Solution Exploration using Multi-Objective Genetic Algorithm for Determining Experiment Candidate

Lorenzo Perino
Doshisha University
Graduate School of Life and Medical Sciences
Kyoto, Japan
lorenzo@mis.doshisha.ac.jp

Akihiro Fujii
Doshisha University
Graduate School of Life and Medical Sciences
Kyoto, Japan

Tsuyoshi Waku
Doshisha University
Graduate School of Life and Medical Sciences
Kyoto, Japan

Akira Kobayashi
Doshisha University
Faculty of Life and Medical Sciences
Kyoto, Japan

Satoru Hiwa
Doshisha University
Faculty of Life and Medical Sciences
Kyoto, Japan

Tomoyuki Hiroyasu
Doshisha University
Faculty of Life and Medical Sciences
Kyoto, Japan

ABSTRACT

Solving problems involves the following two phases. In the first phase, detail of the problem are determined and the solution is found according to these condition. In the second phase, the results are used to narrow down any remaining ambiguities in the problem. In terms of the flow of a river, the former lies upstream of solving the problem while the latter lies downstream. Multiobjective genetic algorithms (GAs) is are able to find possible solution sets involving trade-offs among several different objective functions. In this study, we use a multiobjective GA to grasp variety types of solutions, not solve the problem directly. In other words, we use it for the upstream problem-solving step. As a case study of using multiobjective GAs to explore solutions, we identify cancer cells where the Nrf3 transcription factor is active and consider the problem of determining which genes to focus on in experiments based on that information. In this case, we selected gene candidates that are likely to be associated with Nrf3 activity and experiments (which previously had to be carried out exhaustively) are currently being carried out to confirm these results.

CCS CONCEPTS

• **Applied computing** → **Computational genomics**; *Recognition of genes and regulatory elements*; *Biological networks*; Transcriptomics;

GECCO '18 Companion, July 15–19, 2018, Kyoto, Japan

ACM ISBN 978-1-4503-5764-7/18/07…$15.00
https://doi.org/10.1145/3205651.3208318

KEYWORDS

Multiobjective genetic algorithms, solution exploration, Nrf3 tanscription factor

ACM Reference Format:
Lorenzo Perino, Akihiro Fujii, Tsuyoshi Waku, Akira Kobayashi, Satoru Hiwa, and Tomoyuki Hiroyasu. 2018. Solution Exploration using Multi-Objective Genetic Algorithm for Determining Experiment Candidate. In *GECCO '18 Companion: Genetic and Evolutionary Computation Conference Companion, July 15–19, 2018, Kyoto, Japan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3205651.3208318

1 INTRODUCTION

When solving a given decision-making problem, if it can be formulated as an optimization problem, solution candidates can be found via numerical computation; we can then use the optimization results to defined in detail. In such cases, it can be helpful to select several solution candidates and analyze them to obtain more information about the problem. This kind of so-called upstream analysis can enable us to derive a more detailed specification of the target problem.

In the engineering design field, many design exploration frameworks have been introduced [3, 9]. Obayashi et al. introduced one that used a multiobjective genetic algorithm (GA) to find several candidates for the optimal design at an early stage of the optimization process [11]. In their framework, the optimal design candidates are derived as a Pareto solution set and are then analyzed further to identify the target problem's characteristics and a strategy to improve the optimization process and hence derive better solutions. After this upstream analysis step, the design specification is confirmed and the design process is continued to take advantage of optimal design found.

Hiwa et al. generalized this idea and extended it to decision-making problems [7, 8]. In their approach, the candidate Pareto solutions are generated by multiobjective optimization and the details of the problem are defined by using principal component analysis to extract the solutions' meaningful features and data clustering to classify them. This type

of upstream problem-solving process is called "solution exploration" in this study.

This approach can be applied to biomedical science questions such as finding the physiological relationship between the transcription factor that activates malignant cancer cells and the subsequent cell proliferation and cancer. Here huge numbers of experiments often have to be carried out to identify genes that are highly correlated with the appearance of the target transcription factor because researchers initially have little information about which cancer cells or genes should be considered. Such experiments are extremely costly. Here, we assume that the root of this problem lies in the upstream decision-making process for deciding which of the many possible experimental candidates to consider.

For such problems, a reasonable approach is to conduct general-purpose experiments including all the candidates to increase the likelihood of identifying promising candidates. However, the effectiveness of this approach tends to decline as the number of initial candidates increases. There is a trade-off between the versatility of the experimental method and the usefulness of its results (i.e., the probability of finding meaningful results).

The first step in dealing with this issue is to confirm this trade-off by modeling the problem as a multiobjective optimization problem and solving it using a suitable algorithm. Once the Pareto solution set has been derived, we can determine good candidates for further experiments or promising experimental approaches by extracting the candidates (i.e., candidate samples or sample characteristics) that appear most frequently.

In this paper, we apply this solution exploration method to the problem of deciding which samples to prioritize for further analysis in laboratory experiments. Here, we focus on the problem of determining the cancer cells associated with the Nrf3 transcription factor. This factor activates malignant cancer cells and causes cell proliferation [2, 12] and is attracting attention as a new target for cancer treatment. If particular genes are highly correlated with Nrf3 activity, they are likely to influence the onset of cancer.

Conventional studies require large numbers of laboratory experiments to check for correlations between Nrf3 and different genes in various types of cancer cells. Our goal in applying a solution exploration approach to this problem is therefore to narrow down the list of candidate cells and genes to reduce the number of laboratory experiments required. In particular, we focus on the gene expression correlations between Nrf3 and the TGF-β/SMAD signal factors [5, 13], a group of genes related to tissue fibrosis and cancer metastasis, and identify the cells that should be preferentially investigated.

2 SOLUTION EXPLORATION FRAMEWORK

Figure 1 illustrates the solution exploration process we apply to find promising experimental candidates.

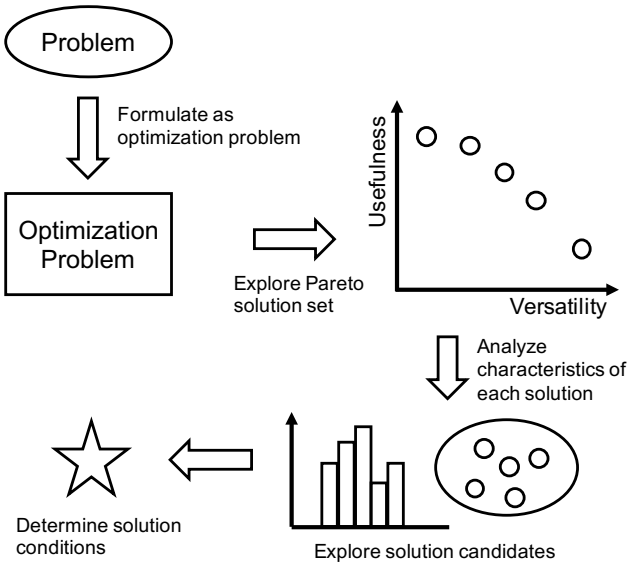


Figure 1: Solution Exploration Framework

Some trade-offs are involved in initially determining the problem conditions. For example, in product design, reducing the production cost and improving the product's quality or performance are often competing factors. In our current problem of selecting candidates for laboratory experiments, there is a trade-off between the versatility of the experimental approach and the usefulness of the results. Our solution exploration framework involves the following three steps: 1) search using a multiobjective GA, 2) analyze the resulting Pareto solution set, and 3) select the candidates. The simplest way to analyze the characteristics of the Pareto solutions is to identify which elements appear most frequently, as these are likely to be important for the target problem.

In the next section, we apply this framework to concrete experimental problems and explain the process in detail.

3 APPLICATION TO CANCER CELL DETERMINATION PROBLEM

Using the solution exploration framework described in the previous section, we now propose a method of determining cells where the Nrf3 expression level is highly correlated with those of the signal factors.

3.1 Nrf3 transcription factor

Transcription factors are protein groups that bind to specific DNA sequences. They are classified as either activators (which increase the expression of particular genes) or repressors (which decrease the expression levels). They regulate the process of binding to regulatory regions, such as promoters and enhancers that control transcription and transcribing genetic information from DNA into RNA.

The Nrf3 gene belongs to the stress response-related transcription factor group called the Cap'n'collar (CNC) family.

This family consists of six transcription factors: p45/NF-E2, Nrf1, Nrf2, Nrf3 (which acts as a transcription activator), and Bach1 and Bach2 (transcription repressors). Nrf3 is known to be highly expressed in various cancer cells and is also believed to activate malignant cancer cells and cause cell proliferation. In lung cancer-derived cells, proteolytic enzymes are activated when Nrf3 is highly expressed, which may bring about carcinogenesis by degrading the tumor suppressor gene. The role of this Nrf3-controlled gene expression mechanism in cancer development has attracted significant attention from researchers.

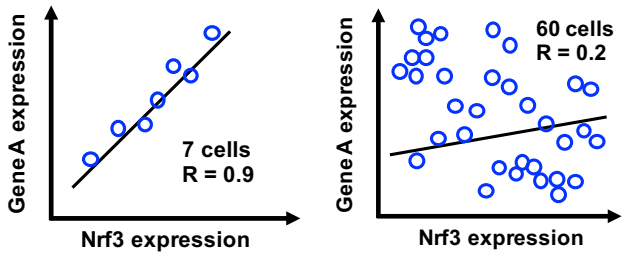
Transforming growth factor β (TGF-β) suppresses the growth of many types of cells, such as epithelial cells, and can lead to cell cancerization if it fails to suppress cell growth. In addition, it can induce epithelial to mesenchymal transitions that can cause tissue fibrosis and allows epithelial cells to be converted into mobile mesenchymal cells that can invade cancer cells and cause metastasis [10].

Three other types of proteins have structures similar to that of TGF-β: TGF-β1, 2, and 3. These activate two types of serine threonine kinase-type receptors, TGF-βR1 and -βR2, resulting in the activation of so-called SMAD signal factors. When TGF-βR1 activates SMAD2 and SMAD3 due to the action of TGF-β, SMAD4 is also bound, forming an SMAD complex. This then translocates into the nucleus and binds to the DNA and various transcription factors, thereby regulating the transcription of numerous genes. Here, we focus on the expression of Nrf3 and its correlations with those of seven genes, namely, GDF5, SMAD3, SMURF1, SMURF2, TGF-β2, TGF-βR2, and TGIF, which are involved with the TGF-β/SMAD signaling mechanism. If these correlations are high in cancer cells, then Nrf3 may be an important factor in the TGF-β/SMAD signaling mechanism.

3.2 Problem definition

DNA microarrays [1, 14] are used to analyze genes, and in this case, they were used to generate gene expression data for various cancer cells. Using these data, we can identify the cancer cells with high correlations between Nrf3 expression and those of the other genes and then select the cells where these genes were most active. However, although methods that target many types of cancer cells in this way are more versatile, the expression correlations tend be lower because cells not associated with Nrf3 are also included in the sample set, as illustrated in Figure 2.

On the other hand, if we look at a narrower range of cancer cell types, we may miss groups of cancer cells showing high correlations. Previous studies have explored which genes are highly correlated with Nrf3 using a cell set called the National Cancer Institute (NCI)-60 cancer cell line panel, which comprises 60 different human cancer cell lines (e.g., leukemia, malignant melanoma, colon, central nervous system, lung, ovary, breast, prostate, and kidney). While it is not necessary to analyze all 60 cell lines, there has been no way to reduce the number considered because there was no way to specify which cells were important for detecting the most significant correlations between Nrf3 and other genes.



(a)Large sample case (b)Small sample case

Figure 2: Gene correlations using (a) a small number of carefully selected cells and (b) all cells.

We used our solution expression framework to overcome this problem. Specifically, we focused on the correlations between Nrf3 and the genes representing the TGF-β/SMAD signal factors associated with tissue fibrosis and cancer metastasis, extracting the set of cancer cells with strong correlations via multiobjective optimization.

To achieve this, we formulated this as the optimization problem of deciding which of the 60 cancer cell lines showed the highest correlations. In this problem, each of 60 decision variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$ ($x_k \in \{0, 1\}, d = 60$) indicates whether or not one of the NCI-60 cell lines is selected when searching for the optimal group of cells. For each decision variable, the value 1 means that the corresponding cancer cell is "of interest," while 0 means the opposite.

In this problem, to aim is to obtain the optimal group of cancer cells with the most significant correlations between Nrf3 expression and those of the TGF-β/SMAD signal factors. We therefore calculated the Pearson correlation coefficients R between Nrf3 and the other genes and utilized its square as the objective function. Figure 3 illustrates the objective function evaluation procedure.

There is, however, a trade-off between the number of cells selected and the probability of finding potential gene expression correlations. In order to investigate this trade-off, we used the number of cells selected as a second objective function. Both objective functions were maximized by the optimization process.

The resulting multiobjective optimization problem was formulated as follows:

$$\begin{aligned} \text{maximize} \quad & f_1(\boldsymbol{x}) = \{R(\boldsymbol{E}_{\text{Nrf3}}(\boldsymbol{x}), \boldsymbol{E}_{\text{A}}(\boldsymbol{x}))\}^2 \\ & f_2(\boldsymbol{x}) = ||\boldsymbol{x}||_1 \\ \text{subject to} \quad & x_k \in \{0, 1\} \ (k = 1, \ldots, d) \end{aligned} \tag{1}$$

where $E_{\text{Nrf3}}(\boldsymbol{x})$ and $E_{\text{A}}(\boldsymbol{x})$ denote the expression levels of Nrf3 and the target gene A, respectively, calculated for the sample set $\boldsymbol{x}$.

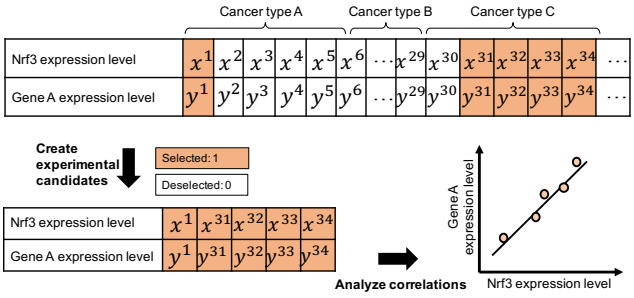


Figure 3: Objective function evaluation procedure

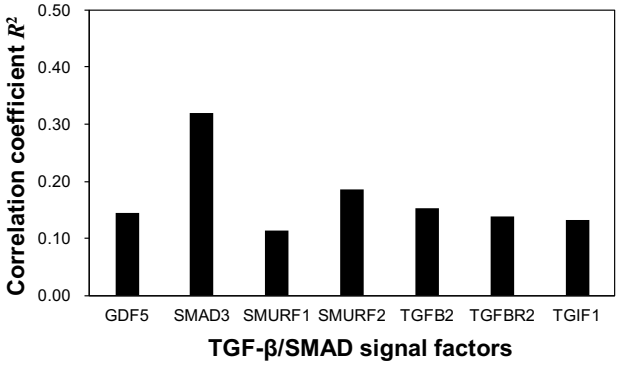


Figure 4: Correlation between Nrf3 and the TGF-β/SMAD signal factors, calculated using all 60 cell lines in NCI-60

4 NUMERICAL EXPERIMENTS

4.1 Baseline method

In this paper, we focus on finding correlations between the Nrf3 expression level and those of the seven TGF-β/SMAD signal factors in cancer cells to evaluate the effectiveness of the proposed method. The seven genes are GDF5, SMAD3, SMURF1, SMURF2, TGF-β2, TGF-βR2, and TGIF. In the baseline approach, all 60 cells in the NCI-60 panel were utilized for the correlation analysis.

4.2 Proposed method

Here, we solved the multiobjective optimization problem using the nondominated sorting GAs II (NSGA-II) [4], adopting the binary genotype representation, uniform crossover (crossover rate = 1.0), and bit-flip mutation (mutation rate = 1/chromosome length). Table 1 summarizes the parameters used. We carried out 10 independent runs for each signal factor gene. The algorithm was implemented using the Distributed Evolutionary Algorithm in Python library [6]. Since each of the seven genes was optimized separately, we obtained seven Pareto solution sets.

Table 1: NSGA-II parameters used

Parameter	Value
Population size	100
Chromosome length	60
Number of generations	200
Crossover rate	1.0
Mutation rate	1/60

5 RESULTS AND DISCUSSIONS

5.1 Baseline method

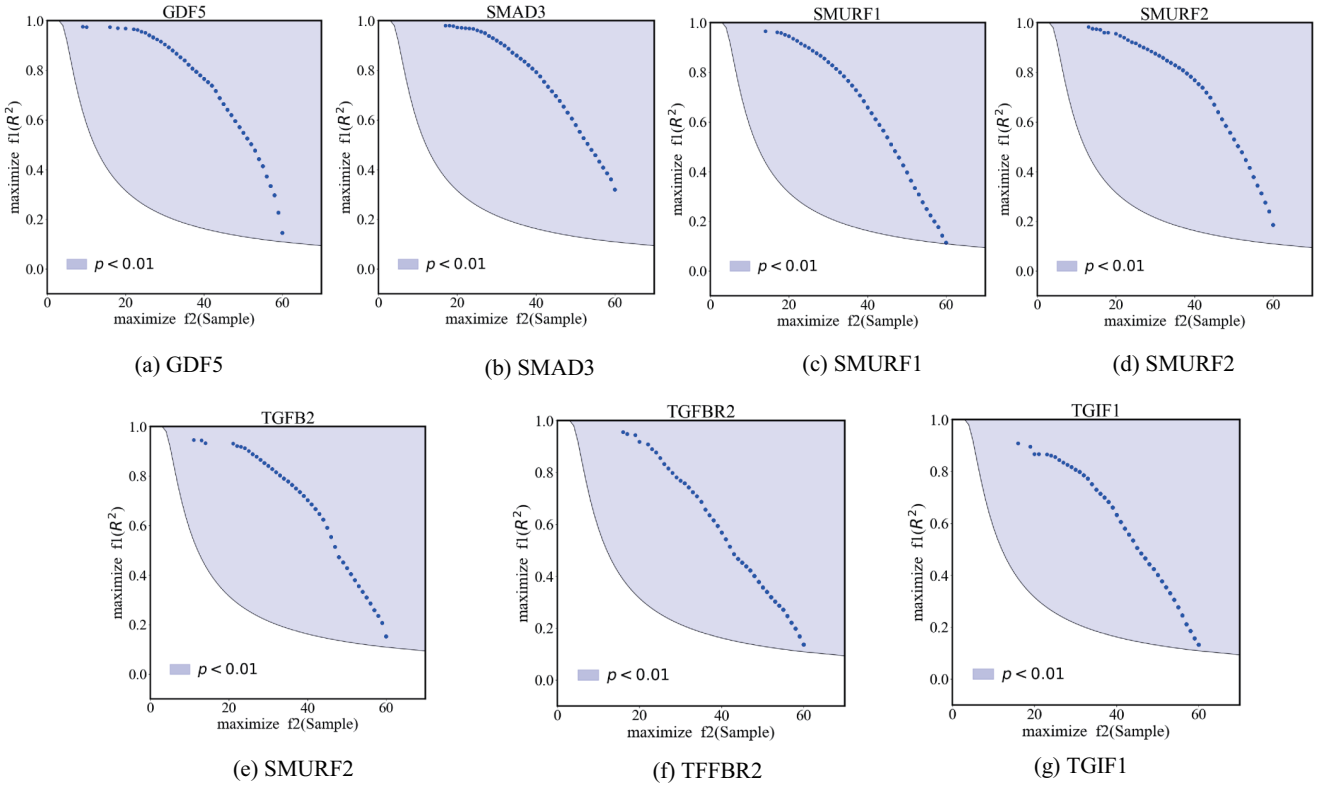
Figure 4 shows the results of the baseline analysis, which calculated the correlations between the Nrf3 expression and those of the other seven genes for all the cancer cell lines in NCI-60. The R2 value was less than 0.4 for all seven genes, meaning it was unable to find any strong correlations between Nrf3 and the TGF-β/SMAD signal factors.

5.2 Proposed method

Figure 5 shows the Pareto solution sets (dots) obtained by NSGA-II for the seven genes. The purple regions indicate where the correlation between Nrf3 and the target gene in each Pareto solution (corresponding to a set of cancer cells) is significant. This indicates that all of the obtained Pareto solutions lie in reliable areas. Furthermore, we found higher correlations between Nrf3 and the seven genes than with the baseline method. From a biological viewpoint, it is important to obtain higher correlation levels and to understand why these strong correlations are present. We therefore investigated which of the cancer cell lines were most frequently selected for the optimized sample set. After discussions with biological experts, we counted how frequently the cell lines were selected for all seven Pareto sets combined. We only counted solutions with $R^2 \geq 0.9$ because we were aiming to discover which cells were most likely to show that Nrf3 is involved in the TGF-β/SMAD signaling mechanism. Figure 6 shows the total number of times each NCI-60 cell line was selected for a Pareto solution set, in descending order of the number of times selected.

The top three cell lines selected for the Pareto sets were as follows: ME_MDA_MB_435 (24 times), BR_T47D (23 times), and ME_MALME_3M (21 times). All of which were sampled from melanomas and breast cancers. Since these cell types can be regarded as likely to exhibit high correlations between Nrf3 expression and those of multiple TGF-β/SMAD signaling factors, we believe that Nrf3 affects the TGF-β/SMAD signaling mechanism and is also associated with the incidence of melanoma and breast cancer.

In contrast, the BR_HS578T breast cancer-derived cells, CNS.SF_295 central nervous system-derived cells, and LE.RPMI_8226 leukemia-derived cells were never selected. This suggests that there are also cancer cells in which Nrf3 expression is not correlated with those of the seven genes, and that Nrf3 is not involved in the TGF-β/SMAD signaling mechanism in these cells. These results will contribute to



(a) GDF5 (b) SMAD3 (c) SMURF1 (d) SMURF2

(e) SMURF2 (f) TFFBR2 (g) TGIF1

Figure 5: Pareto solution sets of seven genes of TGF-β/SMAD signal factor

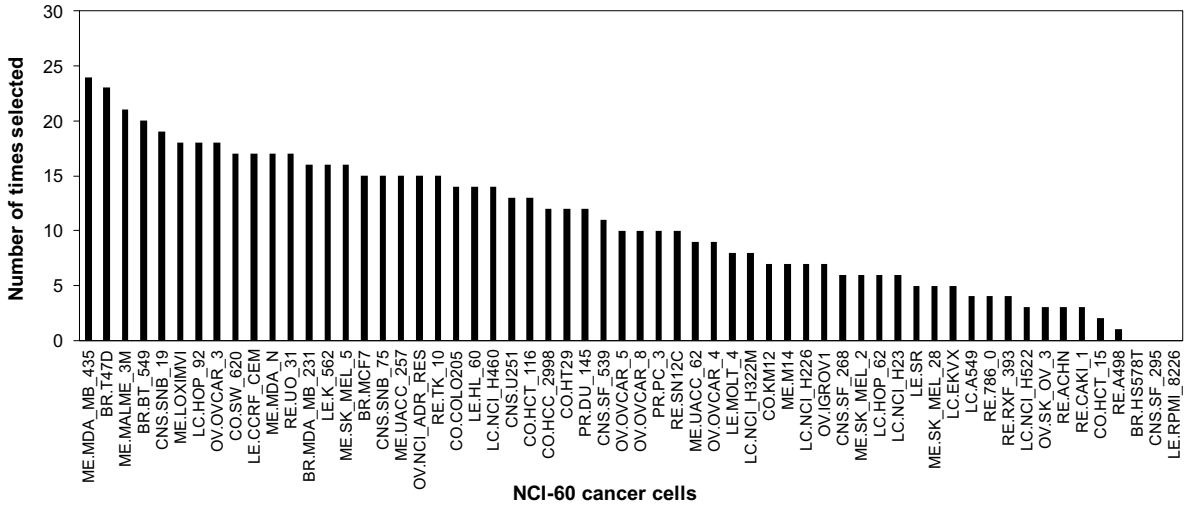


Figure 6: Pareto solution sets (dots) for the seven TGF-β/SMAD signal factors

preventing further laboratory experiments examining cells where there is no Nrf3 association.

Using these results, we were able to obtain experimental candidates involving Nrf3 transcription factor activity. However, our current study has some limitations that will need to be addressed in future work.

First, we have only used NSGA-II in this preliminary study, and the most appropriate parameter values and numbers of iterations and runs required to produce stable results need to be investigated further. In addition, other state-of-the-art algorithms, such as MOEA/D and NSGA-III, may potentially work well, so we should assess which algorithms are suitable for the cancer cell determination problem. Second, biological experts are currently attempting to confirm that the cancel cell candidates determined by our method are actually associated with a combined Nrf3 and TGF-β/SMAD signaling mechanism via laboratory experiments.

6 CONCLUSIONS

The solution exploration framework determines the target problem details based on the solutions produced by multi-objective optimization. In this paper, we have examined the effectiveness of this framework for determining which cancer cell samples involve Nrf3 transcription factor activity. In our numerical experiments, we used the proposed method to automatically extract the cancer cells most likely to yield correlations between Nrf3 expression and those of seven other genes from 60 cell line candidates. Further analysis of the Pareto solution sets then revealed which cancer cells were most likely to feature strong Nrf3 activity. Laboratory experiments are currently being conducted to verify these results.

REFERENCES

[1] H. Causton, J. Quackenbush, and A. Brazma. 2003. *Microarray Gene Expression Data Analysis: A Beginner's Guide*. Wiley. https://books.google.co.jp/books?id=icyw4AiNO88C

[2] Grégory Chevillard and Volker Blank. 2011. NFE2L3 (NRF3): the Cinderella of the Cap 'n' Collar transcription factors. *Cellular and Molecular Life Sciences* 68, 20 (2011), 3337–3348.

[3] Haejin Choi, David L McDowell, Janet K Allen, David Rosen, and Farrokh Mistree. 2008. An inductive design exploration method for robust multiscale materials design. *Journal of Mechanical Design* 130, 3 (2008), 031402.

[4] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation* 6, 2 (2002), 182–197.

[5] Rik Derynck, Julie A Jarrett, Ellson Y Chen, Dennis H Eaton, John R Bell, Richard K Assoian, Anita B Roberts, Michael B Sporn, and David V Goeddel. 1985. Human transforming growth factor-β complementary DNA sequence and expression in normal and transformed cells. *Nature* 316, 6030 (1985), 701.

[6] Félix-Antoine Fortin, François-Michel De Rainville, Marc-André Gardner, Marc Parizeau, and Christian Gagné. 2012. DEAP: Evolutionary algorithms made easy. *Journal of Machine Learning Research* 13, Jul (2012), 2171–2175.

[7] Satoru Hiwa, Tomoyuki Hiroyasu, and Mitsunori Miki. 2014. Design Mode Analysis on a Pareto Solution Set for Decision-Making Support. 2014 (2014), 1–12. https://doi.org/10.2514/1.52081

[8] Satoru Hiwa, Mitsunori Miki, and Tomoyuki Hiroyasu. 2017. Validity of decision mode analysis on an ROI determination problem in multichannel fNIRS data. *Artificial Life and Robotics* 22, 3 (sep 2017), 336–345. https://doi.org/10.1007/s10015-017-0362-5

[9] Sungwoo Jang, Chung Hyun Goh, and Hae-Jin Choi. 2015. Multiphase design exploration method for lightweight structural design: Example of vehicle mounted antenna-supporting structure. *International Journal of Precision Engineering and Manufacturing-Green Technology* 2, 3 (2015), 281–287.

[10] Myriam Labelle, Shahinoor Begum, and Richard O Hynes. 2011. Direct signaling between platelets and cancer cells induces an epithelial-mesenchymal-like transition and promotes metastasis. *Cancer cell* 20, 5 (2011), 576–590.

[11] Seiichiro Morizawa, Taku Nonomura, Shigeru Obayashi, Akira Oyama, and Kozo FujiiI. 2016. Multiobjective Design Exploration of Propeller Airfoils at Low-Reynolds and High-Mach Number Conditions towards Mars Airplane. *Transactions of the Japan Society for Aeronautical and Space Sciences, Aerospace Technology Japan* 14, ists30 (2016), 47–53. https://doi.org/10.2322/tastj.14.Pk_47

[12] Ko Onodera, Jordan A Shavit, Hozumi Motohashi, Masayuki Yamamoto, and James Douglas Engel. 2000. Perinatal synthetic lethality and hematopoietic defects in compound mafG:: mafK mutant mice. *The EMBO journal* 19, 6 (2000), 1335–1345.

[13] Anita B Roberts, Michael B Sporn, Richard K Assoian, Joseph M Smith, Nanette S Roche, Lalage M Wakefield, Ursula I Heine, Lance A Liotta, Vincent Falanga, and John H Kehrl. 1986. Transforming growth factor type beta: rapid induction of fibrosis and angiogenesis in vivo and stimulation of collagen formation in vitro. *Proceedings of the National Academy of Sciences* 83, 12 (1986), 4167–4171.

[14] Dov Stekel. 2003. *Microarray Bioinformatics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511615535