Exploiting medical-expert knowledge via a novel memetic algorithm for the inference of gene regulatory networks

Adrián Segura-Ortiz[1 [0000−0003−2149−5754]], José García-Nieto[1,2, [0000−0003−2985−3480]], and José F. Aldana-Montes[1,2, [0000−0002−2673−9474]]

[1] ITIS Software and Dept. Lenguajes y Ciencias de la Computación, University of Málaga, Spain
[2] Biomedical Research Institute of Málaga (IBIMA), Spain
{adrianseor.99@uma.es,jnieto@uma.es}

Abstract. This study introduces an innovative memetic algorithm for optimizing the consensus of well-adapted techniques for the inference of gene regulation networks. Building on the methodology of a previous proposal (GENECI), this research adds a local search phase that incorporates prior knowledge about gene interactions, thereby enhancing the optimization process under the influence of domain expert. The algorithm focuses on the evaluation of candidate solutions through a detailed evolutionary process, where known gene interactions guide the evolution of such solutions (individuals). This approach was subjected to rigorous testing using benchmarks from editions 3 and 4 of the DREAM challenges and the yeast network of IRMA, demonstrating a significant improvement in accuracy compared to previous related approaches. The results highlight the effectiveness of the algorithm, even when only 5% of the known interactions are used as a reference. This advancement represents a significant step in the inference of gene regulation networks, providing a more precise and adaptable tool for genomic research.

Keywords: Memetic Algorithm · Gene Regulatory Networks · Optimization · Bioinformatics

1 Introduction

In the field of computational biology, the inference of gene regulatory networks (GRNs) has become an indispensable mean to comprehend the mechanisms governing gene expression and their implications in various areas of biomedical research. These networks, which are crucial for understanding biological processes at the molecular level, provide a valuable perspective in the study of diseases [14, 30] and in the development of genetic therapies [29, 36].

However, despite significant advances in this field, the accurate inference of GRNs remains a considerable challenge [15, 23, 34, 42]. There are two main difficulties. The first is the inherent complexity of biological systems [28]. The second

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_1

is the limitations related to the quantity and quality of empirically validated data [6], which are also difficult to properly incorporate into existing methodologies to improve the accuracy of the results. There is a clear need to take advantage of the knowledge that the medical expert and the literature can bring to the partial construction of networks through a priori known interactions.

In response to these challenges, this research proposes an advanced methodology that extends the previous work carried out in GENECI [33]. GENECI has proven effective in addressing the complexity and diversity of networks through the clever consensus of various techniques. Building upon this solid foundation that addresses the first challenge, an additional stage has been integrated to tackle the second drawback, focusing on maximizing the use of known information. This has been approached by designing an adapted additional local search phase, which incorporates prior knowledge about genetic interactions to guide the optimization process, thus allowing for greater precision in the inference of GRNs, through the injection of domain experts' knowledge.

In this domain, it is common for experts to have partial knowledge or hypotheses about specific genetic interactions. This research focuses on the importance of integrating such knowledge into the inference of GRNs. The experimentation in this work is based on the idea of refining and testing the proposal on generic benchmarks with the intention of subsequently validating its application in real-world problems where the complete solution is unknown. This has been conducted by means of a well-grounded set of benchmarks, including DREAM challenges [26] (specifically their 3rd and 4th editions) and the yeast network of IRMA [4]. Results have shown that the application of this approach introduces significant improvements in the inference of GRNs even when a minimal amount of information is used.

This article is organized as follows. The state of the art in this field is presented in Section 2, followed by a detailed description of the approach and methodology in Section 3. Subsequently, the experimentation of this study is presented in Section 4. Conclusions and future lines of work are discussed in Section 5.

2 Related Work

The inference of gene regulatory networks from expression data is a well-studied challenge in computational biology. The literature has explored multiple approaches, including probabilistic graphical models [35], ordinary differential equations (ODEs) [11, 16, 37], and machine learning techniques such as neural networks [10, 12, 20, 39]. Integrative methods combining different omics data types have also been explored [41], along with causality-based approaches [8] and works related to mutual information [38]. The diversity of approaches has led to a wide range of computational techniques aimed at inferring GRNs. Among them, notable for their accuracy and popularity in the literature are ARACNE [25], C3NET [1], CLR [7], GENIE3 [17], and GRNBOOST2 [27].

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_1

In the field of genetic regulatory network inference, seeking a consensus among the results of multiple techniques has been a prominent trend. The DREAM challenge [26] was a significant turning point, demonstrating that combining results from various techniques produces more accurate solutions than individual methods alone [24]. This revelation spurred the exploration of diverse approaches to achieve consensus, such as the analysis of topological features [19], graph mining [18], and evolutionary algorithms [9, 31].

Recent advances reveal novel strategies for achieving consensus among inference techniques, although they still lack a robust methodology tailored to real-world biological networks. EnGRaiN [2] approaches consensus from a mathematical perspective without considering the biological context, while GReNaDIne [32] considers a limited number of techniques with a simple consensus procedure. The challenge of building a weighted and optimal consensus from a set of techniques, taking into account the biological nature of the problem, was addressed in GENECI [33], with results demonstrating a significant improvement in the accuracy of inferred networks.

In the biomedical field, the adoption of memetic algorithms has gained significant traction, demonstrating their versatility and effectiveness in several applications [3, 5, 13, 21, 22, 40]. These algorithms, which combine intensive local search with global evolutionary strategies, have been successfully applied to solve complex problems in this domain. For instance, in [13] the optimization of PPI (Protein-Protein Interactions) network alignment considers both topological structure and sequence similarities, surpassing existing methods in accuracy. Additionally, in protein structure prediction, memetic algorithms have been designed using knowledge from databases to guide the search towards similar native structures, showing promising results comparable to reference prediction methods [5, 21]. In the field of cancer diagnosis, the application of memetic algorithms has demonstrated to enhance the selection of relevant genes by combining local and global search techniques to identify discriminant genes with precision [3].

Finally, the memetic approach has also reached the focus of this work, the reconstruction of GRNs. In [22], an innovative approach is proposed to learn parameters of Recurrent Neural Networks (RNN) and develop an LASSO (Least Absolute Shrinkage and Selection Operator) based framework for the effective reconstruction of GRNs. This method demonstrates superior ability to handle the complexity and sparsity of relationships in real GRNs, outperforming other RNN learning algorithms in large-scale network reconstruction. More recently, in [40], a memetic algorithm is proposed for inferring sparse GRNs using Maximum Entropy Probability Models (MEPMs). This approach addresses the problem from a multi-objective optimization perspective, considering maximum entropy and MEPM constraints as separate objectives.

Given the statistical rigor demonstrated by the GENECI proposal in its results and considering the validity that the memetic approach has shown in biomedical domain problems, it is more than justified to introduce this approach to address the specific problem of reaching a consensus among several inference techniques for the reconstruction of GRNs.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_1

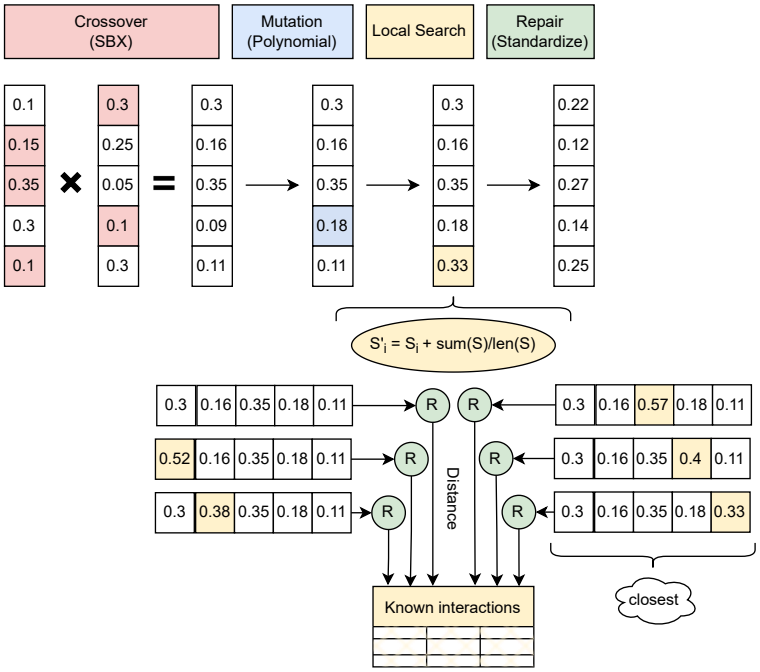

Fig. 1. Succession of phases within the evolutionary process. Individuals are crossed through simulated binary crossover and subsequently subjected to polynomial mutation. Following this, the local search begins where several variations of the individual (encoding a given solution) are compared to select the one whose consensus network is closest to the known interactions. Finally, the individuals are repaired to resume their representation in the form of a weight vector.

3 Proposed Approach

In this article, a memetic algorithm is proposed to optimize the consensus of different techniques for the inference of gene regulation networks. This is based on our previous proposal where an evolutionary process drives this optimization based on the quality and topological characteristics of the networks [33]. This tool has been complemented with a local search phase to guide the optimization process, thanks to prior knowledge of certain gene interactions in the network. This additional phase is located and exemplified in Fig. 1. For a more technical analysis, the pseudocode is set out in Algorithm 1.

The set of candidates subjected to local search is iteratively explored in a loop spanning the length of the individual (line 3 in Algorithm 1). This set comprises the individual provided by the previous phase without any modification (case $i = -1$ in Algorithm 1) and each of the variations resulting from granting an additional vote to each technique (case $i \neq -1$ in Algorithm 1). In other words, the first variation will correspond to adding an additional vote of confidence

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_1

Algorithm 1 Main code of the local search phase

Input Individual sol, Known interactions involved in distance calculation ref.
Output Improved individual $resSol$.

```
1: resSol ← copyOf (sol)
2: minDistance ← inf
3: for i in (−1, len (sol)) do
4:     tmpSol ← copyOf (sol)
5:     if i ≠ −1 then
6:         tmpSol[i] += sum (sol) /len (sol)
7:     RepairSolution (tmpSol)
8:     net ← GetNetwork(tmpSol)
9:     distance ← Distance (net, ref)
10:    if distance < minDistance then
11:        minDistance ← distance
12:        resSol ← tmpSol
13: return resSol
```

to the first technique, quantified as the value of one vote in the case that the system is not weighted (case $i = 0$ and line 6 in Algorithm 1). The exact formula for calculating the new value of the technique in the vector is explained and exemplified in Fig. 1.

After generating the candidates, they are repaired and the consensus network derived from each of them is constructed (lines 7 and 8 in Algorithm 1). Finally, the distance of their confidence levels from the known interactions in the network is measured (line 9 in Algorithm 1). The known interactions will usually be assigned a confidence level equal to 1 in the comparison file. However, if the medical researcher wishes to assign a certain probability to their knowledge, any other value between 0 and 1 is accepted. This means that knowledge of a non-existent interaction could also be reflected, but this case is less common.

If the distance is less than the recorded minimum, the current one becomes the new minimum and the best solution is replaced by the current one (lines 10-12 in Algorithm 1). At the end of the loop, the solution with the smallest distance to the reference is returned (line 13 in Algorithm 1).

The distance is calculated as a simple summation of the absolute value differences between the value of the known interactions (usually 1) and the confidence levels assigned by the consensus network for these interactions. However, the possibility that the set of known interactions is a poorly distributed sample that always favors the same technique during the consensus, has been considered. To mitigate this possibility, an additional parameter has been added that defines the interactions that participate in the calculation of the distance on each iteration.

This parameter is exemplified in Fig. 2 by covering its three possible values, namely: the option *all* is contemplated, in which all the known interactions participate in all local searches; the option *some* in which a randomly chosen subset of them participates on each occasion; and finally the option *one* in which only one of the known interactions chosen randomly is used on each local search.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_1

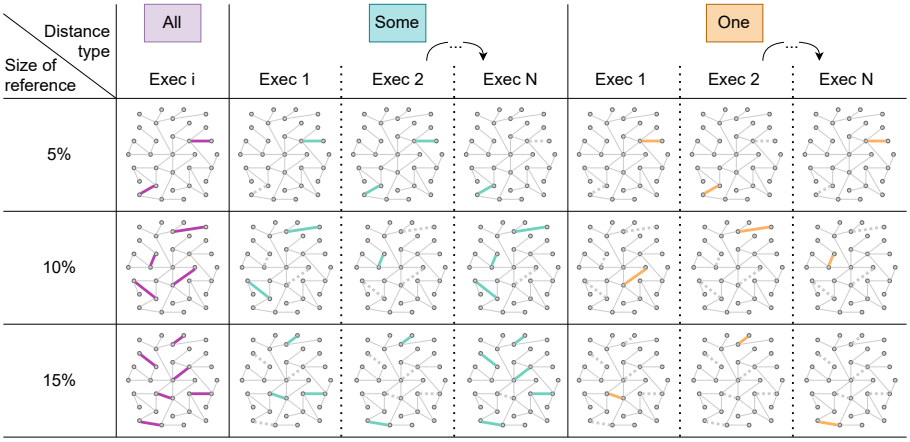

Fig. 2. Examples of interactions involved in the distance calculation in different executions based on the proportion of the gold standard extracted as a set of "known by the expert" interactions (rows) and the type of distance (columns). The case of extracting 5%, 10%, and 15% of the gold standard for the distance types *all*, *some*, and *one* respectively, is shown. As can be observed, all executions take the same reference in the case of *all*, while for *some* and *one*, there is a certain random component that causes differences on each local search.

This local search phase aims at breaking the limitations imposed by GENECI in its aggregate term *Quality*, where techniques whose confidence levels are quite consistent with the remaining ones are somewhat rewarded. Although the consistency of confidence values can increase the reliability of a technique, this strategy sometimes lets certain peculiar interactions that are only inferred by a small subset of techniques slip away. The local search allows for the utilization of prior information to the inference of the network to identify these cases and redirect the evolution of the individuals. It is evident that both strategies are interdependent and must coexist in the evolutionary process, as exceeding the use of previously known information could provoke overfitting.

4 Experimentation

The experimentation addressed in this study employs the academic benchmarks provided by the DREAM challenges [26] (specifically their 3rd and 4th editions) and the yeast network of IRMA [4]. DREAM challenges focused on subnetworks associated with *Escherichia coli* (E. coli) and *Saccharomyces cerevisiae* (yeast) organisms, and includes networks with sizes ranging from 10 to 100 genes. IRMA network comprises 5 genes (CBF1, GAL4, SWI5, GAL80, and ASH1) and encompasses 6 regulatory interactions. These interactions lead to the creation of both "switch on" and "switch off" versions of the network, achieved by cultivating cells in either galactose or glucose conditions, respectively. All these networks

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_1

were also part of the experimentation of GENECI and constitute a total of 27 inference cases. The known interactions of these networks that will guide the evolutionary process have been defined from their gold standards (known solutions information). Specifically, 5% of these references have been extracted for each execution.

The accuracy of the results will be calculated using the AUROC and AUPR metrics, which were set by the DREAM challenges themselves for their competition and make it possible to compare these results with other studies in the literature. Other metrics such as F1-Score and MCC are not considered, as the use of the chosen benchmark standards is deemed sufficient to cover this study.

This section presents the parameter configuration of the proposed method and the subsequent rigorous comparison with regard to GENECI.

4.1 Parameter settings

Given that this proposal partially follows the evolutionary process of GENECI, which is in fact common in standard EA settings, it has been decided to keep as much as possible the parameter setting that was configured in the experimentation of its corresponding article, hence allowing a fair comparison. Therefore, the default settings of simulated binary crossover (with a probability of 0.9), polynomial mutation (with a probability of $1/n$, where n is the number of techniques to be consolidated), and repair based on vector standardization have been established. However, for the additional phase proposed in this work, it remains to determine the probability with which the local search is carried out (which is independent of the crossover and mutation probability) and the way the information from the known interactions is used for the calculation of the distance.

To find the most suitable values for these two parameters, all possible combinations between their values have been considered. For the probability of the local search, the candidate values 0.1, 0.25, 0.4, and 0.55, have been defined. And for the type of distance, the already discussed options of *all*, *some*, and *one*.

Each combination of parameters has been tested with 15 independent executions for each network considered in this study. Afterwards, the performance of each solution was calculated using the AUROC and AUPR metrics with regard to the gold standards. For each network and combination of parameters, the median of their precision values was extracted, which finally allowed the calculation of a Friedman statistical ranking with Holm's non-parametric tests.

The results are shown in Table 1 for the AUPR metric and in Table 2 for the AUROC metric. It can be seen how the winning combination for both cases is the one that always takes into account all the known interactions in the distance calculation and with a higher probability of local search. That is, the combination that employs to a greater extent the external information provided. However, rigorous statistical significance cannot be attributed to this victory since only in one case does it meet the established threshold of $p < 0.05$.

A point to consider regarding the lack of statistical significance is that academic problems have a relatively small network size, sometimes around 10 nodes. This causes the difference between taking all or only a subset of interactions for

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_1

Table 1. Friedman mean rank with Holm's adjusted p values (0.05) for AUPR. Several distance (D) and local search probability (P) configurations are compared based on the AUPR metric. For this purpose, 15 independent runs of each configuration were performed and the median of them (Median) was rescued. After running Friedman's statistical ranking (second column), the winner (highlighted in bold with $*$) is taken as a reference to measure statistical significance against the rest using Holm's nonparametric tests (third column).

AUPR		
Algorithm	$Friedman'sRank$	$Holm'sAdj - p$
***Median D-all P-0.55**	**4.88889**	-
Median D-one P-0.25	5.90741	0.725979
Median D-one P-0.1	5.96296	0.725979
Median D-all P-0.25	6.03704	0.725979
Median D-some P-0.4	6.24074	0.673303
Median D-all P-0.4	6.53704	0.465230
Median D-some P-0.1	6.62963	0.456477
Median D-one P-0.55	6.75926	0.396552
Median D-some P-0.25	6.90741	0.341178
Median D-one P-0.4	6.92593	0.341178
Median D-some P-0.55	7.00000	0.314504
Median D-all P-0.1	8.20370	0.008033

Table 2. Friedman mean rank with Holm's adjusted p values (0.05) for AUROC. The procedure and nomenclature are identical to those in Table 1.

AUROC		
Algorithm	$Friedman'sRank$	$Holm'sAdj - p$
***Median D-all P-0.55**	**5.53704**	-
Median D-some P-0.1	6.24074	1.99405
Median D-one P-0.25	6.42593	1.99405
Median D-one P-0.4	6.42593	1.99405
Median D-all P-0.25	6.46296	1.99405
Median D-one P-0.55	6.46296	1.99405
Median D-one P-0.1	6.59259	1.99405
Median D-all P-0.1	6.61111	1.99405
Median D-all P-0.4	6.74074	1.99405
Median D-some P-0.4	6.77778	1.99405
Median D-some P-0.25	6.79630	1.99405
Median D-some P-0.55	6.92593	1.72664

distance calculation to rely on a couple of interactions, which does not allow for a significant statistical conclusion. However, there is an observable trend towards providing more accurate solutions when the available information is maximized simultaneously through probability and the method of distance calculation.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_1

Regarding the other combinations, another factor that cannot be measured and may have affected the results should be taken into account, granting better precision to combinations with less use of information and worsening the results of others that made greater use of it. In each execution, to form the set of known interactions, a random 5% of the network's gold standard was extracted. Although the number of reference interactions was the same in all executions, their informational value is not necessarily equivalent. That is, the knowledge about the existence of certain interactions may be more valuable than that of others. This is an unpredictable and inevitable fact, since eliminating randomness and establishing fixed reference relationships could bias the results even more.

In the context of the academic networks employed in this study, it is logical to consider extending the winning combination and adding a higher probability of local search to further improve precision levels. However, it should be noted that in such academic problems, the temporal expression levels are simulated from a predefined set of interactions, which ultimately represents the gold standard of the problem. This means that whenever known interactions are added from this gold standard, information from the optimal solution is being shared. This is not the case with real-world networks, and even less so with networks that are intended to be inferred (e.g. in vivo experiments that are not performed yet). In other words, in the cases for which this proposal is intended, the information provided could form part of a good solution known to the domain expert, i.e. a set of interactions that effectively provides a logical explanation of what happens to the gene expression levels during the experiment. However, this may not be the only possible explanation, and there may be other similar alternatives that fit the scenario better. If such information is consistently favored with high probability, it could disturb the direction in which the population evolves during the algorithm execution. Nevertheless, keeping these interactions in mind regularly can bring the population closer to a high-potential zone without condemning the evolution to a possible local minimum.

Given that the optimal solution for these real-world networks intended to be inferred is unknown, the deviation that can be caused by overusing local search could be critical. Therefore, in this case, the most intelligent stance is caution rather than blindly parameterizing in full this proposal based on simulated problems without this broader perspective.

Furthermore, even in academic data where the information injected into the local search is part of the optimal solution to the problem, there is a certain risk that a poorly distributed sample of known interactions may end up diverting the evolution of individuals. The deterioration that these cases can cause to the accuracy of the results increases with the probability of local search. Therefore, once again, setting certain limits is a good practice to maintain a balance that ensures the proposal's security.

Therefore, despite the lack of rigorous statistical significance, the combination of distance *all* and probability 0.55 is chosen as the winner, as it has obtained the first position in the ranking for both precision metrics.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_1

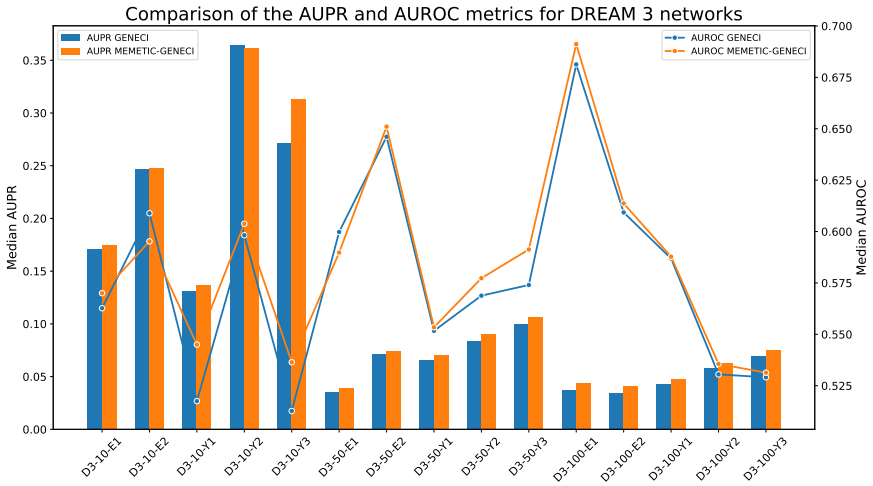

Fig. 3. Comparison of the AUROC and AUPR performance metrics for the GENECI (in blue) and MEMETIC-GENECI (in orange) algorithmic proposals on each of the networks belonging to the third edition of the dream challenges (horizontal axis). For identification, the challenge prefix (D3) is followed by the size of the network (10, 50 or 100) and finally the initial of the organism on which it is based (Y: Yeast, E: E. coli). The bars indicate the medians of the AUPR values and the lines with markers represent the medians of the AUROC values for each network. The AUPR and AUROC values are displayed on separate vertical axes due to their different measurement scales, reserving the left axis for AUPR and the right axis for AUROC.

4.2 Comparison with GENECI

After configuring the parameters of the memetic algorithm, this section quantifies the improvement achieved by this proposal after adding the additional phase of local search. To this end, the precision results presented in the original GENECI article [33] are compared with those obtained by the best parameter combination seen in the previous section. Specifically, for each network and precision metric, the median of GENECI's executions is compared with the median of the executions of the current proposal. This comparison has been decided to be represented visually for editions 3 and 4 of the DREAM challenges (see Figs. 3 and 4 respectively) and presented quantitatively in Table 3 for the IRMA yeast network.

In Fig. 3, it can be observed that the median accuracies of the solutions from the approach in this work surpass, in most cases, the accuracies provided by the original version of GENECI. Upon closer examination, it is noticed that there is a certain relationship between the size of the networks and the stability of this improvement. That is, for larger networks, the enhancement provided by the additional phase of this approach is more robust and decisive. However, in the case of small networks, more varied differences are observed between the

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_1

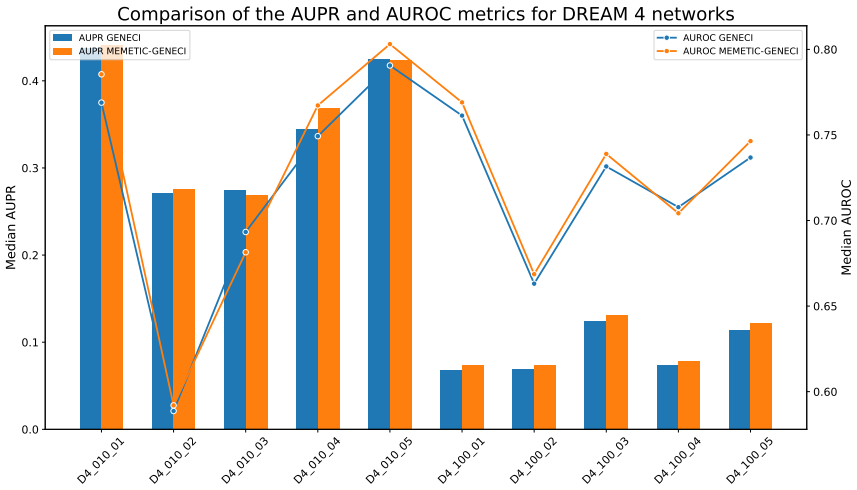

Fig. 4. Comparison of the AUROC and AUPR performance metrics for the GENECI (in blue) and MEMETIC-GENECI (in orange) algorithmic proposals on each of the networks belonging to the fourth edition of the dream challenges (horizontal axis). The nomenclature and interpretation of the graph are identical to those in Fig. 3.

two algorithms, with ties or even a slight lead of the original version appearing in certain cases. This, in a way, validates the choice of the application domain selected for this proposal which, despite being tested on simulated networks, is intended for inferring real-world networks with significantly larger sizes.

Regarding the instability observed for small-sized networks, it is worth mentioning that these cases have a higher probability of obtaining a poorly distributed sample, as the samples have very few interactions and therefore a good representation is not achieved in any case. Therefore, the instability observed in these cases confirms what was previously mentioned in the parameterization, as even with the introduction of correct interactions, a bad sample can divert the proper evolution of the population. However, thanks to the caution and balance achieved in the parameterization, the impact of these exceptional and indetectable cases a priori is quite moderate on the accuracy of the solutions. It is possible to guide and influence the evolution of individuals without completely damaging their convergence.

In Fig. 4, the precision levels of both proposals for networks from DREAM 4 are compared. In this plot, the connection between the size of the networks and the stability of the improvement provided by the local search phase is once again confirmed. Additionally, in this subset of networks, the correlation between both metrics is observed in greater detail. That is, both metrics seem to simultaneously show the same degree of improvement in most cases. This adds a certain reliability to the proposal of this work.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_1

Table 3. Accuracy values for IRMA networks. In this table, a gene network is contemplated for each pair of columns, where in each row the AUPR and AUROC values are provided for each algorithm.

Técnica	**IRMA_switch-off**		**IRMA_switch-on**	
	AUROC	**AUPR**	**AUROC**	**AUPR**
Median GENECI	0.8611	0.7865	0.8889	0.75
Median MEMETIC-GENECI	0.8611	0.7865	0.8939	0.7549

Finally, in Table 3, the precision levels for the yeast network of IRMA are presented. In this case, given that it is such a small network with such a high initial precision level, the margin for improvement is minimal. Additionally, the information available in the set of known interactions is extremely limited, around 1 interaction (the minimum allowed). Nevertheless, a subtle improvement has been achieved in the "switch-on" version, maintaining exactly identical values for the "switch-off" instance. The fact that identical values are obtained is due to the small size of the network, causing precision values to be quite staggered.

After analyzing all the sets of networks, it can be checked how the memetic proposal surpasses GENECI in the majority of cases. To provide greater rigor to this comparison, the Wilcoxon test has been calculated, which has provided a p-value of 2.468690e-03 for AUROC and 1.592934e-05 for AUPR. That is, the improvement in the precision of the results is statistically significant.

The ability to achieve statistically significant improvements with such a restricted sample of known interactions (5% of the gold standard) highlights the algorithm's efficacy in integrating and maximizing the informational value of a limited data set. This is especially crucial in the field of computational biology, where the complete and accurate availability of data can be a constant challenge.

It is worthy to note that thanks to the precautions taken during parameterization, this proposal has demonstrated robustness and reliability. During the experimentation, the subset of interactions designated to form the reference in the local search phase was chosen randomly. This random choice has led to the emergence of poorly distributed samples that could disturb the optimization of the population. However, it has been shown that the impact on the deterioration of accuracy has been minimal in these exceptional cases.

Furthermore, it is also important to comment that this proposal has managed to improve results in a set of extensively worked and studied benchmarking networks, whose margin for improvement was initially very limited. The algorithm's ability to find and exploit areas for improvement in these networks indicates its potential to inject the knowledge provided by the expert and maximize its use to discover novel insights in the data.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_1

5 Conclusions and Future Work

This work presents a novel memetic algorithm for the inference of gene regulatory networks (GRNs), that incorporates a local search phase to leverage prior knowledge of gene interactions. This model was applied to a set of networks widely used as benchmarks in the field, which consists of several DREAM challenge networks and the IRMA yeast network. Finally, a 5% of the known interactions of the gold standards were extracted to feed the local search phase of the algorithm, which modifies the individuals to approximate their consensus networks to the known interactions. Results demonstrate a statistically significant improvement in the inference accuracy compared with the previous GENECI model.

The significance of these findings lies in the algorithm's ability to effectively utilize minimal prior knowledge to guide the evolution of gene regulatory network inferences, offering a more precise and adaptable tool for genomic research. This advancement is particularly relevant in the context of computational biology, where the accurate inference of GRNs is crucial for understanding complex biological processes and diseases at the molecular level.

Looking ahead, one promising direction involves dissecting the aggregate terms of the consensus optimization model into multiple objectives, which could enable a more nuanced optimisation process that better captures the complexity of biological networks. In addition, it is essential to evaluate both the original and improved algorithms against a broader academic benchmark. Such extended testing could facilitate more substantial progress towards applications in real-world networks, where the complexities and scale of the data present unique challenges and opportunities for advancing the field of bioinformatics.

Acknowledgments. This work has been partially funded by grant (funded by MCIN/AEI/10.13039/501100011033/) PID2020-112540RB-C41, AETHER-UMA (A smart data holistic approach for context-aware data analytics: semantics and context exploitation). Adrián Segura-Ortiz is supported by Grant FPU21/03837 (Spanish Ministry of Science, Innovation and Universities).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Altay, G., Emmert-Streib, F.: Inferring the conservative causal core of gene regulatory networks. BMC systems biology **4**(1), 1–13 (2010)
2. Aluru, M., Shrivastava, H., Chockalingam, S.P., Shivakumar, S., Aluru, S.: Engrain: a supervised ensemble learning method for recovery of large-scale gene regulatory networks. Bioinformatics **38**(5), 1312–1319 (2022)
3. Baliarsingh, S.K., Muhammad, K., Bakshi, S.: Sara: a memetic algorithm for high-dimensional biomedical data. Applied Soft Computing **101**, 107009 (2021)
4. Cantone, I., Marucci, L., et al.: A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. Cell **137**(1), 172–181 (2009)

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_1

5. Correa, L., Borguesan, B., Farfan, C., Inostroza-Ponta, M., Dorn, M.: A memetic algorithm for 3d protein structure prediction problem. IEEE/ACM transactions on computational biology and bioinformatics **15**(3), 690–704 (2016)
6. Escorcia-Rodríguez, J.M., Gaytan-Nuñez, E., et al.: Improving gene regulatory network inference and assessment: The importance of using network structure. Frontiers in Genetics **14**, 1143382 (2023)
7. Faith, J.J., Hayete, B., Thaden, J.T., et al.: Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS biology **5**(1), e8 (2007)
8. Finkle, J.D., Wu, J., Bagheri, N.: Windowed granger causal inference strategy improves discovery of gene regulatory networks. Proceedings of the National Academy of Sciences **115**, 2252–2257 (2018)
9. Fujii, C., Kuwahara, H., Yu, G., et al.: Learning gene regulatory networks from gene expression data using weighted consensus. Neurocomputing **220**, 23–33 (2017)
10. Gan, Y., Hu, X., Zou, G., et al.: Inferring gene regulatory networks from single-cell transcriptomic data using bidirectional rnn. Frontiers in Oncology **12** (2022)
11. García-Nieto, J., Nebro, A.J., Aldana-Montes, J.F.: Inference of gene regulatory networks with multi-objective cellular genetic algorithm. Computational Biology and Chemistry **80**, 409–418 (2019)
12. Ghazikhani, A., Akbarzadeh, T., Monsefi, R.: Genetic regulatory network inference using recurrent neural networks trained by a multi agent system. 2011 1st International eConference on Computer and Knowledge Engineering (ICCKE) (2011)
13. Gong, M., Peng, Z., Ma, L., Huang, J.: Global biological network alignment by using efficient memetic algorithm. IEEE/ACM transactions on computational biology and bioinformatics **13**(6), 1117–1129 (2015)
14. Han, P., Gopalakrishnan, C., Yu, H., Wang, E.: Gene regulatory network rewiring in the immune cells associated with cancer. Genes **8**(11), 308 (2017)
15. Hillerton, T., et al.: Fast and accurate gene regulatory network inference by normalized least squares regression. Bioinformatics **38**(8), 2263–2268 (2022)
16. Hurtado, S., Garcia-Nieto, J., Navas-Delgado, I., Nebro, A.J., Aldana-Montes, J.F.: Reconstruction of gene regulatory networks with multi-objective particle swarm optimisers. Applied Intelligence **51**, 1972–1991 (2021)
17. Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., Geurts, P.: Inferring regulatory networks from expression data using tree-based methods. PloS one **5**(9), e12776 (2010)
18. Jiang, H., Turki, T., Zhang, S., Wang, J.T.: Reverse engineering gene regulatory networks using graph mining. In: International Conference on Machine Learning and Data Mining in Pattern Recognition. pp. 335–349. Springer (2018)
19. Khojasteh, H., Khanteymoori, A., Olyaee, M.H.: Engrnt: Inference of gene regulatory networks using ensemble methods and topological feature extraction. Informatics in Medicine Unlocked **27**, 100773 (2021)
20. Kizaki, N., et al.: The inference method of the gene regulatory network with a majority rule. Nonlinear Theory and Its Applications, IEICE **6**, 226–236 (2015)
21. de Lima Corrêa, L., Dorn, M.: A multi-population memetic algorithm for the 3-d protein structure prediction problem. Swarm and Evolutionary Computation **55**, 100677 (2020)
22. Liu, L., et al.: Reconstructing gene regulatory networks via memetic algorithm and lasso based on recurrent neural networks. Soft Computing **24**, 4205–4221 (2020)
23. Liu, W., Jiang, Y., Peng, L., Sun, X., Gan, W., Zhao, Q., Tang, H.: Inferring gene regulatory networks using the improved markov blanket discovery algorithm. Interdisciplinary Sciences: Computational Life Sciences pp. 1–14 (2022)

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_1

24. Marbach, D., Costello, J.C., Küffner, R., et al.: Wisdom of crowds for robust gene network inference. Nature Methods 2012 9:8 **9**, 796–804 (7 2012)
25. Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., et al.: Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In: BMC bioinformatics. vol. 7, pp. 1–15. Springer (2006)
26. Meyer, P., Saez-Rodriguez, J.: Advances in systems biology modeling: 10 years of crowdsourcing dream challenges. Cell Systems **12**(6), 636–653 (2021)
27. Moerman, T., Aibar Santos, S., Bravo González-Blas, C., Simm, J., Moreau, Y., Aerts, J., Aerts, S.: GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics **35**(12), 2159–2161 (11 2018)
28. Narasimhan, S., Rengaswamy, R., Vadigepalli, R.: Structural properties of gene regulatory networks: Definitions and connections. IEEE/ACM Transactions on Computational Biology and Bioinformatics **6**(1), 158–170 (2009)
29. Nazarieh, M., Wiese, A., Will, T., Hamed, M., Helms, V.: Identification of key player genes in gene regulatory networks. BMC Systems Biology **10**, 1–12 (2016)
30. Parikshak, N.N., et al.: Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders **16**(8), 441–458 (2015)
31. Peignier, S., Sorin, B., Calevro, F.: Ensemble learning based gene regulatory network inference. In: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI). pp. 113–120 (2021)
32. Schmitt, P., et al.: Grenadine: A data-driven python library to infer gene regulatory networks from gene expression data. Genes **14**(2) (2023)
33. Segura-Ortiz, A., García-Nieto, J., et al.: Geneci: A novel evolutionary machine learning consensus-based approach for the inference of gene regulatory networks. Computers in Biology and Medicine **155**, 106653 (2023)
34. Skok Gibbs, C., et al.: High-performance single-cell gene regulatory network inference at scale: the inferelator 3.0. Bioinformatics **38**(9), 2519–2528 (2022)
35. Watanabe, Y., Seno, S., Takenaka, Y., Matsuda, H.: An estimation method for inference of gene regulatory net-work using bayesian network with uniting of partial problems. BMC Genomics **13**, S12 (2012)
36. Wijst, M.G.V.D., Vries, D.H.D., Brugge, H., Westra, H.J., Franke, L.: An integrative approach for building personalized gene regulatory networks for precision medicine. Genome Medicine 2018 10:1 **10**, 1–15 (12 2018)
37. Wu, J., Zhao, X., Lin, Z., Shao, Z.: Large scale gene regulatory network inference with a multi-level strategy. Molecular BioSystems **12**, 588–597 (2016)
38. Yang, B., Xu, Y.: Reconstructing gene regulation network based on conditional mutual information. Proceedings of the 2017 International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2017) (2017)
39. Yasuki, H., Kikuchi, M., Kurokawa, H.: Inferring method of the gene regulatory networks using neural networks adopting a majority rule. The 2011 International Joint Conference on Neural Networks (2011)
40. Yin, F., Zhou, J., Xie, W., Zhu, Z.: Inferring sparse genetic regulatory networks based on maximum-entropy probability model and multi-objective memetic algorithm. Memetic Computing **15**(1), 117–137 (2023)
41. Zarayeneh, N., Ko, E., Oh, J.H., Suh, S.C., Liu, C., Gao, J., Kim, D., Kang, M.: Integration of multi-omics data for integrative gene regulatory network inference. International Journal of Data Mining and Bioinformatics **18**, 223 (2017)
42. Zhao, M., He, W., Tang, J., Zou, Q., Guo, F.: A hybrid deep learning framework for gene regulatory network inference from single-cell transcriptomic data. Briefings in Bioinformatics **23**(2), bbab568 (2022)

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_1