# Optimization of Rifamycin B Fermentation in Shake Flasks Via a Machine-Learning-Based Approach

**Prashant M. Bapat, Pramod P. Wangikar**

*Department of Chemical Engineering, Indian Institute of Technology, Bombay, Powai, Mumbai 400 076, India; telephone: +91 22 2576 7232; fax: +91-22-2572 6895; e-mail: pramodw@iitb.ac.in*

*Received 5 June 2003; accepted 15 December 2003*

*Published online 19 February 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/bit.20056*

**Abstract:** Rifamycin B is an important polyketide antibiotic used in the treatment of tuberculosis and leprosy. We present results on medium optimization for Rifamycin B production via a barbital insensitive mutant strain of *Amycolatopsis mediterranei* S699. Machine-learning approaches such as Genetic algorithm (GA), Neighborhood analysis (NA) and Decision Tree technique (DT) were explored for optimizing the medium composition. Genetic algorithm was applied as a global search algorithm while NA was used for a guided local search and to develop medium predictors. The fermentation medium for Rifamycin B consisted of nine components. A large number of distinct medium compositions are possible by variation of concentration of each component. This presents a large combinatorial search space. Optimization was achieved within five generations via GA as well as NA. These five generations consisted of 178 shake-flask experiments, which is a small fraction of the search space. We detected multiple optima in the form of 11 distinct medium combinations. These medium combinations provided over 600% improvement in Rifamycin B productivity. Genetic algorithm performed better in optimizing fermentation medium as compared to NA. The Decision Tree technique revealed the media–media interactions qualitatively in the form of sets of rules for medium composition that give high as well as low productivity. 

**Keywords:** media balancing; decision tree; fermentation

## INTRODUCTION

Rifamycin B, a prominent member of the ansamycin family of antibiotics, is produced by the actinomycete *Amycolatopsis mediterranei* (Sensi and Thiemann, 1967). Its synthetically modified derivatives, such as rifampicin, are used clinically in the treatment of tuberculosis, leprosy, and AIDS-related mycobacterial infections (Sepkowitz et al., 1995). The potent antibacterial activity of the rifamycins is due to their specific inhibition of bacterial DNA-dependent RNA polymerases, enzymes that are responsible for mRNA biosynthesis in the cell. However, they have relatively low activity against eukaryotic RNA polymerases (Hartmann et al., 1967). The global demand for Rifamycin is increasing tremendously and a need exists for cost-effective production of Rifamycin. One of the strategies to achieve this goal is to carry out optimization of medium. Although Rifamycin was discovered in 1959, information on medium optimization is sparse. The Box–Wilson statistical approach was used by Sensi and Thiemann (1967) to evaluate effects of interactions between the various nutrients on antibiotic production. They reported a twofold increase in Rifamycin B productivity to d 0.57 g $L^{-1}$. Venkateswarlu et al. (1999) studied production of Rifamycin B with *Amycolatopsis mediterranei* (MTCC14) using carbon and locally available cheaper nitrogen sources. The Rifamycin B productivity was improved from 0.65 g $L^{-1}$ to 1.4 g $L^{-1}$. Statistically based experimental methods are known for medium optimization and have been demonstrated to be successful in the context of several different products. Examples include optimization of Lovastatin productivity using response surface methodology (Chang et al., 2002) or citric acid fermentation using the Box–Behnken approach (Chen, 1994). A limitation of these methods is that they can handle a small number of variables and at a small number of levels—usually at only two levels. This is due to the exponential increase in the number of experiments required when the number of variables increases (Marteijn et al., 2003). In addition, these methods require preliminary screening to reduce the number of variables. Here, we apply three machine-learning approaches, Genetic algorithm (GA), Neighborhood analysis (NA), and Decision Tree algorithm (DT) for optimization of Rifamycin B productivity. The first two approaches were used as guided search tools for medium optimization, while decision tree provided rules for synthesizing medium composition that give high or low productivity. Some of these approaches have already been used in biotechnological applications. Weuster-Botz and Wandrey (1995) used GA to optimize a 14-component medium for the production of formate dehydrogenase. Marteijn et al. (2003) used GA for optimizing 11-component medium for insect cell feeding. Genetic algorithm based medium optimization starts

*Correspondence to:* Pramod P. Wangikar

with a randomized population of medium compositions represented as chromosomes with each gene representing a medium component. The initial population is evaluated for a fitness function. The fittest individuals are selected and become parents whereby new child chromosomes are generated by crossover and mutation operation on parent's chromosomes. This next generation of medium composition is evaluated for fitness function and the process of selection, crossover, and mutation continues until a convergence criterion is satisfied (Forrest, 1993; Holland, 1992; Kennedy and Krouse, 1999).

Bradamante et al. (2002) used a decision tree algorithm to determine key variables that correlated with the lovastatin fermentation. A decision tree algorithm (Quinlan, 1986) is used for identifying the most important variable(s) and variable attributes in terms of their correlation with the process outcome. At each node of the tree, the examples being considered are separated into groups. This is a binary split for quantitative data or potentially multileaved for qualitative information. In deciding at which variable and which particular value of that variable to make the branch, the information content of every possible resulting group is calculated. The information content is determined by using Shannon's entropy formula for the group of examples at each node as well as the groups resulting for each candidate split. The variable, which maximizes the change of the information content resulting from that particular split is selected as the node for the branch plot. The procedure is then repeated for the remaining variables and examples.

Neighborhood analysis was used by Golub et al. (1999) to classify and predict different types of cancers. Genetic algorithm and Neighborhood analysis based strategies have their own advantages. Genetic algorithm as a global optimization method cannot be locked in a local optimum. In addition, it would be possible to locate multiple optimal medium combinations via GA. Neighborhood analysis is a classification method that we modified to be used as a ''guided local optimum search method.''

We present results for the optimization of Rifamycin B production via GA, NA, and DT with a nine-component medium and report a 10-fold increase in Rifamycin B productivity. The Decision Tree technique provided us with rules for medium composition, which gave high/low productivity.

## MATERIALS AND METHODS

### Materials

Rifamycin B standard was obtained from Lupin Pharma Ltd. (Mumbai, India). Trimethyle amine, HPLC grade acetonitrile, and ortho phosphoric acid were purchased from Merck (Darmstadt, Germany). Medium components were purchased from Hi-Media (Mumbai, India) and General Foods (Indore, India).

### Inoculum Preparation and Fermentation

The Rifamycin B-producing strain, *A. mediterranei* S699, a gift from Professor Giancarlo Lancini at the former Lepetit Laboratories (Geranzano, Italy), was obtained from Prof. Heinz Floss (Washington University, USA). The strain was cultivated and maintained as reported previously (Kim et al., 1996). Inoculum was prepared as reported by Richard and Lancini (1975). Fermentation was performed in 250-mL flasks with single baffle (working volume 40 mL). The initial medium composition of each shake flask was achieved by mixing suitable media components in distilled water. After adjusting the pH to 7.0 with 1N NaOH, each flask was sterilized by autoclaving at 121°C for 15 minutes. Fermentation was initiated by inoculating 10% (v/v) *A. mediterranei* culture (72 h old seed) in the medium. Flasks were shaken at 240 Rpm at 28°C (Kim et al., 1996). After 9 days of incubation, the Rifamycin B was measured using HPLC.

### HPLC Assay of Rifamycin B

*Amycolatopsis mediterranei* culture broths were mixed 1:1 with butanol and shaken for 4 h at room temperature (150 rpm). After centrifugation (10 min, 5000*g*) the butanol phase was separated and 100-μL sample was used for HPLC analysis (Stratmann et al., 2002). The HPLC consisted of Merck Model L-7100 HPLC pump, a manual Rheodyne 7125 injection valve equipped with a 20-μL loop and a 5-μm particle-size Lichrosphere® 100 RP18e column (250 × 4 mm I.D., Merck, Darmstadt, Germany). Mobile phase consisted of solvent A (acetonitrile) and solvent B (0.55% orthophosphoric acid, pH adjusted to 4.5 with triethyleamine). The solvent gradient program first consisted of a linear gradient from 0% A to 50% A in 5 min, followed by 2 min isocratic period, then 5 min linear gradient to 100% A, 3 min isocratic period, followed by 2 min linear gradient to 0% A. The separated components were detected using a Merck L - 7420 UV detector (Merck KgaA, Darmstadt, Germany) at wavelength 254 nm.

### Machine-Learning-Based Optimization Strategy

#### *Medium Components*

The parameter range for the medium component was selected based on previous reports (Kim et al., 1996; Richard and Lancini, 1975; Sensi and Thiemann, 1967; Venkateswarlu et al., 2000) as follows: dextrose (45–210 g $L^{-1}$); soy flour (12–54 g $L^{-1}$); ammonium sulphate (3.25–14.5 g $L^{-1}$); calcium carbonate (3.75–16.5 g $L^{-1}$); potassium phosphate (0.325–1.5 g $L^{-1}$); magnesium sulphate (0.325–1.5 g $L^{-1}$); ferrous sulphate (0.00325–0.015 g $L^{-1}$); zinc sulphate (0.0165–0.075 g $L^{-1}$); cobalt chloride (0.001-0.0045 g $L^{-1}$). Each medium component range is divided into five different concentration levels (33%, 66%, 100%, 125%, 150%). For example, 45 g $L^{-1}$ dextrose is 33% and

210 g $L^{-1}$ is 150%. The concentration of nutrients was kept above 0% as it is observed that all the nine components are important for Rifamycin B productivity. Thus, a search space of $5^9$ different medium compositions was generated. From this, 38 combinations were selected randomly, which constituted the first generation for GA- as well as NA-based optimization (Fig. 1). Each experiment was performed in duplicate to check reproducibility. The optimal medium combinations obtained from GA and NA results were tested in triplicate.

## Genetic Algorithm

We used the Rifamycin B titer at the end of the batch as the fitness function. Each medium composition was represented by a chromosome made of nine different genes with each gene representing a different medium component. Each of the medium components was varied at five distinct levels. From the initial population of 38 medium combinations, high-producing combinations (Rifamycin B > 2.8 g $L^{-1}$) were selected and passed on to the next generation. The next generation of medium composition was obtained by carrying out a single point crossover between the parent chromosomes. Crossover was allowed only at the boundary of genes to avoid absurd values for medium component levels. To counter premature convergence that may result in finding local optima, point mutation was used with the probability of 1% bit change of a chromosome. The offsprings generated were tested experimentally and the whole process continued for four generations (Fig. 1).

## Neighborhood Analysis

Neighborhood analysis is a classification technique, which we modified to perform guided local search for medium optimization and to obtain the classifying medium components (Fig. 1). The 38 medium compositions of the first generation were divided into two classes; high producing (titer > 2.8 g $L^{-1}$) and low producing (titer < 2.8 g $L^{-1}$) respectively. To use this technique as a guided local search, parameter range was selected based on the average of medium components from the high-producing class. Each medium component was varied over 10 distinct levels within 10% of the average value (±10% of the average medium composition). Thus, the local search space consisted of $10^9$ possible medium compositions. Out of this, 10 media were selected randomly and tested for productivity. This procedure was repeated for the next four generations and same the cutoff criterion was applied throughout the experiments.

Classifying medium components were obtained by using the following methodology. Each medium component was represented by an impact vector $v\ (m) = (m_1, m_2, \ldots, m_n)$, where $m_i$ denoted the impact level of media m in $i^{th}$ sample in the initial set $S$ of samples. A class distinction was represented by an idealized impact pattern $c = (c_1, c_2, \ldots, c_n)$, where $c_i = +1$ or 0 according to whether the $i^{th}$ sample belongs to class 1 (good productivity) or class 2 (poor productivity).

*Relationship Between Media and Class.* One can measure ''correlation'' between medium components and a class distinction in variety of ways. We used a measure of correlation proposed by Golub et al. (1999). In this, a measure of correlation, $P(m, c)$ that emphasizes the ''signal

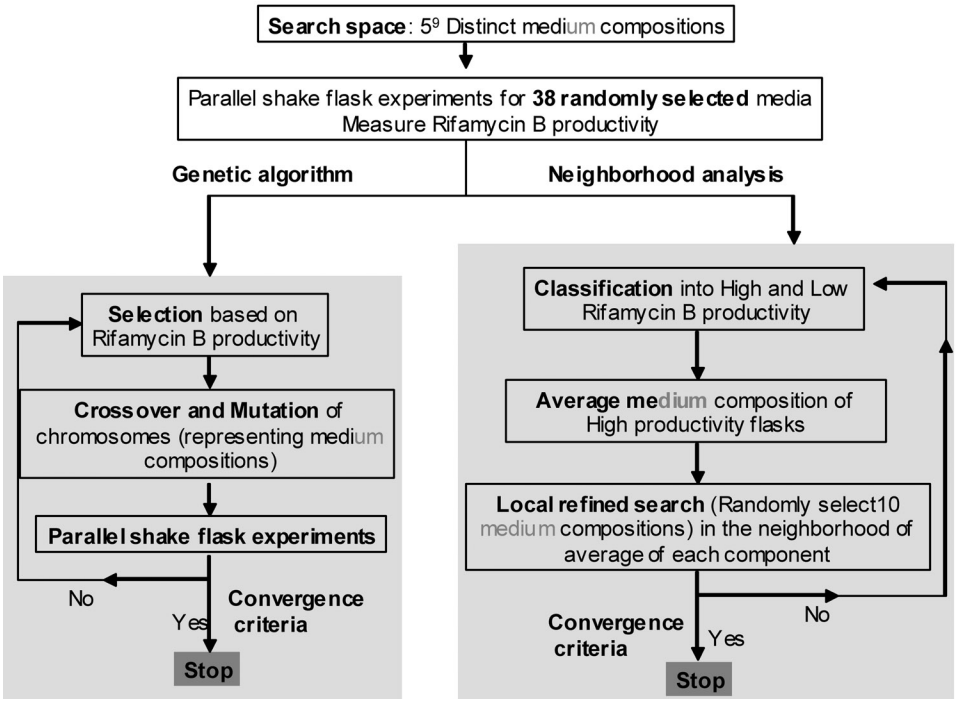


**Figure 1.** Strategy for machine learning based optimization of medium composition for Rifamycin B fermentation.

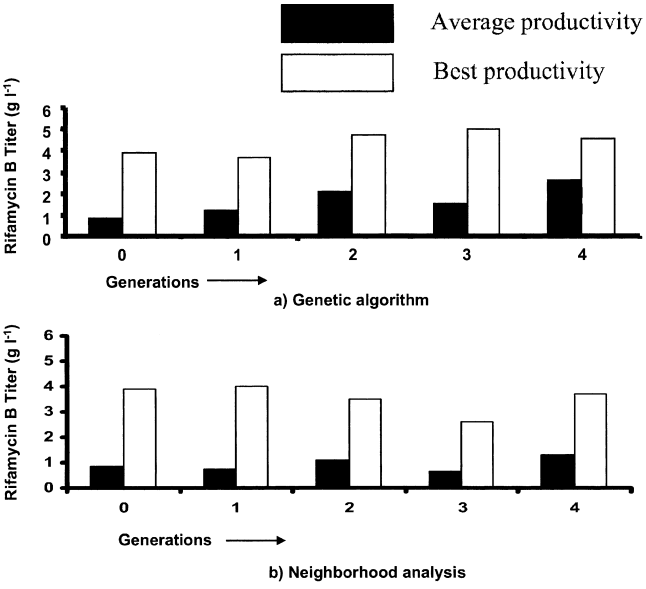


**Figure 2.** Performance of machine learning based optimization of Rifamycin B productivity (a) Genetic algorithm, and (b) Neighborhood analysis.

to noise'' ratio in using a given medium component as a predictor. Let $[\mu_1(m), \sigma_1(m)]$ and $[\mu_2(m), \sigma_2(m)]$ denote the means and standard deviations (*SD*s) of the levels of medium component m for the samples in high and low class, respectively.

$$P(m,c) = [\mu_1(m) - \mu_2(m)]/[\sigma_1(m) + \sigma_2(m)] \quad (1)$$

which reflects the differences between the classes relative to standard deviation within the classes. A large value of $|P(m, c)|$ indicates a strong correlation between the medium component and class distinction.

### Decision Tree Technique

For the Decision Tree technique, data generated from GA and NA (total 178 batches) was used. Response (productivity) was converted into a cell array of high (titer > 2.8 g $L^{-1}$) and low productivity (titer < 2.8 g $L^{-1}$). A matrix of 178 · 1 (for cell array) and matrix of 178 · 9 (for medium components) was generated using Decision Tree toolbox from MATLAB (Mathworks, Natick, MA). The set of rules generated from Decision Tree analysis were validated experimentally.

## RESULTS AND DISCUSSION

The reported yields of Rifamycin B are 2.45 g $L^{-1}$ in 12 days with barbital using *Amycolatopsis mediterranei* VA18 strain (Venkateswarlu et al., 2000) and 2.20 g $L^{-1}$ without bar-

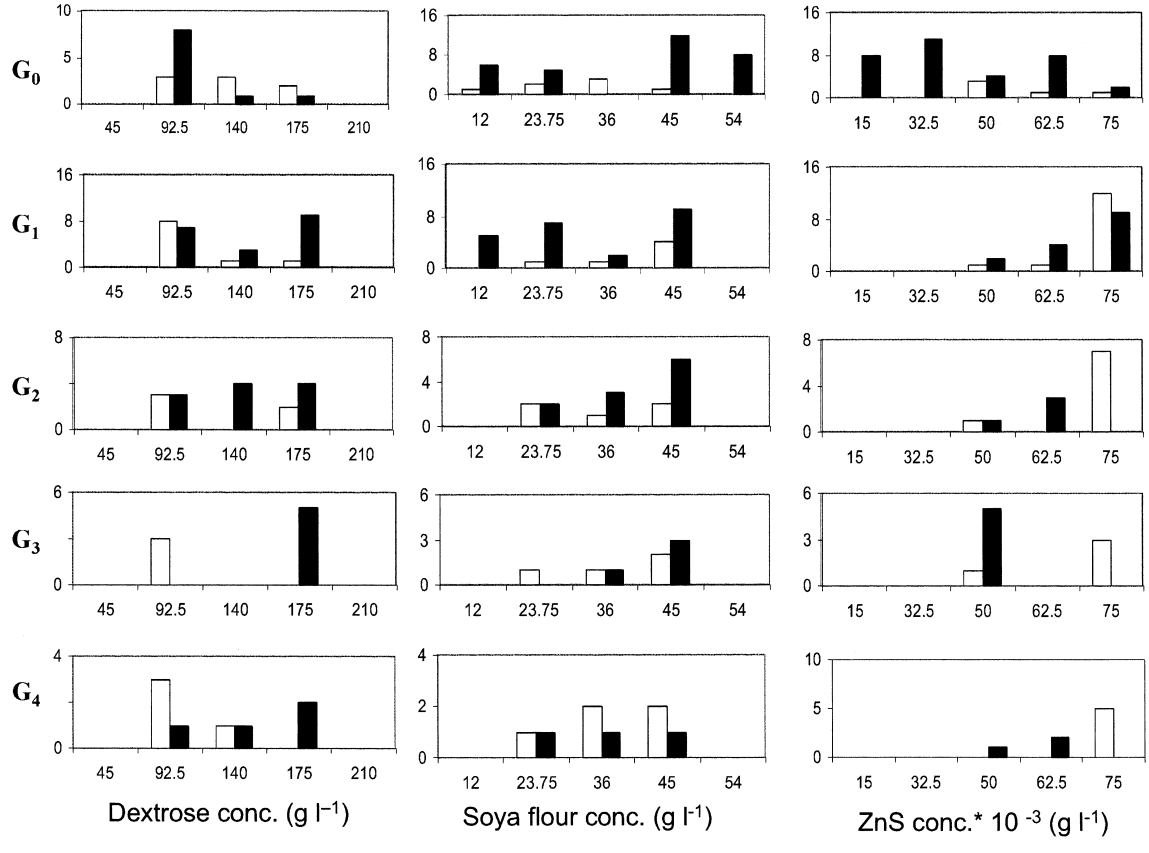


**Figure 3.** Analysis of convergence of medium components. Low productivity (■), High productivity (□). X-axis: Number of medium compositions; Y-axis: Concentration range of medium component.

bital using *Amycolatopsis mediterranei* ATCC 21789 strain (Richard and Lancini, 1975). We used the *Amycolatopsis mediterranei* S699 strain that does not require barbital (Yu et al., 2001). The reported yield for this strain is 0.50 g $L^{-1}$ (personal communication, Prof. Heinz Floss). The medium component concentration range was taken from literature. The nine media components were varied over five levels (33%, 66%, 100%, 133%, 150%) to generate a search space of $5^9$ distinct medium combinations. We used machine learning techniques for improving the Rifamycin B productivity and for understanding interactions among medium components. Here we search a small fraction of $5^9$ search space in an iterative manner where each round consists of parallel shake-flask experiments.

## Optimization Via Machine Learning Based Approach

### *Genetic Algorithm*

A total of five populations (including the initial random population) were generated through GA resulting in 100 shake-flask experiments. Starting from maximum productivity of 3.9 g $L^{-1}$ at generation ''0,'' a maximum productivity of 5.0 g $L^{-1}$ was achieved in the third generation that could not be improved further in the fourth generation. Average productivity increased up to the second generation, due to an increase in the number of healthy individuals (medium compositions with titer above 2.8 g $L^{-1}$). However, in the third generation the average productivity dropped significantly, indicating negative interaction between the medium components. The average productivity increased again in the fourth generation with 50% of the medium compositions showing Rifamycin B titer above the cut off criteria of 2.8 g $L^{-1}$ (Fig. 2a).

Genetic algorithm does not provide clear rules or interactions between medium components. However, it was of interest to see if a given medium component converges to a narrow range over the four generations. In generation ''0,'' an experimental plan generated at random covers the whole parameter range (medium components at five different concentration levels; Fig. 3). However, in the subsequent generations, high productivity was obtained only at one or two levels. For example, dextrose at 92.5 g $L^{-1}$ and 140 g $L^{-1}$ gave better productivity compared to the initial range of 45 g $L^{-1}$ to 210 g $L^{-1}$ while soy flour gave better productivity at three concentrations 23.75, 36.0 and 45.0 g $L^{-1}$ than the initial range of 12 g $L^{-1}$ to 54 g $L^{-1}$. Likewise, multiple distinct values were permitted for each of the medium components in the high producing medium composition. Zinc sulphate was an exception to this rule, which gave high productivity at single start medium concentration at 0.075 g $L^{-1}$ indicating non-interaction with other medium components. Components having more than one concentration level indicate multiple optimal concentration of that medium component. We obtained a total of seven distinct medium compositions from GA, which provide high Rifamycin B productivity and that are different from previously reported media (Table I).

### *Neighborhood Analysis*

We used NA as a strategy for guided search and optimization. At each generation, we labeled the flasks as high producer (titer > 2.8 g $L^{-1}$) or low producer (titer

**Table I.** Optimal media composition obtained via machine learning based techniques.[a] Five rounds of Genetic algorithm (GA) and Neighborhood analysis (NA) were carried out.

| | Initial Media Composition (g $L^{-1}$) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Media composition | Dextrose | Soy flour | Ammonium sulphate | Calcium carbonate | Potassium phosphate | Magnesium sulphate | Ferrous sulphate | Zinc sulphate | Cobalt chloride | Rifamycin B productivity after 9 d (g $L^{-1}$) | Optimization methodology |
| 1 | 159 | 33 | 10.7 | 14.1 | 1.12 | 0.90 | 0.011 | 0.059 | 0.0036 | 3.5 ± 0.30 | |
| 2 | 175 | 12 | 14.4 | 16.5 | 1.50 | 0.33 | 0.013 | 0.063 | 0.0045 | 3.9 ± 0.41 | Neighborhood |
| 3 | 175 | 45 | 12.0 | 16.5 | 1.25 | 1.25 | 0.013 | 0.075 | 0.0045 | 4.0 ± 0.15 | analysis[b] |
| 4 | 163 | 31 | 10.8 | 14.8 | 1.13 | 0.94 | 0.011 | 0.063 | 0.0040 | 3.6 ± 0.25 | |
| 5 | 175 | 45 | 12.0 | 16.5 | 1.25 | 1.25 | 0.013 | 0.075 | 0.0020 | 4.6 ± 0.33 | |
| 6 | 175 | 45 | 12.0 | 16.5 | 1.25 | 1.25 | 0.013 | 0.075 | 0.0030 | 4.6 ± 0.30 | |
| 7 | 92 | 45 | 3.2 | 13.8 | 0.66 | 0.66 | 0.015 | 0.075 | 0.0030 | 4.1 ± 0.50 | Genetic |
| 8 | 140 | 36 | 9.6 | 11.0 | 1.00 | 1.00 | 0.010 | 0.050 | 0.0030 | 3.5 ± 0.29 | algorithm[c] |
| 9 | 92 | 45 | 3.2 | 13.8 | 0.66 | 1.25 | 0.013 | 0.075 | 0.0045 | 5.0 ± 0.36 | |
| 10 | 92 | 24 | 6.3 | 16.5 | 1.00 | 1.00 | 0.013 | 0.075 | 0.0045 | 4.2 ± 0.21 | |
| 11 | 140 | 15 | 9.5 | 11.0 | 1.00 | 1.00 | 0.010 | 0.050 | 0.0040 | 4.8 ± 0.18 | |

[a]Different media compositions were generated through GA and NA algorithms. These media were sterilized and inoculated with 10% inoculum. Rifamycin B titer was measured periodically using HPLC. For more information see Materials and Methods.

[b]Neighborhood analysis: A guided local search algorithm was applied to find optimum media composition for Rifamycin B fermentation. Four distinct media compositions were obtained through the analysis.

[c]Genetic algorithm: A global search algorithm was used to search multiple optimum media compositions in defined search space. The best titers from seven distinct media compositions are shown.

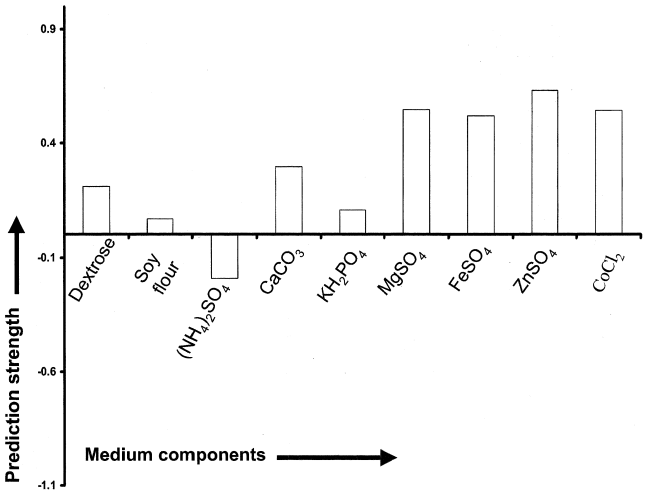


**Figure 4.** Medium classifiers for high Rifamycin B productivity obtained in neighborhood analysis. Prediction strength of medium component indicates the level of correlation of the component level with class distinction (high and low productivity). X-axis: Medium components; Y-axis: Prediction strength.

$< 2.8\ L^{-1}$). Neighborhood analysis reveals the average of the different medium compositions in the good batches and prediction strength ($P(m, c)$) of each medium component. Neighborhood analysis was performed to search the parameter space within $\pm 10\%$ of the average value for each component within the high producer. We performed experiments for four generations of NA to optimize Rifamycin B fermentation medium locally. At each generation, the average medium composition and prediction strength was recalculated based on the shake-flask results obtained until that generation. Out of 78 compositions tested, 11 compositions yielded Rifamycin B productivity above 2.8 g $L^{-1}$. The average productivity remained unchanged around 1 g $L^{-1}$ in all the generations indicating high number of low producing batches in each generation (Fig. 2b). Although average productivity was low, some medium compositions produced titer around 4.0 g $L^{-1}$ (Table I). These medium compositions were different from those reported earlier or those obtained via GA. The analysis of medium component as a class predictors reveals that zinc sulphate and magnesium sulphate have strong correlation in class distinction followed by ferrous sulphate, cobalt chloride, calcium carbonate, and dextrose while potassium phosphate, soy flour, and ammonium sulphate have the least correlation in class distinction (Fig. 4).

### *Decision Tree Technique*

We used the Decision Tree method for obtaining rules for designing optimal medium. Multiple distinct optimal medium compositions point toward possible interactions among the medium components. For example, if medium component A is high, medium component B must be low or if medium component A is low, medium component C must be low and so on. By using the data of 178 batches generated from GA and NA optimization rounds, a decision tree was generated. In the explored search space, three sets of rules were obtained for medium compositions that led to

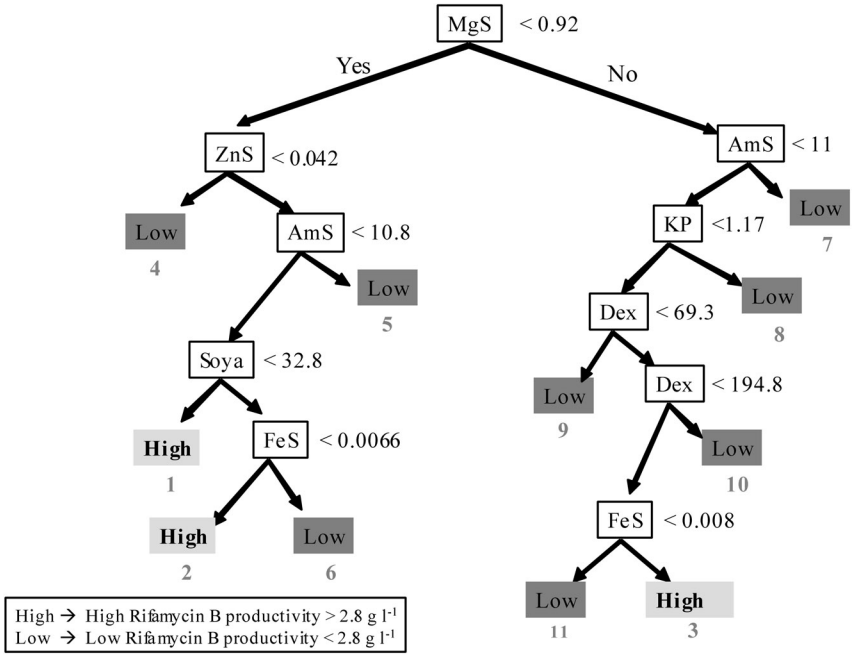


**Figure 5.** Decision tree generated using experimental results of 178 medium compositions. The value at the right side of each node indicates concentration of the respective component in g $L^{-1}$. The medium composition, which falls at or below this value, is directed towards the left side and above this range to right side. The number given below each level indicates the respective rule number, and is referred to in Table II.

**Table II.** Decision tree generated rules for medium components for optimal Rifamycin B productivity.

| Rule no.[b] | Rules | Medium composition (g $L^{-1}$) based on the rule set[a] | | | | | | | | | Rifamycin B productivity after 9 d (g $L^{-1}$) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Dextrose (Dex) | Soy flour (Soya) | Ammonium sulphate (AMS) | Calcium carbonate | Potassium phosphate (KP) | Magnesium sulphate (MgS) | Ferrous sulphate (FeS) | Zinc sulphate (ZnS) | Cobalt chloride | |
| 1 | Mgs < 0.92; ZnS > 0.042; AMS < 10.8; Soya < 32.8 | 92.0 | **19.0** | **5.00** | 13.8 | 0.66 | **0.26** | 0.013 | **0.060** | 0.0045 | 4.1 ± 0.20 |
| 2 | Mgs < 0.92; ZnS > 0.042; AMS < 10.8; Soya < 32.8; FeS > 0.0066 | 92.0 | **43.2** | **5.00** | 13.8 | 0.66 | **0.26** | **0.0026** | **0.060** | 0.0045 | 4.0 ± 0.25 |
| 3 | Mgs > 0.92; AMS < 11; KP < 1.17; Dex (69.3 – 194.8); FeS > 0.008 | **112** | 45.0 | **5.00** | 13.8 | **0.528** | **1.00** | **0.0104** | 0.075 | 0.0045 | 4.6 ± 0.15 |
| 4 | Mgs < 0.92; ZnS < 0.042 | 92.0 | 45.0 | 3.20 | 13.8 | 0.66 | **0.26** | 0.013 | **0.014** | 0.0045 | 2.9 ± 0.10 |
| 5 | Mgs < 0.92; ZnS > 0.042; AMS > 10.8 | 92.0 | 45.0 | **11.6** | 13.8 | 0.66 | **0.26** | 0.013 | **0.060** | 0.0045 | 2.3 ± 0.22 |
| 6 | Mgs < 0.92; ZnS > 0.042; AMS < 10.8; Soya > 32.8; FeS > 0.0066 | 92.0 | **43.2** | **5.00** | 13.8 | 0.66 | **0.26** | **0.013** | **0.060** | 0.0045 | 2.2 ± 0.30 |
| 7 | Mgs > 0.92; AMS > 11 | 92.0 | 45.0 | **11.6** | 13.8 | 0.66 | **1.00** | 0.013 | 0.075 | 0.0045 | 2.3 ± 0.12 |
| 8 | Mgs > 0.92; AMS < 11; KP > 1.17 | 92.0 | 45.0 | **5.00** | 13.8 | **1.20** | **1.00** | 0.013 | 0.075 | 0.0045 | 1.7 ± 0.30 |
| 9 | Mgs > 0.92; AMS < 11; KP < 1.17; Dex < 69.3 | **37.0** | 45.0 | **5.00** | 13.8 | **0.528** | **1.00** | 0.013 | 0.075 | 0.0045 | 0.0 |
| 10 | Mgs > 0.92; AMS < 11; KP < 1.17; Dex > 194.8 | **210** | 45.0 | **5.00** | 13.8 | **0.528** | **1.00** | 0.013 | 0.075 | 0.0045 | 1.4 ± 0.22 |
| 11 | Mgs > 0.92; AMS < 11; KP < 1.17; Dex (69.3 – 194.8); FeS < 0.008 | **112** | 45.0 | **5.00** | 13.8 | **0.528** | **1.00** | **0.0052** | 0.075 | 0.0045 | 2.8 ± 0.28 |
| 12 | Control | 92.0 | 45.0 | 3.20 | 13.8 | 0.66 | 1.25 | 0.013 | 0.075 | 0.0045 | 5.0 ± 0.30 |

[a]Concentration of components (bold font) was based on rules generated through Decision Tree (DT) analysis.

[b]Rule set 1 to 3 are for high productivity and 4 to 11 are for low productivity. The levels of parameter, which do not appear in DT rules, where kept at the levels of the control, which is row number 9 from Table I or row number 12 from this table.

high productivity (Fig. 5). As an example, Rule one states that, if magnesium sulphate (MgS) is below 0.925 g $L^{-1}$ then ammonium sulphate (AMS) and soy flour (Soya) must be below 10.8 g $L^{-1}$, 32.8 g $L^{-1}$, respectively, and zinc sulphate (ZnS) must be above 0.042 g $L^{-1}$. A total of 11 rule-sets were generated through the decision tree algorithm (Table II). One medium combination was generated based on each rule-set for experimental validation of the rules. Levels of the undefined components (not provided in decision tree rules) were kept the same as that in the control (combination No. 9 from Table I). The experimental values of the Rifamycin B titer for the 11 medium combinations correlate well with the productivity class (high and low) predicted by the DT rules. For example, the first three combinations, which are based on rules for high productivity, yield a titer of above 4.0 g $L^{-1}$. Thus, our results (Table II) reveal that decision tree analysis can be used for formulating optimal medium.

## CONCLUSION

Three machine-learning approaches (GA, NA, and DT) were tested for Rifamycin B fermentation medium development. To the best of our knowledge, NA was used for the first time to optimize fermentation medium and other two approaches (GA and DT) were used for the first time for optimization of Rifamycin B fermentation. Genetic algorithm was applied as a global search algorithm. We demonstrated the existence of more than one optimal solution. Although the multiple distinct medium compositions provide comparable yields of Rifamycin B, economic and other constraints may dictate the choice to one composition over the other in a commercial production plant. Genetic algorithm performed better than NA in the optimization process. Using our methodology, a 10-fold improvement in the Rifamycin titer for *Amycolatopsis mediterranei* S699 strain was obtained. In addition, the titer was 2 times higher than the best reported Rifamycin titer for any strain with a reduced fermentation time of 9 days from 12 days. We observed that production of Rifamycin B is maximum on the ninth day after which it declines. In comparison to statistical approaches for media optimization, fewer experiments were necessary to find an optimum composition. For example, in the search space of $5^9$ (1,953,125) medium compositions, we tested 100 compositions for GA compared to the 512 ($2^9$) required for response surface methodology. These machine-learning-based approaches can be

applied to media formulation of other industrially important products as well.

We thank Prof. Heinz Floss (Washington University, USA) for providing Rifamycin B producing strain *Amycolatopsis mediterranei* S699 and Lupin Pharma Ltd. (India) for providing the Rifamycin B and Rifamycin O standards.

## References

Bradamante S, Barenghi L, Beretta G, Bonfa M, Rollini M, Manzoni M. 2002. Production of Lovastatin examined by an integrated approach based on chemometrics and DOSY-NMR. Biotechnol Bioeng 80: 589–593.

Chang Y-N, Huanga J-C, Lee C-C, Shih I-L, Tzengd Y-M. 2002. Use of response surface methodology to optimize culture medium for production of lovastatin by *Monascus ruber*. Enzyme Microbial Technol 30:889–894.

Chen H-C. 1994. Response-surface methodology for optimizing citric acid fermentation by *Aspergillus foetidus*. Proc Biochem 29:399–405.

Forrest S. 1993. Genetic algorithms: Principles of natural selection applied to computation. Science 261:872–878.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286:531–537.

Hartmann G, Honikel K, Knusel F, Nuesh J. 1967. The specific inhibition of DNA directed RNA synthesis by rifamycin. Biochem Biophys Acta 145:843–844.

Holland J. 1992, July. Genetic algorithms. Scientific American J 267: 66–72.

Kennedy M, Krouse D. 1999. Strategies for improving fermentation medium performance: A review. J Industrial Microbiol Biotechnol 23:456–475.

Kim CG, Kirschning A, Bergon P, Zhou P, Su E, Sauerbrei B, Ning S, Ahn Y, Breuer M, Leistner E, Floss HG. 1996. Biosynthesis of 3-amino-5-hydroxybenzoic acid, the precursor of mC7N units in ansamycin antibiotics. J Am Chem Soc 118:7486–7491.

Marteijn RCL, Jurrius O, Dhont J, de Gooijer CD, Tramper J, Martens DE. 2003. Optimization of a feed medium for fed-batch culture of insect cells using a genetic algorithm. Biotechnol Bioeng 81:269–278.

Quinlan JR. 1986. Induction of decision trees. Machine Learning 1: 81–106.

Richard WJ, Lancini G. 1975. Production of rifamycin B. US patent no. 3871965.

Sensi P, Thiemann JE. 1967. Production of rifamycins. Prog Ind Microbiol 6:21–59.

Sepkowitz KA, Rafalli J, Riley L, Kiehn TE, Armstrong D. 1995. Tuberculosis in the AIDS era. Clin Microbiol Rev 8:180–199.

Stratmann A, Schupp T, Toupet C, Schilling W, Oberer L, Traber R. 2002. New insights into rifamycin B biosynthesis: Isolation of proansamycin B and 34a-deoxyrifamycin W as early macrocyclic intermediates indicating two separated biosynthetic pathways. J Antibiotics 55:396–406.

Venkateswarlu G, Murali Krishna PS, Rao VL. 1999. Production of rifamycin using *Amycolatopsis mediterranei* (MTCC14). Bioproc Eng 20:27–30.

Venkateswarlu G, Murali Krishna PS, Sharma G, Rao VL. 2000. Improvement of rifamyicin B fermentation using mutant strains of *Amycolatopsis mediterranei*. Bioproc Eng 23:315–318.

Weuster-Botz D, Wandrey C. 1995. Medium optimization by genetic algorithm for continuous production of formate dehydrogenase. Proc Biochem 30:563–571.

Yu T-W, Muller R, Muller M, Zhang X, Draeger G, Kim C-G, Leistner E, Floss HG. 2001. Mutational analysis and reconstituted expression of the biosynthetic genes involved in the formation of 3-amino-5-hydroxybenzoic acid, the starter unit of rifamycin biosynthesis in *Amycolatopsis mediterranei* S699. J Biol Chem 276:12546–12555.