarding the design principles of a multi-objective evolutionary algorithm.

Keywords: Real-valued multi-objective landscape · Molecular landscape · Analysis concept · MOEA design

1 Introduction

The design of a Multi-Objective Evolutionary Algorithm (MOEA) for a specific class of optimization problems requires the knowledge of the landscape characteristics [1] to tune the algorithm for an increased search performance. The use of MOEAs for molecule optimization has increased significantly, but the general understanding of the molecular landscape properties with the aim of designing an appropriate MOEA to search the molecular space is missing [2]. The analysis of the landscape structure provides information about the landscape characteristics and difficulties of molecular optimization problems. This information provides a better insight into the composition of a search performance optimized MOEA regarding a particular type of algorithm, the types of variation operators and the selection pressure for a suitable balance of global and local search behavior. Nevertheless, it is known that the fitness landscape structure influences the EA performance and various techniques for statistical analysis as qualitative technique and information analysis as quantitative technique are proposed

A.M. Mora and G. Squillero (Eds.): EvoApplications 2015, LNCS 9028, pp. 897–909, 2015.
DOI: 10.1007/978-3-319-16549-3_72

(see [3] for an overview) to analyze single-objective fitness landscapes. In the case of multi-objective landscapes, the important landscape structures modality, ruggedness, correlation and plateaus have to be generalized or defined in a more sophisticated manner. A respectable amount of work has been done in the area of multi-objective landscape analysis for combinatorial optimization problems, e.g. assignment problems [1,4,5]. In [4], potentially useful indicators are discussed characterizing multi-objective landscapes: the modality is characterized by the number and distribution of global optima. From the multi-objective point of view, these global optima are a set of non-dominated or Pareto optimal solutions. The fitness distance correlation (FDC) in the case of single-objective landscapes is a correlation coefficient that indicates the distance between a set of local optima to the nearest global optima. Due to the generalization to multi-objective landscapes, non-dominated solutions (NDS) are considered as global optima. The critical point of this notion is that each of the NDS is the optimum of one single-objective function and therefore, the correlation between the NDS is not necessarily resembling to the correlation between different local optima of a single-objective function. However, the correlation between NDS provides useful information referring to a search process of MOEA moving along the Pareto front. Additionally, concepts are introduced by Garrett, which define other distance indicators: the Euclidean distance between the solutions or mathematically spoken between the fitness vectors or the angle between these fitness values as an alternative distance indicator. These proposed metrics have not been investigated empirically or theoretically so far. An intuitive definition for landscape ruggedness is also given by Garrett. The autocorrelation of the random walks between known Pareto optimal solutions is used to investigate the path ruggedness between the Pareto optima. A further elaboration of these concepts or empirical investigations are missing by today. [4]

A traditional and systematical molecular landscape analysis is presented in [2,8]. Herein, the purpose of the molecular landscape analysis is defined by the examination of the common principle -molecular structure similarity is often related to similar molecule properties- for the underlying optimization problem. Four molecular functions are analyzed separately according to the landscape properties modality, ruggedness, neutrality, local optima and basins. This landscape analysis is based on random walks of length 100 and 500 over a search space with a complexity of 23^5 feasible peptides (5-mer peptides consisting of 23 amino acids). Feasible solutions are character strings of length 5, neighboring solutions differs in exactly one amino acids and one-point mutation is used as moving operator to explore the neighborhood.

In this work, a concept for multi-objective landscape analysis is proposed with the aim of analyzing a multi-objective molecular landscape (MOML), which involves the ideas of Garrett. The analysis results are used for design considerations of a MOEA with optimized search performance for the purpose of multi-objective biochemical optimization. This concept is based on the considerations of important properties modality, correlation, ruggedness and plateaus on real-valued multi-objective landscape. Sophisticated definitions of these landscape properties are presented and discussed. Furthermore, indicators are proposed for this purpose. These techniques are simple to calculate and most important - independent of the optimization problem dimension. This concept is applied on a generic three- and four-dimensional biochemical optimization problem and the results are interpreted according to the guidance of the search process of a MOEA.

2 A Concept for MOML Analysis

The evolution of a landscape analysis concept in general requires the determination of the fitness landscape components: the components of a fitness landscape are a set of genotypes (configuration of the solutions), the fitness functions which evaluate the genotypes and the genetic operators, which represent the move operator to explore the neighborhood. Stadler presented the formal description of the landscape composition [6]:

Definition 1. *A landscape consists of three ingredients: a set X of configurations; a notation X of the neighborhood, the nearness, distances or accessibility on X; and a fitness function $f : X \rightarrow \mathbb{R}$.*

A landscape analysis starts by specifying metrics that characterize the geometric properties. The selection of suitable metrics depends on the organization of the configuration space X and has to take account of the optimization problem. Reidys and Stadler [7] stated three distinct approaches for the organization of X:

1. transition probabilities are used to describe the movement from one configuration to another. The process is describable by Markov chains and is especially applied in the case of combinatorial optimization problems.
2. in the field of computer sciences, genetic operators (mutation or recombination) are usually used as move operators to create new configurations.
3. rigorous mathematical analysis is performed by specified metrics or topologies on X.

The set X comprises all feasible peptides and is given by a character string. According to [8,9], the neighborhood of a configuration is explored by one-point mutation as move operator and neighbored configurations are differing by exactly one amino acid or a character in the MOEA terminology. The one-point mutation is used as move operator for an insight into the mutation potential of a MOEA and to avoid highly differing consecutive configurations, which are potentially produced by a recombination operator. Small changes in the configurations provide information about the effectiveness of the local search of a MOEA. The organization of the configuration set refers to the second approach of Reidys and Stadler as the other approaches are unsuitable: The use of Markov chains is not advisable because of the general difficulty to efficiently design high complex spaces [10], which usually occurs in molecular spaces. Furthermore, X allows no mathematical definitions of metrics of topologies.

Modality. The modality or the investigation of the optima density is examined based on measurements of the random walk part consisting only of the NDS or the individuals of the first front. The modality requests information about the number of NDS, a potential clustering of these or otherwise a large distribution over the MOML. For this purpose, the individuals of the random walk are ranked into fronts. For an Optima Distribution Analysis (ODA), the average Euclidean distance d_{ODA} between all possible combinations of non-dominated fitness values $\boldsymbol{x}_i$ is determined:

$$d_{ODA} = \frac{1}{K} \sum_{i,j} d_{ij} \text{ with } d_{ij} = |\boldsymbol{x}_i - \boldsymbol{x}_j| \text{ for } i,j = 1, ..., M \text{ and } i < j, \quad (1)$$

where M is the number of fitness vectors in the first front and $K = \binom{M}{2}$ the number of all possible combinations of differences d_{ij}. The value of d_{ODA} is a measure for the central tendency of the non-dominated solution diversity. Otherwise, d_{ODA} globally seen as mean value has its limitation in the case of extremal boundary values. Therefore, the diversity of the NDS is quantified via the average distance of all distances d_{ij}:

$$d_{MAD} = \frac{1}{K} \sum_{i,j} |d_{ij} - \bar{d}| \text{ with } i,j = 1, ..., M \text{ and } i < j. \quad (2)$$

with $\bar{d} = d_{ODA}$. The higher the diversity values, the wider is the spread of the NDS over the search space. In the case that the range of the objective function values are differing drastically, the use of the normalized Euclidean distance is advisable. Another indicator for the distribution of NDS is the measurement of the so-called 'beeline' between two consecutive NDS along the random walk path. Therefore, the magnitude of the beeline between two consecutive fitness vectors $\boldsymbol{x}_{i+1}$ and $\boldsymbol{x}_i$ is determined and is set in relation to $\bar{c} = \sum_{i=1,...,N-1} |\boldsymbol{y}_{i+1} - \boldsymbol{y}_i|$, the average Euclidean distance between two consecutive fitness vectors $\boldsymbol{y}_{i+1}$ and $\boldsymbol{y}_i$ of the random walk with N as the number of random walk steps to classify the distribution tendency:

$$b_i = \frac{|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i|}{\bar{c}} \text{ with } i,j = 1, ..., M - 1, \quad (3)$$

where x_i are ordered according to their occurring in the random walk. A low number of b_i indicates that the corresponding distance between the two consecutive NDS is relatively small compared to the average distance of all consecutive distances of the random walk.

Correlation. The correlation is a measure for the relationship between two configurations in the landscape. A correlation analysis of the single fitness functions provides some information about the correlation tendency of the corresponding fitness values. In the case of MOMLs, the correlation between the single molecular fitness functions is of great interest as the high correlation between two time series of different fitness functions theoretically reduces the optimization problem dimension and therefore the problem difficulty. The correlation matrix is a suitable analysis technique for this purpose:

$$M_{corr} = \begin{pmatrix} 1 & corr(f_1, f_2) & ... & corr(f_1, f_k) \\ corr(f_2, f_1) & 1 & ... & corr(f_2, f_k) \\ \vdots & \vdots & \ddots & \vdots \\ corr(f_k, f_1) & corr(f_k, f_2) & ... & 1 \end{pmatrix}, \tag{4}$$

where M_{corr} is symmetrical and consists of the Pearson correlation coefficients of the fitness function f_i and f_j:

$$corr(f_i, f_j) = \frac{\sum_{i=0}^{n} (f_i - \bar{f}) \cdot (f_j - \bar{f})}{\sigma_{f_i} \cdot \sigma_{f_j}} \tag{5}$$

In this context, the correlation coefficients lie in a range of $[-1; 1]$, where negative value symbolize a potential anti-proportional linear relationship and a positive value a possible proportional linear relationship. Furthermore, no or at least a low correlation is given by $|corr(x, y)| < 0.3$. A weak correlation is given by $0.3 \leq |corr(x, y)| \leq 0.8$ and $|corr(x, y)| > 0.8$ indicates a high linear correlation.

Ruggedness. The ruggedness refers to the relationship between each configuration and its neighbors. A landscape is said to be rugged if it reveals high varying fitness values, the greater the fitness differences the more rugged is the landscape. From this point of view, the analysis technique for MOML ruggedness is based on the difference vectors determined between each two consecutive fitness vectors of the random walk. A measure for the variation of the fitness vector values is the magnitude of the absolute value calculated of the difference vectors. The absolute value of the difference vectors provides an insight in the magnitude of differences between the single molecular fitness functions. A closer consideration of the absolute values as a measure for fitness difference leads to the insight that this value does not take account of the fitness variation of the single molecular fitness functions in the sense that potentially only a few of these fitness functions are responsible for a high absolute value. Furthermore, another view on the absolute value reveals that it is no indicator for the direction of the single molecular function moving and therefore no indicator for the increase, decrease or stagnation of the different fitness functions. These considerations lead in conclusion to a definition of ruggedness: a real-valued multi-objective fitness landscape is regarded as rugged if the single fitness functions are moving differently with high fitness differences. As a consequence, this landscape is regarded as smooth if all fitness functions are moving equally or only a very few of these functions are directed differently and with small fitness differences.

The information about the single fitness function directions are provided by the difference vectors between the consecutive fitness vectors. A suitable indicator for the direction of the difference vectors is the angle between the difference vectors as - in general - an angle between vectors is an indicator for similarity [9]:

$$similarity(\boldsymbol{x}, \boldsymbol{y}) = cos(\theta) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{|\boldsymbol{x}| \cdot |\boldsymbol{y}|} \tag{6}$$

For the angle between two consecutive difference vectors, which provide information about the relative position of three consecutive fitness vectors and therefore of the fitness variance direction of a configuration and its neighbors along the random walk, the following geometrically interpretation is stated: an angle of 0 refers to two vectors pointing in the same direction. This implies, that the single fitness values of the consecutive random walk steps, which define these two difference vectors, are all positioned in the same direction. Otherwise, an angle of more than 90 indicates a moving to a large part of single fitness function in different directions. In the case of stagnating objective function values, the difference vector is the zero vector and the angle is not defined. In that case, the angle is set to 0.

Hence, to gain an insight into the potential ruggedness and structure of a MOML, the angle between every two consecutive difference vectors is calculated. Furthermore, the length of the random walk path consisting of the difference vectors, of which two forming an particular angle $\measuredangle(x_{i+1}, x_i) = a$ with $a \in [0; 180]$ is determined to gain information about the magnitude of fitness differences. This path length allows no statistically reasonable interpretation as these values depend on the subspace dimension of the search space covered by the random walk steps. Therefore, this path length is set in relation to the number of random walk steps. The fitness vectors have been normalized to ensure comparative values.

$$p_{length} = \frac{\sum |x_{i+1}| + |x_i|}{2 \cdot (N-1)}. \tag{7}$$

Plateaus. Another important structure of a landscape are the plateaus. The number and size distribution of plateaus are investigated by neutrality measures [2]. In MOML, plateaus are characterized in two different aspects: firstly, plateaus are characterized according to the stagnation of all objective functions values over several steps of the time series and secondly - in a more global view - according to the number of consecutive time series steps in the same Pareto front. The plateau characterization in the sense of objective function stagnation is determined via:

$$|x_{i+1} - x_i| \leq 1 \text{ for } i = 1, ..., N-1. \tag{8}$$

3 Computational Landscape Analysis and Discussion

Simulation Onsets. Short peptide sequences of a length of 20 consisting of 20 canonical amino acids constitute the search space. Therefore, the search space has a complexity of 20^{20} feasible solutions and is further discrete for the proposed physiochemical fitness functions as there are real-valued solution vectors which have no corresponding configurations in the search space. The MOML analysis is performed via random walks with one-point mutation to investigate the neighbored molecular landscape. The mutation of the same amino acid is excluded to avoid a stagnation of the random walk. The start configuration of

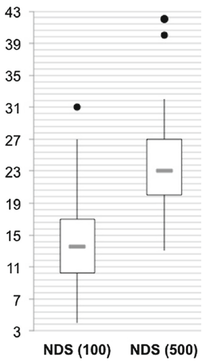

Fig. 1. No. of NDS (3D-MOML).

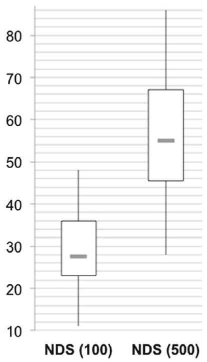

Fig. 2. No. of NDS (4D-MOML).

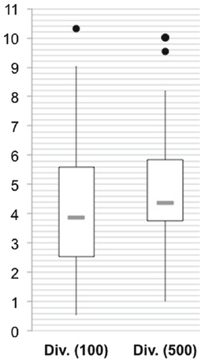

Fig. 3. Diversity of NDS (3D-MOML).

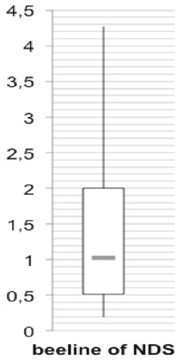

Fig. 4. Beeline of NDS (3D-MOML).

the random walk is initialized randomly. The phenotypes of the MOML are real-valued vectors of length k according to the number of objectives. Random walks of length 100 and 500 are performed and the consecutive real-valued vectors of each random walk are termed time series. For statistic reason, these random walks are repeated at least 30 times.

Physiochemical Properties. Three of the four physiochemical fitness functions are provided by the BioJava library [11]. BioJava is a Java tool that provides different physiochemical property data as well as a module for sequence alignment for peptides and proteins composed of the 20 canonical amino acids. The Needleman Wunsch algorithm (NMW) provided by BioJava is used as global sequence alignment to a pre-defined reference peptide. The optimal alignment is found in a quantitative way by assigning scores for matches, mismatches and gaps. NMW uses different scoring models. Here, the BLOcks SUbstitution Matrix (BLOSUM 100) is used with the percentage identity of 100 [12]. Two further physio-chemical functions are utilized of BioJava: the Molecular Weight (MW) is computed by the sum of the mass of each amino acid plus a water molecule. The Instability Index (InstInd) of a peptide is calculated by the summation of the Dipeptide Instability Weight Values (DIWV) of each two consecutive amino acids in the peptide sequence. The summarized value is normalized then by the peptide length. The fourth physio-chemical function is the Hydrophilicity (Hydro), which is calculated by the method of Hopp and Woods [13]: hydrophilic parts of a peptide are determined by a sliding window of a fixed size over the sequence and averaging the corresponding amino acid scales. Here, the window size is the entire peptide length. All this fitness functions act comparatively to a pre-defined reference peptide and have to be minimization for optimization. The fitness function NMW, MW and Hydro constitute the 3D-MOP. The 4D-MOP has to optimize NMW, MW, Hydro and InstInd.

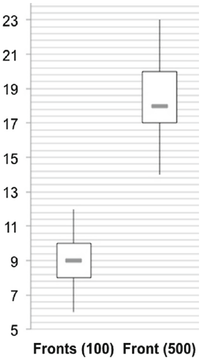

Fig. 5. Number of fronts (3D-MOML).

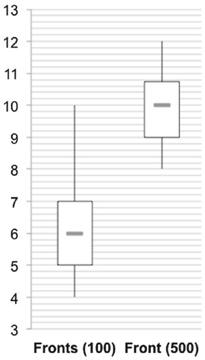

Fig. 6. Number of fronts (4D-MOML).

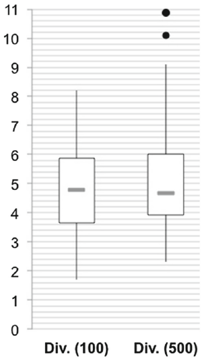

Fig. 7. Diversity of NDS (4D-MOML).

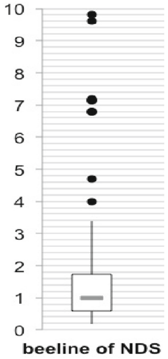

Fig. 8. Beeline of NDS (4D-MOML).

Experiments. In a first step, the modality is investigated for the 3D- and 4D-MOP. Therefore, the number of individuals in the first front, the total number of detected fronts, the diversity of the individuals and the relational beeline is determined by random walks of length 100 or 500 respectively. Figures 1 and 2 depict the number of NDS (NDS) detected in the time series of length 100 and 500. Figure 1 reveals that 50 % of the NDS are in a inter-quartile range determined by 10 % to 17 % of the Random Walk Length (RWL). An increase of the RWL (right boxplot of Fig. 1) results in an increase of the NDS round about 83.9 %. The number of NDS in the time series of the 4D-MOML is on average significantly higher than for the 3D-MOML (Fig. 2). The inter-quartile range of the time series of length 100 (Fig. 2) is determined by 23 % to 36 % of the RWL. A comparison of the time series of length 100 in Figs. 1 and 2 reveals an increase of the NDS round about 53 % in the case of 4D-MOML. An increase of the RWL from 100 to 500 (right boxplot of Fig. 2) results in an increase of NDS round about 84, 2 %, this value is comparable to the results of the 3D-MOML. Concluding, the investigation of larger times series reveals a larger number of NDS, but this increase is of a lower level than the increase of the RWL. Further, the 4D-MOML provides a significantly higher number of NDS than the 3D-MOML. This effect is due to the lower front diversity in the case of 4D-MOML. Figures 5 and 6 depict the front diversity of the 3D- and 4D-MOML in the time series of length 100 and 500. Figure 5 reveals that 50 % of the front numbers in the time series of length 100 are in the inter-quartile range determined by 8 and 10 fronts. An increase of the RWL to 500 results in an increase of the detected front number round about 104 % referring to the results of the time series of a length of 100. The front diversity is significantly lower in the case of the 4D-MOML (Fig. 6).The increase of the RWL from 100 to 500 (right boxplot in Fig. 6) results in an front diversity increase of round about 52,3 %. This percentage increase is only a half of the average increase observed in the 3D-MOML. This is a consequence of the fact that the average number

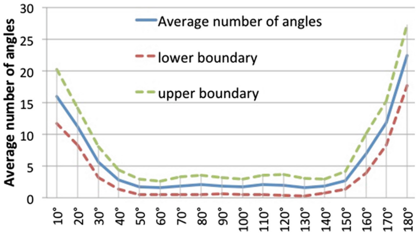

Fig. 9. Average number of angles between two consecutive difference vectors (3D-MOML).

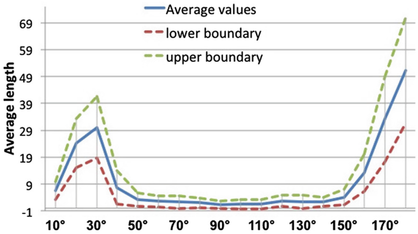

Fig. 10. Average length of the two consecutive difference vectors enclosing a particular angle (3D-MOML).

of NDS in the random walks of a length of 100 is significantly higher than in the case of the 3D-MOML, but the increase of the NDS number by an increase of the RWL is comparable. Therefore, the increase of the front diversity by an increase of the RWL is significantly lower.

Figures 3 and 7 present the spread of the NDS diversity in the time series of length 100 and 500. In general, the diversity d_{MAD} of the NDS - computed by Eq. (2) - is of the same level for the 3D- as well as 4D-MOML. Only, the diversity of the NDS in the 3D-time series of length 100 reveals a tendency for lower diversity values. Figures 4 and 8 depict the average rational beeline (Eq. (3)) between each consecutive NDS in time series of length 100. The boxplots reveal that some of the NDS are clustered and others are positioned in a wide range of distance: 50 % of the relational beeline values are between $0,5$ and 2 in the case of the 3D-MOML (Fig. 4), which indicates that the distance between the corresponding consecutive NDS is more than half $(0,5)$ or twice (2) of the average distance between all consecutive solutions time series. The relational beeline values of the 4D-MOML are between $0,6$ and $1,7$, which indicate that the distance between the consecutive NDS is more than a half $(0,6)$ or more than one and a half $(1,7)$ of the average distance between all consecutive solutions (Fig. 8). However, Fig. 8 reveals some outliers up to a value of 10. This indicates that some distances between the NDS are significantly higher.

The correlation matrix of the physio-chemical functions is given by (Eq. (4)):

$$M_{corr} = \begin{pmatrix} 1 & 0.047 & 0.252 & 0.09 \\ \cdots & 1 & -0.014 & -0.032 \\ \cdots & \cdots & 1 & -0.266 \\ \cdots & \cdots & \cdots & 1 \end{pmatrix}. \tag{9}$$

The matrix entries reveal no linear relationship between the time series of each two molecular fitness functions: there is a weak relationship between NMW and MW (Eq. (9): $corr(f_1, f_3) = 0.252$) as well as InstInd and Hydro (Eq. (9)): $corr(f_3, f_4) = -0.266$) and no correlation between the other combinations. As a consequence, the dimension of a MOML constituted of these four molecular functions is equal to the number of participating objective functions.

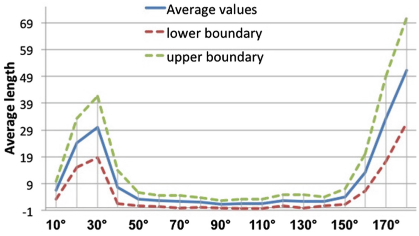

Fig. 11. Average number of angles between two consecutive difference vectors (4D-MOML).

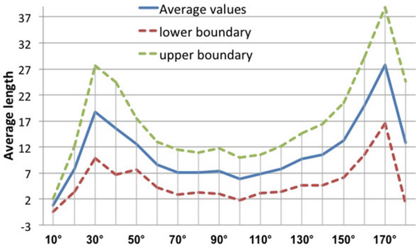

Fig. 12. Average length of the consecutive difference vectors enclosing a particular angle (4D-MOML).

Figures 9 and 10 depict the average number of angles and the average path length with a particular bending of the 3D- time series - categorized in intervals of ten degree on the x-axis. The depicted upper and lower boundaries mark the one-sigma interval. The highest number of angles are detected in the interval of $[170; 180)$ (Fig. 9). This indicates that the difference vectors are oppositely directed and the single objective functions are increasing, decreasing or stagnating over three steps of the time series in very different manners. Exemplary spoken: one objective function increases from one time series step to the next one and decreases afterwards. The second function is exactly moving the other way around and the third function is stagnating from the first to the second solution and increasing or decreasing afterwards. This reveals that the landscape is very rugged along a large number of random walk steps. The second highest number of angles is in the interval of $[0; 10)$. This indicates that the difference vectors are similarly directed and the single objective functions are increasing, decreasing or stagnating in a similar manner. Exemplary spoken: one of the objective functions is stagnating over three time series steps and the other two functions are increasing or decreasing over these three steps. The number of angles in the interval of $[40; 150)$ is almost stable. In general, the larger the angle the larger the number of objective functions revealing oscillating moving behavior in a different manner over three time series steps.

A similar pattern is achieved by calculating the average path length with a particular bending provided by the difference vectors which enclose particular angles (Fig. 10). The highest length is achieved in the interval $[170; 180)$ indicating large differences between the single molecular function values with mainly oscillating behavior. The second highest length is achieved in the intervals $[10; 30)$ indicating large differences between the solutions of the time series mostly positioned in the same direction. The length of the difference vectors enclosing angles in the interval $[40; 150)$ are small and reveal slight changes of the single objective function values. Figures 11 and 12 depict the corresponding results for the 4D-MOML. Compared to the results of the 3D-MOML, the maxima of average number of angles and the average path length are in the intervals $[20; 30)$ and $[160; 170)$ and the values of the interval $[40; 160)$ are higher in the

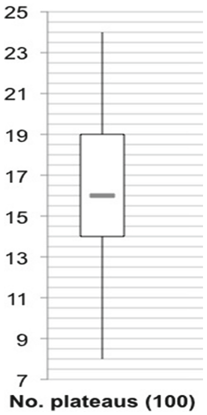

Fig. 13. No. of 3D front plateaus (100).

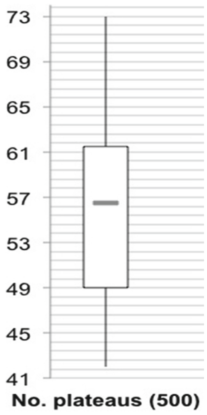

Fig. 14. No. of 3D front plateaus (500).

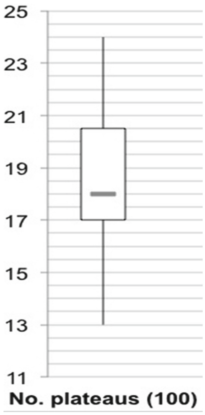

Fig. 15. No. of 4D front plateaus (100).

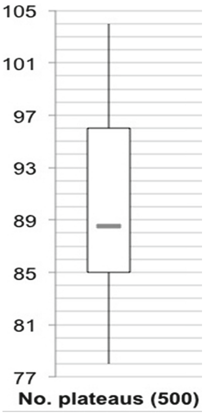

Fig. 16. No. of 4D front plateaus (500).

case of the 4D-MOML. This is a consequence of the fact that the probability of the four objective functions moving similarly or oscillating simultaneously is lower than for a lower number of objective functions. Plateaus are a structural property that provides some information about clustered similar qualified solutions. Firstly, plateaus are identified in MOML by consecutive equal or nearly equal fitness values for each molecular function (see Eq. (8)). In the 30 random walks of length 100 on the 3D-MOML, 20 plateaus have been identified totally: two plateaus of each two consecutive equal fitness values have been identified in five random walks, a plateau of three consecutive equal fitness values have been found in one random walk and the remaining 9 plateaus have been identified in different random walks each consisting of two consecutive equal fitness values. Only 8 plateaus have been detected in the corresponding random walks, each of length 2. Figures 13, 14 and 15, 16 depict the number of front-plateaus in the 3D- and 4D-MOML detected in 30 random walks of length 100 and 500. Front-plateaus are more globally characterized by consecutive time series steps in the same Pareto front. In general, an increase of the time series length results in an increase of the plateaus for the 3D- and 4D-MOML. The number of plateaus is always higher in the case of the 4D-MOML, a consequence of the lower front diversity. Thus, the increase of the plateau number is significantly slower than the increase of the time series length. This is once more a consequence of the higher front diversity in larger time series. In the case of 3D-MOML: 14, 5 % and 7 % of the plateaus detected in the time series of length 100 and 500 are first front plateaus. The average plateau sizes are 2, 31 and 2, 18. The average plateaus size of the first front plateaus are on average larger with 2, 7 and 2, 3 in the time series of length 100 and 500. In the case of 4D-MOML: 36, 5 % and 12, 9 % are first front plateaus in the time series of length 100 and 500. Compared to the 3D-MOML, these percentage increases are a consequence of the lower front diversity of the 4D-MOML. The average plateau sizes are $3,04$ and $2,42$. The average plateaus size of the first front plateaus are $3,08$ and $2,75$ in the time series of length 100 and 500.

4 Conclusions and Future Work

The results of the 3D- and 4D-MOML analysis provide some important hints regarding the design of a MOEA: the 3D- and 4D-MOML are very rugged and no significant structure is discernible according to the distribution of the NDS over the landscapes. The 3D-MOML reveals a higher front diversity and therefore fewer solutions are in the optimal front compared to the 4D-MOML. Further, the average first-front-based plateau size is accordingly smaller. These facts make the 3D-MOML more rugged and therefore more challenging for a MOEA than the 4D-MOML. In general, the significant number of front plateaus in both MOMLs require a specific balance of global and local search behavior of the MOEA: the variation operators of the MOEA have to support a global search in the first generations of the MOEA to tap potential high quality solutions widely spread over the landscape. In the later generations, a more local search behavior of the MOML supports the search process in the neighborhood of the previously detected high quality solutions. The 4D-MOML reveals a higher number of NDS caused by the lower front diversity which requires far-reaching differentiation of the NDS. The most intuitive way to perform this differentiation is by assistance of the selection procedure. A strategy providing a good differentiation is an indicator-based selection strategy. The increase of the RWL and therefore of the investigated MOML does not result in a corresponding increase of NDS in both MOMLs. Further, the NDS are unevenly distributed over the search space. These facts indicate that an increase of the population size does not result in highly improved MOEA performance from a statistical point of view.

The optimization results verifying the 3D-MOP difficulties compared to the 4D-MOP are the topic of future work. Apart from these two biochemical MOP, the generality of this concept is validated on classical MOP as another topic of future work.

References

1. Merz, P., Freisleben, B.: Fitness landscape analysis and memetic algorithms for the quadratic assignment problem. IEEE Trans. Evol. Comput. **4**(4), 337–352 (2000)
2. Emmerich, M., Lee, B.V.Y., Render, A., Faddiev, E., Kruisselbrink, J., Deutz, A.H.: Analyzing molecular landscapes using random walks and information theory. Chem. Cent. J. **3**(1), 20 (2009)
3. Merkuryeva, G., Bolshakovs, V.: Benchmark fitness landscape analysis. Int. J. Simul. Syst. Sci. Technol. **12**(2), 38–45 (2011)
4. Garrett, D., Dasgupta, D.: Multi-objective landscape analysis and the generalized assignment problem. In: Maniezzo, V., Battiti, R., Watson, J.-P. (eds.) Learning and Intelligent Optimization. LNCS, vol. 5313, pp. 110–124. Springer, Heidelberg (2007)

5. Knowles, J.D., Corne, D.W.: Towards landscape analysis to inform the design of a hybrid local search for the multi-objective quadratic assignment problem. In: HIS, pp. 271–279 (2002)
6. Stadler, P.M.: Fitness Landscape. Lecture Notes in Physics, vol. 585. Springer, Heidelberg (2002)
7. Reidys, C.M., Stadler, P.F.: Combinatorial landscape. SIAM Rev. **44**, 3–54 (2002)
8. Lee, B.V.Y.: Analyzing Molecular Landscapes using Random Walk and Information Theory. LIACS, University of Leiden, Masterthesis (2009)
9. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson Addison Wesley, Boston (2006)
10. Ceperly, D., Chen, Y., Crain, R.V., Meng, X., Mira, A., Rosenthal, J.: Challenges and advances in high dimensional and high complexity monte carlo computation and theory. In: Proceedings of the Workshop at the Banff International Research Station for Mathematical Innovation Discovery (2012)
11. BioJava, version 3.0.8. http://biojava.org/wiki/Main_Page
12. Henikoff, S., Henikoff, J.G.: Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA **89**(22), 10915–10919 (1992)
13. Hopp, T.P., Woods, K.R.: A computer programm for predicting protein antigenic determinants. Mol. Immunol. **20**(4), 483–489 (1983)