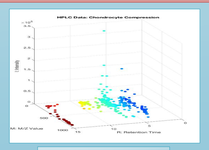
Deducing Active Metabolic Pathways From Untargeted Metabolomics Data ----Formulating the Pathway Recovery Problem

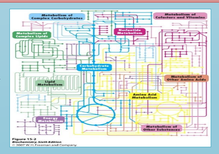
Daniel Salinas, Xuying Wang, Brendan Mumey at Montana State University

Introduction

Metabolic data requires integrating both by its subject and its artifacts. Metabolism is operated on a range of thousands of compounds and reactions. These operations are divided into functional units known as pathways. However, these units do not operate independently. This reflects the complexity of metabolic data. Since metabolic data is generally interpreted in terms of active pathways, it is therefore necessary to generate the untargeted data into active pathways expressed in a set of metabolites. This poster will explore ways to mine the untargeted metabolic data gathered from the June lab at Montana State University into active pathways that can be easily interpreted by biologists.



Figur1: untargeted data



Figur2: active pathways

Pathway Recovery From Untargeted Data

Untargeted Data:

- A metabolite ionizes to an m/z value within a certain tolerance, and not all of these ions strike the LC-MC detector at the same time.
- A given m/z value can be identified with a set of possible metabolites.
- Different metabolites that share a m/z value differ in their retention times.

Typically, we solve this situation by running a targeted analysis:
A known quantity of each metabolite is placed in a sample and the resulting range of m/z values and retention times that correspond to the expected intensity are recorded. However, doing this analysis for all metabolites are not always possible or practical.

Formulating pathway recovery form untargeted data:

We came up with a different approach called database lookup. This way of analyzing untargeted data tries to directly investigate using the m/z values for pathway extraction. Our pathway recovery starts by querying METLIN with the untargeted data, and then querying KEGG with the result of the METLIN look-up. The whole process is represented in Figure3 .

Figure 3

Formulating Pathway Recovery Using Set Cover

Set-Covering problem

Set-Covering Problem consists of a finite set U called the universe, and a family F of subsets of U, such that every element of U belongs to at least one subset of F.

$$U = \bigcup_{s \in F} s$$

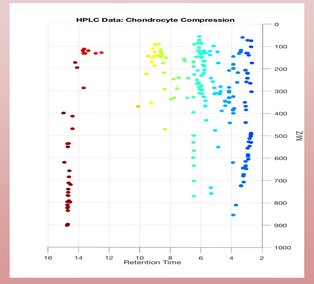
We say a subset of $S \in F$ covers all elements in U. Our goal is to find a minimum size subset $C \subseteq F$ whose members covers all of U.

$$U = \bigcup_{s \in C} S$$

The cost of the Set-Covering is the size of C, which defines as the number of sets it contains and we want $|c|$ to be minimum.

Metabolites, Retention Time & m/z Value

- Retention time can be used to estimate how many different metabolites correspond to the same m/z value.
- Metabolites are often clustered by retention time, and the number of incidents in these clusters is most likely the minimum number of the different metabolites corresponding to an m/z value (Figure 4).
- In Figure 4, there two retention times that share the same m/z value, that means at least two metabolites shared the same m/z value.



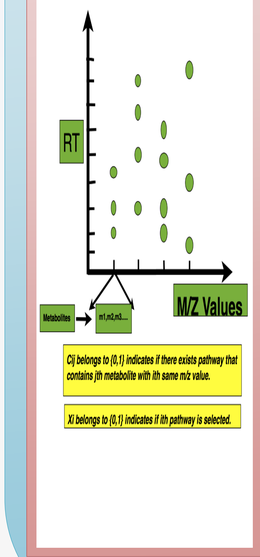
Pathway Recovery Set-Covering Problem

- Colored Set-Covering Problem consists of a finite set U called the universe. Let U be a set of m/z values. $\boldsymbol{U = \{u_1, u_2, \ldots, u_i, u_n\}}$. U also has an associated function $\boldsymbol{f}$: $\boldsymbol{U \rightarrow Z}$, where $\boldsymbol{f(u_i) = r_i}$ and $\boldsymbol{r_i}$ the number of distinct colors (retention time) that are associated with $\boldsymbol{u_i}$(m/z value).
- A family F of subsets of U, such that every element of U belongs to at least one subset of F. A subset $\boldsymbol{S \in F}$. $\boldsymbol{S = \{s_1, s_2, \ldots, s_i, s_m\}}$. S also has an associated function $\boldsymbol{g_i}$: $\boldsymbol{s_i \rightarrow C}$, where C is the set of colors.
- Each Element in $\boldsymbol{s_i}$ has a color and for any $\boldsymbol{u_i \in s_j \cap s_k}$, $g_j(\boldsymbol{u_i})$ may or may not equal $\boldsymbol{g_k(u_i)}$.
- The pathway recovery set-covering problem is to find a minimal cover of U such that $|\bigcup g_i(\boldsymbol{u_i})|$ over all $\boldsymbol{s_j} \in C$ equals $\boldsymbol{f(u_i)}$.
- f: $U \rightarrow \mathrm{Z} \times \mathrm{Z}$, where $\boldsymbol{f(u_i)}$ are the lower and upper bounds on the number of colors in a cover.

Integer Linear Programming

Formulation of Pathway Recovery Problem

An integer linear program is an optimization problem with linear constraints. We defined an ILP that encodes a related pathway recovery set-covering problem, then we adapted both formulations and coded it using Cplex.



The linear Program is defined as:
Objective: minimize $\sum x_i$
(minimize the pathways that can cover all of m/z values)
Subject to:

$$v_1 \leq \chi_1 + x_2 + .. + \leq v_2;$$

(The lower bound and upper of metabolites for a m/z value)

$$\chi_{ij} \leq c_{ij};$$

(If a pathway is selected, then a color is assigned to it)

$$c_{ij} \leq x_{ia} + x_{ib};$$

(If a color is chosen, then a pathway has been chosen to cover it)

Where $x_i \in \{0,1\}$ indicates if ith pathway is selected or not; $c_{ij}(\mathrm{i}, \mathrm{j} \in \mathrm{Z}) \in \{0,1\}$ indicates if there exists a selected pathway that contains jth metabolite with ith m/z value; $(v_1, v_2 \in Z)$ upper bound and lower bound of metabolites for a m/z value.

Data Testing

- The code includes a test case where (f7,f8,f9,f10) and (f5,f6) are valid covers, but clearly the second one is smaller. Our Cplex code can find the smaller cover as an optimal solution.
- Another test that we did was to test whether the code finds the smallest feasible cover or not. The result we got is that if the constraints are set up right then the Cplex code can always find the feasible cover.

Acknowledgements and References

- I would like acknowledge Dr. Mumey and Ph.D. candidate Daniel Salinas for all their support.
- My work was part of the 2017 Summer Research Experience for Undergraduates at Montana State University sponsored by the National Science Foundation
- The main references for this poster come from Daniel Salinas's comprehensive article on "Deucing Active Metabolic Pathways from Untargeted Metabolomics Data"