Inverse Design of Large Molecules using Linear Diophantine Equations

Shawn Martin[1] (smartin@sandia.gov), W. Michael Brown[1], Derick Weis[2], Donald Visco[2], John Kenneke[3], and **Jean-Loup Faulon**[4]

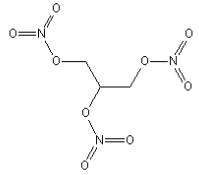
[1]Sandia National Laboratories, Albuquerque, NM; [2]Tennesse Technological University, Cookeville, TN; [3]U.S. Environmental Protection Agency, National Exposure Research Laboratory, Ecosystems Research Division, Athens, GA; and [4]Sandia National Laboratories, Livermore, CA.

Abstract

We have previously developed a method for the inverse design of small ligands. This method can be used to design novel compounds with optimized properties, such as drugs. A key step in our method involves computing the Hilbert basis of a system of linear Diophantine equations. In our previous application, the ligands considered were small peptide rings, so that the resulting system of Diophantine equations was relatively small and easy to solve. When considering larger molecules, however, the Diophantine system is larger and more difficult to solve. Here we present a method for reducing the system of Diophantine equations before they are solved, allowing the inverse design of larger compounds.

Signature

Our method for the inverse design of small molecules is based on a fragmental descriptor called *signature*. Signature encodes molecular structure by counting the occurrences of fragments in a molecule. The molecular signature encoding of nitroglycerin is shown below.



Nitroglycerin
3 o_(n_c_)
6 o=(n_)
3 n_(o_o=o=)
5 h_(c_)
2 c_(o_h_h_c_)
1 c_(o_h_c_c_)

Constraint Equations

Signature can also be used to reverse engineer molecular structures. This is done by deriving constraint relations that must be present between fragments in order that the fragments may be combined to form a molecule. These constraints consist of a *graphicality* equation and multiple *consistency* equations. The graphicality equation assures that the molecular fragments can be combined to form a connected molecular graph and assumes the form

$$\sum_{i\geq 2}(i-2)n_i - n_1 + 2 = 2z,$$

where n_i is the number of vertices of degree i (number of atoms connected to i other atoms), and z is a non-negative integer.

The consistency equations assure that the molecular fragments can be re-connected such that the molecular bonds are consistent. Here we show a consistency equation which guarantees that the number of bonds of type O → C must be equal to the number of bonds of type C → O for nitroglycerin.

3 o_(n_c_) = 2 c_(o_h_h_c_) + 1 c_(o_h_c_c_)

Reductions

We reduce the constraint equations using three simple linear transformations. To describe these transformations, suppose we have m equations and n variables. We write our Diophantine system as $\mathbf{A}^0\mathbf{x}^0 = \mathbf{b}$, where $\mathbf{A}^0_{m\times n} = (a^0_{ij})$, $\mathbf{x}^0_{n\times 1} = (x^0_j)$, $\mathbf{b}_{m\times 1} = (b_i)$, with a^0_{ij}, b_i integer and x^0_j non-negative integer. We use the superscript notation to denote steps in our reduction, never exponentiation.

In our first reduction, we eliminate equations of the form

$$x^0_j = \sum_{k\neq j} a^0_{ik}x^0_k \qquad (1)$$

where $a^0_{ik} \geq 0$ for $k \neq j$. To eliminate an equation of this form, we replace any occurence of x^0_j in $\mathbf{A}^0\mathbf{x}^0 = \mathbf{b}$ with the corresponding sum $\sum_{k\neq j} a^0_{ik}x^0_k$. We can then eliminate both the variable x^0_j and the equation $x^0_j = \sum_{k\neq j} a^0_{ik}x^0_k$ to obtain a reduced system $\mathbf{A}^1\mathbf{x}^1 = \mathbf{b}$.

Our next transformation is achieved by considering equations of the form

$$2x^p_j = \sum_{k\neq j} a^p_{ik}x^p_k \qquad (4)$$

where $a^p_{ik} \geq 0$ for $k \neq j$. In this case, we observe that $a^p_{ik} > 1$ can be replaced by the remainder of a^p_{ik} divided by 2, provided that x^p_j is adjusted appropriately.

Finally, it often occurs that $\mathbf{A}^q$ has a few identically zero columns after the previous reductions, and even some repeated columns. Identically zero columns represent free variables, which can be removed, and repeated columns represent groups of variables that occur together in every equation. These variable groups can be replaced by single variables and recovered later by solving equations with the form

$$\sum_{i_c} x^q_{i_c} = x^{q+1}_j, \qquad (6)$$

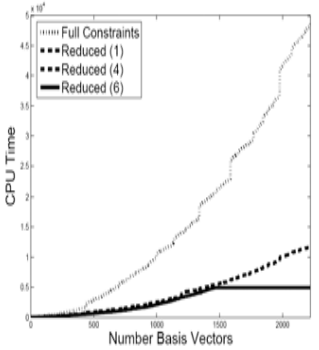
where the sum is over the only the indices i_c corresponding to a specific set of repeated columns.

Diophantine Solver

To solve the reduced system $\mathbf{A}^r\mathbf{x}^r = \mathbf{b}$, we use the Contejean-Devie Diophantine solver. This solver produces a *Hilbert basis* $\mathbf{H}^r$ for the system $\mathbf{A}^r\mathbf{x}^r = \mathbf{b}$. This basis consists of a minimal set of solutions to $\mathbf{A}^r\mathbf{x}^r = \mathbf{b}$ such that any other solution can be obtained via non-negative integer linear combinations of the solutions in $\mathbf{H}^r$. To obtain the basis H for the original system $\mathbf{A}^0\mathbf{x}^0 = \mathbf{b}$, we perform a sequence of transformations which include the addition of unit vectors for any free variable previously eliminated as well as new minimal solutions for any repeated columns that were removed; replacing the variables x^{q+1}_j with the various possibilities for $x^q_{i_{c_1}}$; and using linear transformations to reverse the operations in (1).

Results

We first applied our algebraic reduction to the constraint equations previously derived for peptide rings in an LFA-1/ICAM-1 study. We next applied our reduction to the inverse design of -secretase inhibitors for Alzheimer's disease. This dataset consisted of 61 compounds with varying IC values. Finally, we applied our reduction to the design of non-toxic but still effective conazole fungicides. These 27 fungicides were obtained from the Enivronmental Protection Agency's Persistent, Bioaccumulative, and Toxic (PBT) Profiler database (www.pbtprofiler.net), each with a corresponding fish chronic toxicity value (ChV).



	Original	Reduced	% Reduction
Peptides			
vars	49	34	30.6
eqs	24	11	54.2
cpu time	–	–	90.0
γ-secretase			
vars	68	49	27.9
eqs	21	9	57.1
cpu time	–	–	98.0
Conazoles			
vars	91	64	29.7
eqs	29	15	48.2
cpu time	–	–	94.1

Discussion

We have proposed a simple method for reducing a linear system of homogeneous equations when using the signature molecular descriptor for inverse design of chemicals. We have tested the reduction on three datasets, including a set of ICAM-1 inhibitory peptides, a set of γ-secretase inhibitors, and a set of conazole fungicides. On these three datasets we achieved an average reduction of 29.4% in the number of variables and 53.2% in the number of equations, resulting in an average reduction in computation time of 94.0%. This increase in efficiency allows us to use the signature descriptor to design large molecules, previously impossible with our technique.

Acknowledgements

This work was funded by interagency agreement (IAG) DW89921601 between the Environmental Protection Agency (EPA) and Sandia National Laboratories. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a LockheedMartin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.