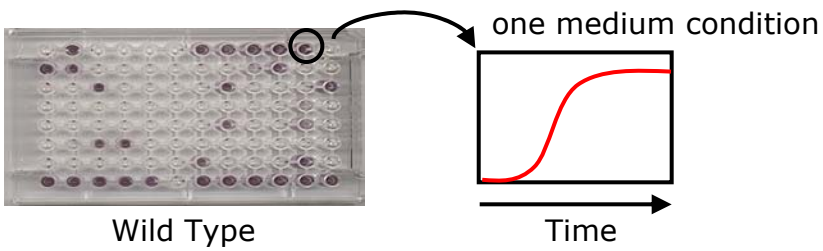
Phenotype Profiling of Single Gene Deletion Mutants of E.coli Using Biolog Technology

Yukako Tohsato and Hirotada Mori

1) *Department of Bioscience and Bioinformatics, Ritsumeikan University*
2) *Graduate School of Biological Sciences, Nara Institute of Science and Technology*

Phenotype MicroArray

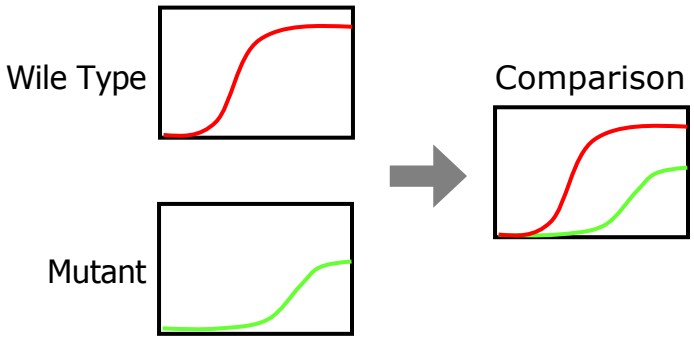
- Phenotype MicroArray (PM) technology is designed to test a large number of cellular phenotypes simultaneously (Bochner et al., 2001)
- The system allows monitoring of cellular respiration during cell growth on 96-well microtiter plates under a maximum of 1920 different medium conditions by colorimetrically detection of generation of purple colored Formazane from Tetrazolium dye corresponding to the intracellular reducing state by NADH simultaneously.



No.2

Wile Type V.S. Mutant

- Compare Single Gene Deletion Mutant to Wild Type to determine gene function



No.3

What is clarified by PM?

- Genotype → Phenotype → Metabolic Pathway
- Knock out a gene → Which phenotypes change ?

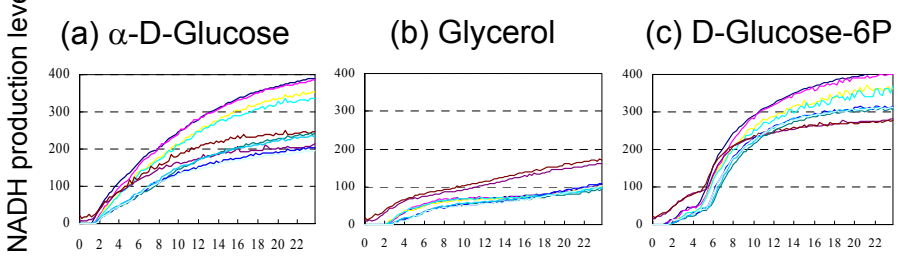
- In this study, analysis using PM data was performed to discover new alternative pathways and identify functions of genes for which the functions have yet to be determined.

No.4

Our Approach and Method

Materials

- *Escherichia coli* K-12
  - 1 host strain (BW25113)
  - 204 mutants (Keio collection library (Baba et al., 2006))
- Growth conditions : 1920 （ex. Suger, Nitrogen , Drug etc）
- Measuring time： 15-minute intervals over 24-hour period.
  - Host strain ：10 times, Mutants ： 2 times



Web application for PM

- PM data consist of enormous amount observation points.
- It made both Excel and me overflow frequently!
- So, we are constructing web applications for PM data to search and analysis.
  - Acknowledgement
    - Yusaku Mazaki
    - Tomohiro Fujita

No.7

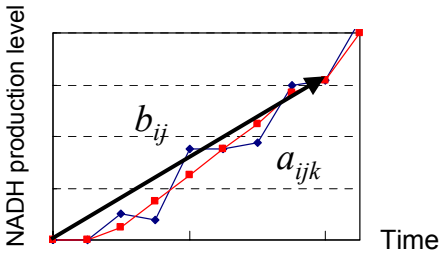
Our Approach

- We report the results obtained by applying the proposed method to PM data from wild-type and 45 single gene deletion strains.
  - The strains related to central metabolism.
  - Selected 288 medium condition of carbon and nitrogen sources.
- Our methods
  - Vectorization of raw data
  - Hierarchical Clustering

No.8

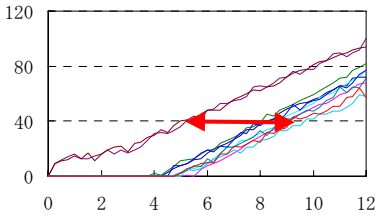
Vectorization (1)

- The original raw data less than a threshold were substituted with zero (zero-substitution).
- The data were smoothed by taking an average of consecutive nine observation points (two hour). $a_{ijk} = \frac{1}{9}\sum_{k'=k}^{k+8} x_{ijk'}$
- Each well is expressed with its maximum slope b_{ij}.
- PM data for each strain can be considered as 288-dimension vector data.

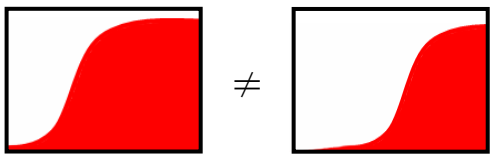


Why did we select maximum slopes ?

- Growth time shift.



  - Maximum values ?
  - Areas ?

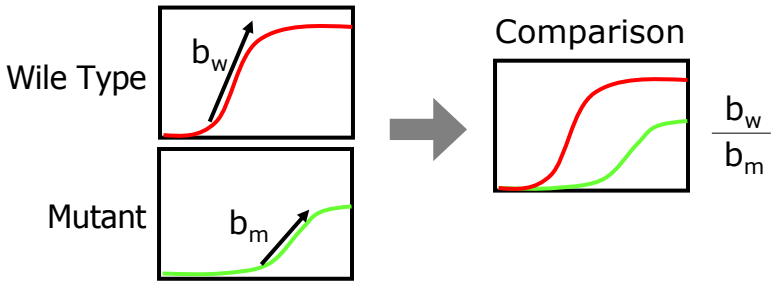


  - Maximum slopes !

No.10

Vectorization (2)

- Calculated the ratio between vector data of mutant and wild.



- The ratio data are converted to a data of 0s, -1s and 1s by setting thresholds for the vector ratios 1.2 and 0.8.

$$v_k = (v_{k1}, v_{k2}, \ldots, v_{k288})$$

  - +1 indicates that the gene deletion activate the respiratory activity.
  - -1 indicates that the gene deletion repress the respiratory activity.

No.11

Clustering Methods

- Manhattan distance $d(v_1, v_2) = \sum_{k=1}^{n} |v_{1k} - v_{2k}|$
  - Distance between Knockout strain and other one
    - The degree of similarity using the distance tends to become larger for pairs of vector data that are less similar.
- Ward's hierarchical method
    - Less susceptible to noise and outliers as compared to other hierarchical clustering methods.

No.12

Assignment of condition and P-values to clusters

- Calculated a P-value for each experimental condition (Tavazoie et al., 1999).

$$P_{\pm 1} = 1 - \sum_{i=0}^{k-1} \frac{\binom{C}{i}\binom{G-C}{n-i}}{\binom{G}{n}}$$

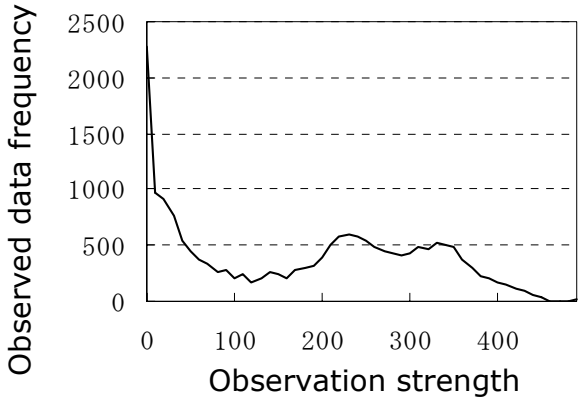
  - *G* is the number of all strain data.
  - *C* is the number of the selected group of strains.
  - *n* is the number of strains with a value of +1 (or -1).
  - *k* is the number of strains with a value of +1 (or -1) within the selected strain group.

No.13

Results !

Selection of threshold for zero-substitution

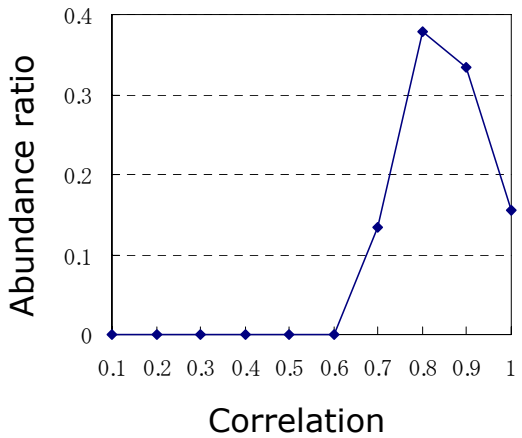
- The value of 100 was set as the threshold for zero-substation.
  - Low observation strength may lead to unstable experimental measurement.



No.15

Repeatability of the PM data

- Calculated Pearson correlation coefficients between PM data for 10 trials in the wild-type.
  - 0.67 ～ 0.96 (Ave. 0.81)

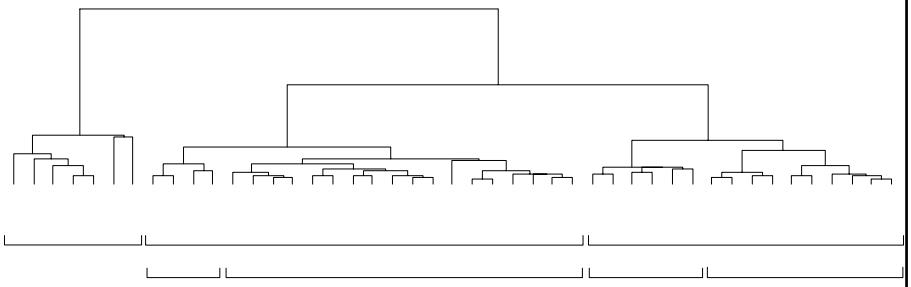


High repeatability

No.16

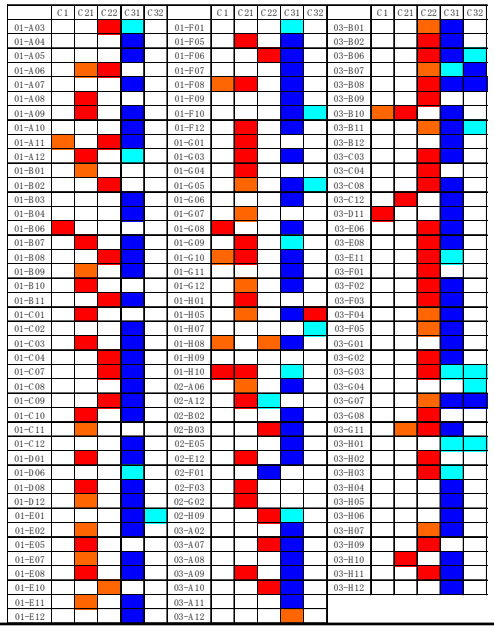
Clustering Result

- 45 single-gene-knockout mutants in central metabolism under 288 condition.
- Three major clusters C_1 to C_3 were obtained.



No.17

Characteristic growth condition



Phenotype profiles of these five clusters.

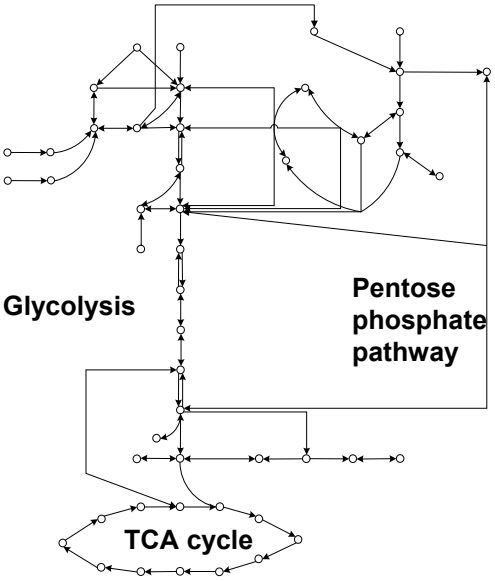
- P_{+1} <= 0.05
- P_{+1} <= 0.1
- P_{-1} <= 0.05
- P_{-1} <= 0.1

C_2 group activated cellular respiratory activity.

C_3 group repressed cellular respiratory activity.

No.18

Distribution of gene-knockout affects



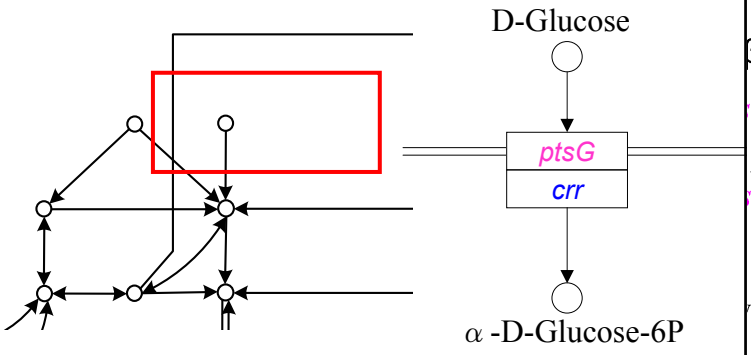
Cluster C_1: green
Cluster C_{21}: red
Cluster C_{22}: pink
Cluster C_{31}: blue
Cluster C_{32}: light blue

- The mutants of cluster C_1 are located at the early stage of the glycolysis.
- Four mutants in cluster C_{31} are closely related to the TCA cycle.

No.19

Analysis result for Phoshotransferase system

- PtsG and Crr form enzyme II complex as PTS (phosphotransferase) system .
- However, deletion of *ptsG* and *crr* genes affect opposite direction in phenotype profiles.

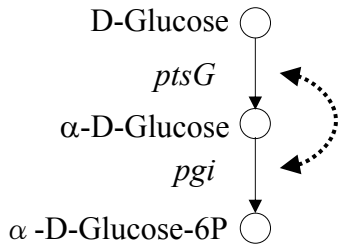


- Crr might function as switching for further steps after transportation of Glucose.

No.20

Phenotype Similarity and Pathway Distance (1)

- What kind of relationship exist between them ?
- Minimal pathway distance : The number of the compound on shortest path between given genes.



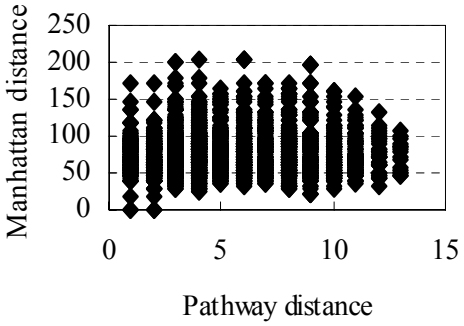
Between *ptsG* and *pgi*
Distance: 1

- Major metabolic path data is represented one adjacency matrix of a directed graph.
- The step between the two compounds in the same metabolic map can be extracted using shortest paths algorithms.

No.21

Phenotype Similarity and Pathway Distance (2)

- We calculated the minimal pathway distance for all strain pairs whose knockout genes are involved in central metabolism.
- For established pairs, phenotypic similarity were determined.



- The results showed no correlation between them!

No.22

Conclusion

- We performed to analyze further insight into central metabolic pathway network to PM data.
  - Medium conditions that activate or repress cellular respiratory activities for the different strain groups were identifies.
  - These results suggested the possibility of metabolism steps with unknown bypass route.
- However, our proposal methods have insufficient sensitivity to continue to identify functions of genes of uncertain function of to analysis for further large-scale data.
  - Robustness
    - Alternative passes
    - Bypass route
    - Unknown passes

No.23

Future works

- Computational method for prediction about bound strength among known reactions.
- Double gene knockout experiments.
- Combination PM and another high-throughput data.

No.24

Thank you very much for your attention.