A Survey of Applications of Multi-Objective Evolutionary Algorithms in Biotechnology

1st Carlos Felipe Coello Castillo
Posgrado en Ciencias Naturales e Ingeniería
Universidad Autónoma Metropolitana, Unidad Cuajimalpa
Av. Vasco de Quiroga 4871, Col. Santa Fe Cuajimalpa
Delegació Cuajimalpa, Ciudad de México, 05348, MEXICO
carlos.coello@cua.uam.mx

2nd Carlos A. Coello Coello
Departamento de Computación
CINVESTAV-IPN (Evolutionary Computation Group)
Av. IPN No. 2508, Col. San Pedro Zacatenco
Mexico City, Mexico
carlos.coellocoello@cinvestav.mx

***Abstract*—This paper presents a survey of applications of multi-objective evolutionary algorithms in several biotechnology areas. The application areas covered in the survey include: molecular docking, metabolic engineering, synthetic biology, optimization of industrial bio-processes and data processing for bioinformatics (which covers multiple sequence alignment and feature selection and classification for diagnosis of diseases). In the final part of the paper, some potential areas for future research are briefly discussed.**

***Index Terms*—Multi-objective optimization, biotechnology, multi-objective evolutionary algorithms, applications.**

I. Introduction

Multi-objective optimization problems are very common in real-world applications, since many of them have two or more (often conflicting) objectives that we aim to optimize at the same time [1]. Multi-objective evolutionary algorithms originated in the 1980s [2] and in the last 25 years, they have become a frequent tool to tackle complex and challenging multi-objective optimization problems in a variety of domains (see for example [3], [4]).

Biotechnology is a discipline that has become increasingly popular due to its many applications in our everyday life that go from the production of bactericides to the generation of vaccines. Although for several years biology has been an area with a limited number of applications of multi-objective evolutionary algorithms (see for example [5]), in recent years, an increasing number of applications have been reported in biotechnology. In this paper, we provide a survey of applications of multi-objective evolutionary algorithms in biotechnology, including a proposed taxonomy as well as some possible areas of future research in the area.

II. Multi-Objective Evolutionary Algorithms

Currently, we have three main families of multi-objective evolutionary algorithms (MOEAs):

1) **Pareto-based MOEAs:** In these approaches, a procedure called *nondominated sorting* or *Pareto ranking* is adopted to rank solutions. The core idea is to identify all the Pareto optimal solutions in the population of a MOEA and to give them the same probability of being selected. Clearly, this selection probability will be higher than the one corresponding to dominated solutions. Additionally, these approaches normally have another mechanism called *density estimator* which aims to prevent convergence to a single solution. These approaches originated in the mid-1990s and were very popular for several years. The main inconvenience of these approaches is their scalability limitation in objective function space. Representative approaches from this family are: the Nondominated Sorting Genetic Algorithm-II [6] (NSGA-II) and the Strength Pareto Evolutionary Algorithm 2 [7] (SPEA2).
2) **Indicator-based MOEAs:** They adopt a performance indicator to select solutions instead of Pareto optimality. The popularity of these MOEAs has increased in recent years, mainly because of their robustness. Their main inconvenience is that the most commonly adopted performance indicator for these approaches is the hypervolume, which has a very high computational cost for problems having more than six objectives [8]. Representative approaches of this family are: the Indicator-Based Evolutionary Algorithm (IBEA) [9] and the S Metric Selection Evolutionary Multiobjective Algorithm [10] (SMS-EMOA).
3) **Decomposition-based MOEAs:** These approaches transform a multi-objective optimization problem into several single-objective optimization problems which are then solved simultaneously to generate the nondominated solutions of the original problem. The use of decomposition methods relies on a scalarizing function, but they work properly even if the Pareto front is nonconvex or disconnected. When using decomposition-based MOEAs, neighborhood search is used to solve simultaneously all the single-objective optimization problems generated from the transformation. The main known limitation of this sort of approach is that scalarizing functions assume that the Pareto front fits within a simplex and, if that is not the case, then no reliable approximation can be produced. The most representative approach of this family is: the Multi-Objective Evolutionary Algorithm based on Decomposition (MOEA/D) [11].

III. A Taxonomy of Applications

Based on the current applications of MOEAs reported in the specialized literature, we propose the following taxonomy:

- Molecular docking
- Metabolic engineering
- Synthetic biology
- Optimization of industrial bio-processes
- Data processing for bioinformatics

In the following subsections, we will review some representative publications corresponding to each of the categories from the above taxonomy. We will also add a brief description of each topic, since this may be necessary for readers who are not familiar with biotechnology.

A. Molecular Docking

Molecular docking consists in locating an appropriate position and orientation for docking a small molecule (ligand) to a larger receptor molecule and it plays a crucial role in the computer-aided design of drugs. The goal of molecular docking is to find an optimized conformation between the ligand (L) and the receptor (R) that results in a minimum binding energy. The high complexity of this optimization problem has motivated the use of metaheuristics to solve it. In fact, the problem can be posed both as a single-objective optimization problems or as a multi-objective optimization problem. If treated as a single-objective problem, the objective to optimize is usually the final binding energy.

Grosdidier et al. [12] were apparently the first to propose the use of two objectives: the first uses the sum of intermolecular and intramolecular energy, neglecting solvent effects and is used to drive the search towards local minima and it was adopted before its efficiency and speed. The minima obtained with this first objective are then exposed to a second, more selective (and computationally expensive) objective, which includes the solvation free energy. So, the authors adopt a lexicographic ordering approach [1] since the objectives are considered in a sequential manner and not simultaneosuly. The authors also stored (in some sort of tabu list) the previously visited unfavorable docking poses, so that they are not revisited. Additionally, they performed the sampling with operators that combine a global and a local search of the conformational space. Some of these operators are semi-stochastic and deal with rotations and translations. Other operators, which are called "smart operators", aim to cross energy barriers by traversing the search space in a deterministic way. This approach was validated using 37 crystallized protein-ligand complexes featuring 11 different proteins. No details about the evolutionary algorithm adopted were provided by the authors.

Janson et al. [13] proposed an approach called clustMPSO, which is based on a multi-objective particle swarm optimizer (MOPSO). This approach simultaneously optimizes the intermolecular and the intramolecular energy. For evaluating the energy, the authors use the binding free energy function which is provided in Autodock 3.5. This approach was able to provide a more diverse set of possible docking conformations than the Lamarckian Genetic Algorithm combined with simulated annealing that is incorporated in Autodock 3.5. They also proposed an approach for the prediction of a docking trajectory.

These same objectives (i.e., the intermolecular and the intramolecular energy) are considered by Garcia-Godoy et al. [14] who present a comparative study in which they included: two variants of the NSGA-II, the speed-constrained multi-objective particle swarm optimizer (SMPSO) [15], the third version of generalized differential evolution (GDE3) [16], MOEA/D and SMS-EMOA. Results are also compared with respect to the Lamarckian Genetic Algorithm combined with simulated annealing that is incorporated in Autodock 3.5. SMPSO had the best overall performance in terms of both convergence and diversity.

Oduguwa et al. [17] adopted three objectives: (1) the intramolecular energy, (2) the intermolecular energy and (3) the shape of the macromolecule. The performance of three MOEAs was compared: NSGA-II, the Pareto Archived Evolution Strategy (PAES) [18] and the Strength Pareto Evolutionary Algorithm (SPEA) [19]. Three complexes from the protein Data Bank were adopted to validate the results. The authors reported that PAES had the best overall performance.

Boisson et al. [20] adopted two criteria: (1) the intermolecular energy and (2) a stability criterion which is based on an entropy calculus on a ligand/receptor complex. This calculation requires a sampling of neighbor complexes. The authors adopted the platform called Multi-Objective Evolving Objects [21] (MOEO) and took two MOEAs included there: NSGA-II and IBEA. Results were compared in terms of the Root Mean Square Deviation (RMSD). Domain-specific mutation operators were adopted by the authors, and a parallel implementation provided by the platform adopted was properly exploited. The authors reported that IBEA obtained better results than NSGA-II.

Sandoval-Perez et al. [22] adopted NSGA-II and considered two objectives: (1) the energy contributions from the covalent bonds between the atoms (bonding terms), and (2) an objective related to the molecular interactions where a covalent bond does not occur (e.g., electrostatic attractions, repulsion forces and van der Waals forces). The authors incorporated a method, based on angles [23], to select a single solution from the "knee" of the Pareto front. The authors argued that the use of a MOEA allowed them to identify molecular complexes with 3D structures relatively close to the ones reported for the analyzed structures. Also, their proposed approach was able to find good complexes when the ligand had a high number of rotatable bonds, which they reported as remarkable, since other available methods had problems in those situations and this approach even outperformed the single-objective solutions generated by Autodock 3.5.

Gu et al. [24] considered three scoring functions: (1) the force-field-based, (2) the empirical-based and (3) the knowledge-based. However, all of them were combined in a linear aggregating function which was optimized by a (single-objective) genetic algorithm. A comprehensive evaluation of the proposed approach (called MoDock) showed the benefits of using a multi-objective strategy, since it produced nearly 70% good docking solutions.

López-Camacho et al. [25] considered two objectives: (1) the Root Mean Square Deviation (RMSD) difference in the coordinates of ligands and (2) the intermolecular energy. The performance of four MOEAs was compared: NSGA-II, SMPSO, GDE3 and MOEA/D. The authors reported that SMPSO had the best overall performance.

García Nieto et al. [26] considered two objectives: (1) minimize the intermolecular energy and (2) minimize the Root Mean Square Deviation (RMSD) between the atom coordinates of the co-crystallized and the predicted ligand conformations. The authors adopted several types of MOPSOs that used different archiving and leader selection strategies. These approaches were compared using 75 molecular instances from the Protein Data Bank database (PDB) characterized by different sizes of HIV-protease inhibitors. Their results indicated that SMPSOhv [27] and MPSO/D [28] showed the best overall performance.

García-Godoy et al. [29] provided not only a comprehensive review on the use of MOEAs in molecular docking, but also presented a study in which three objectives were considered: (1) the intermolecular energy, (2) the intramolecular energy and (3) the RSMD. Three MOEAs were compared: SMPSO, MOEA/D and MPSO/D. The authors reported that SMPSO had the best overall performance in terms of both convergence and diversity.

Recent work on molecular docking has focused more on single-objective formulations of the problem, but using hybrids between evolutionary algorithms and mathematical programming techniques (see for example [30] in which the proposed approach tackles two important problems: (1) the disruptive effect of crossover operators and (2) the high dimensionality of the docking problem when the chosen ligand has many rotatable bonds).

It would be interesting to incorporate geometric constraints to avoid ignoring possible good locations. These constraints would allow the exploration of cavities that are not accessible to the solvent or whose shape can produce collisions between the ligand and the receptor [22]. Also, the use of machine learning techniques (adopting multi-objective versions of the problem) for drug discovery is another promising research area [31]. High scalability and accuracy of the optimization process are also issues that deserve further research.

B. Metabolic engineering

Metabolic engineering refers to the intentional modification of a cellular metabolism for the production of desired compounds. With the advent of recombinant DNA technology, a variety of organisms can be manipulated (e.g., bacteria, fungi, pland and animal cells) to produce several important industrial products such as amino acids, biofuels, polymers and recombinant proteins among many others.

Maia et al. [32] used SPEA2 and NSGA-II for the *in silico* multi-objective optimization of mutant strains. The objectives considered were: (1) maximizing the production of some compound and (2) maximizing the biomass (i.e., keeping the organism viable). The authors used as a case study the production of succinic acid, adopting *E. coli*. Results were compared with respect to the use of a single-objective formulation of the problem in which evolutionary algorithms and simulated annealing were adopted as search engines. The main advantage of using MOEAs in this case was that the authors were able to produce different trade-offs between a desired compound production and the viability of the strain measured by a biomass flux.

Patané et al. [33] proposed a multi-objective formulation of the metabolic engineering problem in which the production of one or more metabolites of interest and the production of biomass are the objectives to be optimized. They used NSGA-II combined with *global sensitivity analysis* and *robustness analysis*. The authors compared their results with respect to those obtained by other (single-objective) metaheuristics in two problems: 1) the overproduction of 1,4-butanediol in *Escherichia coli* and 2) the overproduction of fatty acid using the Redirector framework for *enzymes-up and down-regulation*. The results reported by the authors indicated that the NSGA-II was able to obtain more efficient designs (in terms of metabolite of interest production) at a lower cost than when using other metaheuristics. Additionally, the authors indicated that the most valuable outcome of their study was the set of Pareto optimal solutions produced, which allows to explore the different trade-offs between synthetic and biological objectives, which is something essential for industrial purposes.

Patané et al. [34] proposed the multi-objective metabolic engineering (MOME) algorithm, which is based on the use of the NSGA-II. MOME is tailored for the analysis of metabolic networks and is applied to the problem associated with the overproduction of ethanol in *flux balance analysis* (FBA) models of: (1) *S. aureus*, (2) *S. enterica*, (3) *Y. pestis*, (4) *S. cerevisiae*, (5) *C. reinhardtii*, and (6) *Y. lipolytica*. The use of a MOEA in this problem allowed the authors the explore the trade-offs between the production rate of ethanol and the modelled organism biological objective. The use of a MOEA allowed the authors to identify sets of key genetic manipulations which lead to strains overproducing ethanol (the overproduction is of more than 830%) with a sensible growth (as predicted by the FBA model). This reduces the knock-out cost and allows for an increased biomass production that would allow a sustained industrial process. The authors also used clustering to map the relationship between phenotype and genotype, with the aim of identifying patterns on knocked-out genes from the Pareto optimal strains. Additionally, they analyzed information on the essential genes and other constraints on the growth rate and the external simulated rich media, so that more realistic scenarios could be simulated. This led to a maximum increase in the ethanol production of around 195%.

Fan et al. [35] used multi-objective differential evolution to improve the accuracy of Genome-Scale Metabolic Model (GSMM) simulations of cell metabolism results. The objectives considered were: maximum specific growth rate, minimum ATP production, minimum NADH production and minimum NADPH production. Their study focused on the simulation of *Aspergillus niger*, which is an important filamentous fungi that is widely used in the production of organic acids and enzymes because of its good ability to express and secrete proteins. Industrial enzymes produced by *A. niger* have played a major role in industries such as brewing and fermentation. The authors adopted a simple method (based on locating the "knee" of the Pareto front) to select a single solution. The results reported produced a significant improvement in the accuracy of simulating *Aspergillus niger*.

Several authors have developed novel multi-objective models that have been solved using mathematical programming techniques (e.g., the ϵ-constraint method) and that could clearly be solved using MOEAs (see for example [36] in which the authors proposed a model with four objectives for cancer metabolism: (1) maximization of biomass synthesis, (2) maximization of ATP production, (3) minimization of total abundance of metabolic enzymes and (4) minimization of total carbon uptake). Other authors have considered dynamic multi-objective optimization models, for example, to identify the combination of targets (i.e., enzymatic modifications) and to determine the degree of up- or down-regulation that must be performed on them (see for example [37] in which a large-scale metabolic model of Chinese Hamster Ovary cells is used for antibody production in a fed-batch process).

C. Synthetic biology

Synthetic biology refers to the re-design of organisms for useful purposes by engineering them to have new (useful) abilities. Researchers from this area aim to harness the power of nature to solve a wide variety of problems in medicine, manufacturing and agriculture.

Boada et al. [38] proposed a multi-objective optimization tuning framework to obtain a set of model-based guidelines for the selection of the kinetic parameters required to build a biological device (for which a certain behavior is expected). The authors applied this methodology to the design of a genetic incoherent feed-forward circuit showing adaptive behavior. Two objectives were considered: (1) sensitivity, which is defined as the ratio between the absolute total variation of the output signal and the variation of the input signal, and (2) precision, which is defined as the inverse of the normalized output error. The search engine adopted is called spMODE, and is based on the use of differential evolution. Additionally, they filtered out the nondominated solutions obtained in order to select a *robust* configuration. The authors indicated that their proposed multi-objective framework was able to generate effective guidelines to tune biological parameters so as to achieve a certain (desired) circuit behavior.

González Sánchez et al. [39] used the multi-objective shuffle frog leaping algorithm (MOSFLA) to maximize the expression levels of proteins. MOSFLA combines parallel searches, multiple operators and memetic strategies. In this case, a solution represents an encoded protein. Three objectives were considered: (1) maximize the minimum Codon Adaptation Index value of a solution (the aim is to use the codons with the highest fequency values), (2) find the pair of CDSs that contain more identical subsequences (same subsequences in the same positions), by maximizing the minimum Hamming distance between two CDSs and (3) decrease the length of repeated or common substrings occurring between a pair of CDSs. Results were compared with respect to Tadei et al. [40] who used a multi-objective genetic algorithm adopting the same objectives. Results indicated that MOSFLA was able to outperform the approach from Tadei et al. [40] with respect to three performance indicators: hypervolume, set coverage and maximum spread.

Boada et al. [41] used a MOEA for parameter identification of biological models. Two objectives were considered, related to the minimization of the error between: (1) the time-course plate reader experimental observations and open-loop model predictions and (2) the steady-state flow cytometry experimental observations and close-loop model predictions. The authors adopted spMODE, which is based on the use of differential evolution. Also, they incorporated a multi-criteria decision making stage and used visualization tool diagrams to correlate design objectives with decision variables. The proposed approach produced good identification results.

Gaeta et al. [42] developed an open-source Python software called *Multi-Objective Optimisation algorithm for DNA Design and Assembly* (MOODA) for the design and assembly of DNA molecules. MOODA takes as its input an annotated DNA sequence, and it optimizes it with respect to the user-defined objectives. Four objectives were considered: (1) the GC contents of each designed DNA fragment must be within the limits specified by a given DNA synthesis, (2) maximize the use of frequent codons, (3) maximize the blocks of homogenous size and (4) minimize the number of blocks required for the assembly. MOODA is based on the NSGA-II, but incorporates specialized edit and assembly operators. The edit operators are local search procedures (four edit procedures were defined by the authors to edit both DNA sequences and blocks). This approach was able to find near-optimal manufacturable designs for arbitrary long and complex DNA molecules.

D. Optimization of industrial bio-processes

There are several industrial bio-processes in which it is possible to improve productivity by optimizing certain task. Most of them have multi-objective formulations that allow finding solutions that are more realistic and/or appropriate.

Sarkar and Modak [43] used a multi-objective genetic algorithm in two case studies. In the first of them, they used a lysine fermentation model. The authors considered this as a multi-objective control problem in which both productivity and the yield are to be maximized. In the second case study, they considered the optimal nutrient and inducer feeding strategy for the fed-batch production of induced foreign protein using recombinant bacteria. Again, this was considered as a multi-objectie control problem in which they wanted to maximize the amount of protein while minimizing simultaneously the volume of inducer. In both cases, the authors adopted the NSGA-II. In these two case studies, the authors reported the advantages of finding a variety of trade-off solutions and argued about the importance of providing the decision maker with more immediate information about the optimal operating regime of the process.

Al-Siyabi et al. [44] used multi-objective differential evolution to improve the microbial fermentation to produce lysine. As in [43], the authors considered two objectives: the maximization of the productivity and the yield. The authors adopted two versions of multi-objective differential evolution called MODE III [45] and Harmonic MODE [46] and reported better results with the second of them. Constraints were handled using both penalty methods and tournament selection. In their analysis of results, the authors indicated that for a constant feeding rate, the solutions obtained using the Harmonic MODE algorithm were uniformly distributed, while MODE III produced solutions that were highly crowded towards higher productivity and slightly scattered wth respect to yield. For a varied feeding rate, the results obtained with the Harmonic MODE algorithm were widely spread and showed good convergence.

Since the approval of the first monoclonal antibody (mAb) in 1986, they have been frequently adopted for biotherapeutics research and development for the treatment of several human diseases including cancer, arthritis, Alzheimer and metabolic and transmissible diseases such as Covid-19 and HIV. Consequently, there has been an increasing demand for mAbs, which are mainly produced in a fed-batch bioreactor with temperature and media composition as key control variables. Kumar et al. [47] developed a mathematical model that incorporates the effect of temperature, biomass, glucose, protein, and lactate concentration on mAb productivity. The parameter values of such model were estimated using particle swarm optimization to minimize the normalized error function. Then, they adopted a multi-objective model in which the objectives were to maximize the mAb production while minimizing the reactor run time. The goal was to find the optimal feed strategy and temperature shift time while operating the reactor in a sequential combination of the batch, fed-batch and perfusion modes. The authors adopted the NSGA-II and reported an estimation of a 5% increase in mAb prouction for a reactor run time of 15 days.

Villegas-Quiceño et al. [48] used a multi-objective genetic algorithm to improve process productivity in plant cell suspension cultures of *Thevetia preuviana*. The two objectives considered were: maximize the biomass growth rate and minimize substrate consumption. The authors extracted (by observation) a single solution, located in the "knee" of the Pareto front and conducted an experimental validation of it. Their experiments corroborated that this solution was indeed capable of increasing the productivity in terms of metabolite production for the plant cell considered in the study.

E. Data Processing for Bioinformatics

The potential benefit of multi-objective optimization in bioinformatics has been recognized for a long time (see for example [49]). We will discuss next two particular application areas of MOEAs within bioinformatics: multiple sequence alignment and feature selection and classification for diagnosis of diseases.

1) Multiple Sequence Alignment: A problem that is particularly interesting because of its complexity is multiple sequence alingment (MSA), which is the process of aligning three or more biological sequences (DNA, RNA, protein). MSA has a variety of applications in computational biology, including: protein structure predictions, biological function analyses and phylogenetic modeling. MSA is known to be an NP-complete optimization problem in which the time complexity of finding an optimal alignment grows exponentially with the number of sequences and their lengths. Additionally, it is also important to provide an efficient method to measure alignment accuracy, but there is no consensus on how to do it, and several scores have been proposed for that purpose. Examples of scores are: STRIKE [50] (which estimates the molecular contacts from protein structures to calculate alignment accuracies), totally conserved columns (TCC) percentage, and gaps and non-gaps percentage.

Ortuño et al. [51] proposed an approach (based on the NSGA-II) called *Multiobjective Optimizer for Sequence Alignments based on Structural Evaluations* (MO-SAStrE). In this case, three objectives were optimized: (1) STRIKE score, (2) non-gaps percentage and (3) totally conserved columns. The authors showed that their proposed approach outperforms other aligners (in a statistically significant way), including: ClustalW, Multiple Sequence Alignment Genetic Algorithm (MSA-GA), PRRP, DIALIGN, Hidden Markov Model Training (HMMT), Pattern-Induced Multi-sequence Alignment (PIMA), MULTIALIGN, Sequence Alignment Genetic Algorithm (SAGA), PILEUP, Rubber Band Technique Genetic Algorithm (RBT-GA) and Vertical Decomposition Genetic Algorithm (VDGA).

Rubio-Largo et al. [52] proposed H4MSA, which is based on the Shuffled Frog-Leaping Algorithm, for solving the MSA problem. Two objectives were considered: the weighted sum-of-pairs function with affine gap penalties (WSP) and the number of totally conserved columns score. Results were compared with respect to 16 well-known aligners and six tailored genetic algorithms. The authors reported that the results obtained with H4MSA had a remarkable accuracy.

Zambrano-Vega et al. [53] proposed a more efficien and effective version of MO-SAStrE [51], called M2Aligh, which is based on the NSGA-II and can be executed in parallel in multi-core systems. The authors used the same objectives as in [51]. M2Align was able to outperform MO-SAStrE in the STRIKE and TCC scores, but also obtained results with significant time reductions using up to 20 cores.

There are several more papers available on multiple sequence alignment which are not included here due to space constraints (see for example [54]).

2) Feature selection and classification for diagnosis of diseases: Several problems in medicine involve the diagnosis of a disease based on a number of tests done on the patients. Improvements in technology have caused the creation of very large databases which makes very difficult to discover meaningful relationships buried in data. In recent years, machine learning has been used for this task, and the use of MOEAs has allowed to perform more complex and accurate classifications.

Valenzuela et al. [55] used the NSGA-II for optimizing the volumes of interest (VOIs) to extract three-dimensional textures from Magnetic Resonance Images in order to diagnose Alzheimer's Disease (AD), Mild Cognitive Impairment converter, Mild Cognitive Impairment nonconverter and Normal subjects. The idea was to use the MOEA to search for small regions in the brain that are related to AD, since this can lead to a better diagnosis. Two objectives were considered: (1) minimize the complexity (number of VOIs) and (b) maximize the accuracy of the classifier (they adopted one-versus-all support vector machine classifier). The authors reported obtaining excellent results in multi-class classification, with accuracies of up to 94.4%, while also extracting significant information on the location of the most relevant points of the brain.

Wang et al. [56] adopted a multi-objective particle swarm optimization-based hybrid algorithm (MOPSOHA) for cancer subtype diagnosis. The authors considered four objectives: (1) accuracy, (2) the number of features, and two entropy-based measures: (3) relevance and (4) redundancy. The proposed approach was tested with 41 cancer datasets including thirty-five cancer gene expression datasets and six independent disease datasets. MOPSOHA had a high subtype discrimination power for cancer subtype diagnosis.

The identification of biomarkers is essential for the diagnosis and prognosis of certain diseases, such as cancer. The purpose of gene selection is to find the minimum number of genes that can properly classify (i.e., tumour or normal) a sample with a high accuracy. Then, the selected genes can be studied as potential biomarkers. Coleto-Alcudia and Vega-Rodríguez proposed a two-step gene selection method. The first step consisted of a filtering process of the most relevant genes of a gene expression dataset. In this step, the authors combined three feature selection methods commonly adopted in gene selection. Since the gene selection process itself involves two objectives (minimize the number of selected genes and maximize the classification accuracy), the second step is performed with the Artificial Bee Colony based on Dominance (ABCD) algorithm which is a Pareto-based version of the ABC algorithm [57]. The authors concluded that their proposed approach is effective in gene selection for the identification of cancer biomarkers from RNA-seq data.

Singh and Singh [58] proposed a stacking-based evolutionary ensemble learning system called *NSGA-II-Stacking* for predicting the onset of Type-2 diabetes mellitus within five years. In the proposed approach, four different types of learners are adopted as base learners which are trained with five boostrapped samples generated by crossvalidation. Then, the NSGA-II is used to select models from 20 trained base learners. Two objectives were considered: maximizing the classification accuracy and minimizing the assembly complexity. The authors indicated that their proposed approach outperformed several individual machine learning approaches and conventional ensemble approaches, achieving an accuracy of up to 83.8%.

IV. Future Research Areas

In the applications reviewed, more work is still required as described next:

- **Parallel computing:** Applications in biotechnology are computationally demanding and the need for parallel computing is evident. Although some researchers have developed parallel implementations of their MOEAs (see for example [53]), the development of more efficient parallel implementations of MOEAs is clearly required [59].
- **Incorporation of preferences:** In most applications of MOEAs in biotechnology, researchers are interested in extracting a single solution from the Pareto front. In fact, many researchers suggest extracting the solution at the "knee" of the Pareto front, since it represents the best possible compromise (particularly when dealing with only two objectives). However, there is a wide variety of methods to incorporate user's preferences that could be used instead [60].
- **Improvement of the current databases:** There is an evident need to consolidate and curate many of the databases that are required for some applications in biotechnology, since many of them are incomplete or require of some additional processing. This would allow a more extensive application of MOEAs in biotechnology.
- **Big data:** Some areas such as systems biology and omics require huge amounts of data and there is a clear need for computational techniques that allow a more efficient processing of such massive volumes of information. The use of multiobjective optimization in big data analytics has been already suggested (see for example [61]) and could bring important benefits.
- **Machine learning:** Machine learning has been used in a wide variety of biotechnology applications [62] and it is possible to find intersections between machine learning and multi-objective optimization in both directions. We can either use multi-objective concepts to produce new machine learning models or we can use machine learning techniques that can improve the performance of multi-objective evolutionary algorithms. In the first case, multi-objective optimization has already shown to produce benefits to machine learning techniques (e.g., by producing solutions or hyperparameters that produce more balanced classifiers in terms of accuracy and computational cost [63]). In the second case, there are already practical applications of the use of machine learning techniques to aid multi-objective optimizers (see for example [64]).

V. Conclusions

Like many other application areas, biotechnology has a great potential for the use of multi-objective evolutionary algorithms. This area contains a wide number of complex and computer demanding problems in which it is desirable to optimize two or more objectives simultaneously. Researchers in this area also have a great interest in selecting a single solution out of the many that conform the Pareto front approximation generated by the multi-objective optimizer, mainly because such solution can be compared with respect to the output of the single-objective optimizers previously used in this domain, but also because for validation purposes (in an experimental way). After reviewing several application areas within biotechnology that have benefitted from the use of MOEAs, we provided a few possible paths for future research. Finally, we believe that it is important to develop tools that facilitate the integration and application of MOEAs in biotechnology. This would help to extend the use of these techniques in this domain.

References

[1] C. A. Coello Coello, G. B. Lamont, and D. A. Van Veldhuizen, *Evolutionary Algorithms for Solving Multi-Objective Problems*, 2nd ed. New York: Springer, September 2007, iSBN 978-0-387-33254-3.

[2] J. D. Schaffer, "Multiple Objective Optimization with Vector Evaluated Genetic Algorithms," in *Genetic Algorithms and their Applications: Proceedings of the First International Conference on Genetic Algorithms*. Lawrence Erlbaum, 1985, pp. 93–100.

[3] S. Srinivasan and S. Ramakrishnan, "Evolutionary multi objective optimization for rule mining: a review," *Artificial Intelligence Review*, vol. 36, no. 3, pp. 205–248, October 2011.

[4] M. Bhuvaneswari, Ed., *Application of Evolutionary Algorithms for Multi-objective Optimization in VLSI and Embedded Systems*. India: Springer, 2015, iSBN 978-81-322-1957-6.

[5] A. López Jaimes and C. A. Coello Coello, "An Introduction to Multi-Objective Evolutionary Algorithms and some of Their Potential Uses in Biology," in *Applications of Computational Intelligence in Biology: Current Trends and Open Problems*, T. Smolinski, M. G. Milanova, and A.-E. Hassanien, Eds. Berlin: Springer, 2008, pp. 79–102, iSBN 978-3-540-78533-0.

[6] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA–II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, April 2002.

[7] E. Zitzler, M. Laumanns, and L. Thiele, "SPEA2: Improving the Strength Pareto Evolutionary Algorithm," in *EUROGEN 2001. Evolutionary Methods for Design, Optimization and Control with Applications to Industrial Problems*, K. Giannakoglou, D. Tsahalis, J. Periaux, P. Papailou, and T. Fogarty, Eds., Athens, Greece, 2001, pp. 95–100.

[8] N. Beume, C. M. Fonseca, M. Lopez-Ibanez, L. Paquete, and J. Vahrenhold, "On the Complexity of Computing the Hypervolume Indicator," *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 5, pp. 1075–1082, October 2009.

[9] E. Zitzler and S. Künzli, "Indicator-based Selection in Multiobjective Search," in *Parallel Problem Solving from Nature - PPSN VIII*, X. Y. et al., Ed. Birmingham, UK: Springer-Verlag. Lecture Notes in Computer Science Vol. 3242, September 2004, pp. 832–842.

[10] N. Beume, B. Naujoks, and M. Emmerich, "SMS-EMOA: Multiobjective selection based on dominated hypervolume," *European Journal of Operational Research*, vol. 181, no. 3, pp. 1653–1669, 16 September 2007.

[11] Q. Zhang and H. Li, "MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 6, pp. 712–731, December 2007.

[12] A. Grosdidier, V. Zoete, and O. Michielin, "EADock: Docking of Small Molecules Into Protein Active Sites With a Multiobjective Evolutionary Optimization," *PROTEINS: Structure, Function, and Bioinformatics*, vol. 67, no. 4, pp. 1010–1025, June 2007.

[13] S. Janson, D. Merkle, and M. Middendorf, "Molecular docking with multi-objective Particle Swarm Optimization," *Applied Soft Computing*, vol. 8, no. 1, pp. 666–675, January 2008.

[14] M. J. García-Godoy, E. Lopez-Camacho, J. García-Nieto, A. J. Nebro, and J. F. Aldana-Montes, "Solving Molecular Docking Problems with Multi-Objective Metaheuristics," *Molecules*, vol. 20, no. 6, pp. 10 154–10 183, June 2015.

[15] A. J. Nebro, J. J. Durillo, J. Garcia-Nieto, C. A. Coello Coello, F. Luna, and E. Alba, "SMPSO: A New PSO-based Metaheuristic for Multi-objective Optimization," in *2009 IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making (MCDM'2009)*. Nashville, TN, USA: IEEE Press, March 30 - April 2 2009, pp. 66–73, iSBN 978-1-4244-2764-2.

[16] S. Kukkonen and J. Lampinen, "GDE3: The third Evolution Step of Generalized Differential Evolution," in *2005 IEEE Congress on Evolutionary Computation (CEC'2005)*, vol. 1. Edinburgh, Scotland: IEEE Service Center, September 2005, pp. 443–450.

[17] A. Oduguwa, A. Tiwari, S. Fiorentino, and R. Roy, "Multi-Objective Optimisation of the Protein-Ligand Docking Problem in Drug Discovery," in *2006 Genetic and Evolutionary Computation Conference (GECCO'2006)*, M. K. et al., Ed., vol. 2. Seattle, Washington, USA: ACM Press. ISBN 1-59593-186-4, July 2006, pp. 1793–1800.

[18] J. D. Knowles and D. W. Corne, "Approximating the Nondominated Front Using the Pareto Archived Evolution Strategy," *Evolutionary Computation*, vol. 8, no. 2, pp. 149–172, 2000.

[19] E. Zitzler and L. Thiele, "Multiobjective Evolutionary Algorithms: A Comparative Case Study and the Strength Pareto Approach," *IEEE Transactions on Evolutionary Computation*, vol. 3, no. 4, pp. 257–271, November 1999.

[20] J.-C. Boisson, L. Jourdan, E.-G. Talbi, and D. Horvath, "Single- and Multi-Objective Cooperation for the Flexible Docking Problem," *Journal of Mathematical Models and Algorithms*, vol. 9, pp. 195–208, 19 February 2010.

[21] A. Liefooghe, M. Basseur, L. Jourdan, and E.-G. Talbi, "ParadisEO-MOEO: A Framework for Evolutionary Multi-objective Optimization," in *Evolutionary Multi-Criterion Optimization, 4th International Conference, EMO 2007*, S. Obayashi, K. Deb, C. Poloni, T. Hiroyasu, and T. Murata, Eds. Matshushima, Japan: Springer. Lecture Notes in Computer Science Vol. 4403, March 2007, pp. 386–400.

[22] A. Sandoval-Perez, D. Becerra, D. Vanegas, D. Restrepo-Montoya, and F. Nino, "A Multi-objective Optimization Energy Approach to Predict the Ligand Conformation in a Docking Process," in *Genetic Programming, 16th European Conference, EuroGP 2013*, K. Krawiec, A. Moraglio, T. Hu, A. Şima Etaner-Uyar, and B. Hu, Eds. Vienna, Austria: Springer. Lecture Notes in Computer Science Vol. 7831, April 3-5 2013, pp. 181–192.

[23] J. Branke, K. Deb, H. Dierolf, and M. Osswald, "Finding Knees in Multi-Objective Optimization," in *Parallel Problem Solving from Nature - PPSN VIII*. Birmingham, UK: Springer-Verlag. Lecture Notes in Computer Science Vol. 3242, September 2004, pp. 722–731.

[24] J. Gu, X. Yang, L. Kang, J. Wu, and X. Wang, "MoDock: A Multi-Objective Strategy Improves the Accuracy for Molecular Docking," *Algorithms for Molecular Biology*, vol. 10, February 18 2015, article number: 8.

[25] E. López-Camacho, M. J. García-Godoy, J. García-Nieto, A. J. Nebro, and J. F. Aldana-Montes, "A New Multi-objective Approach for Molecular Docking Based on RMSD and Binding Energy," in *Algorithms for Computational Biology. Third International Conference, AlCoB 2016, Proceedings*, M. Botón-Fernández, C. Martín-Vide, S. Santander-Jiménez, and M. A. Vega-Rodríguez, Eds. Trujillo, Spain: Springer. Lecture Notes in Computer Science Vol. 9702, 21-22 June 2016, pp. 65–77, iSBN 978-3-319-38826-7.

[26] J. García Nieto, E. López-Camacho, M. J. García-Godoy, A. J. Nebro, and J. F. Aldana-Montes, "Multi-Objective Ligand-Protein Docking with Particle Swarm Optimizers," *Swarm and Evolutionary Computation*, vol. 44, pp. 439–452, February 2019.

[27] A. J. Nebro, J. J. Durillo, and C. A. Coello Coello, "Analysis of Leader Selection Strategies in a Multi-Objective Particle Swarm Optimizer," in *2013 IEEE Congress on Evolutionary Computation (CEC'2013)*. Cancún, México: IEEE Press, 20-23 June 2013, pp. 3153–3160, iSBN 978-1-4799-0454-9.

[28] C. Dai, Y. Wang, and M. Ye, "A New Multi-Objective Particle Swarm Optimization Algorithm Based on Decomposition," *Information Sciences*, vol. 325, pp. 541–557, December 20 2015.

[29] M. J. García-Godoy, E. López-Camacho, J. García-Nieto, J. Del Ser, A. J. Nebro, and J. F. Aldana-Montes, "Bio-inspired optimization for the molecular docking problem: State of the art, recent results and perspectives," *Applied Soft Computing*, vol. 79, pp. 30–45, June 2019.

[30] J. Ji, J. Zhou, Z. Yang, Q. Lin, and C. A. Coello Coello, "AutoDock Koto: A Gradient Boosting Differential Evolution for Molecular Docking," *IEEE Transactions on Evolutionary Computation*, vol. 27, no. 6, pp. 1648–1662, December 2023.

[31] M. García-Ortegón, G. N. Simm, A. J. Tripp, J. M. Hernández-Lobato, A. Bender, and S. Bacallado, "DOCKSTRING: Easy Molecular Docking Yields Better Benchmarks for Ligand Design," *Journal of Chemical Information and Modeling*, vol. 62, pp. 3486–3502, 2022.

[32] P. Maia, I. Rocha, E. C. Ferreira, and M. Rocha, "Evaluating evolutionary multiobjective algorithms for the in silico optimization of mutant strains," in *2008 8th IEEE International Conference on BioInformatics and BioEngineering*. Athens, Greece: IEEE, 8-10 October 2008, iSBN 978-1-4244-2844-1.

[33] A. Patané, A. Santoro, J. Costanza, G. Carapezza, and G. Nicosia, "Pareto Optimal Design for Synthetic Biology," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 9, no. 4, pp. 555–571, August 2015.

[34] A. Patané, G. Jansen, P. Conca, G. Carapezza, J. Costanza, and G. Nicosia, "Multi-objective optimization of genome-scale metabolic models: the case of ethanol production," *Annals of Operations Research*, vol. 276, pp. 211–227, 2019.

[35] X. Fan, J. Zhou, J. Xia, and X. Yan, "Genome-Scale Metabolic Model's multi-objective solving algorithm based on the inflexion point of Pareto front including maximum energy utilization and its application in *Aspergillus niger* DS03043," *Biotechnology & Bioengineering*, vol. 119, no. 6, pp. 1539–1555, June 2022.

[36] Z. Dai, S. Yang, L. Xu, H. Hu, K. Liao, J. Wang, Q. Wang, S. Gao, B. Li, and L. Lai, "Identification of Cancerassociated metabolic vulnerabilities by modeling multi-objective optimality in metabolism," *Cell Communication and Signaling*, vol. 17, 2019, article number: 124.

[37] A. F. Villaverde, S. Bongard, K. Mauch, E. Balsa-Canto, and J. R. Banga, "Metabolic engineering with multi-objective optimization of kinetic models," *Journal of Biotechnology*, vol. 222, pp. 1–8, 20 March 2016.

[38] Y. Boada, G. Reynoso-Meza, J. Picó, and A. Vignoni, "Multi-objective optimization framework to obtain model-based guidelines for tuning biological synthetic devices: an adaptive network case," *BMC Systems Biology*, vol. 10, 2016, article number: 27.

[39] B. González Sánchez, M. A. Vega-Rodríguez, and S. Santander-Jiménez, "Multi-objective memetic meta-heuristic algorithm for encoding the same protein with multiple genes," *Expert Systems with Applications*, vol. 136, pp. 83–93, 1 December 2019.

[40] Goro Terai and Satoshi Kamegai and Akito Taneda and Kiyoshi Asai, "Evolutionary design of multiple genes encoding the same protein," *Bioinformatics*, vol. 33, no. 11, pp. 1613–1620, June 2017.

[41] Y. Boada, A. Vignoni, and J. Picó, "Multiobjective Identification of a Feedback Synthetic Gene Circuit," *IEEE Transactions on Control Systems Technology*, vol. 28, no. 1, pp. 208–223, January 2020.

[42] A. Gaeta, V. Zulkower, and G. Stracquadanio, "Design and assembly of DNA molecules using multi-objective optimization," *Synthetic Biology*, vol. 6, no. 1, pp. 1–9, 11 October 2021.

[43] D. Sarkar and J. M. Modak, "Pareto-optimal Solutions for Multi-objective Optimization of Fed-batch Bioreactors Using Nondominated Sorting Genetic Algorithm," *Chemical Engineering Science*, vol. 60, no. 2, pp. 481–492, January 2005.

[44] B. Al-Siyabi, A. M. Gujarathi, and N. Sivakumar, "Harmonic multi-objective differential evolution approach for multi-objective optimization of fed-batch reactor," *Materials and Manufacturing Processes*, vol. 32, no. 10, pp. 1152–1161, 22 February 2017.

[45] A. M. Gujarathi, A. H. Motagamwala, and B. V. Babu, "Multiobjective Optimization of Industrial Naphtha Cracker for Production of Ethylene and Propylene," *Materials and Manufacturing Processes*, vol. 28, no. 7, pp. 803–810, July 3 2013.

[46] V. L. Huang, P. Suganthan, A. K. Qin, and S. Baskar, "Multiobjective Differential Evolution with External Archive and Harmonic Distance-Based Diversity Measure," Nanyang Technological University, Singapore, Tech. Rep., 2005.

[47] D. Kumar, N. Gangwar, A. S. Rathore, and M. Ramteke, "Multi-objective optimization of monoclonal antibody production in bioreactor," *Chemical Engineering and Processing - Process Intensification*, vol. 180, October 2022, article number: 108720.

[48] A. P. Villegas-Quiceño, J. P. Arias-Echeverri, D. Aragón-Mena, S. Ochoa-Cáceres, and M. E. Arias-Zavala, "Multi-objective optimization in biotechnological processes: application to plant cell suspension cultures of *Thevetia peruviana*," *Revista Facultad de Ingeniería, Universidad de Antioquia*, no. 87, pp. 35–40, 2018.

[49] J. Handl, D. B. Kell, and J. Knowles, "Multiobjective Optimization in Bioinformatics and Computational Biology," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 2, pp. 279–292, April-June 2007.

[50] C. Kemena, J.-F. Taly, J. Kleinjung, and C. Notredame, "STRIKE: evaluation of protein MSAs using a single 3D structure," *Bioinformatics*, vol. 27, no. 24, pp. 3385–3391, December 2011.

[51] F. M. O. no, O. Valenzuela, F. Rojas, H. Pomares, J. P. Florido, J. M. Urquiza, and I. Rojas, "Optimizing multiple sequence alignments using a genetic algorithm based on three objectives: structural information, non-gaps percentage and totally conserved columns," *Bioinformatics*, vol. 29, no. 17, pp. 2112–2121, September 2013.

[52] Álvaro Rubio-Largo, M. A. Vega-Rodríguez, and D. L. González-Álvarez, "A Hybrid Multiobjective Memetic Metaheuristic for Multiple Sequence Alignment," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 4, pp. 499–514, August 2016.

[53] C. Zambrano-Vega, A. J. Nebro, J. García-Nieto, and J. F. Aldana-Montes, "M2Align: parallel multiple sequence alignment with a multi-objective metaheuristic," *Bioinformatics*, vol. 33, no. 19, pp. 3011–3017, October 2017.

[54] B. Chowdhury and G. Garai, "A review on multiple sequence alignment from the perspective of genetic algorithm," *Genomics*, vol. 109, no. 5-6, pp. 419–431, October 2017.

[55] O. Valenzuela, X. Jiang, A. Carrillo, and I. Rojas, "Multi-Objective Genetic Algorithms to Find MostRelevant Volumes of the Brain Related to Alzheimers Disease and Mild Cognitive Impairment," *International Journal of Neural Systems*, vol. 28, no. 9, 2018, article number: 850022.

[56] Y. Wang, Z. M. an Ka-Chun Wong, and X. Li, "Evolving Multiobjective Cancer Subtype Diagnosis From Cancer Gene Expression Data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 6, pp. 2431–2444, November-December 2021.

[57] D. Karaboga and B. Basturk, "A Powerful and Efficient Algorithm for Numerical Function Optimization: Artificial Bee Colony (ABC) Algorithm," *International Journal of Global Optimization*, vol. 39, pp. 459–471, 2007.

[58] N. Singh and P. Singh, "Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 1–22, January-March 2020.

[59] J. G. Falcon-Cardona, R. H. Gomez, C. A. C. Coello, and M. G. Castillo Tapia, "Parallel Multi-Objective Evolutionary Algorithms: A Comprehensive Survey," *Swarm and Evolutionary Computation*, vol. 67, December 2021, article Number: 100960.

[60] H. Wang, M. Olhofer, and Y. Jin, "A Mini-Review on Preference Modeling and Articulation in Multi-Objective Optimization: Current Status and Challenges," *Complex & Intelligent Systems*, vol. 3, no. 4, pp. 233–245, December 2017.

[61] C. Barba-González, J. García-Nieto, A. J. Nebro, and J. F. Aldana-Montes, "Multi-objective Big Data Optimization with jMetal and Spark," in *Evolutionary Multi-Criterion Optimization, 9th International Conference, EMO 2017*, H. Trautmann, G. Rudolph, K. Klamroth, O. Schütze, M. Wiecek, Y. Jin, and C. Grimme, Eds. Münster, Germany: Springer. Lecture Notes in Computer Science Vol. 10173, March 19-22 2017, pp. 16–30, iSBN: 978-3-319-54156-3.

[62] A. Holzinger, K. Keiblinger, P. Holub, K. Zatloukal, and H. Müller, "AI for life: Trends in artificial intelligence for biotechnology," *New Biotechnology*, vol. 74, pp. 16–24, May 2023.

[63] Q. Qu, Z. Ma, A. Clausen, and B. N. Jorgensen, "Comprehensive Review of Machine Learning in Multi-objective Optimization," in *2021 IEEE 4th International Conference on Big Data and Artificial Intelligence (BDAI'2021)*. Qingdao, China: IEEE Press, 2-4 July 2021, pp. 7–14, iSBN 978-1-6654-4843-7.

[64] Z. Wang, J. Li, G. P. Rangaiah, and Z. Wu, "Machine learning aided multi-objective optimization and multi-criteria decision making: Framework and two applications in chemical engineering," *Computers & Chemical Engineering*, vol. 165, September 2022, article No. 107945.