Sunil Template

Contents

1 Multi objective evolutionary algorithms applied to Protein Structure Prediction Problem **1**
Author Name1, Author Name2, Author Name3, and Author Name4
1.1 Introduction 1
1.2 Protein folding 1
1.3 Multi-objective Optimization 3
1.3.1 Non-dominated sorting Genetic Algorithm II 4
1.3.2 IBEA (Indicator-Based Evolutionary Algorithm) . . . 4
1.4 Proposed method 6
1.5 Experiments 9
1.5.1 Results for the first approach using the MOEAs without modifications 10
1.6 Conclusion 10

Bibliography **11**

Chapter 1

Multi objective evolutionary algorithms applied to Protein Structure Prediction Problem

Author Name1

Affiliation text1

Author Name2

Affiliation text2

Author Name3

Affiliation text3

Author Name4

Affiliation text4

1.1 Introduction 1
1.2 Protein folding 1
1.3 Multi-objective Optimization 3
1.3.1 Non-dominated sorting Genetic Algorithm II 4
1.3.2 IBEA (Indicator-Based Evolutionary Algorithm) 4
1.4 Proposed method 6
1.5 Experiments 9
1.5.1 Results for the first approach using the MOEAs without modifications 10
1.6 Conclusion 10

1.1 Introduction

1.2 Protein folding

We will briefly recall some of the main biological concepts related to the protein folding problem that are relevant to our discussion.

Proteins are macromolecules made out of twenty different amino acids, also referred to as residues. An amino acid has a peptide backbone and a distinctive side chain group. The peptide bond is defined by an amino group and a carboxyl group connected to an alpha carbon to which a hydrogen and side chain group are attached.

Amino acids are combined to form sequences which are considered the primary structure of the peptides or proteins. The secondary structure is the locally ordered structure brought about via hydrogen bounding mainly within the peptide backbone. The most common secondary structure elements in proteins are the alpha helix and the beta sheet. The tertiary structure is the global folding of a single polypeptide chain.

Under specific conditions, the protein sequence folds into a unique native 3-d structure. Each possible protein fold has associated energy. The *thermodynamic hypothesis* states that the native structure of a protein is the one for which the free energy achieves the global minimum. Based on this hypothesis, many methods that search for the protein native structure define an approximation of the protein energy and use optimization algorithms that look for the protein fold that minimizes this energy. These approaches mainly differ in the type of energy approximation employed and in the characteristics of the protein modeling.

The achievement of the protein native structure is the result of the so-called protein folding process. The laws that govern protein folding are unknown. Therefore a number of ideas have emerged that try to answer this question: how do amino acid sequences specify proteins 3-d structure?

There are two main approaches to protein folding, commonly referred as the "classical" and "new" views. The "classical" view considers folding as a defined sequence of states leading from the unfolded to the native state. This sequence is called the pathway [16]. In the "new" view approach, folding is seen as the progressive organization of an ensemble of partially folded structures through which the protein passes on its way to the folded structure [14]. This approach emphasizes the idea of each state being an ensemble of rapidly inter-converting conformations. One of the main differences between both approaches is that the "new" view allows for a more heterogeneous transition state than the "classical" view, which concentrates on a single, well-defined folding pathway [1].

Figure 1.2 shows one schematic representation of the "classical" (left) and "new" (right) views of protein folding. In the figure, each possible protein configuration is represented as a circle, and arrows represent possible transitions between configurations. In both approaches, the native state (filled circle) is achieved when the energy is minimized.

SinglePointCrossover TwoPointsCrossover MultiPointCrossover (para crossovers)

BitFlipMutation LoopMoveOperator LocalMoveOperator SegmentMutationOperator OppositeMoveOperator (para mutacoes)

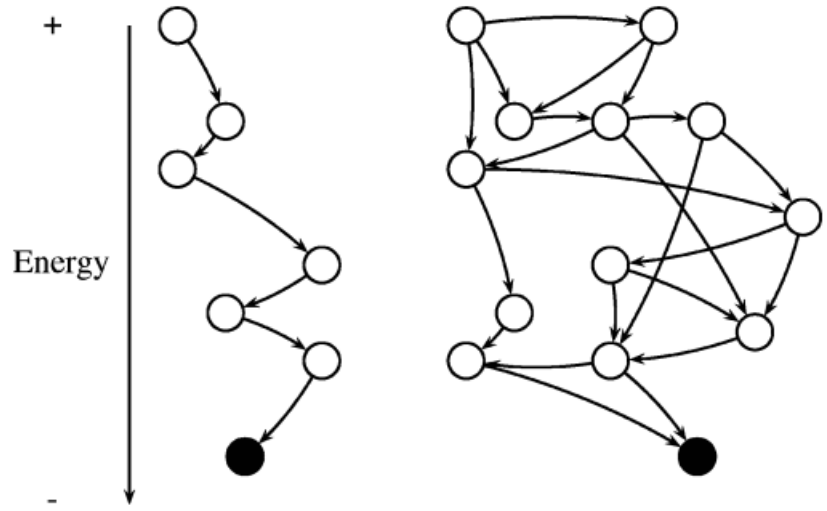

FIGURE 1.1: Schematic representation of the classical (left) and new (right) views of protein folding.

1.3 Multi-objective Optimization

Evolutionary Algorithm (EA) is a optimization and search technique, highly parallel, inspired by the Darwinian principle of natural selection and genetic reproduction. The nature principles that inspire the EAs are simple. According to the theory of C, Darwin, the principle of natural selection favors individuals with high fitness, therefore, with high probability of reproduction. Individuals with more descendants have more chance to perpetuate their genetic code in future generations. The genetic codes is what gives the identity of each individual and are represented in the chromosomes. These principles are used in the construction of computational algorithms, that searches for better solutions given a specific problem by the evolution of a population of solutions coded in artificial chromosomes – data structures used to represent a feasible solution for a given problem in the algorithm execution [15].

Real world problems commonly have multiple objectives to minimized/maximized and are present in most knowledge areas. To optimize multi objective problems, are considered two or more objectives witch usually are conflicting. To these problems is impossible to find one unique solution. A set of solutions is reached evaluating the Pareto dominance relation [3] between the solutions. The main objective is to find the solutions that are non-dominated by any other. A solution dominates other, if and only if, it was

better in at least one of the objectives, without being worst in any of the objectives. The set of non-dominated solutions constitutes the Pareto Front. Finding the the real Pareto Front is a NP-hard problem [10], this way, the objective is to find a good approximation of this front.

Multi-Objective Evolutionary Algorithms (MOEAS) are extensions of EAs to multi objective problems that applies the concepts of Pareto dominance to create different strategies to evolve and diversify the solutions. In this work were used two MOEAs: NSGAII [8] and IBEA [22].

1.3.1 Non-dominated sorting Genetic Algorithm II

The main characteristic of this algorithm is a strong elitism mechanism, classifying at each generation every solution in different fronts according with the non-dominance relation (line 15 of Algorithm 1). After the classification, solutions from the first front, are non-dominated by any other solution. Solutions from the second front are dominated only by the solutions of the first front, and so on. For solutions of the same front, the algorithm uses a Crowding Distance operator to calculate how distant are the neighbors of a given solution (line 19 of Algorithm 1). Solutions with high values of Crowding Distance have priority, because they will contribute more to the population's diversity. The binary tournament selects solutions from the small front with the higher values of Crowding Distance. A new population is generated using the crossover and mutation operators (line 25 of Algorithm 1).

1.3.2 IBEA (Indicator-Based Evolutionary Algorithm)

In the multi-objective optimization context, optimizing consists in find a front with a good approximation to the true Pareto front. However, there is no general definition about what is the true Pareto front. This way, indicators have been used to evaluate the quality of a approximation front. The *hypervolume* is a example of indicator to the evaluation and comparison of fronts.

The IBEA is an algorithm that considers the optimization by the use of quality indicators. The indicator is the way used to evaluate the non-dominated set of solutions [9]. To use the IBEA it is necessary define which indicator will be used to associate each ordered pair of solutions to a scalar value. One of the most used indicators is the *hypervolume* due to its capacity of evaluate the convergence and diversity at the same time of the search process [12].

$$F(x_i) = \sum_{x_j \in (P - x_i)} -e^{\frac{-I_{Hy}(x_j, x_i)}{k}} \tag{1.1}$$

For the IBEA fitness calculation (Equation 1.1), k is a parameter commonly used with a value of 0.05. The value for $F(x_i)$ corresponds to a quality

Algorithm 1 NSGAII

```
1: N ← Population Size
2: T ← Max evaluations
3: P_0 ← CreatePopulation(N);
4: CalculateFitness(P_0);
5: FastNonDominatedSort(P_0);
6: Q_0 ← 0
7: while Q_0 < N do
8:     Parents ← BinaryTournament(P_0);
9:     Children ← CrossoverMutation(Parents);
10:    Q_0 ← Children
11: end while
12: CalculateFitness(Q_0);
13: t ← 0
14: while t < T do
15:    R_t ← P_t ∪ Q_t;
16:    Fronts ← FastNonDominatedSort(R_t);
17:    P_{t+1} ← 0
18:    i ← 0
19:    while P_{t+1} + Front_i < N do
20:        CrowdingDistanceAssignment(Front_i);
21:        P_{t+1} ← P_{t+1} ∪ Front_i
22:        i ← i + 1
23:    end while
24:    CrowdingDistanceSort(Front_i);
25:    P_{t+1} ← P_{t+1} ∪ Front_i[1 : (N − P_{t+1})]
26:    Parents ← BinaryTournament(P_{t+1});
27:    Q_{t+1} ← CrossoverMutation(Parents);
28:    t ← t + 1
29: end while
30: return P ← Set of non-dominated solutions.
```

loss measure of the approximation to the Pareto front if the solution x_i was removed of the population [9], based on the value of the quality indicator I_{Hy}, in this case, the *hypervolume*. Based on the fitness calculation described above, the basic IBEA algorithm consists in iteratively do the selection (line 10 of Algorithm 2), crossover, mutation (line 11 of Algorithm 2) and environment selection, removing the worst individual from the population and updating the values of fitness of the remaining individuals (lines 4 to 8 of Algorithm 2).

[1] *Hypervolume*: Proposed quality indicators used in the study of [23], denoted as the "size of the covered search space". This indicator has two important advantages in relation to others [21]: 1 - Sensitive to any kind of improvement in the approximation set in relation to other set. 2 - As result of 1, the indicator guarantee that for any approximation set A that has high values of hypervolume, also has all the solutions of the true Pareto front.

Algorithm 2 IBEA

```
1: N ← Population Size
2: T ← Max Evaluations
3: k ← Scale factor of Fitness
4: P ← CreatePopulation(N);
5: m ← 0
6: CalculateFitness(P);
7: while m ≥ T or other stop criterion is reached do
8:     P̄ ← BinaryTournament(P);
9:     P ← CrossoverMutation(P̄);
10:    m ← m + 1
11:    while Size(P) > N do
12:        x* ← WorstIndividualByFitness();
13:        RemoveFromPopulation(x*, P);
14:        CalculateFitness(P);
15:    end while
16: end while
17: return P ← Set of non-dominated solutions
```

1.4 Proposed method

Two multi-objective approaches were designed in this chapter, using the MOEAs (NSGAII and IBEA) described on subsection 1.3. The relative representation was chosen to represent the chromosomes. Integer vectors are used whereas the genes specifies which direction, relative to the previous residue, should be placed the next residue. The genes can assume only tree values (0,1,2), 0 indicates that next residue should be placed at right from the previous one, 1 indicates that the next residue should be placed at left from the previous and 2 indicates that the next residue should be placed in front of the previous one. Figure 1.2 shows an example of a hypothetical chromosome and the path generated by it in the 2D lattice.

The first approach consisted on applying two well-known state of art MOEAs (IBEA and NSGAII) to the PSP using the HP-2D model. The genetic operators used by IBEA and NSGAII algorithms, in this approach, were only the single point crossover in the case of crossover and bit flip mutation for mutation operator. It was decided to use only single point crossover and bit flip mutation because this combination of operators presented better results in previous experiments within the PSP problem and the HP-2D model.

In the case of the second approach the IBEA and NSGAII algorithms were modified in order to improve their results when compared with the first approach. The modifications implemented will be described next:

- **Pool of operators**: As mentioned in section X, the use of traditional

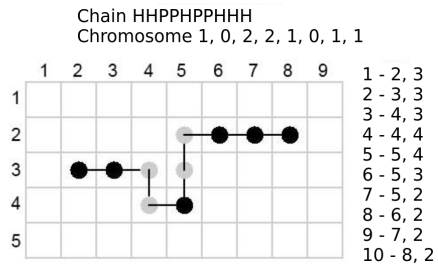

FIGURE 1.2: Example of a conformation generated by a chromosome with relative representation.

operators usually does not guide the search to prominent regions of the search space of the HP-2D model. In order to supplement the MOEAs, a pool of operators were designed based on the literature and are presented on table X. For every mating the crossover and mutation operators are selected randomly from the pool of operators and than applied. Also the crossover and mutation operators are always applied differently from the first approach which uses a probability of occurrence. The pool of operators will be described next:

- Single Point Crossover (1x): A single point on both parent individuals is selected. All data beyond that point in either individual will be swapped between the two parent individuals. Resulting in two distinct child individuals [11].
- Two Points Crossover (2x): Two points are selected on both parent individuals. Everything between the two points is swapped between the parents. Building two new distinct individuals [11].
- Multi Points Crossover (MPX): The MPX operator is similar to 2X, but the number of points, c, is a function of the sequence length, n, given by $c = int(n \times 0.1)$ [6]. The MPX operator is usually used to promote structural diversity by performing a random shuffle between individuals, although not as uniform crossover.
- Bit Flip Mutation (BFM): The BFM operator selects one random gene from a parent individual and changes it to other value. Resulting in one new individual [11].
- Loop Move Mutation (LMM): The LMM operator swaps the directions between two randomly chosen consecutive genes. This operator introduces a corner movement [4]. Figure 1.4 presents a example of application of this operator.
- Loop Move Mutation (LoMM): This operator is similar with LMM however exchanges directions between genes that are five positions apart on the sequence creating a loop movement. Both LMM and

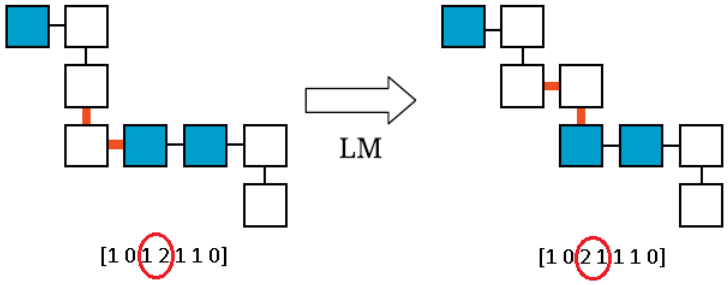

FIGURE 1.3: Example of application of the LMM operator. The genes from the red circle of leftmost figure were swapped resulting in the rightmost figure.

LoMM are useful to generate modifications on compact structures [7].

- Segment Mutation (SM): This operator changes a random number of consecutive genes (from two to seven) into new random directions. This operator introduces large conformational changes and has a high probability of creating collisions, in order to avoid too much invalid solutions the repair mechanism is applied on the generated child [7].
- Opposite Mutation (OM): This operator changes a random number of consecutive genes to its inverse position. In the case of the relative representation to the HP-2D model, only left and right can be mapped to its inverse.

- **Backtrack Initialization**: Traditionally the initial population of solutions are generated randomly in the presented MOEAs. This have a great potential of generating a large number of invalid solutions for the PSP problem within the HP-2D model. Solutions that are not self-avoiding walk (SAW) are said to contain collisions. If the initial population is fully generated randomly the evolutionary algorithms will spend time processing invalid solutions until getting good results. In order to subdue this problem a backtrack strategy should be applied. In this approach 20 percent of the initial population were generated using the backtrack initialization.

For both approaches the same objectives were used and will be explained at next:

- **Energy value**: This is the main objective and consist in the energy of given protein conformation. The goal is to minimize the energy value and it is calculated as described in section X. This objective guides the

search progress towards regions that the energy associate to the protein conformations are minimal. Thus, achieving protein conformations which are closest to native structure of a protein.

- **Distance between the two farthest residues**: This is a secondary objective and it was inspired by the related work X. The motivation for this objective is that more compact conformations tend to have more hydrophobic contacts which means a lesser energy value. The distance between two residues is calculated using the Euclidean distance.

The relative representation it is subject to generation of invalid solutions using the HP-2D model. A solution is considered invalid when the solution does not perform a self-avoiding walk (SAW) as mentioned before. In other words a invalid solution is when a given residue collides with another already placed on the lattice. A simple mechanism for repairing these situations was developed, the code can be seen in the Algorithm 3.

Algorithm 3 Mechanism to repair infeasible solutions

1. Obtains the direction that next residue should be placed.
2. Verifies if this direction will cause collision.
3. If the a collision is detected, a new direction is used.
4. Repeat the step 2 and 3, until be possible to place the next residue, or if all directions were tested and cause collisions.
5. If was possible to place the next residue, the mechanism achieved success, if not, the solution is considered infeasible and it will be penalized in the evaluation process.

This mechanism was implemented because in previous experiments was observed that the number of infeasible solutions was too big. It is necessary mention that the even with the mechanism to repair solutions, there are still infeasible solutions because the mechanism can not always repair. Thus, infeasible solutions are penalized by subtracting the number of collisions to the quantity of topological neighbors. This mechanism is used by the evaluation process of both approaches described before (IBEA and NSGAII without any major modifications and the same algorithms with the modifications mentioned).

To evaluate and compare the performance of multi-objective algorithms, quality indicators are commonly used. In this study was used the hypervolume indicator, which considers the volume of the search space dominated by the known pareto front [24] of an algorithm. Higher hypervolume value means that the quality is better than one lesser hypervolume value.

1.5 Experiments

This section presents the set of experiments designed to evaluate the performance of the approaches introduced in section 1.4. The HP sequences used in the experiments are shown in table 1.1, those instances have been used in previously works such as [2, 18, 20, 5, 17, 19, 13]. The values presented in table 1.1 correspond to the sequence identifier, the size of aminoacid sequence, the best known solutions ($H(x*)$) for the HP-2D model and the sequence itself.

TABLE 1.1: HP instances used in the experiments. The search space of each instance is 2^n where n is the size of the instance.

inst.	size	$H(\mathbf{x}^*)$	sequence
$s1$	20	-9	$HPHPPHHPHHPHPHHPPHPH$
$s2$	24	-9	$HHPPHPPHPPHPPHPPHPPHPPHH$
$s3$	25	-8	$PPHPPHHP^4HHP^4HHP^4HH$
$s4$	36	-14	$P^3HHPPHHP^5H^7PPHHP^4HHPPHPP$
$s5$	48	-23	$PPHPPHHPPHHP^5H^{10}P^6$ $HHPPHHPPHPPH^5$
$s6$	50	-21	$HHPHPHPHPH^4PHP^3HP^3HP^4$ $HP^3HP^3HPH^4\{PH\}^4H$
$s7$	60	-36	$PPH^3PH^8P^3H^{10}PHP^3$ $H^{12}P^4H^6PHHPHP$
$s8$	64	-42	$H^{12}PHPH\{PPHH\}^2PPH\{PPHH\}^2$ $PPH\{PPHH\}^2PPHPHPH^{12}$
$s9$	85	-53	$H^4P^4H^{12}P^6H^{12}P^3H^{12}P^3$ $H^{12}P^3HP^2H^2P^2H^2P^2HPH$
$s10$	100	-48	$P^6HPH^2P^5H^3PH^5PH^2P^4H^2$ $P^2H^2PH^5PH^{10}PH^2PH^7$ $P^{11}H^7P^2HPH^3P^6HPH$
$s11$	100	-50	$P^3H^2P^2H^4P^2H^3PH^2PH^2PH^4$ $P^8H^6P^2H^6P^9HPH^2PH^{11}P^2$ $H^3PH^2PHP^2HPH^3P^6H^3$

1.5.1 Results for the first approach using the MOEAs without modifications

The first experiment consists on executing the MOEAs (NSGAII and IBEA) without any modifications on each sequence.

1.6 Conclusion

Bibliography

[1] David Baker. A surprising simplicity to protein folding. *Nature*, 405:39–42, 2000.

[2] Ugo Bastolla, Helge Frauenkron, Erwin Gerstner, Peter Grassberger, and Walter Nadler. Testing a new monte carlo algorithm for protein folding. *arXiv preprint cond-mat/9710030*, 1997.

[3] Henri Joseph Léon Baudrillart. *Manuel d'économie politique*. Guillaumin et cie, 1872.

[4] Andrea Bazzoli and Andrea GB Tettamanzi. A memetic algorithm for protein structure prediction in a 3d-lattice hp model. In *Applications of Evolutionary Computing*, pages 1–10. Springer, 2004.

[5] Carlos Cotta. Protein structure prediction using evolutionary algorithms hybridized with backtracking. In *Artificial Neural Nets Problem Solving Methods*, pages 321–328. Springer, 2003.

[6] Fábio L Custódio, Hélio JC Barbosa, and Laurent E Dardenne. Investigation of the three-dimensional lattice hp protein folding model using a genetic algorithm. *Genetics and Molecular Biology*, 27(4):611–615, 2004.

[7] Fábio Lima Custódio, Helio JC Barbosa, and Laurent Emmanuel Dardenne. A multiple minima genetic algorithm for protein structure prediction. *Applied Soft Computing*, 15:88–99, 2014.

[8] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *Evolutionary Computation, IEEE Transactions on*, 6(2):182–197, 2002.

[9] Elliackin Messias do Nascimento Figueiredo, Teresa Bernarda Orientadora Ludermir, and Carmelo José Albanez Coorientador Bastos Filho. Algoritmo baseado em enxame de partículas para otimização de problemas com muitos objetivos. 2013.

[10] Carlos M Fonseca, Joshua D Knowles, Lothar Thiele, and Eckart Zitzler. A tutorial on the performance assessment of stochastic multiobjective optimizers. In *Third International Conference on Evolutionary Multi-Criterion Optimization (EMO 2005)*, volume 216, page 240, 2005.

[11] John H Holland. Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence. 1975.

[12] Hisao Ishibuchi, Noritaka Tsukamoto, and Yusuke Nojima. Evolutionary many-objective optimization. In *Genetic and Evolving Systems, 2008. GEFS 2008. 3rd International Workshop on*, pages 47–52. IEEE, 2008.

[13] Neal Lesh, Michael Mitzenmacher, and Sue Whitesides. A complete and effective move set for simplified protein folding. In *Proceedings of the seventh annual international conference on Research in computational molecular biology*, pages 188–195. ACM, 2003.

[14] J. N. Onuchic and P. G. Wolynes. Theory of protein folding. *Current Opinion in Structural Biology*, 14:70–75, 2004.

[15] Marco Aurélio Cavalcanti Pacheco. Algoritmos genéticos: princípios e aplicações. *ICA: Laboratório de Inteligência Computacional Aplicada. Departamento de Engenharia Elétrica. Pontifícia Universidade Católica do Rio de Janeiro. Fonte desconhecida*, 1999.

[16] V. S. Pande, A. Y. Grosberg, T. Tanaka, and D. S. Rokhsar. Protein folding pathways: Is a 'new view' needed? *Current Opinion in Structural Biology*, 8(1):68–79, 1998.

[17] Roberto Santana, Pedro Larranaga, and José A Lozano. Protein folding in 2-dimensional lattices with estimation of distribution algorithms. In *Biological and Medical Data Analysis*, pages 388–398. Springer, 2004.

[18] Alena Shmygelska, Rosalia Aguirre-Hernandez, and Holger H Hoos. An ant colony optimization algorithm for the 2d hp protein folding problem. In *Ant Algorithms*, pages 40–52. Springer, 2002.

[19] Alena Shmygelska and Holger H Hoos. An improved ant colony optimisation algorithm for the 2d hp protein folding problem. In *Advances in Artificial Intelligence*, pages 400–417. Springer, 2003.

[20] Ron Unger and John Moult. Genetic algorithms for protein folding simulations. *Journal of molecular biology*, 231(1):75–81, 1993.

[21] Eckart Zitzler, Dimo Brockhoff, and Lothar Thiele. The hypervolume indicator revisited: On the design of pareto-compliant indicators via weighted integration. In *Evolutionary multi-criterion optimization*, pages 862–876. Springer, 2007.

[22] Eckart Zitzler and Simon Künzli. Indicator-based selection in multiobjective search. In *Parallel Problem Solving from Nature-PPSN VIII*, pages 832–842. Springer, 2004.

[23] Eckart Zitzler and Lothar Thiele. Multiobjective optimization using evolutionary algorithmsa comparative case study. In *Parallel problem solving from naturePPSN V*, pages 292–301. Springer, 1998.

[24] Eckart Zitzler, Lothar Thiele, Marco Laumanns, Carlos M Fonseca, and Viviane Grunert Da Fonseca. Performance assessment of multiobjective optimizers: an analysis and review. *Evolutionary Computation, IEEE Transactions on*, 7(2):117–132, 2003.