The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

Solving Drug Enzyme-Target Identification Problem Using Genetic Algorithm

Yu-Ying Zhao Xue-Mei Ning

College of Science, Beijing Forestry University, Beijing, 100083, China
Emails: zhyuying@bjfu.edu.cn, nxm@amss.ac.cn

Abstract In this paper, a genetic algorithm is proposed to solve the drug enzyme-target identification problem and a model of a more complex case of this problem is considered.

Keywords Drug Design; Drug Target Identification; Genetic Algorithm; Metabolic Network

1 Introduction

Drugs play a key role to cure a disease. However it is well known that every drug has side effects. How to decrease the side effects of drugs is an important problem for the drug-design. In the post-genomic era, drug research focuses more on identification of biological targets (gene products, such as enzymes or proteins) for drugs, which can be manipulated to produce the desired effect (of curing a disease) with minimum disruptive side effects [1][2][7][8][10]. An effective model based on the metabolic networks can help accomplish this work.

In the metabolic networks of organisms, enzymes catalyze reactions and then the reactions can produce metabolites (compounds) that are necessary for life. When enzyme malfunctions some certain compounds will accumulate probably and may result in disease [9]. Such compounds are termed as target compounds and others non target compounds [7][8]. The target compounds can be eliminated effectively by the inhibition of some enzymes. However, the inhibition of some enzymes may stop some non target compounds also. This will damage the metabolic network. Then given a metabolic network and a set of target compounds, how to identify the optimal set of enzymes whose inhibition eliminates the target compounds and incurs minimum damage is a valuable problem for drug design.

In [6], the drug enzyme-target identification problem was considered by introducing a graph model for metabolic networks. In order to solve this optimization problem, Sridhar proposed two algorithms: a branch and bound algorithm and an iterative algorithm. The branch and bound algorithm can find the optimal solution, but this algorithm may become computationally infeasible for very large metabolic networks. The iterative algorithm can arrive at a suboptimal solution within reasonable time bounds. Due to the limitations of the branch and bound algorithm and the iterative algorithm, a genetic algorithm is designed in this paper.

In the model considered by Sridhar, the evaluation of the damage to a metabolic network is based on the assumption that all the enzymes and the compounds are of equal importance in the metabolic network. However this is unnaturally because the enzymes and compounds play different roles in the metabolic network. According to the importance of the enzymes and the compounds we give a different way to evaluate the damage on a network, and so the drug enzyme-target identification problem is defined differently in this paper. Furthermore the genetic algorithm proposed can be generalized to this complex case if only the fitness value is modified simply.

Our experiments on the human metabolic network demonstrate that the proposed algorithm can accurately identify the target enzymes for known successful drugs fast and the model proposed in this paper get the same optimal solution in [7][8].

The rest of the paper is organized as follows. Section 2 formally defines the drug enzyme-target identification problem and our model based on a complex case of this problem. Section 3 presents the genetic algorithms for the proposed problems. Section 4 discusses experimental results. Section 5 concludes the paper.

2 Problem Definition

In the metabolic networks of organisms, enzymes catalyze reactions and then the reactions can produce metabolites (compounds) that are necessary for life. When a target compound needs to be stopped, all the reactions that can produce it must be inhibited. And a reaction can be inhibited only if one reactant is stopped. Given a metabolic network and a set of target compounds, how to identify the optimal set of enzymes whose inhibition eliminates the target compounds and incurs minimum damage is the drug enzyme-target identification problem. If the enzymes and the compounds are considered to be of equal importancečňthe damage can be evaluated by the number of non target compounds that the inhibition of some set of enzymes stops, this problem can be simply defined as follows [7] [8]:

Problem 1: Given a large metabolic network and a set of target compounds, identify the optimal set of enzymes whose inhibition eliminates all the target compounds and stops the minimum number of non target compounds.

Considering the enzymes and the compounds play different roles and then may have different importance in the metabolic network, the damage on the network can be evaluated differently and the drug enzyme-target identification problem can be defined differently correspondingly. Up to now, there are several measures of the nodes in complex networks, including degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, and so on [4]. However there appears to be no compelling evidence at the current time that the more complex centrality measures perform any better as indicators of essentiality than simple degree centrality. Then the degree centrality is adopted to measure the importance of the enzymes and compounds in the metabolic network in this paper. Concretely, all the enzymes and the compounds in the network are attached a weight to denote their importance. The weight of an enzyme relating to its degree can be calculated by the following equations:

$$w_{ei} = \frac{d_{ei}}{\max\limits_{i} d_{ei}}$$

Where w_{ei} denotes the weight of enzyme i and d_{ei} denotes the degree of it, and $\max_i d_{ei}$ denotes the max value of all the d_{ei}. Similarly, the weight of a compound can be calculated as follows:

$$w_{ej} = \frac{d_{ej}}{\max\limits_{j} d_{ej}}$$

Under this assumption, the drug enzyme-target identification problem can be defined as the following model.

Problem 2: Given a large metabolic network and a set of target compounds, identify the optimal set of enzymes whose inhibition eliminates all the target compounds and incurs minimum damage to the rest of the network, where the damage is evaluated by the weighted sum of the stopped enzymes and compounds in the rest of the network.

3 Genetic Algorithm

In this section, we develop genetic algorithms for the proposed problem 1 and problem 2. Genetic algorithm is an adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetic. It is started from a random initial guess solution and attempt to find one that is the best under some criteria and conditions [3].

In the genetic algorithm for our problems, a gene, corresponding to an enzyme, is presented by a binary value which denotes an enzyme exists or not, 1 denotes existence and 0 the otherwise, when some enzymes are inhibited in the network. And a chromosome, corresponding to a solution to the problem, is a set of genes which includes all of the enzymes. A population is a set of chromosomes which are produced in different generation. Operations are defined for a population such as crossover operation, mutation operation and selection operation. These operations are all related to the fitness value of each chromosome that is an important factor in genetic algorithms.

The fitness value defines a score which gives each chromosome a probability to be chosen for breeding or to live. When some target enzymes are inhibited, some reactions will not happen and some compounds will be stopped correspondingly. The fitness function should evaluate the target enzyme sets. In the algorithm for problem 1, the fitness value is calculated as follows:

$$fitnessvalue = \max\{\varepsilon \sum_{i=1}^{|E|} x_{ei} + \sum_{j=1, j\notin TC}^{|C|} x_{cj} - \sum_{ck\in TC} x_{ck}, 0\},$$

where $x_{ei}, i = 1,2,...,|E|$ is presented by a binary value which denotes an enzyme exists or not after some enzymes are inhibited in the metabolic network, and $x_{cj}, j = 1,2,...,|C|$ is presented by a binary value which denotes a compound can be produced or not when some enzymes are inhibited in the network. TC denotes the set of the target compounds. The drug enzyme-target identification problem aims to maximize the value $\sum_{j=1,j\notin TC}^{|C|} x_{cj} - \sum_{ck\in TC} x_{ck}$. In order to identify the optimal enzyme-target set, the fitness value of a solution (some enzyme-target set) is calculated by adding $\varepsilon \sum_{i=1}^{|E|} x_{ei}$ to $\sum_{j=1,j\notin TC}^{|C|} x_{cj} - \sum_{ck\in TC} x_{ck}$. Where is a small number such as 0.000001. For *problem 2*,

the fitness value is calculated by the following equation:

$$fitnessvalue = \max\{\varepsilon \sum_{i=1}^{|E|} w_{ei}x_{ei} + \sum_{j=1, j\notin TC}^{|C|} w_{cj}x_{cj} - \sum_{ck\in TC} w_{ck}x_{ck}, 0\},$$

where w_{ei} is the weight of the enzymes, w_{cj} and w_{ck} are the weights of the compounds.

In each generation, the crossover operation chooses two chromosomes in the population (every chromosome is chosen with a probability proportional to its fitness value) and a random point, and then each of the two chromosomes are sliced into two parts and those parts are exchanged to generate two new chromosomes. For every chromosome, the mutation operation chooses a point randomly and changes its value from 0 to 1 or otherwise. After crossover and mutation operations, a new population is created and the next operation is to select chromosomes continued alive or eliminable in the next generation. This selection operation is based on the fitness value.

The genetic algorithm for the two problems is described detailed in the following:

Input: the metabolic network and the target compound set;

Parameters: *NumGene* is the number of generations, P_k is the set of the chromosomes in the k-th generation, $SizeP_k$ is the number of elements in P_k;

Step 0: Generate an initial population P_1 randomly and compute the fitness value of every chromosome, $k = 1$;

Step 1: If $k > NumGene$, go to step output, else the next population P_{k+1} is generated from population P_k:

(1) Crossover operation for a number of chromosomes in population P_k, this leads to an enlargement of P_k.

(2) Mutation operation on the chromosomes that were created during crossover. Compute the fitness value and find the best chromosome based on the fitness value.

(3) Sort chromosomes based on the fitness value and select $SizeP_{k+1}$ chromosomes from the first in the sorted list.

$k := k + 1$;

Output: the best chromosome that corresponds to the optimal solution computed.

4 Experimental Results

In this section, the paper will present computational results from running the two genetic algorithms presented in this paper. Experiments are deployed on an AMD Athlon 64 processor with 2.0 GHz clock speed and 448 MB main memory. The computational experiments aim to test the efficiency of the genetic algorithms and study the difference of *problem 1* and *problem 2*.

In order to test the efficiency of the proposed genetic algorithms, we use the drugs at the database in KEGG [5] as our benchmarks. KEGG contains a database of known drug molecules along with the enzymes they inhibit and their therapeutic category. Here we use D03080 and D02562, which have been used in [6] [7], as the data information to test our algorithms.

Drug D03080 appears in several networks, including the arachidonic network (hsa 00590) given in Figure 1. This drug inhibits enzyme E1.13.11.34 to stop the compounds C02165, C02166, C05951 and C05952 and it also stops the non-target compounds

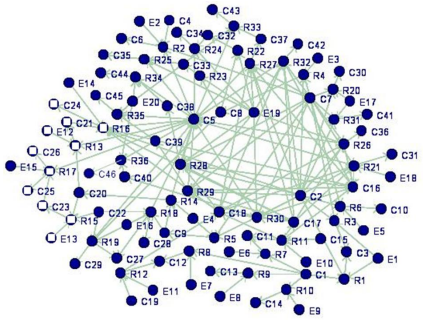

Figure 1: The arachidonic network (hsa00590). E denotes enzyme, R denotes reaction and C denotes compounds. The hollow nodes are stopped in our computational results.

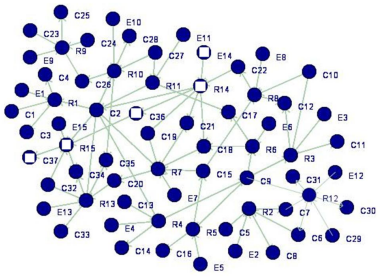

Figure 2: The histidine network (hsa00340). E denotes enzyme, R denotes reaction and C denotes compounds. The hollow nodes are stopped in our computational results.

C05356, C04805, C00909, C04853 and C14827. However when we consider compounds C02165, C02166, C05951 and C05952 as the target compounds, the genetic algorithm for problem 1 finds that the optimal target enzyme set should be E3.3.2.6 and E4.4.1.20 and the non-target compound is C04853. This result is same to the result got in [7].

Drug D02562 exists in several networks, one is the histidine metabolic network (hsa00340) given in Figure 2. This drug inhibits E.1.4.3.4 to eliminate the target compounds C05827 and C05828. When given the target compounds C05827 and C05828 and the associated network hsa00340, the genetic algorithm can find the same results.

These results indicate that our proposed genetic algorithm can solve problem 1 efficiently.

The same results as above are gotten when the genetic algorithm for problem 2 is implemented on drugs D03080 and D02562. From Figure 1 and 2, we can see that the degrees of the stopped enzymes and compounds are all small. This suggests that when the importance of the enzymes and the compounds are evaluated by the degree centrality problem 1 and problem 2 are equal for the two networks considered. But this may not be the case if other target compounds or other networks are considered.

Additionally, all the experiments are finished within few milliseconds.

5 Conclusion

In this paper, we proposed a genetic algorithm to solve the drug enzyme-target identification problem. Secondly, we introduced the degree centrality to evaluate the importance of the nodes in the metabolic network and put forward a new problem. Thirdly, for the new problem we devised a genetic algorithm.

Our experiments on the known drugs show that the genetic algorithm can solve the drug enzyme-target identification problem efficiently. The results also suggest that problem 1 and problem 2 will get the same optimal solution for the two networks considered if the importance of the nodes in the network is evaluated by the degree centrality. However this will be not the case if other networks are considered or other centralities are adopted. This needs further consideration.

Acknowledges

We would like to thank professor Xiang-Sun Zhang, Zhen-Ping Li, Rui-Sheng Wang and the anonymous reviewers for their help.

References

[1] C Smith. Hitting the target. Nature, Mar 2003, 422:341-347.

[2] J Drews. Drug discovery: A historical perspective. Science, Mar 2000, 287(5460):1960-1964.

[3] Kimmo Nieminen and Sampo Ruuth. Genetic algorithm for finding a good first integer solution for MILP. http://www.doc.ic.ac.uk/research/technicalreports/2003/DTR03-4.pdf

[4] Mason, O. and Verwoerd, M. Graph theory and networks in biology. Systems Biology, IET, 2007, 1:89-119.

[5] M Kanehisa, S Goto, S Kawashima and A Nakaya. The KEGG database at GenomeNet. Nucleic Acids Res., 2002, 30(1):42-46.

[6] Padmavati Sridhar, Tamer Kahveci, and Sanjay Ranka. Opmet: A metabolic network-based algorithm for optimal drug target identification. Technical report, CISE Department, University of Florida, Sep 2006.

[7] Padmavati Sridhar, Tamer Kahveci, and Sanjay Ranka. An iterative algorithm for metabolic. network-based drug target identification. PSB, 2007, 12:88-99.

[8] Padmavati Sridhar, Bin Song, Tamer Kahveci, and Sanjay Ranka. Mining metabolic networks for optimal drug targets. PSB, 2008, 13:291-302.

[9] R. Surtees and N. Blau. The neurochemistry of phenylketonuria. European Journal of Pediatrics, 2000, 159(S2):109-113.

[10] T. Takenaka. Classical vs pharmacology in drug discovery. BJU International, Sep 2001,88(2):7-10.