NRC Publications Archive Archives des publications du CNRC

Data integration and knowledge discovery in life sciences

Famili, Fazel; Phan, Sieu; Fauteux, François; Liu, Ziyang; Pan, Youlian

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. /
La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.

Publisher's version / Version de l'éditeur:

IEA/AIE'10 Proceedings of the 23rd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, 3, pp. 102-111, 2010-06-01

NRC Publications Record / Notice d'Archives des publications de CNRC:

<https://nrc-publications.canada.ca/eng/view/object/?id=4f1db006-8c54-4f32-84a3-a0c049c4f617>

<https://publications-cnrc.canada.ca/fra/voir/objet/?id=4f1db006-8c54-4f32-84a3-a0c049c4f617>

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at

<https://nrc-publications.canada.ca/eng/copyright>

READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site

<https://publications-cnrc.canada.ca/fra/droits>

LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at

PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

Data Integration and Knowledge Discovery in Life Sciences

Fazel Famili¹, Sieu Phan¹, Francois Fauteux¹,
Ziying Liu¹, Youlian Pan¹

¹ Knowledge Discovery Group, Institute for Information Technology, National Research
Council Canada, 1200 Montreal Road, Ottawa, Ontario, K1A 0R6, Canada
{Fazel.Famili, Sieu.Phan, Francois.Fauteux, Ziying.Liu, Youlian.Pan}@nrc-cnrc.gc.ca

Abstract. Recent advances in various forms of omics technologies have generated huge amount of data. To fully exploit these data sets that in many cases are publicly available, robust computational methodologies need to be developed to deal with the storage, integration, analysis, visualization, and dissemination of these data. In this paper, we describe some of our research activities in data integration leading to novel knowledge discovery in life sciences. Our multi-strategy approach with integration of prior knowledge facilitates a novel means to identify informative genes that could have been missed by the commonly used methods. Our transcriptomics-proteomics integrative framework serves as a means to enhance the confidence of and also to complement transcriptomics discovery. Our new research direction in integrative data analysis of omics data is targeted to identify molecular associations to disease and therapeutic response signatures. The ultimate goal of this research is to facilitate the development of clinical test-kits for early detection, accurate diagnosis/prognosis of disease, and better personalized therapeutic management.

Keywords: Data integration, Knowledge Discovery, Integrated Omics.

1 Introduction

“Omics” refers to the unified study of complex biological systems characterized by high-throughput data generation and analysis [1]. Bioinformatics methods for data storage, dissemination, analysis, and visualization have been developed in response to the substantial challenges posed by the quantity, diversity and complexity of omics data. In parallel, there has been an explosion of online databases and tools for genome annotation and for the analysis of molecular sequences, profiles, interactions, and structures [2, 3]. The next major challenge is to develop computational methods and models to integrate these abundant and heterogeneous data for investigating, and ultimately deciphering complex phenotypes.

DNA sequencing methods and technologies have evolved at a fast pace since the release of the first genome sequence of a free-living organism in the mid 90’s [4]. There are currently over 350 eukaryotic genome sequencing projects [5]. Full genome sequencing has been completed in several organisms, including animal, plant, fungus and protist species. Genome annotation is a good example of the integration of multiple computational and experimental data sources [6]. In the human genome, the ENCyclopedia Of DNA Elements (ENCODE) aims to provide additional resolution, and to identify all functional elements, including all protein-coding and non-coding genes, *cis*-regulatory elements and sequences mediating chromosome dynamics [7].

Transcriptomics is also an advanced and relatively mature area, for which consensus data analysis methods are emerging [8]. Recent and versatile technologies, including high-density whole-genome oligonucleotide arrays [9] and massively parallel sequencing platforms [10] will likely improve the reliability and depth of genomic, epigenomic and transcriptomic analyses.

Notwithstanding the innovative aspects of technologies and sophisticated computational methods developed, each omics approach has some inherent limits. Ostrowski and Wyrwicz [11] mention that the input data for integration must: *i*) be complete, *ii*) be reliable and *iii*) correlate with the biological effect under investigation.

Although providing a reasonably exhaustive coverage of expressed genes, transcriptome analysis is not particularly reliable, and findings often need to be confirmed by additional experimental validation, *e.g.* using reverse transcription polymerase chain reaction (RT-PCR). Another common, legitimate concern with transcriptome analysis is that levels of mRNAs do not necessarily correlate with the abundance of matching gene product(s), and may only reveal gene regulation at the level of transcription. Proteomics, on the other hand, provides a more accurate picture of the abundance of the final gene products, but current methods, even with the latest high resolution technologies [12], are still associated with relatively low sensitivity in protein identification, and dubious reproducibility [13]. The analysis of transcriptomic and proteomic data is an example where data integration has the potential of improving individual omics approaches and overcoming their limitations [14].

Omics data that can eventually be integrated for the comprehensive elucidation of complex phenotypes include functional gene annotations, gene expression profiles, proteomic profiles, DNA polymorphisms, DNA copy number variations, epigenetic modifications, etc.

In the Knowledge Discovery Group at the Institute for Information Technology of the National Research Council Canada, one of our goals is to develop methods and tools for the integration of omics data. We aim to develop general methods with applications in diverse fields, *e.g.* the identification of biomarkers and targets in human cancer and the identification of genes associated with quantitative traits to be used in selection and engineering of crop plants. In this paper, different projects from our group are reviewed and approaches are illustrated with applications on biological data.

2 Integrative approach to informative gene identification from gene expression data

2.1 A multi-strategy approach with integration of prior knowledge

Owing to its relatively low cost and maturity, microarray has been the most commonly used technology in functional genomics. One of the fundamental tasks in microarray data analyses is the identification of differentially-expressed genes between two or more experimental conditions (*e.g.* disease vs. healthy). Several statistical methods have been used in the field to identify differentially expressed genes. Since different methods generate different lists of genes, it is difficult to

determine the most reliable gene list and the most appropriate method. To retain the best outcomes of each individual method and to complement the overall result with those that can be missed using only one individual method, we have developed a multi-strategy approach [15] that takes advantage of prior knowledge such as GO annotation, gene-pathway and gene-transcription-factor associations.

Fig. 1 is an overview of the proposed multi-strategy approach with prior knowledge integration. Microarray data are first passed through a basic data preprocessing stage (background correction, normalization, data filtering, and missing value handling). The next step is to apply different experimental methods to obtain lists of differentially expressed genes. We then establish a confidence measure to select a set of genes to form the *core* of our final selection. The remaining genes in the lists form the *peripheral* set which is subject to exclusion or inclusion into the final selection by similarity searches between the *peripheral* and the *core* lists. The similarity searches are based on prior knowledge such as *i*) biological pathways (based *e.g.* on the KEGG database) or *ii*) biological function or process (based on GO annotations) or *iii*) regulation by similar mechanisms (based on common transcription factors).

Depending on the context, there is a variety of ways to define the confidence measure to form the *core* and *peripheral* sets of genes. A unanimous voting scheme could define a simple confidence measure, under which the *core* consists of genes that are identified by all methods that were applied. Another, less stringent voting scheme is to define the *core* as the genes that are selected by more than one method.

Fig. 1. The multi-strategy approach with prior knowledge integration

In this section we describe the application of the proposed methodology to identify defense response genes in plants [16]. When a plant is infected by a biotrophic pathogen, the concentration of salicylic acid (SA) rises dramatically and massive changes in patterns of gene expression occur. The accumulation of SA is needed to trigger different signaling events, which can also be triggered by exogenous spraying of SA on the plants, even in the absence of

pathogen. To help establish the role of various transcription factors involved in disease resistance [17, 18] and the regulation of defense genes, a set of microarray experiments were conducted using four genotypes of *Arabidopsis thaliana*: Columbia wild-type Col-0, mutant *npr1*, double mutant in all Group I TGA factors, *tga1 tga4*, and triple mutant in all Group II TGA factors, *tga2 tga5 tga6*. Triplicate samples were collected before, 1 and 8 hours after SA treatment. Affymetrix *Arabidopsis* ATH1 (20K probe sets) microarray platform was used in this study. A total of 36 arrays were hybridized.

Wild-type (Col-0)			Mutant <i>npr1</i>			Mutant <i>tga1-4</i>			Mutant <i>tga2-5-6</i>		
0H	1H	8H	0H	1H	8H	0H	1H	8H	0H	1H	8H
rep1	rep1	rep1	rep1	rep1	rep1	rep1	rep1	rep1	rep1	rep1	rep1
rep2	rep2	rep2	rep2	rep2	rep2	rep2	rep2	rep2	rep2	rep2	rep2
rep3	rep3	rep3	rep3	rep3	rep3	rep3	rep3	rep3	rep3	rep3	rep3

Fig. 2. 36 Affymetrix ATH1 (20K probe sets) microarrays

To identify SA-induced genes we identify the differentially-expressed genes for the following 7 pairs of conditions

- wild-type @ 8h vs. wild-type @ 0h
- *npr1* @ 8h vs. wild-type @ 0h
- *tga1-4* @ 8h vs. wild-type @ 0h
- *tga2-5-6* @ 8h vs. wild-type @ 0h
- *npr1* @ 8h vs. wild-type @ 8h
- *tga1-4* @ 8h vs. wild-type @ 8h
- *tga2-5-6* @ 8h vs. wild-type @ 8h

The background subtracted data were processed through global quantile normalization across 36 arrays and filtering. The final list contains 10256 genes. The following is the detail of applying the multi-strategy methodology:

- Four methods were used:
 - M1: t-test with fold-change set to 2 and p-value 5%
 - M2: SAM with fold-change set to 2 and FDR 5%
 - M3: Rank-Products (RP) with FRD set to 5%
 - M4: fold-change with threshold set at 1.5
- Confidence measure: majority voting model, i.e., genes that were identified by more than one method.
- Gene recruitment mechanisms (similarity search):
 - Genes in the *peripheral* set that participate in the same biological pathway as some in the *core* set
 - Genes in the *peripheral* set that have similar promoter characteristics as some in the *core* set.

The methodology identified a list of 2303 *core* genes and a list of 3522 *peripheral* genes. Through the similarity search with the aid of prior knowledge, we were able to identify an additional 408 genes from KEGG pathway search, and 198 genes from transcription factor binding site search. The recruitment algorithms uncovered many

important SA mediated and other defence genes that could have been missed if single analysis method were used (a partial list is shown in Fig. 3).

	SA Mediated Response	Defense Response	Immune Response	External Stimulus Response
At3g46590	TF			
At3g11820	PWY	PWY	PWY	PWY
At4g11820	PWY	PWY		PWY
At3g52400	PWY	PWY	PWY	PWY
At1g74840	TF			
At5g67300	TF			
At4g31550		TF		
At3g63180		TF		
At5g46450		TF		
At4g19510		TF		
At3g14595		TF		
...				

Fig. 3. The prior knowledge search algorithms recruited additional important SA mediated and other defence genes.

2.2 Transcriptomics and proteomics integration: a framework using global proteomic discovery to complement transcriptomics discovery

The transcriptome data generated by microarray or other means can only provide a snapshot of genes involved in a physiological state or a cellular phenotype at mRNA level. Advances in proteomics have allowed high-throughput profiling of expressed proteins to elucidate the intricate protein-protein interactions and various biological systems dealing with the downstream products of gene translation and post-translational modification. Genome-wide analysis at the protein level is a more direct reflection of gene expression. The consistency between proteomics and transcriptomics can increase our confidence in identification of biomarkers; differences can also reveal additional post-transcriptional regulatory or recover the missed genes by other reasons to complement transcriptomics discovery. In this section we exploit proteomics and transcriptomics in parallel. We developed a framework using global proteomics discovery, as shown in Figure 4, to enhance the confidence in transcriptomics analysis and to complement the discovery with genes that could have been missed for various reasons such as chemical contamination on the arrays, analysis threshold settings, or genes that were not spotted on the original platform.

Fig. 4. A transcriptomics and proteomics complementary discovery framework

In the following, we describe an application of the above framework to an experiment to identify EMT-related (Epithelial-to-Mesenchymal Transition) breast cancer biomarkers. The microarray experiments were performed by Biotechnology Research Institute, NRC. This dataset was generated by exposing JM01 mouse cell line [19] to a treatment with the Transforming Growth Factor (TGF- β) for 24 hours. TGF- β induces an Epithelial-to-Mesenchymal Transition in these cells, a phenomenon characterized by significant morphology and motility changes, which are thought to be critical for tumour progression. The transcriptome changes after 24 hours in TGF- β treated vs. non-treated control cells were monitored using Agilent 41K mouse genome array (four technical replicates). The microarray analysis was done using a multi-strategy approach as described in [15]. The proteomics experiments were performed by Institute for Biological Sciences, NRC [20]. The global proteomics discovered 13 proteins, which are induced by TGF- β , and are mapped to the corresponding transcriptomics results. As shown in Table 1, the discovery of this group of genes (Clu, Fn, Itga5, Acpp, Itgb5, Itga6, and Tacstd2) by transcriptomics approach were also confirmed by the proteomics approach. The proteomics approach identified additional genes (Actg1, Hnmpu, Ubqln2, Pttglip, Ldlr and Itgb4) that were missed from the transcriptomics approach.

Table 1. Complementary discovery through global proteomics in JM01 data.

Gene symbol	Proteomics results	Swiss-Prot	Transcriptomics results
Up-regulated genes			
Clu	x	Q06890	x
Fn1	x	P11276	x
Itga5	x	P11688	x
Acpp	x	NP_997551.1	x
Itgb5	x	O70309	x
Hnrnpu	x	NP_058085.1	
Actg1	x	P63260	
Down-regulated genes			
Tacstd2	x	NP_064431.2	x
Itga6	x	Q61739	x
Ubqln2	x	Q9QZM0	
Pttg1ip	x	Q8R143	
Ldlr	x	P35951	
Itgb4	x	NP_598424.2	

3 Integrative data analysis for gene set biomarker and disease target discovery

Complex human diseases occur as a result of multiple genetic alterations and combinations of environmental factors. The integrative analysis of multiple omics data sources is a promising strategy for deciphering the molecular basis of disease and for the discovery of robust molecular signatures for disease diagnosis/prognosis [11]. Oncology is the field of research where biomarker discovery is most advanced. Various types of biomarkers have been identified for use in prognostic and eventually provide patients with personalized treatment [21, 22]. Emerging themes in cancer research include *i*) the exploitation of panels of biomarkers for successful translation of discoveries into clinical applications [23] and *ii*) the understanding of cancer at the pathway-level [24]. We believe that there is a considerable potential for disease target discovery and biomarker identification in shifting from single gene to metabolic pathway (or other functional gene set) analysis of omics data. Moreover, having identified disease targets in their biological context may be useful in further steps of the commercialization process.

Gene Set Analysis (GSA) approaches intend to identify differentially expressed sets of functionally related genes [25]. Most GSA methods start with a list of differentially expressed genes, and use contingency statistics to determine if the proportion of genes from a given set is surprisingly high [26]. Gene Set Enrichment

Analysis is a popular alternative [27]. It tests whether the rank of genes ordered according to P-values differs from a uniform distribution. Goeman and Bühlmann recently reviewed existing GSA methods, and strongly recommended the use of self-contained methods [28]. We are currently developing and testing statistical methods for GSA analysis of cancer expression profiles. The Kyoto Encyclopedia of Genes and Genomes (KEGG) [29] and the Gene Ontology (GO) [30] are used to group genes into sets, and differential expression is assessed for gene sets rather than individual genes. Future developments will include integration of pathway and ontology knowledge in combination with transcriptomics and proteomics analysis of tumor samples.

4 Conclusion

One of the major challenges in dealing with today's omics data is its proper integration through which various forms of useful knowledge can be discovered and validated. In this paper we discussed our attempts in integrating omics data and introduced case studies in which various forms of omics data have been used to complement each other for knowledge discovery and validation. Among methods developed, a novel multi-strategy approach showed some interesting results in the analysis of transcriptomics and proteomics data, which also includes biological experimental validation. Until now, all of our integrated case studies have resulted in interesting discoveries, among which are cases where using a single form of biological data would have resulted in missing some valuable information. This is evident from our transcriptomics/proteomics integration example explained in this paper. Our ultimate goal is to develop platforms that facilitate development of clinical test kits that are based on multiple sources of omics data.

Acknowledgments. The experiments on the effect of salicylic acid on *Arabidopsis thaliana* were conducted by Fobert's Lab at the Plant Biotechnology Institute, NRC. The microarray experiments for JM01 cell lines were conducted by O'Connor-McCourt's Lab at the Biotechnology Research Institute, NRC. The proteomics experiments (mass-spectrometry) for the JM01 were performed by Kelly's Lab at the Institute for Biological Sciences, NRC. We thank them for sharing the data.

References

1. Joyce, A.R., Palsson, B. O.: The model organism as a system: integrating 'omics' data sets. *Nat. Rev. Mol. Cell Biol.* 7, 198-210 (2006)
2. Baxevanis, A.D.: The importance of biological databases in biological discovery. *Curr. Protoc. Bioinformatics* Chapter 1: Unit 1.1 (2009)
3. Galperin, M.Y., Cochrane, G.R.: Nucleic acids research annual database issue and the nar online molecular biology database collection in 2009. *Nucleic Acids Res.* 37, D1-4 (2009)
4. Fleischmann, R.D., Adams, M. D., et al.: Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496-512 (1995)

5. National Center for Biotechnology Information (NCBI): Genome sequencing projects statistics. Retrieved December 6, 2009 from <http://www.ncbi.nlm.nih.gov>.
6. Brent, M.R.: Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nat. Rev. Genet.* 9, 62-73 (2008)
7. ENCODE Project Consortium: The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306, 636-640 (2004)
8. Allison, D.B., Cui, X., et al.: Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.* 7, 55-65 (2006)
9. Mockler, T.C., Chan, S., et al.: Applications of DNA tiling arrays for whole-genome analysis. *Genomics* 85, 1-15 (2005)
10. Shendure, J., Ji, H.: Next-generation DNA sequencing. *Nat. Biotechnol.* 26, 1135-1145 (2008)
11. Ostrowski, J., Wyrwicz, L. S.: Integrating genomics, proteomics and bioinformatics in translational studies of molecular medicine. *Expert. Rev. Mol. Diagn.* 9, 623-630 (2009)
12. Hu, Q., Noll, R. J., et al.: The Orbitrap: a new mass spectrometer. *J. Mass. Spectrom.* 40, 430-443 (2005)
13. Lubec, G. Afjehi-Sadat, L.: Limitations and pitfalls in protein identification by mass spectrometry. *Chem. Rev.* 107, 3568-3584 (2007)
14. Nie, L., Wu, G., et al.: Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications. *Crit. Rev. Biotechnol.* 27, 63-75 (2007)
15. Liu, Z., Phan, S., Famili, F., Pan, Y., Lenferink, A., Cantin, C., Collins, C., O'Connor-McCourt, M.: A multi-strategy approach to informative genes identification from gene expression data. *J.Bioinfo. Comput. Biol.* in press (2010)
16. Phan, S., Shearer, H., Tchagang, A., Liu, Z., Famili, F., Fobert, F., Pan, Y.: *Arabidopsis thaliana* defense gene response under pathogen challenge. The 9th GHI-AGM, Montreal, June 8-10 (2009)
17. Fobert, P., Després, C.: Redox control of systemic acquired resistance. *Curr. Op. Plant Biol.* 8, 378-382 (2005)
18. Kesarwani, M., Yoo, J., Dong, X.: Genetic Interactions of TGA transcription factors in the regulation of pathogenesis-related genes and disease resistance in *Arabidopsis*. *Plant Physiol.* 14, 336-346 (2007)
19. Lenferink, A.E.G., Magoon, J., Cantin, C., O'Connor-McCourt, M.D.: Investigation of three new mouse mammary tumor cell lines as models for transforming growth factor (TGF)- β and Neu pathway signaling studies: identification of a novel model for TGF- β -induced epithelial-to-mesenchymal transition. *Breast Cancer Res.* 6, 514-530 (2004)
20. Hill J.J., Tremblay T.L., Cantin C., O'Connor-McCourt M.D., Kelly J.F., Lenferink A.E.G.: Glycoproteomic analysis of two mouse mammary cell lines during transforming growth factor (TGF)- β induced epithelial to mesenchymal transition. *Proteome Science* 7:2 (2009)
21. Tainsky, M.A.: Genomic and proteomic biomarkers for cancer: a multitude of opportunities. *Biochim. Biophys. Acta* 1796, 176-193 (2009)
22. Chin, L., Gray, J.W.: Translating insights from the cancer genome into clinical practice. *Nature* 452, 553-563 (2008)
23. Ross, J. S.: Multigene classifiers, prognostic factors, and predictors of breast cancer clinical outcome. *Adv. Anat. Pathol.* 16, 204-215 (2009)
24. The Cancer Genome Atlas Research Network: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061-1068 (2008)
25. Dinu, I., Potter, J.D., et al.: Gene-set analysis and reduction. *Brief Bioinform.* 10, 24-34 (2009)
26. Khatri, P., Draghici, S.: Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 21, 3587-3595 (2005)

27. Subramanian, A., Tamayo, P., et al.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* 102, 15545-15550 (2005)
28. Goeman, J.J., Buhlmann, P.: Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 23, 980-987 (2007)
29. Ogata, H., Goto, S., et al.: KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 27, 29-34 (1999)
30. Ashburner, M., Ball, C.A., et al.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25-29 (2000)