Efficient RNAi-Based Gene Family Knockdown via Set Cover Optimization

Wenzhong Zhao ⋆, M. Leigh Fanning and Terran Lane

University of New Mexico, Department of Computer Science
Albuquerque, NM87131, USA

Abstract

RNA interference (RNAi) is a recently discovered genetic immune system with widespread therapeutic and genomic applications. In this paper, we address the problem of selecting an efficient set of initiator molecules (siRNAs) for RNAi-based gene family knockdown experiments. Our goal is to select a minimal set of siRNAs that (a) cover a targeted gene family or a specified subset of it, (b) do not cover any untargeted genes, and (c) are individually highly effective at inducing knockdown. We show that the problem of minimizing the number of siRNAs required to knock down a family of genes is *NP-Hard* via a reduction to the set cover problem. We also give a formal statement of a generalization of the basic problem that incorporates additional biological constraints and optimality criteria. We modify the classical branch-and-bound algorithm to include some of these biological criteria. We find that, in many typical cases, these constraints reduce the search space enough that we are able to compute exact minimal siRNA covers within reasonable time. For larger cases, we propose a probabilistic greedy algorithm for finding minimal siRNA covers efficiently. Our computational results on real biological data show that the probabilistic greedy algorithm produces siRNA covers as good as the branch-and-bound algorithm in most cases. Both algorithms return minimal siRNA covers with high predicted probability that the selected siRNAs will be effective at inducing knockdown. We also examine the role of "off-target" interactions – the constraint of avoiding covering untargeted genes can, in some cases, substantially increase the complexity of the resulting solution. Overall, however, we find that in many common cases, our approach significantly reduces the number of siRNAs required in gene family knockdown experiments, as compared to knocking down genes independently.

Key words: Set Cover, Minimal siRNA Cover, Gene Family Knockdown, siRNA Design, RNAi

⋆ Author to whom all correspondence should be addressed.
Email address: wzhao@cs.unm.edu (Wenzhong Zhao), *leigh@cs.unm.edu* (M. Leigh Fanning), *terran@cs.unm.edu* (Terran Lane).

1 Introduction

RNA interference (RNAi) is a recently discovered posttranscriptional gene silencing (PTGS) mechanism that seems to play both regulatory and immunological roles in the eukaryotic genetic system [1–5]. RNAi has aroused a great deal of excitement in both therapeutic and genomic experimental communities because of its potentials for treatment of a wide spectrum of diseases such as HIV [6,7]; spinocerebellar ataxia type 1 (SCA1) and Huntington's diseases [8]; and certain classes of cancers [2,9,10] as well as its demonstrated use in functional genomic studies via controlled gene knockdown [11–13]. The complete RNAi process is an involved, multi-stage biochemical process. A number of surveys describe it in detail (e.g., [2,5]). The key aspect of RNAi that is relevant to us, and that makes it so therapeutically promising, is that it can be used to selectively knock down the expression of individual genes in a highly target-specific fashion. In particular, by introducing a carefully chosen initiator molecule, known as a short-interfering RNA (siRNA) into a cell, we can induce the cellular machinery into degrading the messenger RNA (mRNA) product of a targeted gene and prevent transcription into protein. Thus, we can suppress the function of a harmful gene (such as an oncogene or the malfunctioning gene underlying a genetic disease such as Huntington's disease). In a functional genomics context, we can perform gene activity studies by comparing the phenotypes of an organism with or without a gene's expression. We give a more detailed description of the RNAi process and the role of siRNAs in Section 2.

To date, RNAi research has almost exclusively focused on single-target studies: selecting siRNA molecules to knock down the function of a single gene. In this paper, we consider the problem of *gene family knockdown*: selecting a *set* of siRNAs so as to simultaneously degrade an entire group of related genes, known as a gene family. This problem arises when we wish to study the roles of multiply redundant genes or when we wish to suppress the activity of an entire gene pathway. Because the efficacy of a specific siRNA in knocking down its target gene is related to its homology to that gene, there is hope that we can knock down a gene family with fewer siRNAs than the naïve "one-for-one" approach. Because synthesis of individual siRNAs can run to hundreds or thousands of dollars, there is advantage to using compact sets of siRNAs for simultaneous gene family knockdown. Alternatively, there is evidence that a carefully selected pool of siRNAs may be more effective at single-gene knockdown than any single siRNA would be alone [13,14]. This indicates that there may be advantage in constructing multiply redundant sets of siRNAs when attempting gene family knockdown. Both tasks are complicated by the need to account for variable efficacy of different siRNA molecules and to avoid "off-target" effects in which the siRNA causes unintended knockdown of an untargeted gene to which it incidentally has high homology.

In this paper, we formalize the gene family knockdown problem in two different methods: first, using a simplistic model of siRNA efficacy and off-target effects, and later employing a framework that allows more sophisticated models of efficacy and incorporates additional pragmatic biological constraints. While siRNA efficacy is currently poorly understood and not well modeled, empirical efficacy data is being accumulated at a rapid rate and we expect the more general model to become applicable in the near future. For the moment, however, we focus on using the currently available, albeit admittedly simple, models of efficacy to develop our gene family knockout framework and perform empirical studies. We show that our formulation leads to an interpretation of this problem as a variant of the classical set cover problem [15] and, therefore, exact optimization of siRNA families is NP-hard. We introduce two optimization algorithms for solution of this problem: an exact, branch-and-bound technique that achieves minimal siRNA pool sizes with high predicted pool efficacy; and a probabilistic greedy algorithm that produces a bounded approximation to the minimal set.

We apply our methods to two different gene families and a number of subsets of each. The first family, the set of Fibrinogen-related protein (FREP) genes from the snail *Biomphalaria glabrata* are medically relevant because this snail is a model organism for infection by the human-affecting parasite *Schistosoma mansoni* [16,17]. The second family, the olfactory genes of nematode *Caenorhabditis elegans*, is an excellent model for studies of chemosensation [38]. It is also one of the largest known gene families that we employ to study the scaling properties of our algorithms. Using these data, along with relevant biological constraints, we find that the exact, branch-and-bound method is effective up to reasonably large gene families. The probabilistic greedy algorithm can reach larger gene families and, in our experiments, often achieves results as good as the branch-and-bound algorithm on the families that they can both handle. Both algorithms return minimal siRNA covers with high probability that those selected siRNAs are functional. We also find that the "off-target" interactions can, in some cases, substantially increase the complexity of the resulting solution. Overall, however, our approach significantly reduces the number of siRNAs required in gene family knockdown experiments, as compared to knocking down genes independently.

The rest of the paper is organized as follows. In Section 2 we give a brief introduction to RNA interference (RNAi), and describe the role of siRNAs and biological constraints for designing functional short interfering RNAs (siRNAs). We introduce necessary notations, and formally define the two siRNA covering problems in Section 3. Section 4 describes an exact, branch-and-bound algorithm and a probabilistic greedy algorithm for the minimal siRNA cover problem. In Section 5 we report our computational results of the two algorithms for knocking down target genes in the two gene families. Finally, we conclude and describe future work in Section 6.

2 RNA Interference via short interfering RNA

In this section, we describe briefly the RNA interference (RNAi) process, the role of short interfering RNAs (siRNAs), and the rules for designing highly functional siRNAs. The basic mechanism of RNAi is considered to be a multi-step process [1,2,5]: (1) gene-specific double-stranded RNAs (dsRNAs) are introduced into a cell; (2) the long dsRNAs are cleaved into short 17-25 nucleotide siRNA segments via the RNAse III-family enzyme DICER; (3) each siRNA subsequently assembles with additional protein components, forming RNA-induced silencing complex (RISC); (4) each siRNA directs the RISC to bind to a complementary mRNA transcript by base pairing interactions between the siRNA antisense strand and the mRNA; and (5) the siRNA/RISC complex, bound to the target mRNA, guides an RNA-dependent RNA polymerase (RdRP) enzyme to cleave and degrade the target mRNA, resulting in suppression of the expression of the targeted gene.

At the heart of the cleavage event is the degree of matching between an siRNA segment and the mRNA, and the success to gene silencing relies highly on the sequence homology between the siRNA antisense strand and the target mRNA. The first two steps in the RNAi mechanism may be bypassed by introduction of synthetic siRNAs directly into cells. Therefore, the use of carefully designed siRNAs becomes a simple and reliable method for silencing targeted genes with high siRNA potency and specificity. Most RNAi researchers select subsequences of a target mRNA as siRNAs [1] . The typical reported length of a functional (or effective) siRNA is between 17 and 25 [18].

Many siRNAs may knock down a specific target mRNA, however, not all of them are equally effective [14,19]: Some repress 90% of gene expression, others 30% or less. During the last few years, a number of different rules have been proposed for designing functional siRNAs [20,21], among which the siRNA rational design rules suggested by Reynolds et. al [21] provide consistent criteria for selecting functional siRNAs. Based on these design rules, several siRNA design tools have been developed [22,23]. In this paper, we use the rational design rules for predicting knockdown efficacy for siRNAs [2] . The rational design rules are briefly described in Table 1, and readers are referred to the original paper for detailed information [21].

Each rule assigns a score to an siRNA, and the *rank* of that siRNA is the sum of scores for all the rules. Thus, ranks are integers between -2 and 10. The ranking provides a prediction of how effective an siRNA will be able to knock

[1] An siRNA can have mismatches with its target mRNA, however, any mismatch between an siRNA and the target mRNA may result in poor knockdown efficacy.
[2] Note that as we gain new understanding regarding the prediction of gene knockdown efficacy for siRNAs, new rules can be incorporated into the program easily.

Table 1 Reynolds' rational design rules [21]

Rule	Description	Score
1	Moderate to low GC content (30-52%)	1 point
2	At least 3 A/U at position 15-19	1 point /per A or U
3	No internal repeats with 3 or more nts	1 point
4	A at position 19	1 point
5	A at position 3	1 point
6	U at position 10	1 point
7	G/C at position 19	-1 point
8	G at position 13	-1 point

down a target gene. siRNAs with rank of 6 or higher are usually considered to be functional (or effective) by Reynolds et. al [21].

3 Notation and Problem Formulation

Because of its generally high specificity to a single target mRNA, RNAi has so far been primarily used to target and knock down the expression of individual genes in isolation. Often, however, it is useful to be able to knock down multiple genes simultaneously. For example, a family of closely related genes may have mutually redundant function; to observe any phenotypic change, it may be necessary to suppress the entire family simultaneously. For single gene knockdown, it usually suffices to select a subsequence of the target mRNA as the initiator siRNA. For families of genes, however, it is less clear how to design an optimal *set* of siRNAs to target the entire family. In this paper, we examine this *gene family knockdown problem*.

Before we formulate the problem of minimizing the number of siRNAs required for gene family knockdown experiments, we first describe the necessary notation.

Let $\Sigma = \{a, c, g, u\}$ be the nucleotide alphabet of mRNA. We denote by Σ^* the set of nucleotide sequences over Σ. An mRNA is a nucleotide sequence $m \in \Sigma^*$. A genome for an organism is a set of nucleotide sequences, with each representing an mRNA, i.e., $T = \{m | m$ is an mRNA in an organism.$\}$. The following definition defines an siRNA.

Definition 1 *Given an mRNA m, a nucleotide sequence s is an* siRNA *for*

m if and only if s is a subsequence of m. An siRNA is called an L-siRNA *if and only if it has length of L nucleotides.*

The following example illustrates how siRNAs can be enumerated from a target mRNA.

Example 1 *Given a target mRNA of N nucleotides, as shown at the bottom of Figure 1, altogether we can enumerate (N-L+1) potential* L-siRNAs *(in this case, L = 21). Two of them are shown at the top of the same figure: siRNA1 has rank of 9 and siRNA2 has rank of 7.*

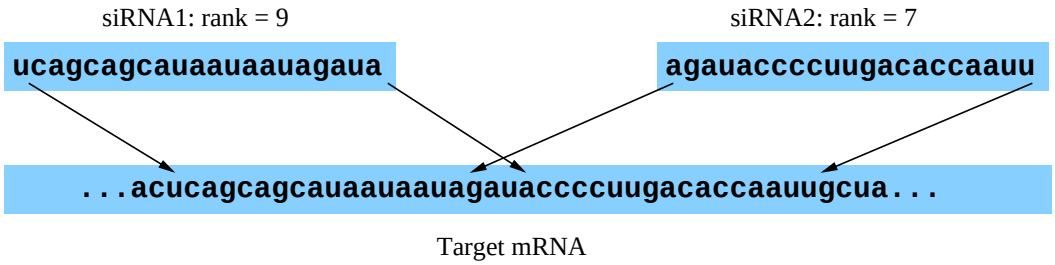

Fig. 1. Relationship between siRNAs and the target mRNA

In order to find the minimum number of siRNAs required to knock down a larger number of genes, we are particularly interested in finding genes sharing common subsequences, which we call a *gene group*. We define gene group as follows:

Definition 2 *Given a set of genes* $G = \{m_1, m_2, \ldots, m_d\}$*, G is an* L-restricted gene group *if and only if there exists at least one nucleotide sequence, s, which is an L-siRNA for all genes in G. We denote* $g(s)$*, i.e.* $g(s) \equiv G$*, the gene group that is covered by L-siRNA s.*

We say that the siRNA *s covers* the gene group G, and that G is the gene group *associated* with s. Next we define an *siRNA cover*.

Definition 3 *Given a set of target genes* $M = \{m_1, m_2, \ldots, m_n\}$*, a collection of L-siRNAs covering all the genes is called an* L-restricted siRNA cover *(siRNA cover for short) for M. The* minimal siRNA cover *for M is the one with the smallest number of distinct siRNAs.*

In the rest of this paper we restrict our search to siRNAs that contain 21 nucleotides (or 21-siRNAs) and that are also subsequences of the target mRNA. As such, we use the terms siRNA and L-siRNA interchangeably.

Here we introduce an instance of siRNA covering problem.

Definition 4 *An* siRNA covering problem instance *is defined as a tuple* $\langle T, M, S, \Gamma, \rangle$*, where* $T = \{m_1, m_2, \ldots, m_N\}$ *is the set of all genes in an organism and, without loss of generality,* $M = \{m_1, m_2, \ldots, m_n\} \subseteq T$ *is the set of tar-*

get genes. $S = \{s_1, s_2, \ldots, s_K\}$ *is the set of siRNAs enumerated from all the genes in* T. Γ *is the set of gene knockdown efficacies for all siRNAs in* S*: They can be computed by either indirect prediction methods (i.e., siRNA ranking) or direct prediction methods (i.e., gene knockdown efficacy predicting).*

Now we are ready to define our siRNA covering problems.

3.1 Minimal siRNA Cover Problem

Here we assume that siRNA silencing efficacy can be captured by siRNA ranking R computed according to the rational design rules described in Table 1. Given an siRNA covering problem instance $\langle T, M, S, R\rangle$, we want to find the minimum set of siRNAs required to knock down all the target genes in M. Formally we define the problem as follows:

Definition 5 *Given an siRNA covering problem instance* $\langle T, M, S, R\rangle$*, the* minimal siRNA cover problem *is to find the minimum set of siRNAs* $S' \subseteq S$ *that covers all the target genes (i.e.,* $\cup_{s\in S'} g(s) = M$*) without hitting any off-target genes (i.e.,* $\cup_{s\in S'} g(s) \cap (T - M) = \emptyset$*). As a secondary criterion, it optimizes for siRNA rank where possible.*

The following lemma shows that the minimal siRNA covering problem reduces to the set cover problem.

Lemma 1 *The minimal siRNA covering problem reduces to the set cover problem if the secondary criterion can be ignored.*

Proof sketch:
After appropriate rearrangement, the covering relationship between K siRNAs and N genes can be represented as a matrix $W' = \{w'_{ij}\} \in \{0, 1\}^{N\times K}$, as shown in Figure 2. Columns represent siRNAs enumerated from all genes in T. Rows represent all genes in the genome (including both target and off-target genes). An entry $w'_{ij} = 1$ in the matrix indicates that the j^{th} siRNA covers the i^{th} gene, while an entry $w'_{ij} = 0$ means that the siRNA does not cover the gene. Each siRNA covers one or more target genes and possibly off-target genes as well. siRNAs without off-target effects are those that do not cover any off-target genes, while siRNAs with off-target effects are those that do cover at least one off-target gene.

First, siRNAs with off-target effects are discarded since the current approach does not allow any off-target reactions. Thus, we remove the right part of the matrix. Assume there are k remaining siRNAs which do not have off-target effects, i.e., $S'' = \{s''_1, s''_2, \ldots, s''_k\}$. Subsequently, we remove all the off-target genes since none of the remaining siRNAs covers them, eliminating the bottom

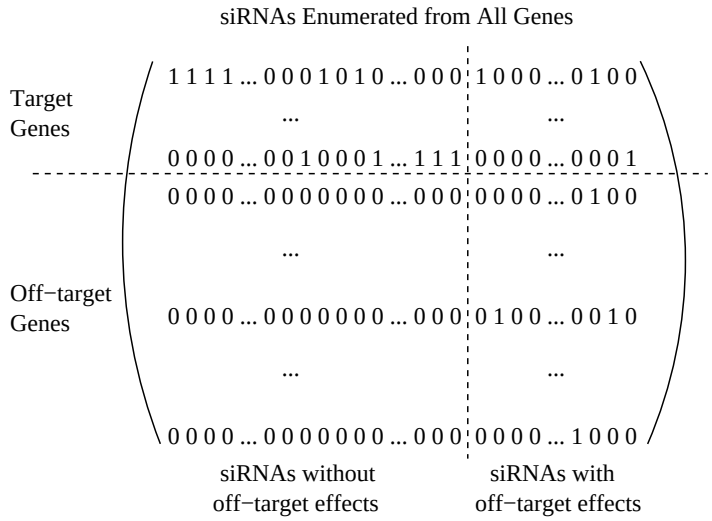

Fig. 2. Covering relationship between siRNAs and genes, including both target genes and off-target genes

part of the matrix as well. After the above operations, only the top left part of the matrix $W = \{w_{ij}\} \in \{0,1\}^{n \times k}$ remains. Therefore, if we ignore the secondary criterion, the above problem becomes: among the siRNAs with no off-target effects, find the minimum set of siRNAs that covers all targeted genes, which is exactly same as the set cover problem. □

Note that the original siRNA cover problem instance $\langle T, M, S, R \rangle$ also reduces to a new instance $\langle M, S'', R \rangle$, which we call *reduced siRNA cover problem instance*, where S'' is the set of siRNAs with no off-target effect. The minimal siRNA cover problem reduces to the set cover problem [24] except that it has a secondary objective - maximize the rank of the selected siRNAs where possible. So the minimal siRNA cover problem is at least as hard as the set cover problem and, therefore, is also an NP-hard problem. Thus in general no polynomial time exact solution exists for the minimal siRNA cover problem unless $P = NP$.

3.2 Weighted Minimal siRNA Cover Problem

In addition to the ranking methods described in the previous section, some researchers have developed algorithms to correlate siRNA gene knockdown *efficacy* with biological features, and have proposed methods for directly predicting gene knockdown efficacy [25,26]. Typically, the knockdown efficacy is defined as the percentage of gene expression being repressed.

Here, instead of using ranking, we utilize the gene knockdown efficacy to represent the effectiveness of how an siRNA will knock down a specific gene.

An siRNA gene knockdown efficacy matrix $E = \{e_{ij}\}$ is computed, where $e_{ij} \in [0, 1]$ is the efficacy with which the j^{th} siRNA covers the i^{th} gene. The problem described below is to find the minimum set of siRNAs that covers all target genes with a specified lower bound of knockdown efficacy, and that covers all off-target genes with a specified upper bound of knockdown efficacy.

Definition 6 *Given an siRNA covering problem instance* $\langle T, M, S, E \rangle$*, the* weighted minimal siRNA cover problem *is to find the minimum set of siRNAs* $S' \subseteq S$ *that covers all the target genes with a lower bound of knockdown efficacy, and that covers all the off-target genes with an upper bound of knockdown efficacy.*

Let x be the assignment (0-1) vector for the set of siRNAs S, where an "1" of the j^{th} element in x indicates that the corresponding siRNA, s_j, is selected. Let $\alpha_i, \beta_i \in [0, 1]$ be the lower bound of knockdown efficacy for i^{th} target genes and the upper bound of knockdown efficacy for $(i - n)^{th}$ off-target genes, respectively. We can write the above problem as the following combinatorial optimization problem.

Objective: $min\ \Sigma_{j=1}^{K} x_j$

such that

$$\Sigma_{j=1}^{K} x_j e_{ij} \geq \alpha_i, \text{ i = 1, 2, ..., n}$$

$$\Sigma_{j=1}^{K} x_j e_{ij} \leq \beta_i, \text{ i = n+1, n+2, ..., N}$$

$$x_j \in \{0, 1\}, \text{ j = 1, 2, ..., K}$$

The first constraint sets the lower bounds of knockdown efficacy for all target genes, while the second constraint sets the upper bounds of knockdown efficacy for all off-target genes. The criterion $x_j \in \{0, 1\},\ j = 1, 2, \ldots, k$ makes it an integer linear programming problem.

The following lemma shows the weighted minimal siRNA cover problem reduces to the standard minimal siRNA cover problem under certain conditions.

Lemma 2 *If* e_{ij} *can only take* $\{0, 1\}$ *values,* $\alpha_i = 1, i = 1, 2, \ldots, n$*, and* $\beta_i = 0, i = n + 1, n + 2, \ldots, N$*, then the weighted minimal siRNA cover problem reduces to the standard minimal siRNA cover problem.*

Proof sketch:
Given: $\Sigma_{j=1}^{K} x_j e_{ij} \leq \beta_i$ and $\beta_i = 0, i = n+1, n+2, \ldots, N$. Since $0 \leq x_j \in \{0, 1\}$ and $0 \leq e_{ij} \in [0, 1]$, then we know $x_j e_{ij} = 0$ for $i = n + 1, n + 2, \ldots, N$ and $j = 1, 2, \ldots, K$. Therefore, for a given j, if there exists any $e_{ij} \neq 0$ and $n + 1 \leq i \leq N$, then $x_j = 0$, eliminating all siRNAs with off-target effects. Then both the objective and the other constraint reduce exactly to the corresponding ones in the standard minimal siRNA cover problem. □

The standard minimal siRNA cover problem is a special case of the weighted version. Therefore, the weighted minimal siRNA cover problem is at least as hard as the standard version, which means no polynomial time exact solution exists, either. Currently we do not have a good solution to this problem, but development of heuristic algorithms for the weighted problem is ongoing.

4 Algorithms for the Minimal siRNA Cover Problem

The set cover problem has been well studied and a number of exact and approximate methods exist for it, including the exact branch-and-bound algorithms [27,28], greedy algorithms [15,29], LP relaxation [30], Lagrangian relaxation [31–33], and genetic algorithms [34,35]. For a detailed analysis of many of these, we refer readers to the recent survey by Caprara et. al [36].

In this section, we describe two algorithms for the minimal siRNA covering problem. We modify the branch-and-bound algorithm to include biological constraint. Inspired by the randomized greedy algorithms [29,37], we propose a probabilistic greedy algorithm for selecting minimal siRNA covers efficiently.

4.1 Data Pre-processing

Data pre-processing is a transformation process that maps the minimal siRNA cover problem into the minimum set cover problem. It starts by enumerating all potential siRNAs from the target genes. Simultaneously, ranks for all the siRNAs are computed using the rational design rules, as shown in Table 1. To avoid cross reactivity, most RNAi researchers use BLAST search to predict whether an siRNA will have off-target effects. However, it has been noted that BLAST is not the best homology detection tool, especially for short sequences, and it may miss some good homologous sequences because its minimum word length is 7. We developed our own algorithm to filter out siRNAs with high probability of cross reaction with genes outside the target set.

4.2 Exact Branch-and-Bound Algorithm

Like the standard set cover problem, the minimal siRNA cover problem can be solved using branch-and-bound techniques. A search tree is constructed by iteratively picking an siRNA and branching on it. At each point the algorithm generates two subtrees with one corresponding to selecting the siRNA and the other to de-selecting the siRNA. When an siRNA is selected, deduction techniques are used to reduce the search space and will be discussed in detail later.

During the search, the algorithm keeps track of the current best cover with the lowest number of siRNAs. When there is a contradiction, i.e., the current path cannot lead to a better cover than the current best one, backtracking is performed.

The following lemma allows us to prune branches which cannot lead to any covers better than the current best one.

Lemma 3 *Consider a reduced siRNA cover problem instance* $\langle M, S, R\rangle$*, and a partial cover* $S' \subset S$ *for* $M' \subset M$*, where* $M' = \cup_{s\in S'} g(s)$*. Let* $M'' = M - M'$ *denote the set of uncovered genes, and* $S'' = S - S' = \{s''_1, s''_2, \ldots, s''_{k'}\}$ *be the unselected siRNAs, sorted in non-increasing order of the number of uncovered genes that an siRNA can cover (i.e.,* $|g(s''_j) \cap M''|$*). Then any complete cover derived from the partial cover* S' *requires at least* $(n' + |M'|)$ *siRNAs, where* n' *is the smallest* n'' *satisfying* $\Sigma_{j=1}^{n''} |g(s''_j) \cap M''| \geq |M''|$*.*

Proof sketch:
Any given unselected siRNA, s''_j, can cover a maximum number of $|g(s''_j) \cap M''|$ uncovered genes. Since the unselected siRNAs are assumed to be sorted in non-increasing order of the number of uncovered genes that an siRNA can cover, any combination of n'' unselected siRNAs can cover at most $\Sigma_{j=1}^{n''} |g(s''_j) \cap M''|$ uncovered genes (the maximum can be reached only when the gene groups covered by the first n'' siRNAs are all disjoint.). Therefore, any complete cover derived from the partial cover S' requires at least n' additional siRNAs where n' is the smallest n'' satisfying $\Sigma_{i=1}^{n''} |g(s''_i) \cap M''| \geq |M''|$. In other words, any complete cover derived from the partial cover S' requires at least $(n' + |M'|)$ siRNAs. □

Algorithm 1 shows the pseudocode for the exact, branch-and-bound process. S is the set of unselected siRNAs, and M is the set of uncovered target genes. The argument *rank* contains ranks for all the siRNAs. x is the current selection of siRNAs and b is the best selection found so far. The algorithm starts with $(S, M, rank, \mathbf{0}, \mathbf{1})$, where $\mathbf{0}$ and $\mathbf{1}$ represent vectors with all 0s and all 1s, respectively.

During the reduction stage, the algorithm discards any dominated siRNA, which covers only a subset of genes also covered by another siRNA. An essential siRNA is an siRNA which is the only one that can cover a particular target gene. We add all essential siRNAs to the solution x since they must be in any complete cover. We update the uncovered genes M accordingly.

After reduction is done, the potential lower bound on the current path is estimated based on Lemma 3. If the current estimate requires more siRNAs than the current best cover b, then we prune the current path. Otherwise, it generates two branches, one for selecting the next siRNA and the other for de-selecting the same siRNA. Simultaneously, the algorithm uses as the

Algorithm 1 Branch-and-Bound Algorithm

```
1  Branch_and_Bound (S, M, rank, x, b) {
2     Reduce S/M and update x
3     Sort the rest of the siRNAs in non-increasing order
           of uncovered genes they can cover
4     Compute the lower bound on the current path based on Lemma 3
5     if (the lower bound > |b| or the low bound > Number_From_Greedy)
6        return b
7     if (|M| = 0)
8        return x
9     s1 = first(S)
10    y = Branch_and_Bound (S - {s1}, M - g(s1), rank, x, b)
11    if (|y| < |b| or (|y| = |b| and
           avgRank(y, rank) > avgRank(b, rank)) )
12       b = y
13    y = Branch_and_Bound (S - {s1}, M, rank, x, b)
14    if (|y| < |b| or (|y| = |b| and
           avgRank(y, rank) > avgRank(b, rank)) )
15       b = y
16    return b
17 }
```

upper bound the number of siRNA in the cover obtained from our greedy algorithm (i.e., the parameter *Number_From_Greedy*) to prune some paths since the minimal cover must be no worse than the one obtained from any greedy algorithm.

Note that this algorithm puts primary emphasis on the size of an siRNA cover (i.e., $|b|$) and secondary emphasis on the average rank of siRNAs (i.e., computed by the method $avgRank() = \Sigma_j rank(s_j)x_j/|x|$) in b. Whenever there is a tie between two siRNA covers, it uses the average rank of siRNA covers as a tie-break, and picks the one with the higher average rank. Therefore, this algorithm produces siRNA covers with high probability that those selected siRNAs are functional. This is the step that differs from the traditional branch-and-bound algorithm.

4.3 Probabilistic Greedy Algorithm

The branch-and-bound algorithm always produces the minimal siRNA cover, however, it requires exponential time in the worst case. In this section, we propose a near-optimal probabilistic greedy algorithm.

The probabilistic greedy algorithm shares some common aspects with the randomized greedy algorithm [29], but it differs in some important ways. The randomized greedy algorithm [29] iteratively selects to the cover an siRNA completely randomly from a subset of unselected siRNAs, while our approach selects each siRNA randomly but according to a selection probability distribution.

Let $f(s)$ be the number of genes which an siRNA s covers, and $h(g)$ be the number of siRNAs which cover a particular gene g. By analyzing the covering relationship matrix (such as the one in Figure 3), we found two important features which imply whether an siRNA is a good pick: $f(s)$ and $min_{g \in g(s)}\ h(g)$. Based on the two features, we define a selection metric for selecting siRNAs, and then construct a subset of the unselected siRNAs as the set of potential siRNAs based on the selection metric and limiting parameter $\alpha \in [0, 1]$. We compute a selection probabilistic distribution over the potential siRNA set.

		Unselected siRNAs					
		1	2	3	4	5	6
Uncovered Genes	1	0	1	1	1	1	0
	2	0	1	1	0	0	0
	3	0	1	0	0	0	1
	4	0	0	0	0	1	0
	5	1	0	0	1	0	1
	6	0	1	0	1	0	1
	7	1	0	0	0	1	1
	8	0	0	1	0	0	1

Fig. 3. Covering relationship between uncovered genes and unselected siRNAs

The selection metric for siRNA, s, found to be the most effective is:

$$m(s) = \frac{f^2(s)}{min_{g \in g(s)}\ h(g)},$$

The selection probability distribution over the set of potential siRNAs S' is defined as:

$$P(s) = \frac{m(s)}{\Sigma_{s \in S'}\ m(s)},$$

where $S' = \{s | s \in S \text{ and } m(s) > \alpha * max(m(s))\}$.

The following example illustrates how we compute the features, metric, and the selection probability distribution, given an siRNA-gene covering relationship matrix.

Example 2 *Consider a covering relationship matrix for uncovered genes and unselected siRNAs, as shown in Figure 3. Let* $\alpha = 0.6$. *First we compute* $f(s)$

and $m(s)$ for all siRNAs $s \in S$. Then we construct the set of potential siRNAs S', i.e., in this case, $S' = \{s_2, s_5, s_6\}$. Finally, we compute the probability distribution $P(s)$ over S'. The final results are shown in Table 2.

Table 2 The metrics and probability distribution of siRNAs

siRNA	$f(s)$	$min_{g \in s}\ h(g)$	$m(s)$	$P(s)$
s_1	2	3	4/3	N/A
s_2	4	2	8	0.27
s_3	3	2	9/2	N/A
s_4	3	3	3	N/A
s_5	3	1	9	0.30
s_6	5	2	25/2	0.43

Algorithm 2 shows the pseudocode for the probabilistic greedy algorithm. S is a set of siRNAs, and M is the set of the target genes. The argument *rank* contains ranks for all the siRNAs.

Unlike the standard greedy algorithm, which always picks the next siRNA covering the largest number of uncovered genes in M, the probabilistic greedy algorithm uses a non-deterministic process. It first computes the selection metrics $m(s)$ for all unselected siRNAs, then constructs a subset of the unselected siRNAs S' based on the metric $m(s)$ and parameter α, and finally randomly picks the next siRNA $s \in S'$ according to the probability distribution $P(s)$.

This procedure is repeated k times from line 6 to 27, and returns the best siRNA cover it finds. Whenever there is a tie between two siRNA covers, the algorithm uses their average ranks as a tie-break, and chooses the one with the higher average rank. Since there are up to $min(|M|, |S|)$ siRNAs in an siRNA cover, thus, the inner *while* loop could have up to $min(|M|, |S|)$ iterations, where each iteration requires $O(|S|)$ time. Therefore, the algorithm requires $O(k \cdot |S| \cdot min(|M|, |S|))$ time. The probabilistic greedy algorithm reduces to standard greedy algorithm, when $\alpha = 1.0$ and $k = 1$.

Like the branch-and-bound algorithm described in Section 4.2, this algorithm puts primary emphasis on the size of an siRNA cover (i.e., $|opt|$) and secondary emphasis on the average rank of siRNAs in *opt*, thus, it also produces siRNA covers with high probability that those selected siRNAs are functional.

Algorithm 2 Probabilistic Greedy Algorithm

```
1  ProbGreedy (S, M, rank) {
2     α = aValue                    //α ∈ [0, 1]
3     k = aValue                    //Number of repetition
4     opt = S
5     iter = 0
6     while (iter < k) {
7        C = ∅
8        while (M ≠ ∅) {
9           MAX = 0
10          for each s ∈ S {
11             Compute metric m(s) //See details in text
12             if (MAX < m(s)) MAX = m(s)
13          }
14          S' = ∅                 //Set of potential siRNAs
15          for each s ∈ S {
16             if (m(s) > α * MAX) S' = S' ∪ {s}
17          }
18          Compute probability distribution P(s) over S'
29          Select s randomly from S' according to P(s)
20          C = C ∪ {s}
21          S = S - {s}
22          M = M - g(s)
23       }
24       Remove redundant siRNAs from C
25       if (|C| < |opt| or (|C| = |opt| and
             avgRank(C, rank) > avgRank(opt, rank) )
26          opt = C
27    }
28    return C
29 }
```

5 Computational Results

Based on the two algorithms described above, we implemented a pure Java program for selecting minimal siRNA covers required for the gene family knockdown experiments.

We apply our methods to two different gene families. The first family, the set of FREP genes from the snail *Biomphalaria glabrata*, is of interest in human immunological studies because both humans and *B. glabrata* may become infected by the parasite *Schistosoma mansoni* [16,17]. Within the snail, infection is detectable via observed expression of several novel genes from the im-

munoglobulin superfamily (IgSF). Fibrinogen-related proteins (FREPs), containing one or two IgSF domains and a fibrinogen (FBG) domain, are seen in response to parasitic infection. Our collaborators in the immunology group of the University of New Mexico Department of Biology are interested in being able to suppress all FREP and FREP-related activity as part of their study of the parasitology of *S. mansoni*. The partially sequenced *B. glabrata* genome is available at NCBI. The second family, the olfactory genes of nematode *Caenorhabditis elegans*, is an excellent model for studies of chemosensation [38]. It is also one of the largest known gene families and we employ it to study the scaling properties of our algorithms. We collected the *C. elegans* cDNA sequences from the Wormbase at Sanger Institute [39].

Different experiments were conducted to examine the potential benefit of our gene family knockdown approach, targeting these two gene families and subsets of each. Examining performance on subsets of known gene families provides two benefits. First, it gives us data on the scalability of our algorithms. And second, it allows us to examine the tradeoff between minimizing the set of siRNAs and avoiding off-target effects. Our current set cover approach to this problem treats off-target interactions as hard constraints, which may render it difficult or impossible to locate a compact set of siRNAs to target a family (e.g., if there are no siRNAs in the target family that do not also occur elsewhere in the genome). When we attempt to target a *subset* of a gene family, the other members of that family become part of the off-target genome, which may increase the complexity of finding a good siRNA set for the targeted part of the family. Experiments of this type allow us to understand the performance of our algorithms under a spectrum of difficulty conditions. Finally, we examine the scalability of our algorithms with off-target constraints disabled altogether to establish a baseline for the growth of covering set sizes and determine whether off-target interactions affect the results at all. The latter experiments demonstrate that off-target effects are significant in all cases and cannot be neglected.

Here we report results of these tests. The actual (and virtual) target gene families used in our experiments are:

- target family 1 contains 13 FREP genes, with one from each FREP subfamily.
- target family 2 contains 54 fibrinogen (FBG) sequences from the FREP family, listed in the article by Zhang et. al [17].
- target family 3 contains a group of 150 somewhat related genes (neglecting olfactory genes that share no siRNAs at all with any other members of the greater olfactory gene family) from the olfactory family.

For target family 1, knockout target set size varied from two FREP genes up to all FREP genes (size 13). For set sizes of 12 and 13, all possible combi-

nations were tested. The remaining were randomly sampled for a total of 50 tests each. We compared results obtained from the three different algorithms: the exact branch-and bound, the probabilistic greedy and Chvatal's standard greedy [15]. Table 3 shows the results obtained from these three algorithms when applied to subsets of the 13-gene FREP family. We can see that the probabilistic greedy algorithm reduces the siRNA cover size by up to 5.3% compared to Chvatal's standard greedy algorithm, and produces siRNAs covers as good as the exact, branch-and-bound algorithm in most cases. It appears that, through randomization, the probabilistic greedy algorithm successfully avoids being trapped in local optima.

Table 3 Average size of siRNA covers for FREP genes from three different algorithms

Algorithm	Average size of siRNA cover											
Number of target genes	2	3	4	5	6	7	8	9	10	11	12	13
Chvatal's Greedy (Sc)	1.8	2.6	3.2	3.8	4.3	5.0	5.4	5.9	6.1	5.9	5.6	5.0
Probabilistic Greedy (Sp)	1.8	2.6	3.2	3.7	4.3	4.8	5.2	5.6	5.9	5.8	5.6	5.0
Branch-and-Bound (Sb)	1.8	2.5	3.2	3.7	4.3	4.8	5.2	5.6	5.9	5.8	5.6	5.0
Percentage Improved (%) [a]	0.0	0.0	0.0	2.7	0.0	4.0	3.8	5.3	3.4	1.7	0.0	0.0

[a] Computed by $100 * (Sc - Sp)/Sp$.

We applied the probabilistic greedy algorithm to the target family 3. Each experiment was run 50 times with randomly sampled genes, and target set size varied from 10 to 150 genes. We examine the scaling effect on both the size of siRNA covers and the average rank of siRNAs in the covers. As shown in Figure 4, siRNA cover size increases with the size of target set, while average rank of siRNAs in covers deceases with the size of the target set. By turning off the off-target filter (a software layer which filters out siRNAs with off-target effects), we can expect a smaller siRNA cover for the same target set. Overall, however, the gene family knockdown approach significantly reduces the number of siRNAs (e.g. up to 52%) required in gene family knockdown experiments as compared to knocking down genes independently, and the selected siRNAs are functional with high predicted probability since the average ranks of siRNAs in those covers range from 7.2 to 8.6 (siRNAs with rank of 6 or high are usually considered to be functional.).

We applied the probabilistic greedy algorithm to the target family 2 as well, and studied off-target effects by comparing the results to those with off-target constraints disabled. The computational results are shown in Figure 5. The size of siRNA covers, when turning off off-target filter, increases with the target set size. However, the size of siRNA covers, when turning on off-target filter, first increases with the size of the target set, and then decreases with the size of the target set after some point. The failure rate (the percentage of target sets for which there do not exist siRNA covers) shows similar trend as the size of target set increases. This is due to the existence of two competing processes: covering genes in the target set and avoiding off-target reactions with genes

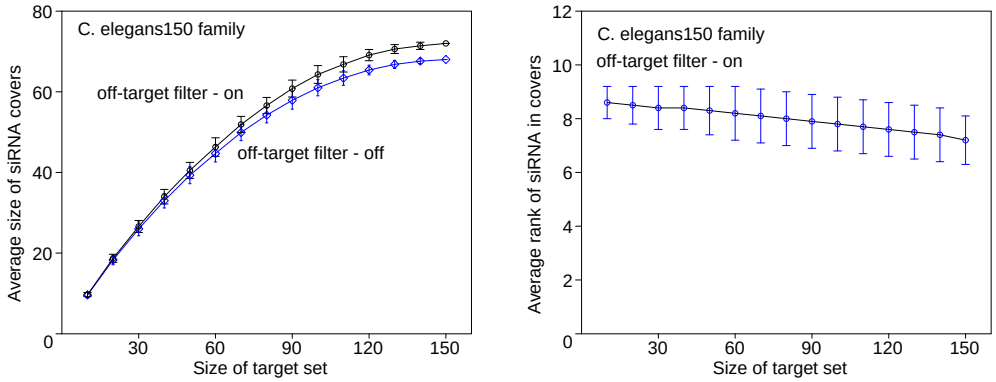

Fig. 4. Average size of siRNA covers and average rank of siRNAs in covers

outside the target set yet in the same family. Similar results were found for target family 1 as well, as shown in Figure 6. Overall, off-target reactions to genes outside the target family affect the selection of siRNA covers, while off-target reactions to genes within the same family (e.g., closely related genes) substantially increase the complexity of the resulting solution.

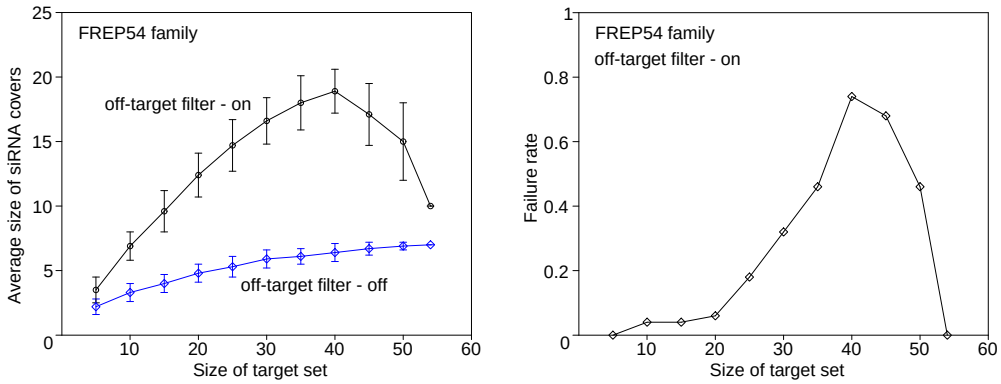

Fig. 5. Effects of off-target reactions on selection of siRNA covers

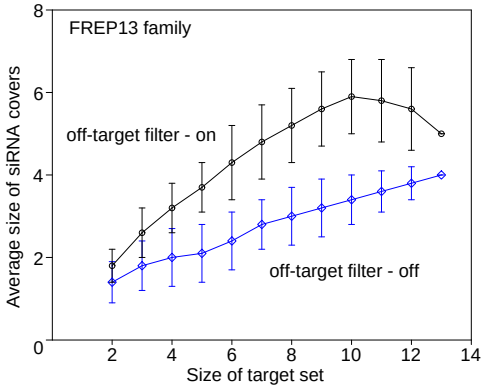

Fig. 6. Average size of siRNA covers

6 Conclusions and Future Work

In this paper, we described the gene family knockdown approach for selecting siRNAs for family knockdown experiments. In many typical cases, this approach significantly reduces the number of siRNAs required in gene family knockdown experiments as compared to knocking down genes one-by-one. The probabilistic greedy algorithm successfully avoids being trapped in local optima, and produces siRNAs covers as good as the exact, branch-and-bound algorithm in most cases. Both algorithms produce siRNA covers with high predicted probability that the selected siRNAs are highly functional. Simultaneously, we found that off-target reactions to genes outside the target family affect the selection of siRNA covers, while off-target reactions to genes within the same family (e.g., closely related genes) substantially increase the complexity of the resulting solution.

There are three major foci in our ongoing and future research: 1) Use soft constraints (i.e., employing similarity measures) to replace the hard constraints (i.e., using exact sequence match) for both selecting siRNAs and detecting off-target reactions; 2) Develop algorithms for solving the weighted minimal siRNA cover problem; 3) Select a minimum set of dsRNAs, instead of an siRNA cover, to cover the target genes such that each dsRNA represents as many siRNAs as possible, and simultaneously minimize off-target effects.

Acknowledgments

The authors thank Si-Ming Zhang and Coenraad Adema for providing insightful information about the FREP gene family and pointing to the olfactory gene family, and Vladamir Vuksan for assistance with data analysis. This work was supported by NIH Grant Number 1P20RR18754 from the Institutional Development Award (IDeA) Program of the National Center for Research Resources.

References

[1] T. Tuschl, RNA interference and small interfering RNAs, Chembiochem 2 (4) (2001) 239–245.

[2] G. J. Hannon, RNA interference, Nature 418 (2002) 244–251.

[3] P. Ahlquist, RNA-dependent RNA polymerases, viruses, and RNA silencing, Science 296 (2002) 1270–1273.

[4] R. H. A. Plasterk, RNA silencing: The genome's immune system, Science (2002) 1263–1265.

[5] N. Agrawal, P. V. N. Dasaradhi, A. Mohmmed, P. Malhotra, R. K. Bhatnagar, K. Mukherjee, S, Rna interference: Biology, mechanism, and applications, Microbiology and Molecular Biology Reviews 67 (4) (2003) 657–685.

[6] J. M. Jacque, K. Triques, M. Stevenson, Modulation of HIV-1 replication by RNA interference, Nature 418 (2002) 435–438.

[7] R. Surabhi, R. Gaynor, RNA interference directed against viral and cellular targets inhibits human immunodeficiency virus type 1 replication, Journal of Virology 76 (24) (2002) 12963–12973.

[8] H. Xia, Q. Mao, S. L. Eliason, S. Q. Harper, I. H. Martins, H. T. Orr, H. L. Paulson, L. Yang, R. M. Kotin, B. L. Davidson, RNAi suppresses polyglutamine-induced neurodegeneration in a model of spinocerebellar ataxia, Nature Medicine 10 (2004) 816–820.

[9] A. Borkhardt, Blocking oncogenes in malignant cells by RNA interference — new hope for a highly specific cancer treatment?, Cancer Cell 2 (3) (2002) 167–168.

[10] S. Barik, Development of gene-specific double-stranded rna drugs, Annals of Medicine 36 (7) (2004) 540–551.

[11] J.-T. Chi, H. Y. Chang, N. N. Wang, D. S. Chang, N. Dunphy, P. O. Brown, Genomewide view of gene silencing by small interfering RNAs, PNAS 100 (11) (2003) 6343–6346.

[12] R. S. Kamath, A. G. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta, A. Kanapin, N. Le Bot, S. Moreno, S. M., W. D. P., P. Zipperlen, J. Ahringer, Systematic functional analysis of the caenorhabditis elegans genome using RNAi, Nature 421 (2003) 231–237.

[13] A. C. Hsieh, R. Bo, J. Manola, F. Vazquez, O. Bare, A. Khvorova, S. Scaringe, W. R. Sellers, A library of siRNA duplexes targeting the phosphoinositide 3-kinase pathway: determinants of gene silencing for use in cell-based screens, Nucleic Acids Research 32 (3) (2004) 893–901.

[14] J. Karpilow, D. Leake, B. Marshall, siRNA: Enhanced functionality through relational design and chemical modification, PharmaGenomics (2004) 32–40.

[15] V. Chvatal, A greedy heuristic for the set covering problem, Mathematics of Operations Research 4 (1979) 233–235.

[16] S.-M. Zhang, C. M. Adema, T. B. Kepler, E. S. Loker, Diversification of ig superfamily genes in an invertebrate, Science 305 (2004) 251–254.

[17] S.-M. Zhang, E. S. Loker, Representation of an immune responsive gene family encoding fibrinogen- related proteins in the freshwater mollusc biomphalaria glabrata, an intermediate host for schistosoma mansoni, Gene 341 (2004) 255–266.

[18] S. M. Elbashir, W. Lendeckel, T. Tuschl, RNA interfernce is mediated by 21- and 22- nucleotide RNAs, Genes and Development 15 (2001) 188–200.

[19] P. Satrom, O. S. Jr., A comparison of siRNA efficacy predictors, Biochemical and Biophysical REsearch Communications 321 (2004) 247–253.

[20] S. M. Elbashir, J. Harborth, K. Weber, T. Tuschl, Analysis of gene function in somatic mammalian cells using small interfering rnas, Methods 26 (2002) 199–213.

[21] A. Reynolds, D. Leake, Q. Boese, S. Scaring, W. Marshall, A. Khvorova, Rational siRNA design for RNA interference, Nature Biotechnology 22 (3) (2004) 326–330.

[22] Y. Ding, C. Y. Chan, C. E. Lawrence, Sfold web server for statistical folding and rational design of nucleic acids, Nucleic Acids Research 32 Web Server issue (2004) W135–W141.

[23] B. Yuan, R. Latek, M. Hossbach, T. Tuschl, F. Lewitter, siRNA selection server: an automated sirna oligonucleotide prediction server, Nucleic Acids Research 32 Web Server issue (2004) W130–W134.

[24] M. R. Garey, D. S. Johnson, Computers and intractability: A guide to the theory of NP-completeness, Freeman and Company, New York, 1979.

[25] K. Q. Luo, D. C. Chang, The gene-silensing efficacy of siRNA is strongly dependent on the local structure of mRNA at the targeted region, Biochemical and Biophysical Research Communications 318 (2004) 303–310.

[26] P. Pancoska, Z. Moravek, U. M. Moll, Efficient RNA interference depends on global context of the target sequence: quantitative analysis of silensing efficacy using Eulerian graph representation of siRNA, Nucleic Acids Research 32 (4) (2004) 1469–1479.

[27] G. D. Micheli, Synthesis and Optimization of Digital Circuits, McGraw Hill, 1994.

[28] E. Balas, M. C. Carrera, A dynamic subgradient-based branch-and-bound procedure for set covering, Operations Research 44 (1996) 875–890.

[29] T. A. Feo, M. G. C. Resende, A probabilistic heuristic for a computationally difficult set covering problem, Operations Research Letters 8 (2) (1989) 67–71.

[30] J. E. Beasley, An algorithm for set covering problems, European Journal of Operational Research 31 (1987) 85–93.

[31] M. L. Fisher, An application oriented guide to lagrangian relaxation, Interfaces 15 (2) (1985) 10–21.

[32] S. Ceria, P. Nobili, A. Sassano, A lagrangian-based heuristic for large-scale set covering problems, Mathematical Programming 81 (1998) 215–228.

[33] A. Caprara, M. Fischetti, P. Toth, A heuristic method for the set covering problem, Operations Research 47 (1999) 730–743.

[34] J. E. Beasley, P. C. Chu, A genetic algorithm for the set covering problem, European Journal of Operational Research 94 (1996) 392–404.

[35] A. V. Eremeev, A genetic algorithm with a non-binary representation for the set covering problem, in: Proc. of Operational Research, Springer-Verlag, 1998, pp. 175–181.

[36] A. Caprara, P. Toth, M. Fischetti, Algorithm for the set covering problem, Annals of Operations Research 98 (2000) 353–371.

[37] T. A. Feo, M. G. C. Resende, Greedy randomized adaptive search procedures, Journal of Global Optimization 6 (1995) 109–133.

[38] E. R. Troemel, Chemosensory signaling in c. elegans, BioEssays 21 (1999) 1011–1020.

[39] The Sanger Institute, From Wormbase - the C. elegans genome database. http://www.wormbase.org/ (February, 2004).