

Omics Data Handling Assignment

1. Introduction

The advent of high throughput technologies has seen an explosion and diversification in biological data. One of the biggest challenges facing omics is integrating data from multi-omic experiments in a meaningful and rigorous way. Open access to this data through ... databases such as NCBI and UniProt has seen an increased use of data driven methods in biological and medical research.

One of the major issues faced by those integrating biological data is the lack of data heterogeneity. There are no standards used, this includes names and symbols for genes and proteins, or ontologies of terms used. Also all the major databases use different unique IDs or accession numbers. An example is gene names, even with guidelines set for a standard nomenclature set as long ago as 1979 alternative gene names are still appearing in literature today. This exasperates the problems of data collation as manual checks need to be made for consistency and redundancy of data from different sources.

The big data era has led to the need for new database methodologies. The well-established relationship databases are replaced by more flexible and scalable NoSQL (not only SQL) databases, Amazon, Google and Facebook all use NoSQL in various types .NoSQL data modelling often starts from the application-specific queries, i.e. “What questions do I have?”, as opposed to relational modelling which is driven by the structure of the available data and asks “What answers do I have?”. NoSQL data structures are less rigid and cope with the fuzziness of biological data and can rapidly adapt to changing requirements.

In this assignment we develop a NoSQL database designed to integrate the data from four groups that were collaborating in analysing an interesting variant of Escherichia coli (E.Coli) M-12 strain M5151. This included transcriptome, proteome and metabolome experimental data consisting of either normalised expression values or ratios.

We designed a database to collate the data so that it could be compared and shared between the groups. This data was then used to detect metabolic pathways that were differentially expressed between the wild-type and mutant stains of E.coli.

2. Methods

The specification for this project was to implement an integrative database where all aspects of the data can be shared between the research groups and meaningful comparisons can be made between the experimental datasets.

Group	Experiment	E.coli strain	Platform	Experimental data
1	Transcriptomics	E.coli M12 Mutant	Affematrix	Normalised quantitative values
2	Transcriptomics	E.coli M12 Wild Type	Incyte	Normalised quantitative values
3	Proteomics	Mutant/Wt	LC-MS	Ratios
4	Metabolomics	Mutant/Wt	MS	Modulated ratios

1. Table 1 Research Groups

Design

As the LSSR database software had been chosen by the groups the data structure was already defined. This data structure is flexible enough to store the minimum data requirements from all four groups.

Molecule data :

- Unique ID
- External cross references, UniProt, NCBI Entrez geneID, KeggID
- Protein and gene name and symbol
- Function, mass, No of amino acids
- Platform used to collect data
- Experiment ID , cross-referenced to the experiment data below.
- Experimental result ,Value or ratio

Experiment data:

- Unique experiment ID, This is cross referenced in the Molecular data structure above
- Name of group/contact
- The type of experiment, i.e. which omics transcriptomics, proteomics or metabolomics ,
- The E.coli strain studied, and type of value ie, normalised quantitative value or ratio data 9 this should include information on the normalisation procedure used.

A NoSQL database structure was used via the “Deploy software” from PADB which was used to implement the database .

Cross-mapping

The first step in data is to cross-map all the molecules to a common well used identifier. In this case each of the first three datasets were mapped to SwissProt/UniProt ID's using UniProt mapping facility. The results were merged into a single mapping reference table which contained all the molecules with their cross references. This was used later to merge the experimental data to the relevant protein information

Data Curation

The UniProt IDs were then used to collect all the required information from the UniProt database, (IDs, gene name/symbol, function etc.). This also included their Entrez gene ID and Kegg ID. A unique molecule identifier was added as the unique key for the database. The Mygene bioconductor package was used in R to get the same information from NCBI, the data was checked for consistency between the sources, and inconsistencies were further investigated with Ecogene.

At this point contaminant molecules that did not originate from E.coli were removed after checking they had no E.coli paralogs using blast. This include human albumin and several proteins from the bottlenose dolphin.

Data Merging

This data was then merged with the results, (values or ratios), for each experiment separately. This resulted in a table with all three datasets with each molecule being uniquely identified (one database ID for each UniProt ID) with all the specified information.

For data set four the metabolite names were mapped to their Kegg ID using the MetaboAnalyst website (Xia et al., 2015) and a unique data base Id added. Some molecules had to be checked manually for inclusion with Kegg ID's being found directly from Kegg.


Database Building

This data was then aligned within a spreadsheet to the correct column as specified by the DeployDB software. A reference file was made to include experimental details for each dataset. The deploy software was then run to implement the database into searchable interlinked webpages. The database was enabled for column removal, column sorting and exporting datasets.

3. Results

The deployed database contained a total count of 211 molecules, with a redundant entry count of 530. Below are screenshots of the various views you can have. The initial view, The search view, a molecule view, study view and the export screen.

Initial View



LSSR - Large Scale Screening Resource tissue listing

Show / hide columns Export table

Show 10 entries Search:

Tissue	Number of entries	Number of merged entries	Number of studies
all entries	530	211	4

Showing 1 to 1 of 1 entries Previous 1 Next

Compile date 02-26-2016 © PADB initiative

Search View

LSSR - Entry list

Select first letter of gene name to jump to the right entry list
[A](#) | [B](#) | [C](#) | [D](#) | [E](#) | [F](#) | [G](#) | [H](#) | [I](#) | [J](#) | [K](#) | [L](#) | [M](#) | [N](#) | [O](#) | [P](#) | [Q](#) | [R](#) | [S](#) | [T](#) | [U](#) | [V](#) | [W](#) | [X](#) | [Y](#) | [Z](#) | [1](#) | [2](#) | [3](#) | [4](#) | [5](#) | [6](#) | [7](#) | [8](#) | [9](#) | [0](#) | [other](#)

A

Entry	Gene/Symbol	Name	Entries	Express prof
ecoDB74	aceE	Pyruvate dehydrogenase E1 component (PDH E1 component) (EC 1.2.4.1)	3	
ecoDB7	aceF	Dihydrolipoaldehyde acetyltransferase component of pyruvate dehydrogenase complex (EC 2.3.1.12) (Dihydrolipoamide acetyltransferase component of pyruvate dehydrogenase complex) (E2)	3	
ecoDB104	acnA	Aconitate hydratase A (ACN) (Aconitase) (EC 4.2.1.3) (Iron-responsive protein-like) (IRP-like) (RNA-binding protein) (Stationary phase enzyme)	3	
ecoDB117	acnB	"Aconitate hydratase B (ACN) (Aconitase) (EC 4.2.1.3) ((2R,3S)-2-methylisocitrate dehydratase) ((2S,3R)-3-hydroxybutane-1,2,3-tricarboxylate dehydratase) (2-methyl-cis-aconitate hydratase) (EC 4.2.1.99) (Iron-responsive protein-like) (IRP-like) (RNA-binding protein)"	3	
ecoDB115	aidB	Putative acyl-CoA dehydrogenase AidB (EC 1.3.99.-)	3	
ecoDB114	alsE	D-allulose-6-phosphate 3-epimerase (EC 5.1.3.-)	3	
ecoDB63	aphA	Class B acid phosphatase (CBAP) (EC 3.1.3.2)	3	
ecoDB106	appC	Cytochrome bd-II ubiquinol oxidase subunit 1 (EC 1.10.3.10) (Cytochrome bd-II oxidase subunit I)	3	
ecoDB130	aqpZ	Aquaporin Z (Bacterial nodulin-like intrinsic protein)	3	

2226768Y
Francesca Young

LSSR - molecule cluster ecoD9B8

Study View

PubMed ID	3				
Authors	CCC				
Title	Mutant/WT				
Journal	Proteomics				

Species	Tissue / Source	Subcell	Disease	Sample digest	Separation	Detection	Quantification
e.coli						LC-MS	

Identified molecules

Show entries
Search:

Molecule ID	Reference	Gene/Symbol	Name	Ratio (disease/control)	Frequency % found in disease	Frequency % found in healthy
ecodb1	FRDA_ECOLI	frdA		-3.431967955		
ecodb10	SDHB_ECOLI	sdhB	Succinate dehydrogenase iron-sulfur subunit (EC 1.3.5.1)	-3.091859804		
ecodb100	PHOQ_ECOLI	phoQ	Sensor protein PhoQ (EC 2.7.13.3) (EC 3.1.3.-) (Sensor histidine protein kinase/phosphatase PhoQ)			
ecodb101	HIPA_ECOLI	hipA	Serine/threonine-protein kinase HipA (Ser/Thr-protein kinase HipA) (EC 2.7.11.1) (Toxin HipA)	-6.037607241		

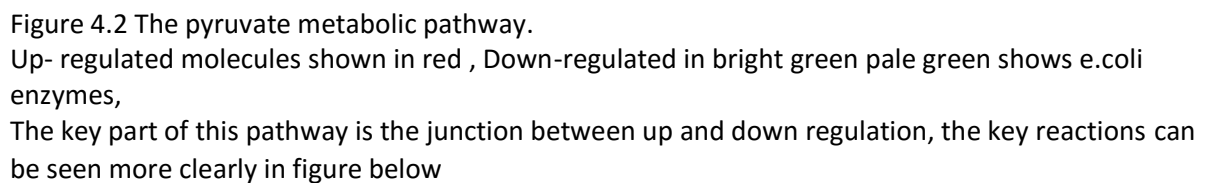
Sorted Data Export View.

LSSR study3

file:///C:/Users/Fran/Deploy/lsr/study/3.htm

Molecule ID	Reference	Gene/Symbol	Name	Ratio (disease/control)	Frequency % found in disease	Frequency % found in healthy
ecodb85	EFTU1_ECOLI	tufA	Elongation factor Tu 1 (EF-Tu 1) (Bacteriophage Q beta RNA-directed RNA polymerase subunit III) (P-43)	9.080153375		
ecodb21	GLK_ECOLI	glk	Glucokinase (EC 2.7.1.2) (Glucose kinase)	7.895604756		
ecodb27	RS4_ECOLI	rpsD	30S ribosomal protein S4	7.77916875		
ecodb22	MNH_ECOLI	mnhH	Divalent metal cation transporter MnhH	7.632719202		
ecodb6	TOP1_ECOLI	topA	DNA topoisomerase I (EC 5.99.1.2) (DNA topoisomerase I) (Omega-protein) (Relaxing enzyme) (Swivelase) (Unwisting enzyme)	7.169629679		
ecodb149	FADL_ECOLI	fadl	3-ketoacyl-CoA thiolase (EC 2.3.1.16) (ACSS) (Acetyl-CoA acyltransferase) (Acyl-CoA ligase) (Beta-ketothiolase) (Fatty acid oxidation complex subunit beta)	7.075331789		
ecodb75	RSE_A_ECOLI	rseA	Anti-sigma-E factor RseA (Regulator of SigE) (Sigma-E anti-sigma factor RseA) (Sigma-E factor negative regulatory protein)	6.752540514		
ecodb18	ENO_ECOLI	eno	Enolase (EC 4.2.1.11) (2-phospho-D-glycerate hydro-lyase) (2-phosphoglycerate dehydratase)	6.415204095		
ecodb109	FTSN_ECOLI	ftsN	Cell division protein FtsN	6.037103875		
ecodb13	OMPT_ECOLI	ompT	Protease 7 (EC 3.4.23.49) (OmpTn) (Outer membrane protein 3B) (Protease A) (Protease VII)	5.97201171		
ecodb9	PFKB_ECOLI	pfkB	ATP-dependent 6-phosphofructokinase isozyme 2 (ATP-PFK 2) (Phosphofructokinase 2) (EC 2.7.1.11) (6-phosphofructokinase isozyme II) (Phosphohexokinase 2)	5.929331306		
ecodb127	DXR_ECOLI	dxr	1-deoxy-D-xylulose 5-phosphate reductoisomerase (DXP reductoisomerase) (EC 1.1.1.267) (1-C-methyl-D-erythritol 4-phosphate synthase)	5.895798881		
ecodb76	SPOT_ECOLI	spot	"Bifunctional (ppGpp synthase/hydrolase) SpoT [Includes: GTP pyrophosphokinase (EC 2.7.6.5) (ppGpp synthase) (ATP-GTP 3'-pyrophosphotransferase) (Stringent response-like protein) (ppGpp synthase II); Guanosine-3',5'-bis(diphosphate) 3'-pyrophosphohydrolase (EC 3.1.7.2) (Penta-phosphate, uranosine, 3'-nucleosubisubstrate)"]	5.799610963		

Figure 4.1 a) Glycolysis b) Fatty Acid Metabolism
Red up-regulated, green down-regulated. Pale green enzymes on E.coli pathways.



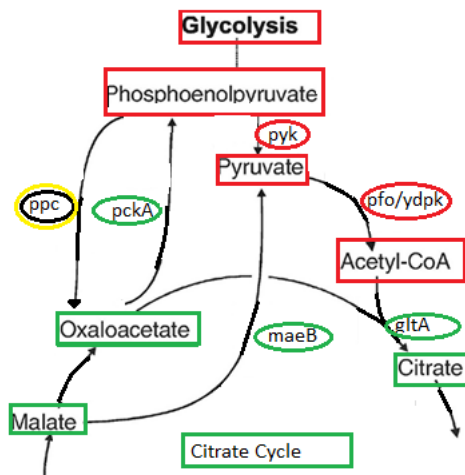


Figure 4.3 TheMissing Gene

E.coli does not have the gene *ps*, coding for the enzyme pyruvate carboxylase, EC.6.4.1.1 which converts pyruvate to oxaloacetate, but does have the gene *ppc*, coding for phosphoenolpyruvate carboxylase, EC.4.1.1.31 (Sauer and Eikmanns, 2005) which converts phosphoenolpyruvate directly to oxaloacetate bypassing pyruvate. In Figures 4.2 and 4.3 above it can be seen that there is no phosphoenolpyruvate carboxylase present in any of the datasets. This indicates there is a high likelihood that this is this gene that is affected, possibly by mutation rendering it inactive or potentially by knockout as part of the investigation.

5. Discussion

Evaluation

Although in this assignment the exercise was hypothetical I will treat it as a prototype design making the assumptions that the requirements would to integrate full datasets and that there would be many more experimental datasets to include.

The database has met the minimum requirements as specified but there are several issues arising from using a software that imposes a structure and view of the data that is less than ideal.

All the information required has been collated and added to the database, it is possible to view the datasets of the individual experiments and to view the results for an individual molecule across the experiments.

One of the main issues for the end user is that the data is presented in a format that is very unclear, this results from coercing the data into a database structure that was designed for much more complex data including information about tissue and disease. Many of the headings do not accurately reflect the nature of the data.

Another specification was to be able to share the data. Although you are able to export the data it is in a less than ideal format as one would expect the groups to want use the results in network or group analysis such as Kegg, David, Genemania or Cytoscape. Ideally the data should be able to be loaded straight into a spreadsheet package such as Excel to manipulate the data into the required input for further analysis.

Adding further datasets would require repeating the process of described in the methods above and re-running the Deploy software. This is a time consuming and involved process with a high learning overhead.

Future Developments

To enable the research groups to be able to add further dataset themselves a number of developments could be implemented.

1. Adapt the deploy software to change the headings and remove many of the unneeded columns from the view. Extend the export to export TDF or CSV files.
2. Develop software that can take the gene IDs and results and automatically get the required information from UniProt and or NCBI. Data inconsistencies could be flagged up and the user be allowed to edit the data.
3. Develop software that can parse this information into the correct format to be deployed, and then run the deploy software to via a user friendly GUI. This could be combined with 2)

Alternative strategies could be used for the database design as for single server local database many of the advantages afforded by NoSQL are lost. One of the biggest advantages are high availability with less down time which are not a major. If it the types of data to be included in the foreseeable future a relational database could be a viable alternative. Although this would be more rigid it would be simpler to implement initially and easier to update as new data was produced.

Conclusion

We have shown combining data from multiple experiments from different omics has potential to increases the power of the results. The development of databases that can usefully store this heterogenic data is an important step in being able to integrate multi-omic data.

The accelerating growth in both volume and types of data produced means the adoption of agreed standards within the biological community is becoming more important to facilitate data reuse and data sharing. It will reduce the overheads in data conversion and coercion and therefore enhance productivity

6. References

- Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* *34*, D354–D357.
- Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* *44*, D457–D462.
- Sauer, U., and Eikmanns, B.J. (2005). The PEP—pyruvate—oxaloacetate node as the switch point for carbon flux distribution in bacteria: We dedicate this paper to Rudolf K. Thauer, Director of the Max-Planck-Institute for Terrestrial Microbiology in Marburg, Germany, on the occasion of his 65th birthday. *FEMS Microbiol. Rev.* *29*, 765–794.
- Xia, J., Sinelnikov, I.V., Han, B., and Wishart, D.S. (2015). MetaboAnalyst 3.0—making metabolomics more meaningful. *Nucleic Acids Res.* gkv380.