Diversity Promoting Strategies in a Multi- and Many-Objective Evolutionary Algorithm for Molecular Optimization

Susanne Rosenthal(✉)

Department of Information Systems, RFH - University of Applied Sciences of Cologne, 50676 Cologne, Germany
susanne.rosenthal2@rfh-koeln.de

Abstract. Computer-aided drug design is an approach to effectively identify and analyse molecules for therapeutic and diagnostic interventions. Generally, libraries with a broad range of compounds revealing a high genetic diversity with an at most similar behavior in bioactivity have to be created. For this purpose, an evolutionary process for multi- and many-objective Molecular Optimization (MO) has been designed and improved during the past decade. Diversity plays a central role in Evolutionary Algorithms (EAs) to prevent premature convergence to suboptimal solutions and several methods to promote diversity on different levels of an EA have been proposed. The aspect of genetic diversity in MO is a further challenge that has to be controlled and promoted by different strategies on various stages of a problem-specific EA. This work presents an application-specific re-interpretation of different diversity aspects on various stages of an EA for MO. A sophisticated survival selection strategy combining a specific ranking method with application-specific diversity promoting technologies is introduced and benchmarked to the recently proposed many-objective evolutionary algorithm AnD on four molecular optimization problems with 3 up to 6 objectives.

Keywords: Genetic dissimilarity · Genotype and fitness diversity · Multi- and many-objective molecular optimization

1 Introduction

Drug discovery for therapeutical and diagnostic entities is a highly complex process and still costly, difficult and time-consuming. The aim of drug discovery is to identify candidate antibodies to disease-relevant targets that are complementary in shape and charge to these targets with which they interact and bind. This process is often a combination of computer techniques, bioinformatic approaches and laboratory experiments to simultaneously improve molecular properties like affinity, selectivity and metabolic stability [1].

For this purpose, a single-objective EA for MO has been evolved revealing exponential fitness improvement of candidate peptides within 10 iterations,

B. Filipič et al. (Eds.): BIOMA 2020, LNCS 12438, pp. 294–307, 2020.
https://doi.org/10.1007/978-3-030-63710-1_23

slowed down to linear fitness improvement afterwards [2]. A sophisticated version of this approach with similar properties for multi-objective MO, termed as COmponent-Specific Evolutionary Algorithm for Molecular Optimization (COSEA-MO), has been reported and benchmarked on a 3- and 4-dimensional physiochemical optimization problems in [3]. The components have been compared to several state-of-the-art components and a fine-tuning of the parameters, number of recombinations and population size, has been performed [3–5]. Furthermore, COSEA-MO has been enhanced for the application on multi- and many-objective MO problems by a winning-score based ranking method as survival selection [6] providing again exponential fitness improvement within 10 iterations. This enhanced version has been evolved under specific conditions:

- provides exponential convergence improvement within 10 iterations on multi- as well as many-objective molecular optimization problems,
- components are parameter-free in the sense that no parameters have to be chosen by the user which have a high impact on the performance,
- the algorithm does not make use of reference points, weight vectors or a division of the search space by hyperboxes, which also have a high impact on the performance and have to be chosen carefully by the user.

The genetic diversity with the meaning of genetic material among the candidate optimized peptides is an important feature and less work has been done so far to control this aspect of diversity in an evolutionary process, especially in the field of MO.

Generally, diversity is the second important aim in evolutionary optimization and is usually addressed in an evolutionary process to prevent premature convergence on suboptimal solutions. Therefore, several diversity strategies are included on different stages in an evolutionary process acting on the three levels genotype, phenotype and fitness with a different impact on the performance [7].

The contribution of this work is a application-specific re-interpretation of diversity promoting aspects in an evolutionary process for multi- and many-objective molecular optimization and an enhancement of this evolutionary process, COSEA-MO, to control diversity among the candidate optimized peptides by a sophisticated selection procedure to identify a significant number of highly qualified peptides with an at most wide range of genetic diversity among themselves. For these issues, the following questions are addressed in this work:

1. What does *diversity* mean in the field of MO?
2. How to address *diversity* on different stages of an evolutionary process?

A sophisticated selection strategy as a linear combination with the terms molecular quality, genetic diversity and dissimilarity is introduced. The molecular quality is measured by a winning-score technique [8]. Hamming distance and a dissimilarity measure based on the matrix of Sneath [9] is used to calculate the diversity of the genetic material. The performance of COSEA-MO with the new selection function is evaluated on four MO problems with 3 to 6 objectives and is compared the recently proposed many-objective evolutionary algorithm AnD (ANgle-based selection and shift-based Density estimation strategy)

[10] with the survival selection principle 'diversity-first-and-convergence-second'. AnD has the same simple framework structure as COSEA-MO, is also independent of problem-specific weight vectors and reference points and outperformed several state-of-the-art many-objective EAs on standard benchmark problems. AnD has been chosen for comparison as it is currently the only state-of-the-art algorithm that is compatible with the second and third condition mentioned above.

The outline of this work is as follows: Sect. 2 gives an overview of preliminary knowledge, re-interprets these general aspects of diversity in the application field of MO and describes related work. Section 3 introduces the proposed approach COSEA-MO with the new survival selection strategy and discusses the methods of diversity promotion on different stages of the algorithm. Section 4 presents the simulation onsets and the experimental results, which are discussed in Sect. 5.

2 Preliminary Knowledge and Related Work

Analyzing an evolutionary process regarding the term diversity, at least three levels are recognizable to promote diversity: genotype, phenotype and fitness. Genotype is the internal representation of an individual in an evolutionary process and is directly manipulated by the evolutionary operators. In the case that the genotype presentation cannot be directly evaluated by the fitness functions, a transformation into a phenotype representation is necessary. In this case, fitness distance measures are also effective measures for genotype and phenotype distance [7].

In the field of MO, individuals are usually represented as amino acids sequences and molecular functions - assuming that approximate molecular fitness functions for property prediction are available - directly work on this representation, therefore genotype and phenotype coincide. The fitness values are real numbers in the so-called chemical space. In [11], chemical space is defined as a N-dimensional Cartesian space in which molecules are mapped using chemoinformatic descriptors, which quantify physical, chemical and topological properties of molecules. The Euclidean distance is an intuitive distance measure to calculate the chemical space diversity based in the descriptor values of two molecules i and j:

$$D_{i,j} = \sqrt{\sum_{k=1}^{N} (d_{i,k} - d_{j,k})^2}$$

On genotype level, the common Hamming Distance is a straightforward metric to evaluate genetic diversity:

$$D_{i,j}^{ham} = \frac{XOR(i,j)}{N},$$

where i and j are two strings of length N, $XOR(i,j)$ is the number of positions that differ in two strings.

The locality principle states that small changes in genotype correspond to small changes in phenotype and result in small variations on fitness level. This principle is not an intrinsic character of the optimization problem, but of the genotype-phenotype-fitness-mapping. Generally, phenotype variation in multi-objective optimization causes more fitness variations because obtaining identical fitness values is less probable in higher-dimensional spaces [7].

This locality principle does not hold in molecular landscapes [12]. The reason for this is a further aspect according to the work of Sneath [9] that has to be considered: A correlation study is performed between changes of amino acids in the chemical structure of a molecule and its impact on the molecule bioactivity. The 20 canonical amino acids are considered in this work evaluating their single influence on the bioactivity of a molecule by a systematical substitution of one or more amino acids. The outcome of this work is a correlation matrix of the canonical amino acids quantifying their dissimilarity (D) or similarity (1-D) respectively to each other. The resemblance of the amino acids is obtained by comparing as many chemical properties as possible. The consequence from this work transferred to the field of MO is that diversity and dissimilarity are two complementary aspects on genotype level which have to be equally considered in MO: two peptides potentially have the same Hamming Distance value but highly differ in their dissimilarity values regarding the varying amino acids and therefore provide highly differing physiochemical fitness values.

Diversity in Evolutionary Algorithms (EAs) is usually quantified in three different ways: firstly, as a distance metric between individuals, secondly as an individual attribute reflecting how far an individual is positioned from the population (individual diversity). Thirdly, the population diversity is defined as the average individual diversity.[7]

It has to be noted that individual and population diversity in EAs usually refers to diversity on fitness level and transferred to MO, individual and population diversity is related to distances in chemical space.

In the related work [13], dissimilarity inspired by biodiversity measures has firstly been applied to address diversity in many-objective evolutionary optimization. A new diversity measure, which is an accumulation of dissimilarity in the population based on an adopted L_p-norm, enhances diversity maintenance in a many-objective evolutionary process. The diversity of a solution is determined by the sum of dissimilarity values to the remaining members of the population. Diversity performance of four popular multi-objective evolutionary algorithms has been improved on four standard benchmark problems with two to ten objectives.

In this work, COSEA-MO with the sophisticated selection strategy is compared to AnD (ANgle-based selection and shift-based Density estimation strategy) [10] in this work. To the best of the authors knowledge, AnD is currently the only available state-of-the-art algorithm that is compatible with the second and third condition mentioned in the introduction and provides a specific diversity promoting strategy within the selection. AnD selects promising individuals from the union of parent and child population for the next iteration with a diversity-first-and-convergence-second principle. In AnD, the well-known vector angle and

shift-based density estimation in the selection process are combined. Angle-based selection is used to identify two individuals with minimal angle. This is by the idea that these individuals represent the search in the same direction and waste computational resources if both individuals survive. The individual with lower shift-based density estimation is deleted in order to ensure convergence. AnD has been compared to seven state-of-the-art MaOEA on a variety of benchmark problems with 5, 10 and 15 objectives and reveals highly competitive performance. AnD is chosen for experimental comparison in this work as it the same simple framework structure like COSEA-MO (Algorithm 1), provides optimized default parameters for the non-expert use and is independent of weight vectors or reference points, which usually have a strong impact on the performance and are usually unknown in real-world applications.

3 Proposed Approach

This section describes an enhanced version of COSEA-MO to promote high genetic diversity and to ensure exponential fitness improvement within 10 iterations at the same time. The framework of COSEA-MO is given in Algorithm 1. The algorithm starts with the random initialization of the start population P_0 of size N. The individuals represent peptides encoded as character strings consisting of 20 different characters symbolizing the 20 canonical amino acids. During the evolution process, an offspring generation Q_t of size N is generated by the variation operators recombination and mutation (*RandomMatingAndVariation*). Then, P_t and Q_t are combined to a population U_t of size $2N$. Finally, a survival selection strategy (*LinearSelection*) is performed to select N individuals of U_t for the next generation P_{t+1}. An overview of diversity-preserving methods on different stages of the evolutionary process is given and the components of COSEA-MO are introduced.

Algorithm 1: Framework of COSEA-MO

Input: Population P_t, population size N, number of optimal solutions m, total number of generations T
Output: Next generation P_{t+1}
1: Random initialization of P_0;
2: **while** $t < T$ **do**
  $Q_t \leftarrow RandomMatingAndVariation(P_t)$;
  $U_t \leftarrow P_t \cup Q_t$;
  $P_{t+1} \leftarrow LinearSelection(U_t)$;
  $t \leftarrow t + 1$;
end

3.1 Diversity Strategies on Different Stages of COSEA-MO

COSEA-MO uses diversity strategies on three stages, firstly in parent selection for recombination, secondly on the stage of variation by guiding the search process with a suitable balance of exploration and exploitation on the basis of deterministic dynamic operators and thirdly by a new sophisticated survival selection strategy: Firstly, three parents are randomly selected from the population P_t for variation. The specific number of parents is motived to ensure a higher genetic diversity of the genetic material in the offspring genotype compared to the common choice of two parents. Secondly, deterministic dynamic variation operators are used for a high explorative search in early generations and a exploitative search in later generations. A linear dynamic recombination operator and an adapted version of the deterministic dynamic mutation operator of Bäck and Schütz [14] are used to generate offspring. The variation rates are adapted dynamically by predefined decreasing functions with the iteration progress: the recombination operator varies the number of recombination points by a linearly decreasing function

$$x_R(t) = \frac{l}{4} - \frac{l/4}{T} \cdot t,$$

where l is the peptide length, T the total number of the generations and t the index of the current generation. The adapted mutation operator determines the mutation probabilities via

$$p_{BS} = (a + \frac{l-2}{T-1}t)^{-1}$$

with $a = 5$. The mutation rates of the traditional operator are reduced by a higher value for a.

Thirdly, a new selection strategy is used in COSEA-MO as survival selection. A fitness value is assigned to each peptide in U_t by a linear combination consisting of a term reflecting the peptide quality, a term for genetic diversity and one for genetic dissimilarity as well as similarity respectively. Peptide quality is measured by a winning-score (WS) value for each peptide relative to the remaining members of the population. The WS method describes the difference between the number of superior and inferior objectives between two individuals: let sup_{ij} be the number of objectives in a solution i that is superior to the corresponding objectives in a solution j while inf_{ij} is the number of objectives in i that is inferior to j. The WS-value of the i-th solution in a population of size N is given by [8]:

$$WS(i) = \sum_{j=1}^{N} w_{ij} \text{ with } w_{ij} = sup_{ij} - inf_{ij}$$

Obviously, it is $w_{ij} = -wji$ and $w_{ii} = 0$. This assignment ensures that solutions with high WS-values are close to the true Pareto front.

The genetic diversity is measured by the traditional Hamming Distance (HD) relative to a predefined reference peptide. The genetic dissimilarity (D) is

calculated averaging the dissimilarity values of a peptide i to a predefined reference peptide r according to the dissimilarity matrix of Sneath

$$D(i) = \frac{1}{l} \sum_{j=1}^{l} D(i_j, r_j),$$

where i_j and r_j refers to the $j-$th amino acid position. Since the amino acids at each position of both peptides are compared, they have to be of the same length l. The values of WS, HD and D are scaled to are range of 0 to 1 ensuring an equal impact on the fitness value.

The selection procedure starts with assigning of a fitness value to each individual of U_t by the following linear combination:

$$F(i) = a \cdot WS(i) + b \cdot HD(i) + c \cdot D(i) + d \tag{1}$$

with the weights a, b, c and d (Table 1). The terms WS, HD and D have to be maximized: peptides with an average high number of superior objectives relative to other members of the population, a high genetic diversity of the material and a high average similarity in bioactivity $(1 - D(i))$ to a reference peptide at the same time are preferred. The peptides with the N-highest fitness values are selected for the next generation.

4 Experimental Studies

The performance of COSEA-MO with different selection configurations are compared to the recently published AnD on four differently dimensional MO problems and are evaluated according to the convergence behavior, diversity in chemical space and average dissimilarity. AnD has the same framework structure as COSEA-MO and the same variation operators are used for a fair comparison of the selection strategies. The different configurations of COSEA-MO are given by different selection function with various weights (Eq. 1). All experiments are implemented in the open source jMetal library 4.5. [15]. Each configuration is run 30 times on each MO problem with 10 iterations and a population size of 100.

Table 1. Applied linear selection functions in COSEA-MO

Abbr.	Weights	Selection by
V1	$a = c = 0.5$, $b = d = 0$	WS value and dissimilarity
V2	$a = 1$, $b = c = d = 0$	Only WS value
V3	$a = b = 0.5$, $c = d = 0$	WS value and Hamming Distance
V4	$a = b = d = 0.333$, $c = -0.333$	WS value, Hamming Distance and similarity

The individuals are 20-mer peptides composed of the 20 canonical amino acids. Short peptides of length 20 are of specific interest because of their favorable properties as drugs.

4.1 Physiochemical Optimization Problems

Four optimization problems (Table 2) with 3 up to 6 objective functions are applied predicting physiochemical peptide properties. The optimization problems comprise molecular properties like charge, solubility in aqueous solutions, molecule size, molecule stability and structure. The six physiochemical functions are generic in the sense that the physiochemical properties are determined by descriptor values of the amino acids in the molecule sequence and are provided by the open source BioJava library [16]. A description of the determination methods and a motivation for the physiochemical function selection is given in [6]: Needleman Wunsch Algorithm (NMW), Molecular Weight (MW), Average Hydrophilicity (Hydro), Instability Index (InstInd), Isoelectric Point (pI) and Aliphatic Index (aI). These six objective functions act comparatively to reflect the similarity of a particular peptide to a pre-defined reference peptide: $f(\text{CandidatePept.}) := |f(\text{CandidatePept.}) - f(\text{ReferencePept.})|$. Therefore, the four objective functions have to be minimized. Furthermore, the objective values are normalized by the theoretical maximal value of each objective.

Table 2. Physiochemical functions of the different optimization problems

Dimension	Abbr	Objective functions
3D	3D-MOP	NMW, MW, Hydro
4D	4D-MaOP	NMW, MW, Hydro, InstInd
5D	5D-MaOP	NMW, MW, Hydro, InstInd, pI
6D	6D-MaOP	NMW, MW, Hydro, InstInd, pI, aI

4.2 Performance Metrics

Three metrics are used to measure convergence, diversity and dissimilarity. These metrics are applied on 20% approximately optimal individuals in each iteration for all configurations. These optimal individuals are determined by WS values in all configurations. The Average Cuboid Volume (ACV) is used to measure the convergence behavior [17]. ACV calculates the averaged spanned space of each solution to an ideal reference point, which is usually known in real-world applications. The ACV indicator is given by

$$ACV = \frac{1}{n}\sum_{i=1}^{n}(\prod_{j=1}^{k}(x_{ij} - r_j)), \tag{2}$$

where n is the number of individuals that are evaluated, k the number of objectives and r_j the ideal point. The lower the ACV values, the better the

convergence behavior since the MO problems have to be minimized. ACV as a simple statistical measure is preferred over traditional convergence metrics since it is independent of Pareto optimal solution sets which are usually unknown in real-world applications, of low computation cost, independent of the problem dimension and relative to the number of solutions allowing a comparison of differently sized solution sets.

A state-of-the-art statistical evaluation method is used to evaluate the diversity performance. The diversity is determined by the standard deviation of the solution set to the gravity point of this set. Therefore, this diversity measure refers to population diversity in chemical space.

The dissimilarity is determined as average dissimilarity of the 20% candidate peptides to a pre-defined reference peptide according to the dissimilarity matrix of Sneath. This measure is a diversity measure on genotype level and a problem-specific measure to evaluate diversity of the genetic material.

4.3 Experimental Results

The performance results of the COSEA-MO configurations V1 - V4 on 3D-MOP for the three indicators are depicted in Fig. 1, 3 and 5, the results of AnD for all test problems are depicted in Fig. 13, 14 and 15. The graphs present the average performance results for 10 iterations including the start population. The overall favorable performance is given by very low ACV results with high diversity and high average dissimilarity values. Generally, the ACV performance results of V1 to V4 are remarkably close especially in the last generations. The configurations V1 to V4 reveal outstanding convergence behavior within 10 iterations by significantly lower ACV values compared to AnD, that does not provide any convergence behavior at all but has the highest diversity values in terms of population diversity in chemical space. Best convergence behavior with the lowest scattering of the ACV results is achieved by V2, WS solely selection, followed by V3. But V2 has also the lowest diversity and dissimilarity results. V4 achieves good convergence results with the best diversity and average dissimilarity values and therefore provides best overall performance.

In 4D-MaOP (Fig. 2, 4, 6), V1 to V4 achieve again outstanding convergence behavior compared to AnD that does not reveal any convergence but has the highest diversity values. V4 provides very good performance results with very good convergence and diversity values as well as high average dissimilarity results. V2 provides again fast convergence in the first generations but with the lowest diversity and dissimilarity results. V1 achieves second best overall performance.

In 5D-MaOP (Fig. 7, 9, 11), V1 to V4 reveal again very good convergence results. Here, AnD also provide a slight convergence improvement, but far from the results of V1 to V4. Diversity values are once again the highest. Best convergence results are achieved by V2 with lowest diversity and dissimilarity results. Both, V1 and V4 provide very good convergency and diversity results. High average dissimilarity results are provided by V1 followed by V3 and V4.

Similar results are observable for 6D-MaOP (Fig. 8, 10, 12): outstanding convergence behavior is achieved by V1 to V4, AnD reveals slight convergence

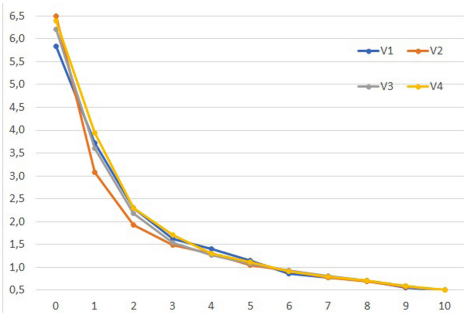

Fig. 1. 3D-MOP: ACV results

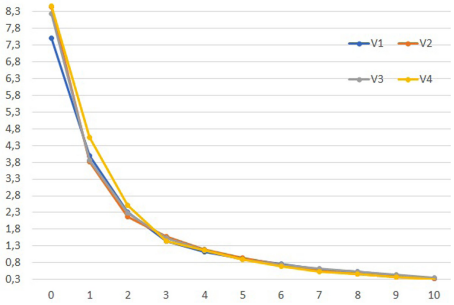

Fig. 2. 4D-MaOP: ACV results

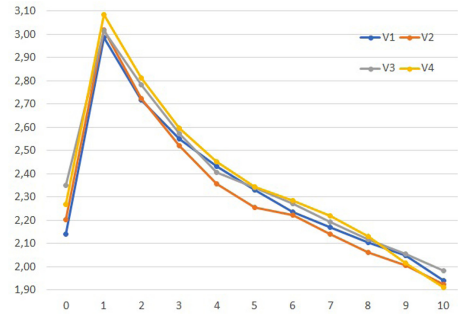

Fig. 3. 3D-MOP: diversity results

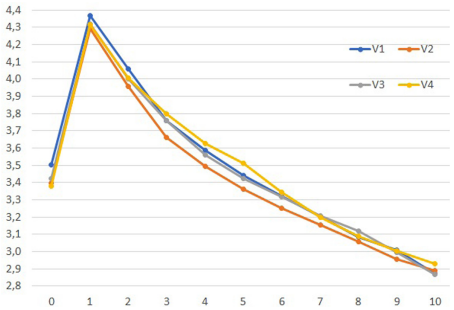

Fig. 4. 4D-MaOP: diversity results

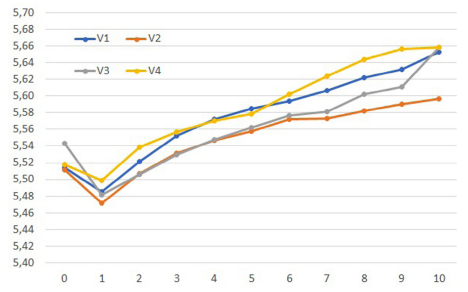

Fig. 5. 3D-MOP: dissimilarity results

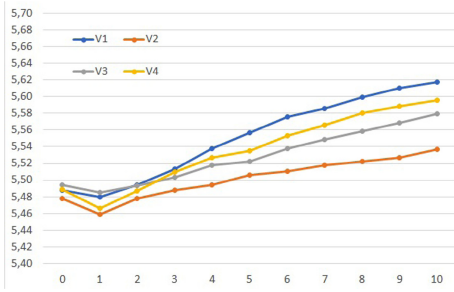

Fig. 6. 4D-MaOP: dissimilarity results

improvement with the highest diversity values. Best convergence behavior is achieved by V4 followed by V1. V4 reveals constantly good diversity and acceptable dissimilarity results.

Summarizing, AnD does only provide slight convergence behavior on 5D- and 6D-MaOP with highest diversity in chemical space. This corresponds to the diversity-first-and-convergence-second principle. It has to be noted that this principle solely act in chemical space. Moreover, AnD seems to be a real

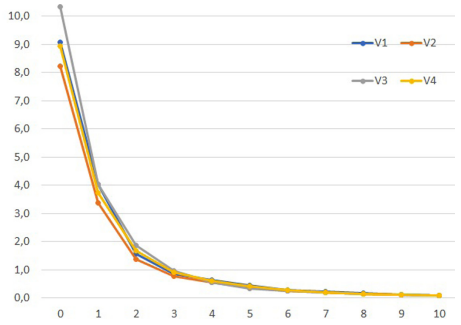

Fig. 7. 5D-MaOP: ACV results

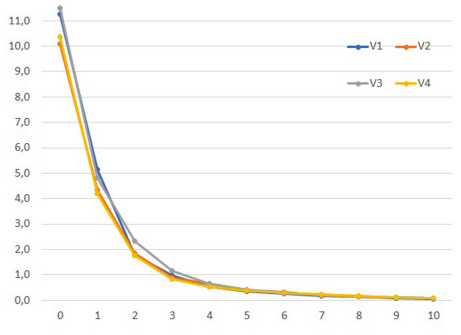

Fig. 8. 6D-MaOP: ACV results

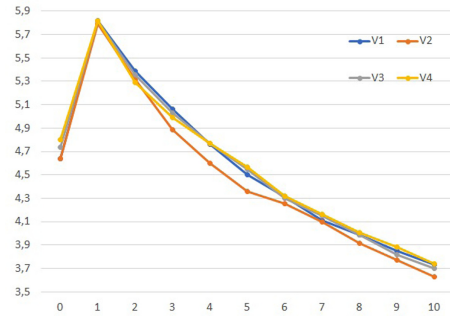

Fig. 9. 5D-MaOP: diversity results

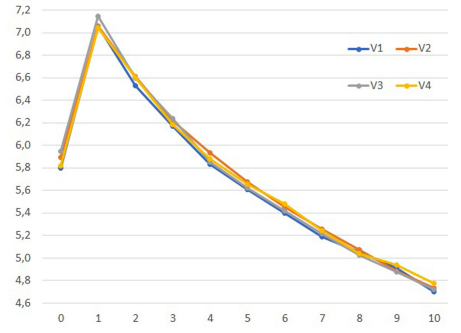

Fig. 10. 6D-MaOP: diversity results

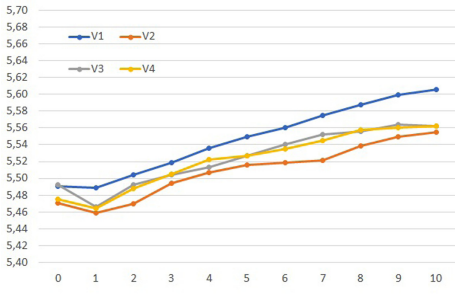

Fig. 11. 5D-MOP: dissimilarity results

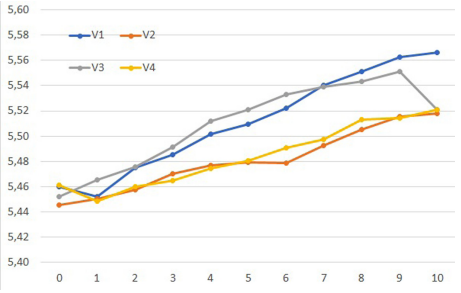

Fig. 12. 6D-MaOP: dissimilarity results

many-objective EA, since no convergence behavior is observable on 3D-MOP and 4D-MaOP. In general, V2 provides the overall best convergence performance but poor diversity and average dissimilarity results caused by the solely WS selection technique. V1 generally provides good convergence, very good average dissimilarity and generally good diversity results, which is caused by equal WS and

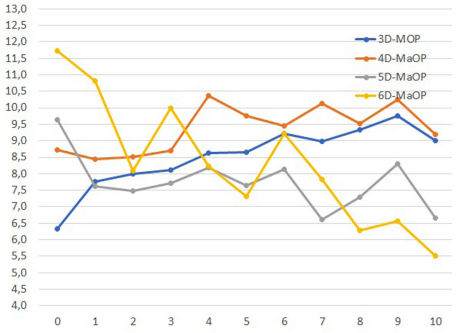

Fig. 13. AnD: ACV results

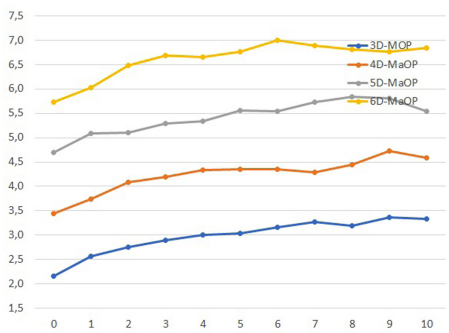

Fig. 14. AnD: diversity results

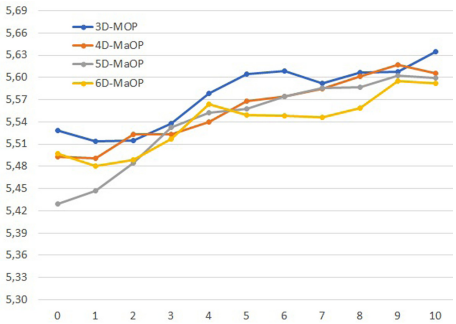

Fig. 15. AnD: dissimilarity results

dissimilarity based selection. V4 achieves good overall performances in all test cases. V4 selects individuals according to the highest WS values, high diversity in genetic material (HD) and high similarity on amino acid level. Since V3 also selects individuals based on WS values and according to high genetic diversity, HD empirically seems to be an important measure to promote diversity in MO. The configuration V4 is preferred as selection strategy due to the complementation of HD with the aspect of amino acid similarity within the selection strategy. The experimental results reveal the identification of highly qualified candidate molecules with a high diversity in genetic material and high average dissimilarity at the same time.

5 Discussion and Conclusion

The aim of MO is the identification of highly qualified candidate peptides according to the physiochemical objectives with a high diversity of the genetic material and comparable bioactivity. Diversity is addressed in EA on different levels with various methodologies and performance measures. In this work, the aspect of

diversity is re-defined and re-interpreted on different stages of a proposed EA for MO. At this point, the issues raised in the introduction have to be focussed: The first issue addresses the re-definition of diversity in MO. Diversity has to be considered on genotype and fitness level and different techniques have to be applied to control and promote diversity on these levels. Two complementary aspects define diversity on gentoype level: firstly, the diversity of genetic material measured by the number of differing amino acids between two molecule sequences and secondly, diversity in terms of amino acid dissimilarity according to Sneath. Both aspects together have an impact on variations in fitness level. Diversity in fitness level is referred to diversity in chemical space and measurable by distance metrics.

Different diversity promoting methods have been included in COSEA-MO on three stages: in parent selection for recombination, in the variation operators and in survival selection. This work also presents a sophisticated selection strategy based on the diversity considerations in MO. Individuals for the next iteration are chosen by a linear combination as selection function with the terms molecular quality, genetic diversity and dissimilarity calculated by a WS method, HD and average dissimilarity of the amino acids relative to a predefined reference peptide. The performance has been compared to the diversity-first-convergence-second selection principle and AnD on four different dimensional MO problems, where diversity refers to diversity in chemical space. The four selection configurations of COSEA-MO clearly outperform AnD in terms of convergence in all test cases which emphasizes the clear and application-specific definition of the term diversity. AnD reveals a slight convergence behavior only in the two higher-dimensional test cases. Especially the COSEA-MO selection configurations with diversity promoting strategies on genotype level provide remarkble results in all test cases.

In future work, a deeper understanding of genotype diversity and amino acid dissimilarity in MO and its impact on molecular landscapes have to be analyzed to control and improve the search behavior in evolutionary strategies. Furthermore, different methodologies have been proposed for sequence alignment and a systematic comparison regarding diversity promoting in evolutionary search prozesses will be focussed.

References

1. Zhou, S.F., Zhong, W.: Drug design and discovery: principles and applications. Molecules **22**(2), 279 (2017)
2. Röckendorf, N., Borschbach, M.: Molecular evolution of peptide ligands with custom-tailored characteristics. PLOS Comput. Biol. **8**(12) (2012). https://doi.org/10.1371/journal.pcbi.1002800
3. Rosenthal, S., Borschbach, M.: Design perspectives of an evolutionary process for multi-objective molecular optimization. In: Trautmann, H., et al. (eds.) EMO 2017. LNCS, vol. 10173, pp. 529–544. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54157-0_36

4. Rosenthal, S., El-Sourani, N., Borschbach, M.: Impact of different recombination methods in a mutation-specific MOEA for a biochemical application. In: Vanneschi, L., Bush, W.S., Giacobini, M. (eds.) EvoBIO 2013. LNCS, vol. 7833, pp. 188–199. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37189-9_17
5. Rosenthal, S., Borschbach, M.: Impact of population size, selection and multi-parent recombination within a customized NSGA-II for biochemical optimization. Int. J. Adv. Life Sci. IARIA **6**(3&4), 310–324 (2014)
6. Rosenthal, S., Borschbach, M.: A winning score-based evolutionary process for multi-and many-objective peptide optimization. IJCC **I**, 49–58 (2019)
7. Squillero, G., Tonda, A.: Divergence of character and premature convergence: a survey of methodologies for promoting diversity in evolutionary optimization. Inf. Sci. **329**, 782–799 (2015)
8. Maneeratana, K., Boonlong, K., Chaiyaratana, N.: Compressed-objective genetic algorithm. In: Runarsson, T.P., Beyer, H.-G., Burke, E., Merelo-Guervós, J.J., Whitley, L.D., Yao, X. (eds.) PPSN 2006. LNCS, vol. 4193, pp. 473–482. Springer, Heidelberg (2006). https://doi.org/10.1007/11844297_48
9. Sneath, P.: Relations between chemical structure and biological activity in peptides. J. Theor. Biol. **12**(2), 157–195 (1966)
10. Liu, Z.Z., Wang, Y., Huang, P.Q.: AnD: a many-objective evolutionary algorithm with angle-based selection and shift-based density estimation. Inf. Sci. **509**, 400–419 (2018)
11. Rupakheti, C., Virshup, A., Yang, W., Beratan, D.: Strategy to discover diverse optimal molecules in the small molecule universe. J. Chem. Inf. Model **55**(3), 529–537 (2015)
12. Emmerich, M., Lee, B., Render, A.: Analyzing molecular landscapes using random walks and information theory. Chem. Central J. **3**, P20 (2009). https://doi.org/10.1186/1752-153X-3-S1-P20
13. Wang, H., Jin, Y., Yao, X.: Diversity assessment in many-objective optimization. IEEE Trans. Cybern. **47**, 1510–1522 (2017)
14. Bäck, T., Schütz, M.: Intelligent mutation rate control in canonical genetic algorithms. In: Raś, Z.W., Michalewicz, M. (eds.) ISMIS 1996. LNCS, vol. 1079, pp. 158–167. Springer, Heidelberg (1996). https://doi.org/10.1007/3-540-61286-6_141
15. Nebro, A., Durillo, J.: jmetal: Metaheuristic Algorithms in Java (2019)
16. BioJava: CookBook, r. http://www.biojava.org/wiki/BioJava
17. Rosenthal, S., Borschbach, M.: Average cuboid volume as a convergence indicator and selection criterion for multi-objective biochemical optimization. In: Emmerich, M., Deutz, A., Schütze, O., Legrand, P., Tantar, E., Tantar, A.-A. (eds.) EVOLVE – A Bridge between Probability, Set Oriented Numerics and Evolutionary Computation VII. SCI, vol. 662, pp. 185–210. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-49325-1_9