Comparison of Current Approaches for Integration of Transcriptomic and Proteomic Data: Advancements In 'Omics Tools 2010-2013

Samir V. Deshpande, Ph.D.[1], Rabih E. Jabbour, Ph.D.[2], Mary M. Wade, Ph.D.[2], Jennifer W. Sekowski, Ph.D.[2]

(1) Science & Technology Corporation, (2) U.S. Army Edgewood Chemical Biological Center

Background

A lingering and difficult problem in modern 'Omics-enabled research is the major bottleneck at the point of integration of data generated by multiple proteomics and transcriptomics approaches. In this work we have identified and compared suitable bioinformatics tools that promise varying degrees of solutions to this problem. We have tools chosen from the currently available pool of open source and proprietary software.

- We compared the bioinformatics tools: STRING, Cytoscape and Ingenuity Pathway Analysis (IPA, from Qiagen) for their quality of pathway description and data integration.
- This review will provide a stress test for the most commonly used bioinformatics tools, derived by mining the current scientific literature. The purpose of this study/survey was to determine the most comprehensive bioinformatics tool(s) for the functionalities such as data integration and co-visualization of peptides/proteins (proteomics) derived from liquid chromatography–tandem mass spectrometry (LC/MS-MS) studies along with transcripts derived from whole transcriptome RNA sequencing (RNA-seq).
- The RNA-seq and proteomics data used for this study was previously published by other researchers and was downloaded from public repository, PRIDE (the PRoteomics IDEntifications database).
- The ability to perform pathway analysis is one of the most important tool requirements, as it plays an important role in understanding the biological impact of gene regulation and protein–protein interactions.

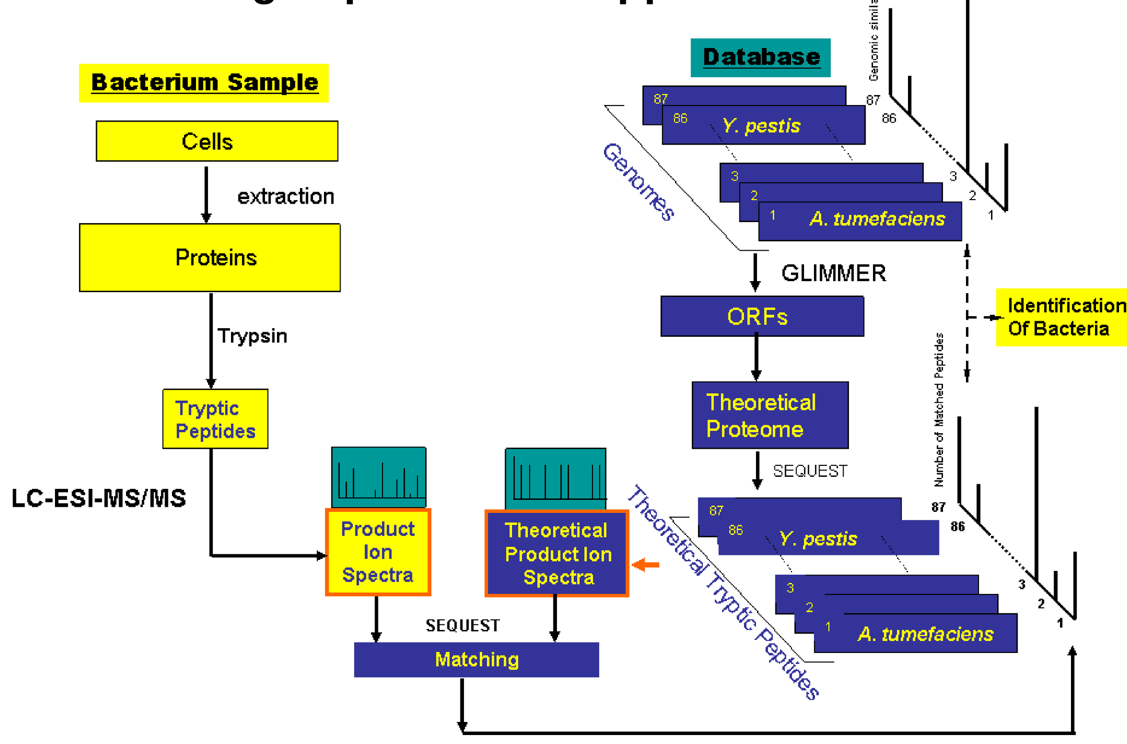
Introduction

- Proteins are key functional molecules in cells and serve as the main link between genotype and phenotype.
- Advances in RNA-seq and mass spectrometry-based proteomics has dramatically improved quality of transcriptomic and proteomic data in recent years.
- Proteomics provides an orthogonal approach to genomic and transcriptomics approaches and can be used to identify and quantify most of the proteins expressed.
- Peptide identification and assignment of protein identifications is carried out by searching databases with peptide mass fingerprinting algorithms; e.g: SEQUEST, Mascot, X!Tandem or OMSSA.
- The advantage of integrating multiple data from genomics, transcriptomics, and proteomics is to increase the coverage and, therefore, the confidence in identification of changes in molecular targets, such as biomarkers.
- This integrative approach allows the identification of peptides derived from novel protein coding genes, splice variants and sequence variations leading to comprehensive characterization of samples.

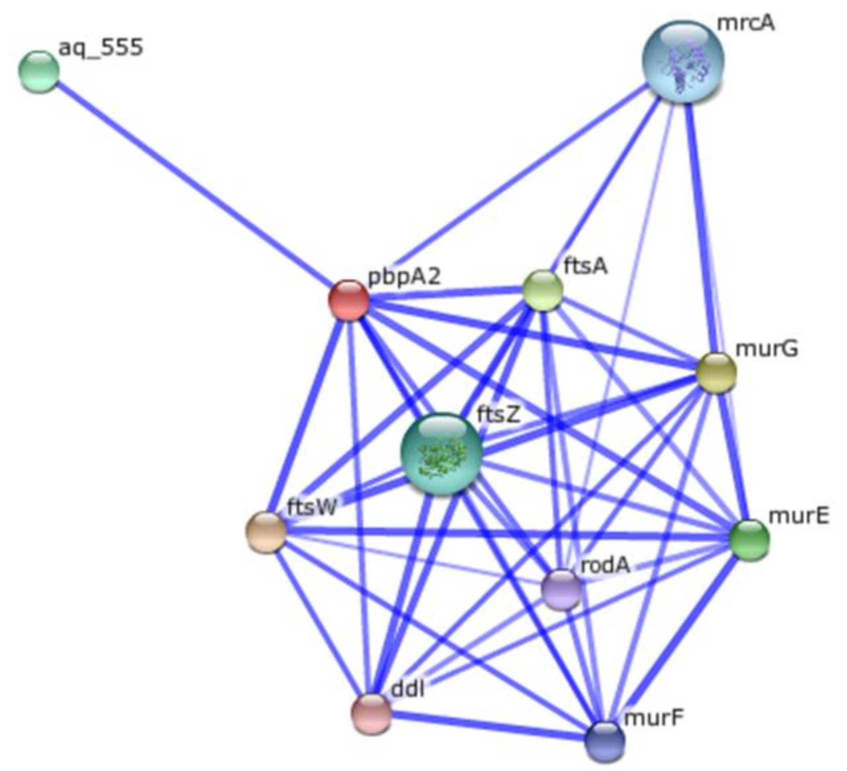
Challenges

- As noted in Peng (2009) and Fernandez (2009), difficult to connect gene changes to peptide/protein changes since connection is not direct.
- Technical challenges such as: dynamic range of detection of proteins, differences in pH, low solubility, hydrophobicity, post-transcriptional modification (PTM) of RNA are not picked up by RNA-Seq.
- Novel alternative splicing events based on the RNA-seq data, are not always easy to confirm using proteomics.
- Integrating data from disparate technologies is risky since data simplification introduces biological biases.
- Many-to-many mapping of genes' identifiers from multiple platforms complicates the integration.
- A single gene loci can produce multiple transcript and protein isoforms through alternative splicing.

Shotgun proteomics approach



pbpA2 penicillin binding protein 2 (578 aa)
(Aquifex aeolicus)



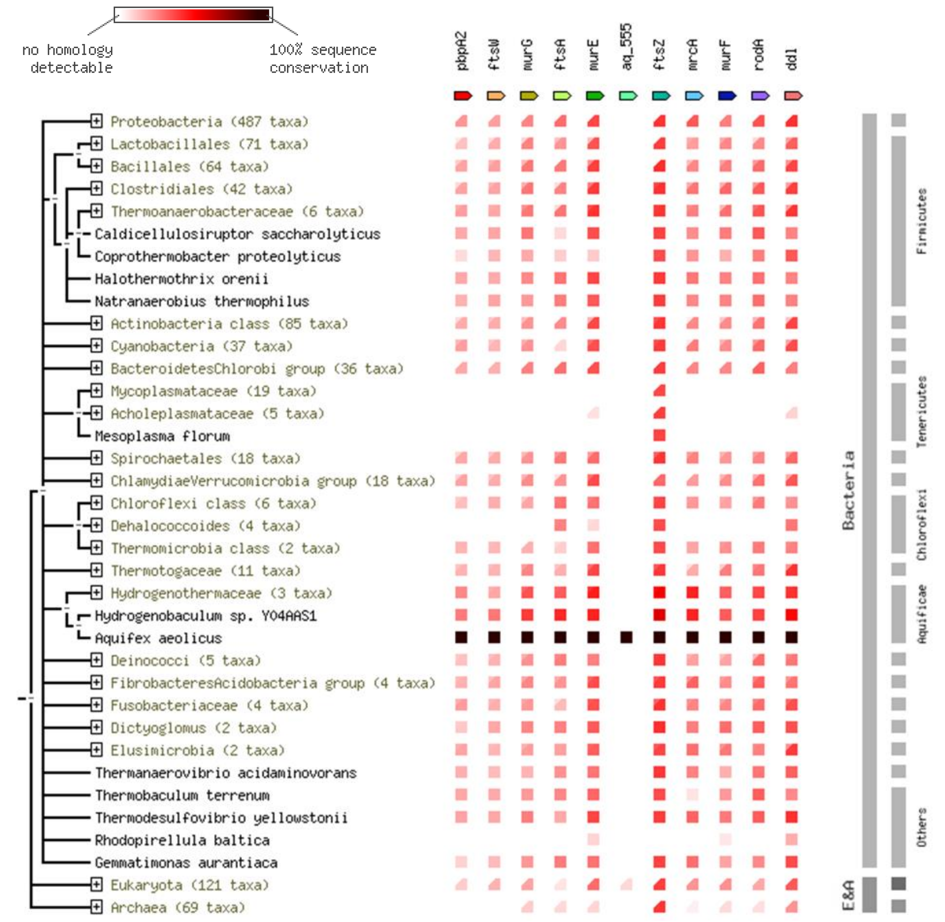
Evidence view for pbpA2; Different line colors represent the types of evidence for the association.

Predicted Functional Partners:

		Score
ftsW	cell division protein FtsW (448 aa)	0.987
murG	undecaprenyldiphospho-muramoylpentapeptide beta-N-acetylglucosaminyltransferase; Cell wall form [...] (344 aa)	0.972
ftsA	cell division protein FtsA (416 aa)	0.920
murE	UDP-MurNac-tripeptide synthetase; Catalyzes the addition of meso-diaminopimelic acid to the nuc [...] (482 aa)	0.876
aq_555	hypothetical protein (267 aa)	0.865
ftsZ	cell division protein FtsZ; This protein is essential to the cell-division process. It seems to [...] (367 aa)	0.856
mrcA	penicillin binding protein 1A (726 aa)	0.845
murF	UDP-MurNAc-pentapeptide sythetase; Involved in cell wall formation. Catalyzes the final step in [...] (445 aa)	0.840
rodA	rod shape determining protein RodA (372 aa)	0.808
ddl	D-alanine-D-alanine ligase; Cell wall formation (By similarity) (291 aa)	0.765

Results

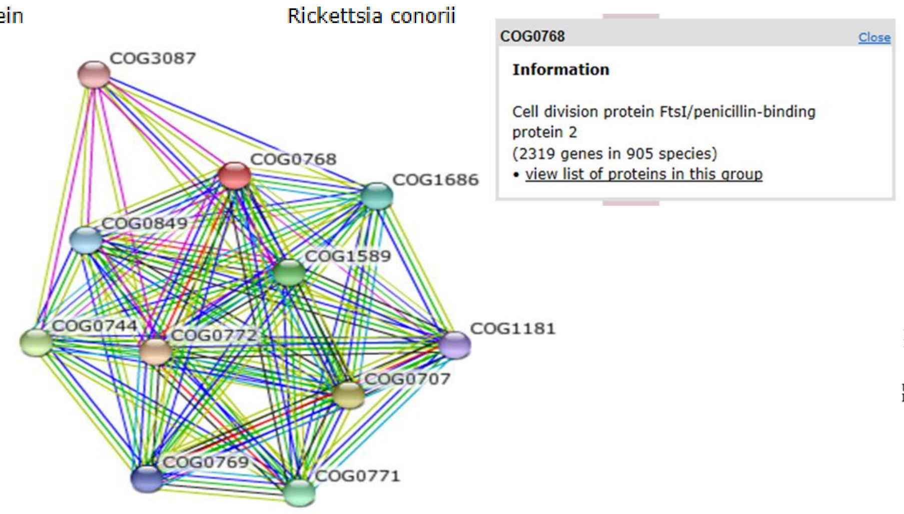
Occurrence view- Where is my protein conserved?



Proteins named 'pbpA2' are linked to the orthologous group 'COG0768'.

COG0768 - Cell division protein FtsI/penicillin-binding protein 2 (2319 proteins in 905 species)

ftsI	penicillin-binding protein PbpA2	Orientia tsutsugamushi Ikeda
RP567	penicillin binding protein (PbpA2)	Rickettsia prowazekii
pbpA2	penicillin-binding protein	Rickettsia felis
pbpA2	penicillin-binding protein	Rickettsia bellii RML369
pbpA2	Penicillin-binding protein	Rickettsia africae
pbpA2	penicillin binding protein 2	Aquifex aeolicus
pbpA2	penicillin-binding protein	Rickettsia conorii



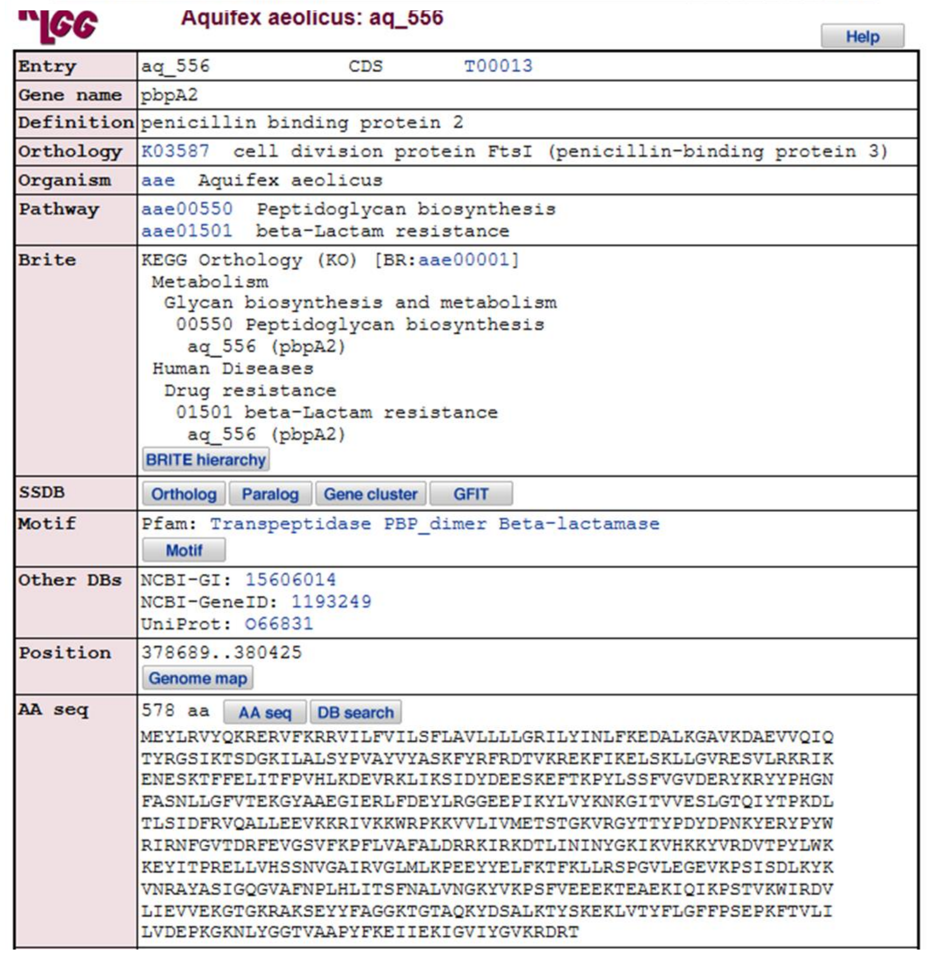
Evidence view COG0768

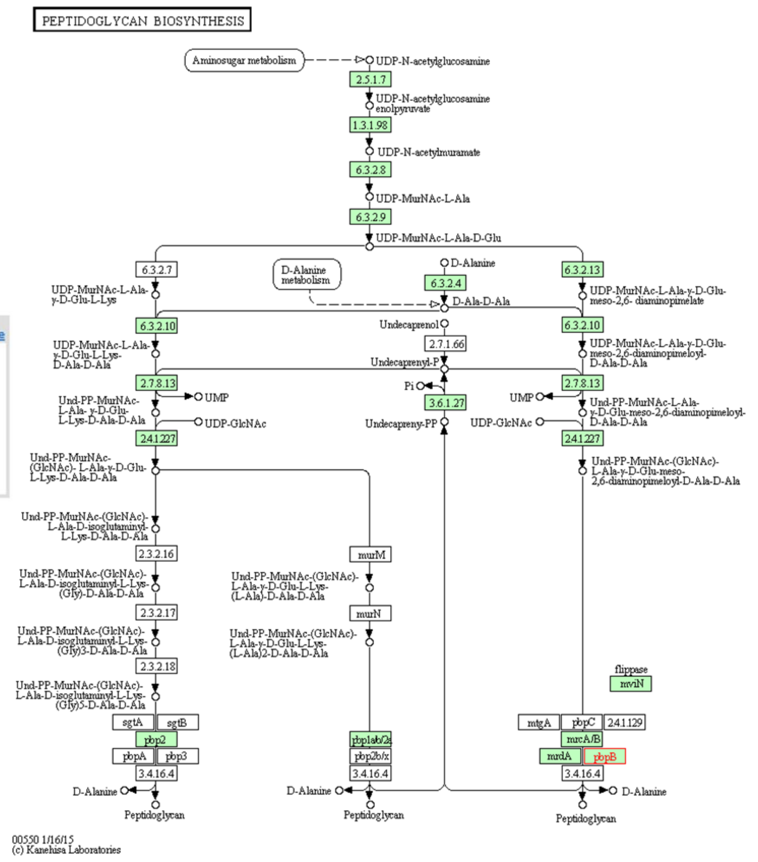
Your Input:

COG0768 Cell division protein FtsI/penicillin-binding protein 2

Predicted Functional Partners:

		Neighborhood	Gene Fusion	Cooccurrence	Coexpression	Experiments	Databases	Textmining	Score
COG0772	Bacterial cell division membrane protein	•	•	•	•	•		•	0.999
COG0707	UDP-N-acetylglucosamine-LPS N-acetylglucosamine transferase	•		•	•		•	•	0.999
COG0744	Membrane carboxypeptidase (penicillin-binding protein)	•		•		•	•	•	0.999
COG1589	Cell division septal protein	•		•		•		•	0.999
COG0771	UDP-N-acetylmuramoylalanine-D-glutamate ligase	•		•	•			•	0.999
COG1686	D-alanyl-D-alanine carboxypeptidase	•		•			•	•	0.998
COG0849	Actin-like ATPase involved in cell division	•		•	•	•		•	0.998
COG0769	UDP-N-acetylmuramyl tripeptide synthase	•		•				•	0.998
COG1181	D-alanine-D-alanine ligase and related ATP-grasp enzymes	•		•		•		•	0.998
COG3087	Cell division protein					•		•	0.997





Conclusions

- STRING and CYTOSCAPE which are both open source software, providing less information than IPA.
- IPA has many additional features with regard to data integration.
- We found no open source tools currently provide a good data integration capability so there appears to be a need for bioinformatics tools to be developed in order to bridge this gap.
- There is a need for continuous development of algorithms and tools to translate multidimensional 'omics data into knowledge that will lead to new discoveries.

References

R. Abersold, M. Mann, "Mass Spectrometry based proteomics", Nature, 422(6928);198-207 (2003).

J. Khatun et.al., "Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions.", BMC Genomics,14,141 (2013).

D. Szklarczyk et al., "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," Nucleic Acids Research 39(suppl 1), D561–D568 (2011).

ZY Peng et. al., "A proteomic study of the response to salinity and drought stress in an introgression strain of bread wheat," Mol Cell Proteomics, 8, 2676-2686 (2009).

C. Fernandez et al., "Omics analyses reveal a potential link between hormone –sensitive lipase and polyamine metabolism," J Proteome Res. 8, 5008-5019 (2009)

M. Eisenacher et al., "Search and decoy: the automatic identification of mass spectra," Methods Mol. Biol. 893, 445–488 (2012)

D. N. Perkins et al., "Probability-based protein identification by searching sequence databases using mass spectrometry data," ELECTROPHORESIS 20(18), 3551–3567 (1999)

Acknowledgements: The authors thank the Defense Threat Reduction Agency/Joint Science and Technology Office for their assistance and funding of this work under Project CB3810. The views expressed in this presentation are those of the authors and do not necessarily reflect the official policy or position of the Department of Defense or the U.S. Government.