Abstract: DNA microarrays are manufactured by synthesizing probes on a solid surface with the help of light and a sequence of lithographic masks. Unintentional illumination can create defects on the microarray due to small dimensions and light properties, but a suitable arrangement of probes can reduce the probability of defects. The problem of designing DNA microarrays is computationally hard and there is no publicly available algorithm that can solve this problem exactly, in polynomial time. This study investigates the suitability of the ant colony optimization (ACO) metaheuristic for finding optimal or at least good microarray designs. This research is based on a MAX-MIN ant system variant that is enhanced with 2-opt local optimization and max-$\kappa$-best pheromone reinforcement strategy. Experiments were conducted on problem instances based on border length and conflict index models. The proposed algorithm found solutions that are better than the best solutions previously published for 10 out of 14 problem instances.

Key words: Biochip design, swarm intelligence, max-$\kappa$-best, MAX-MIN ant system.

1. Introduction

A DNA microarray is a set of probes placed on solid material that can be used for genetic testing. The probes are fragments of DNA made from nucleobases: cytosine, guanine, adenine, and thymine. DNA microarrays, also known as biochips, have applications in genetics, molecular biology, medicine, and pharmacology for detecting genes, diagnosing diseases, inventing drugs, toxicological analyses etc. [1].

In the process of designing DNA microarrays arises a hard combinatorial problem of arranging probes in a way that minimizes the chance of defects and reduces the complexity of masks used in production.

For many optimization problems, efficient exact algorithms are not available. The question whether NP-hard problems can be solved in polynomial time remains unresolved in spite of serious research efforts. Over the past few decades, researchers have developed versatile metaheuristic methods that can often find good solutions for such problems, although they cannot guarantee finding optimal solutions.

In this paper, the possibility of using the ant colony optimization (ACO) metaheuristic for the microarray layout problem is investigated. Ant colony optimization (ACO) is a general metaheuristic inspired by the behavior of biolgical ants. It is a constructive type of a stochastic optimization algorithm whose construction procedure is guided by pheromone trails, and possibly by heuristic information that is specific to a particular type of the optimization problem. The pheromone trails are altered by ants to reflect the experience that the ants have gathered in previous attempts to find optimal solutions of the problem.

Since its introduction, ACO algorithms were found to be well-suited for many challenging computational problems: combinatorial optimization, dynamic and stochastic optimization, multiobjective optimization,
constraint satisfaction, and continuous optimization [2], [3].

The first published ACO algorithm, ant system (AS), proved to be a promising concept, but it could not compete with state-of-the-art algorithms. In attempts to improve algorithmic performance, different variants of ACO algorithms were published in the literature. Elitist ant system (EAS) allowed the best solution found from the beginning of the algorithm to additionally reinforce pheromone trails in order to improve convergence toward more promising areas in the search space[4]. Ant colony system (ACS) uses a rather greedy pseudorandom proportional rule in the solution construction procedure; there are local and global pheromone evaporation, and only one ("the best") solution reinforces the pheromone trails [5]. Rank-based ant system (AS\textsuperscript{rank}) enhances AS by using more elitism in the pheromone reinforcement procedure than EAS. Only a subset of the best solutions found in the current iteration and the best so far solution are used for the pheromone reinforcement procedure [6]. Approximate nondeterministic tree search (ANTS) introduces a unique rule for solution construction and does not perform explicit pheromone evaporation [7]. MAX-MIN ant system uses explicit pheromone bounds and pheromone reininitialization in the case the algorithmic stagnation is detected [8]. Best-worst ant system (BWAS) is similar to MMAS, but in addition, it incorporates mutation of pheromone trails (a concept from evolutionary algorithms) and negative feedback of the worst solution [9]. Three bound ant system (TBAS) uses occasional pheromone contractions instead of regular pheromone evaporations [10]. Among many versions of ACO, MAX-MIN ant system is the most popular owing to good results for many optimization problems; followed by ant colony system. TBAS is a new ACO algorithm experimentally evaluated only on a few types of optimization problems, but it showed very promising results (better than MMAS).

For the purpose of this research, MMAS with 2-opt local optimization and extended pheromone reinforcement strategies were used. Conducted experiments showed that MMAS is well-suited for the microarray layout problem, and for most problem instances the new best solutions were obtained.

2. Microarray Design Problem

One way of producing a DNA microarray is to synthesize probes on a solid surface made of glass or plastic. At the beginning the microarray has empty spots, on which probes are built by adding nucleobase after nucleobase until all probes are completed. Although one probe in not actually one DNA molecule, but rather many copies of the same single-stranded DNA molecule on a particular spot, it is logically simpler to think about the probe as one single-stranded DNA molecule. The probes can be produced by using photolithography similar to the process of producing integrated circuits in microelectronics. Photolithographic masks are used to select spots on which particular nucleobases will be inserted. Using one mask (this is one step of production) only one type of nucleobases: cytosine (C), guanine (G), adenine (A) or thymine (T) can be inserted on unmasked spots. In all other spots on the microarray, which are masked, nucleobases are not inserted. These steps are repeated by using right masks in the right order until all probes are completed.

For example, Fig. 1 shows six masks for producing very simple microarray with 9 spots, organized in 3×3 matrix, where every single-stranded DNA is made from three nucleobases. The probe in the right upper corner (AGT) is produced by inserting A in step 1 (using mask M1), G in step 2, and T in step 5. The insertion of particular nucleobase is done by using light which goes through unmasked spots. Because of very small dimensions and nature of light, the nearby area around selected spots might be unintentionally illuminated, and some nucleobases might be added on wrong probes. It is more likely that this defects will happen around the borders between unmasked and masked spots. By carefully arranging and grouping probes it is possible to minimize defects in production. The goodness of layout might be evaluated by using two evaluation models: border length and conflict index [11].
The total border length for microarray layout is defined with (1) as a sum of border lengths $B_1, \ldots, B_T$, of all masks $M_1, \ldots, M_T$. For one mask $M_t$ border length can be calculating by using (2), where $\text{border}(x, y)$ is equal to 1 if $x$ and $y$ are adjacent side by side or up and down, but not diagonally, and is equal to 0 otherwise. In the case of the microarray layout from Fig. 1, the border lengths for particular masks are: $B_1 = 9$, $B_2 = 7$, $B_3 = 7$, $B_4 = 6$, $B_5 = 8$, $B_6 = 3$, giving the total border length $B = 34$.

Conflict index is a more general model which takes into account additional considerations about defects in microarray production. It is defined for spot $u$ over all masks $M_1$ to $M_T$ by (3), where $\text{masked}(u, t)$ is equal to 1 if spot $u$ is masked by $M_t$, and equal to 0 otherwise. Euclidian distance between spots $x$ and $y$ is denoted as $d(x, y)$. Weighting function $\omega(x, t)$ assigns higher values to spots closer to the center of the microarray and lower values to spots closer to the edges [11]. The total conflict index $C$ is a sum of conflict indexes for all spots of the microarray defined by (4).

The goal of the optimization problem is to find microarray layout that minimize total border length or total conflict index, depending on the chosen model. The both models can be coded in quadratic assignment problem.

\[
B = \sum_{t=1}^{T} B_t 
\]

\[
B_t = \sum_{x \in X, y \in X_t} \text{border}(x, y) 
\]

\[
C(u) = \sum_{t=1}^{T} \left( \text{masked}(u, t) \cdot \omega(z, t) \cdot \sum_{v \in N(u)} \frac{1 - \text{masked}(v, t)}{(d(u, v))^2} \right) 
\]

\[
C = \sum_{u} C(u) 
\]

3. Ant Colony Optimization for Microarray Design Problem

The ant colony optimization is an iterative constructive metaheuristic. In every iteration of the algorithm, $m$ ants independently construct $m$ different solutions. The solution is constructed by adding solution components to a partial solution until the complete solution is constructed. The choice of selecting particular component is affected by pheromone and heuristic values. After the end of each iteration, the pheromone values are updated based on the quality of constructed solutions.

On the high level of abstraction, MAX-MIN ant system metaheuristic can be described with pseudocode
presented in Fig. 2. Precise implementation of particular procedures also depends on type of the problem and design decisions made by a researcher.

```
Initialize()
Loop(UntilStoppingCriteriaAreNotSatisfied){
    AntConstructSolutions();
    LocalOptimization(); //optional
    PheromoneEvaporation();
    PheromoneReinforcement();
}
```

Fig. 2. The pseudocode of MAX-MIN ant system for microarray design problem.

In procedure Initialize() algorithmic parameters are set, memory structures are prepared, pheromone and heuristic values are initialized. Both pheromone and heuristic values are usually organized in matrices, with each element linked to one particular solution component. For MMAS, which uses two pheromone bounds in the pheromone update procedure, the upper pheromone bound is set according to (5), where $\rho$ is evaporation rate parameter and $f(s^*)$ is the fitness of the optimal solution. In practice the optimal solution is not known in advance, but estimation of the optimal solution obtained by a simple method can be used instead.

The lower pheromone bound is set to $\tau_{\min} = \theta \cdot \tau_{\max}$. An appropriate value for parameter $\theta \in (0, 1)$ can be calculated by choosing probability $p_{SH}$ and using approximate expression given by (6). Alternatively, more precise formulas might be used [12], [13]. In expression (6), $|s|$ is the number of components in complete solution $s$, and $n_{avg}$ is the average number of components that are considered in one step of solution construction procedure. The initial pheromone trail value is set equal to upper bound $\tau_{\max}$. When local optimization is applied $\theta$ is usually set to $0.5 \cdot |s|$.

$$\tau_{\max} = \left[ \rho \cdot f(s^*) \right]^{-1}$$  \hspace{1cm} (5)

$$\theta = \left( 1 - \frac{1}{\sqrt{n_{SH}}} \right) \cdot \left[ \frac{(n_{avg} - 1)}{\sqrt{n_{SH}}} \right]^{-1}$$  \hspace{1cm} (6)

After the initialization is completed, inside the loop, the ant constructs solutions, update pheromone trails and optionally some demone actions like local optimization is performed, until stopping criteria is met (e.g. predefined number of iterations or maximum allowed time).

In AntConstructSolutions() procedure, $m$ ants construct solutions independently. A solution is constructed by adding a component by a component into partial solution $s^p$ until the final solution $s$ is constructed. At the beginning of construction, the partial solution $s^p$ is empty. The $C_p$ set contains all solution components that can be considered for adding to a partial solution $s^p$. In the case of assigning probes to spots on a microarray, a solution component is one assignment of a particular probe to a particular spot. For microarray with $n$ spots and $n$ probes, there are $n^2$ possible components.

In one step of AntConstructSolutions() procedure, one component is selected using random proportional rule defined by (7). The probability of adding the component $c_i$ into partial solution $s^p$ depends on the pheromone trail value $\tau_{ci}$ associated with that component, heuristic information for that component $\eta_{ci}$, parameters $\alpha$ and $\beta$, and values associated with all other components that are currently in the set $C_p$.

$$p(c_i \mid s^p) = \frac{\tau_{ci}^\alpha \cdot \eta_{ci}^\beta}{\sum_{c \in C_p} \tau_{c}^\alpha \cdot \eta_{c}^\beta}$$  \hspace{1cm} (7)
When choosing solution component, i.e. probe-spot assignment, first a free spot (or alternatively an unassigned probe) is chosen randomly, and then the assignment is based on random-proportional rule (7).

Performing local optimization is generally optional. By changing the part of solutions, e.g. by removing and adding some solution components, LocalOptimization() procedure is used to improve solutions constructed in AntConstructSolutions() procedure.

In PheromoneEvaporation() procedure all pheromone values are evaporated by multiplying with \((1 - \rho)\) according to expression (8). In the case that some pheromone value \(\tau_c\) becomes lesser than \(\tau_{\text{min}}\), its value is set equal to \(\tau_{\text{min}}\).

\[
\tau_c = (1 - \rho) \cdot \tau_c, \forall c \in C
\]

(8)

In PheromoneReinforcement() procedure it is necessary to choose a solution, according to some strategy, whose components will be rewarded with additional pheromone values. In the case of \(\kappa\)-best strategy the best solution found in last \(\kappa\) iterations is chosen. When max-\(\kappa\)-best strategy is used, then the best solution may be used at most \(\kappa\) iterations [13], [14]. Regardless of selected strategy, for each component \(c\) of chosen solution \(s_{\text{best}}\) the value of corresponding pheromone trail is increased according to expression (9). In the case that some pheromone value \(\tau_c\) becomes larger than \(\tau_{\text{max}}\), its value is set equal to \(\tau_{\text{max}}\).

\[
\tau_c = \tau_c + \left[f(s_{\text{best}})\right]^{-1}, \forall c \in s_{\text{best}}
\]

(9)

4. Experimental Settings and Results

Test instances of microarray layout problem, based on both border length and conflict index models, are publicly available on web page http://www.rahmannlab.de/research/microarray-design along with the best solutions that are known. The shorter names for the problem instances are used in this paper, but the original longer name can be uniquely identified by problem type and problem size. All instances are square-shaped so every solution has eight equivalent representations. There are four possible rotations and also for each rotation it is possible to make a mirror reflection of the solution.

Our implementation of MMAS with 2-opt local optimization and extended reinforcement strategy is done in C++ language. Experiments were conducted on Isabella cluster (isabella.srce.hr), running Linux on machines with various hardware. Local optimization was applied on all solutions constructed in AntConstructSolutions() procedure in the current iteration. For each solution, 2-exchange moves with first improvement were enforced until the solution become 2-optimal. In order to speed up local optimization, data structures with “don’t look bits” were used. Also, look up table with precomputed values of expression (7) were used to speedup solution constructions in AntConstructSolutions() procedure. Algorithmic parameters were set according to recommendations from literature and our own experience [13], [15]. Parameters were set to \(\alpha = 1\), \(\beta = 0\), and number of ants \(m = 20\). Few different values were tried for parameters \(\rho\) and \(\kappa\) (for max-\(\kappa\)-best strategy) in order to tune the algorithm for better performance.

For smaller problem instances (size \(n \leq 81\)) parameter \(\rho\) was tested with values 0.01, 0.05, 0.2 and parameter \(\kappa\) with values 8, 32, \(\infty\). For larger instances \(\rho = 0.01\) and max-8-best were omitted.

The maximum number of iterations was set to 10000, except for instances mlpbl121, mlpci121, and mlpbl144 for which it was set to 20000. For each parameter setting experiment was repeated 40 times.

Results of our experiments based on medians are showed in Table 1 and Table 2. Solution quality is expressed as a relative deviation of a median of solutions from the new best solutions published. It was calculated by (10) where \(med\) is the median value of solutions and \(nbs\) is the new best solutions for a particular instance (for many instances the best solutions were obtained by experiments published in this
The backgrounds of cells in Table 1 are colored according to their values for particular problem instance. The best results are colored with yellow, the worst with red and all the other with shades between yellow and red. For example, in the case of mlpbl64 the best result \((\text{med}_{\text{rel}} = 0.80\%\)\) obtained by max-\(\infty\)-best and \(\rho = 0.01\) is colored yellow and the worst result \((\text{med}_{\text{rel}} = 2.32\%\)\) obtained by max-8-best and \(\rho = 0.01\) is colored red.

By the color distribution on Fig. 1, it can be observed that the smallest tested parameter values \(\rho\) and \(\kappa\) resulted with poorer algorithmic performance, but the best parameter settings depended on problem instance.

In the case conflict index instances, the best performance was observed with max-32-best and \(\rho = 0.2\) for the most instances. For mlpci36 the median of solutions is equal to the new best solution published \((nbs = 168611971)\).

The best results and algorithms previously used to solve instances of microarray layout problem are taken from web page http://www.rahmannlab.de/research/microarray-design and are presented along with our results in Table 3 and Table 4. The previously used algorithms are greedy randomized adaptive search with path relinking (GRASP-PR) [16], heuristic algorithms RTL-1 [17] and RTL-2 [18], and hybrid of genetic algorithm with tabu search (GATS) [19]. The best solutions in some cases were obtained with parameter settings for which algorithmic performance (in terms of median) was not the best. Parameters \(\rho\) and \(\kappa\) from max-\(\kappa\)-best strategy for which MMAS obtained the best solutions are listed in the rightmost column in Table 3 and Table 4.
For instance mlpbl36, MMAS with 2-opt obtained solution equal to the best solution previously published. MMAS with 2-opt local optimization obtained solutions that are better than the best solutions that were published for three instances. In cases of the three largest instances, MMAS with 2-opt obtained better solutions than GRASP-PR, but worse than GATS.

In the case of conflict index based instances of microarray layout problem, MMAS with 2-opt local optimization obtained the new best solutions for all available test instances.

<table>
<thead>
<tr>
<th>instance name</th>
<th>instance size</th>
<th>best solution previously published</th>
<th>best solution found by algorithm</th>
<th>MMAS with 2-opt</th>
<th>best solution</th>
<th>parameters (ρ, κ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mlpbl36</td>
<td>36</td>
<td>3296</td>
<td>-</td>
<td>3304</td>
<td>3352</td>
<td>3296</td>
</tr>
<tr>
<td>mlpbl64</td>
<td>49</td>
<td>4564</td>
<td>4580</td>
<td>-</td>
<td>4660</td>
<td>4564</td>
</tr>
<tr>
<td>mlpbl81</td>
<td>64</td>
<td>6048</td>
<td>6080</td>
<td>-</td>
<td>6200</td>
<td>6048</td>
</tr>
<tr>
<td>mlpbl100</td>
<td>81</td>
<td>7644</td>
<td>-</td>
<td>-</td>
<td>7900</td>
<td>7644</td>
</tr>
<tr>
<td>mlpbl121</td>
<td>100</td>
<td>9432</td>
<td>-</td>
<td>-</td>
<td>9684</td>
<td>9432</td>
</tr>
<tr>
<td>mlpbl144</td>
<td>121</td>
<td>11640</td>
<td>-</td>
<td>-</td>
<td>12032</td>
<td>11640</td>
</tr>
</tbody>
</table>

<p>| Table 4. The Best Solutions Obtained for Instances Based on Conflict Index |</p>
<table>
<thead>
<tr>
<th>instance name</th>
<th>instance size</th>
<th>best solution previously published</th>
<th>best solution found by algorithm</th>
<th>MMAS with 2-opt</th>
<th>best solution</th>
<th>parameters (ρ, κ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mlpcci36</td>
<td>36</td>
<td>169016907</td>
<td>169925219</td>
<td>169016907</td>
<td>169016907</td>
<td>all tested</td>
</tr>
<tr>
<td>mlpcci49</td>
<td>49</td>
<td>237077377</td>
<td>238859844</td>
<td>237077377</td>
<td>236355034</td>
<td>all tested</td>
</tr>
<tr>
<td>mlpcci64</td>
<td>64</td>
<td>326696412</td>
<td>327770071</td>
<td>326696412</td>
<td>325671035</td>
<td>(0.2, 32)</td>
</tr>
<tr>
<td>mlpcci81</td>
<td>81</td>
<td>428682120</td>
<td>434317170</td>
<td>428682120</td>
<td>427582150</td>
<td>(0.05, 32)</td>
</tr>
<tr>
<td>mlpcci100</td>
<td>100</td>
<td>525401670</td>
<td>532573788</td>
<td>525401670</td>
<td>523806646</td>
<td>(0.05, 32)</td>
</tr>
<tr>
<td>mlpcci121</td>
<td>121</td>
<td>658317466</td>
<td>664137090</td>
<td>658317466</td>
<td>657514941</td>
<td>(0.05, 32)</td>
</tr>
<tr>
<td>mlpcci144</td>
<td>144</td>
<td>803379686</td>
<td>813127758</td>
<td>803379686</td>
<td>802219898</td>
<td>(0.05, 32)</td>
</tr>
</tbody>
</table>

5. Conclusion

In this research, the MAX-MIN ant system was adopted for finding optimal designs of DNA microarrays. The basic ACO metaheuristic was enhanced with 2-opt local optimization to improve algorithmic performance. Limited parameter exploration revealed that the best performance for the instances based on border length is often achieved with rather greedy parameters, i.e. max-κ-best strategy and evaporation rate set to 0.2. In the case of the conflict index based instances, the best results are achieved with slightly less greedy parameter settings. For many problem instances, the proposed algorithm successfully obtained solutions that are better than the best solutions that were previously published.

References


Nikola Ivkovic received the MS degree in computer engineering and the Ph.D. degree in computer science from the Faculty of Electrical Engineering and Computing, University of Zagreb. He is a member of the research and teaching staff at the Department of Information Technologies and Computing of the Faculty of Organization and Informatics, University of Zagreb. His research interests include computational intelligence and optimization, parallel programming, formal methods, operating systems, and computer networks.

Marin Golub received the BS degree in 1992 in electrical engineering, MS degree in 1996 and Ph.D. degree in 2001 in computer science, all at the Faculty of Electrical Engineering and Computing, University of Zagreb. Currently he is working as an associate professor at the Department of Electronics, Microelectronics, Computer and Intelligent Systems, Faculty of
Domagoj Jakobovic received the BS degree in December 1996 and the MS degree in December 2001, in electrical engineering. Since April 1997, he is a member of the research and teaching staff at the Department of Electronics, Microelectronics, Computer and Intelligent Systems of the Faculty of Electrical Engineering and Computing, University of Zagreb. He received the PhD degree in December 2005 on the subject of generating scheduling heuristics with genetic programming.