NP COMPLETENESS FOR OPTIMAL ENZYME COMBINATION IDENTIFICATION

TAMER KAHVECI*

Computer and Information Science and Engineering,
University of Florida,
Gainesville, FL, 32611
E-mail: tamer@cise.ufl.edu

We prove that the problem of finding the optimal set of enzymes is NP-complete. The *exact cover by 3-sets (X3S)* can be reduced to the drug target identification problem in polynomial time. We first state the X3S problem, which is NP-Complete.

1. Problem definition

We develop a graph based representation that captures the interactions between reactions, compounds, and enzymes. Our graph representation is a variation of the boolean network model [3,1]. R, C, and E denote the set of reactions, compounds, and enzymes respectively. The vertex set consists of all the members of $R \cup C \cup E$. A vertex is labeled as reaction, compound, or enzyme based on the entity it refers to. Let V_R, V_C, and V_E denote the set of vertices from R, C, and E. A directed edge from vertex x to vertex y is then drawn if one of the following three conditions holds: (1) x represents an enzyme that catalyzes the reaction represented by y. (2) x corresponds to a substrate for the reaction represented by y. (3) x represents a reaction that produces the compound mapped to y.

Figure 1 illustrates a small hypothetical metabolic network. In this figure, C_4 is the target compound (i.e., the production of C_4 should be stopped). In order to stop the production of C_4, R_2 has to be prevented from taking place. The obvious solution is to disrupt one of its catalyzing enzymes (E_2 in this case). Another is by stopping the production of one of its reactant compounds (C_2 or C_3 in this case). If we stop the production of C_2, we need to recursively look for the enzyme which is indirectly responsible for its production (E_1 in this case). Thus, the production of the target compound can be stopped by manipulating either E_1 or E_2.

Figure 1 shows the disruption of E_2 and its effect on the network. Inhibiting E_2 results in the knock out of compounds C_5, C_8 and C_9 in addition to the target compound, C_4. Note that the production of C_7 is not stopped since it is produced by R_1 even after

*to whom correspondence should be addressed

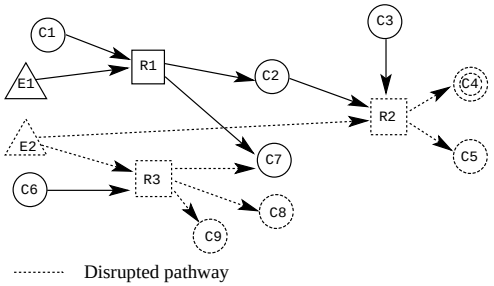

Figure 1. A graph constructed for a hypothetical metabolic network with three reactions R_1, R_2, and R_3, three enzymes E_1, E_2, and E_3, and nine compounds $C_1, \cdots, C_9$. Circles, rectangles, and triangles denote compounds, reactions, and enzymes respectively. Here, C_4 (shown by double circle) is the target compound. Dotted lines indicate the subgraph removed due to inhibition of enzyme E_2.

the inhibition of E_2. We define the number of non-target compounds knocked out as the *damage*, the manipulation of an enzyme set causes to the metabolic network. In this case, the damage of inhibiting E_2 is 3 (i.e., C_5, C_8 and C_9). The damage of inhibition of E_1 is 2 (i.e., C_2 and C_5). The important observation is that E_1 and E_2 both achieve the effect of disrupting the target compound, C_4. Hence, E_1 and E_2 are both potential drug targets. However, E_1 is a better drug-target than E_2 since it causes lesser damage.

Formally the optimal enzyme combination identification problem is: "*Given a set of target compounds T ($T \subset C$), find the set of enzymes X ($X \subseteq E$) with minimum damage, whose inhibition stops the production of all the compounds in T.*"

For simplicity, we assume that the input compounds to all reactions are present in the network and that there are no external inputs. Different enzymes and compounds may have varying levels of importance in the metabolic network. We consider all the enzymes and compounds to be of equal importance. This assumption can be relaxed by assigning weights to enzymes and compounds based on their role in the network. Also, we are not incorporating back-up enzyme activities [2] in this paper. This can be achieved by creating vertices for sets of enzymes in our graph representation. However, we do not discuss these extensions in this paper.

2. NP-completeness of the problem

We prove that the problem of finding the optimal set of enzymes is NP-complete. The *exact cover by 3-sets (X3S)* can be reduced to the drug target identification problem in polynomial time. We first state the X3S problem, which is NP-Complete [?].

> **X3S:** Given a set of n items $X = \{x_1, x_2, \cdots, x_n\}$, where $n = 3m$ is a multiple of 3. Given 3-sets (i.e., sets with three items) c_i, $1 \leq i \leq k$ such that $c_i \subseteq X$ and $\cup_{i=1}^{k} c_i = X$. The X3S problem seeks whether there exists m 3-sets c_i whose union is X.

Figure 2 shows how the X3S problem can be reduced to drug discovery problem in polynomial time. Given an instance of the X3S problem, we map each of the 3-sets c_i to

a compound C_i. We map each of the items x_j to a reaction R_j. We draw a hypothetical edge from C_i to R_j if the 3-set c_i contains x_j, indicating that the compound C_i is used by the reaction R_j. We draw an edge from each of the R_j to a single hypothetical target compound indicating that the target compound is produced by all the k reactions. In the hypothetical metabolic network, each compound C_i corresponding to a 3-set is produced by a single reaction that also produces $k + 1$ other compounds. Each such reaction is catalyzed by a single and unique enzyme.

The set of enzymes whose inhibition eliminates the target compound with minimum damage in the hypothetical network produces the answer to the X3S problem. The sketch of the proof is as follows: In order to eliminate the target compound, all the reactions R_j, $1 \leq j \leq n$, needs to be stopped. This can only be done by eliminating at least one of the input compounds C_i for each R_j. Eliminating each input compound stops three reactions. Thus, the set of compounds that needs to be eliminated should cover the entire reaction set R_j, $1 \leq j \leq n$. The minimality of the set of compounds that needs to be eliminated is dictated by the definition of the damage. Each compound C_i can be eliminated by stopping the reaction that produces it. This can only be done by inhibiting the corresponding enzyme. Stopping each such reaction incurs a damage of $k + 2$. This is because the reaction that produces C_i also produce $k + 1$ additional non-target compounds. Thus, if the solution set to the drug target identification problem contains q enzymes, it incurs a damage of $q(k + 2)$. There exists a solution to the X3S problem if and only if $q = m$. Therefore, since X3S is an NP-complete problem, the drug target identification is an NP-complete problem too.

References

1. S.A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, 1993.
2. M.T.A. Ocampo and W. Chaung et. al. Targeted deletion of mNth1 reveals a novel DNA repair enzyme activity. *Mol Cell Biol.*, 22(17):6111–21, Sep 2002.
3. R. Somogyi and C.A. Sniegoski. Modeling the complexity of genetic networks: Understanding multi-gene and pleiotropic regulation. *Complexity*, 1:45–63, 1996.

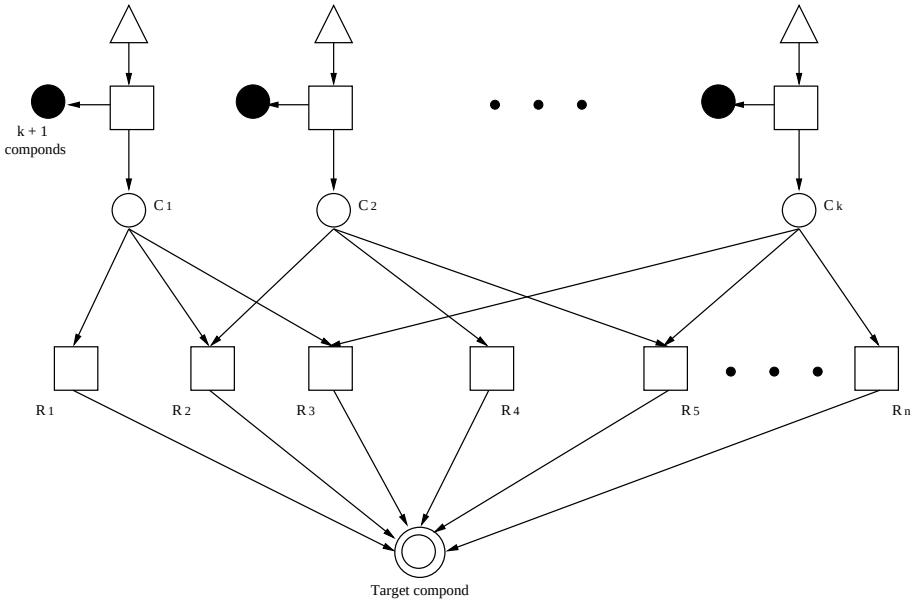

Figure 2. Polynomial time mapping of the X3S problem to the drug target identification problem using metabolic networks. Each black circle denotes a set of $k + 1$ compounds.