Breakthroughs in genomics data integration for predicting clinical outcome

Yves A. Lussier[1-4,*] and Haiquan Li[1]

[1] Department of Medicine, University of Illinois at Chicago, Chicago, Illinois, USA

[2] Department of Bioengineering, University of Illinois at Chicago, Chicago, Illinois, USA

[3] Center for Interventional Health Informatics, University of Illinois at Chicago, Illinois, USA

[4] Cancer Center, University of Illinois, Chicago, Illinois, USA

With the rapid progression of biotechnologies in the last few decades, molecular biology assays have shifted from limited to high throughput. Indeed, we successively witnessed an accelerated development of 'omics technologies at multiple molecular scales: DNA arrays, DNA methylation screening, proteome-wide interaction screening, Chip-Seq, protein array, next generation sequencing, RNA-seq and next generation protein mass-spectrometry. The availability of large repository of 'omics data has stimulated the prolific growth of analytical methods for clinical outcome prediction specialized for one type of 'omics measurement, but comparatively fewer cross-scales ones (2 molecular scales) and very rare multiple scales ones (≥3 molecular scales; "multiscale"). Of note, abundant original cross-scale and multiscale 'omics methods have been developed for identifying gene function [1], novel disease-gene [2, 3] and diseases' biomodules (e.g. biomodules of microRNA-mRNA co-expression [4]). However, these approaches are insufficient for predicting clinical outcome of complex disorders [5, 6]. Here, we provide a framework to illustrate the difficulty of increasing the accuracy of a clinical outcome predictor from multiple scales of 'omics data, and position the significance of Kim et al. [7] that appears in this issue of the Journal of Biomedical Informatics. We also include a brief historical perspective of foundational methodologies in cross-scale and multiscale 'omics analytics (single 'omics analytics are out of scope of this perspective).

Categorization of cross-scale and multiscale analytics

Through complementary or synergistic information between two genome-wide 'omics measures, the finest methodological developments in cross-scale analyses magnify the accuracy of discovery science

* Corresponding author, lussier.y@gmail.com
Yves A. Lussier, MD
Professor of Medicine and of Engineering
Director, Center for Interventional Health Informatics
Assoc. Dir. for Informatics, Cancer Ctr and Ctr for Clinical and Translational Science
The University of Illinois in Chicago
Ph: (312) 355-0478
Fax: (312) 996-5413

and their statistical power on smaller samples size. Furthermore, while cross-scales studies employ linear as well as complex non-linear analytical mathematics, multiscale ones have generally been deceivingly simple in terms of modeling, proposing one integrative approach regardless of the biomedical scale. Indeed, the complexity of independently modeling each combination of two scales and thereafter the higher degree interactions for deep integration of *n* scales has not yet been attempted. Combinatorics informs us that such modeling grows exponentially with scale[†].

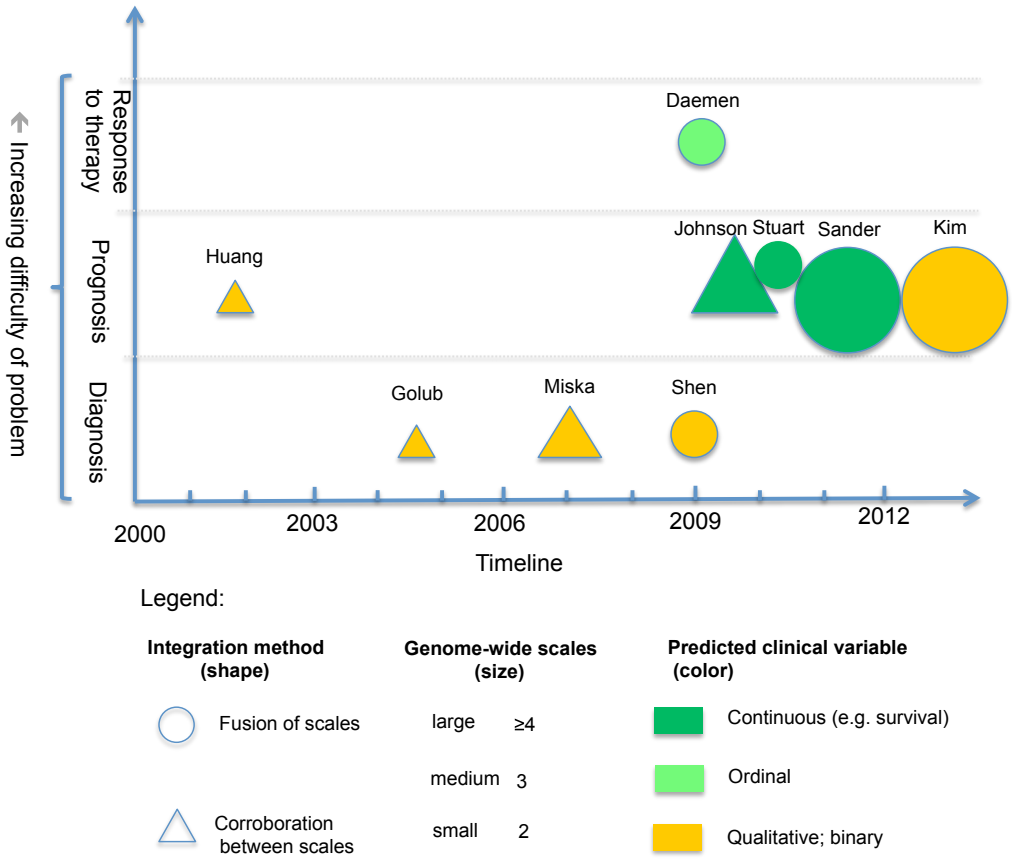

Figure 1: Seminal integrative genomics analytics leading to predicting increasingly difficult clinical outcome problems. Nine seminal methodological studies that led to the use of data fusion across four scales of 'omics data are presented in a timeline (x axis) according to four additional characteristics: the increasing difficulty of the clinical prediction (y axis), the level of data integration (shape), the number of 'omics scales integrated (size), the complexity of the clinical variable to predict (color). Full integration

[†] The number of mathematical models required to specify each interaction between *n* scales of 'omics measurements increases as follows (note that combinations of higher order than binary are included, e.g. ternary, quaternary, etc.):

Count of Interaction Models = $(2^n - n - 1)$.

Thus, for n=3,4, 5 or 6 scales, the number of interaction models required are 4, 11, 26, or 57, respectively.

or fusion of biological data (circles) has, in principle, more predictive accuracy than corroboration between scales (triangles). However, the former are more complex and are discovered later historically (circles follow triangles). Kim's method classifies clinical prognosis by fusing four 'omics scales, and shows that comprehensive multiscale fusion is more accurate than any other partial combination of scales. Of note, we included the earliest manuscripts with fully described methods and comprising two, then three and then four or more genomics scales for each of the three prediction types: (i) diagnosis, (ii) prognosis, (iii) therapeutic response.

Figure 1 provides a *timeline* (*x axis*) of pioneering cross-scale (2 scales) and multiscale (≥3 scales) 'omics studies for predicting clinical outcome. Cross-scales clinical outcome predictors from 'omics data precede the more complex multiscale ones; integration of fewer 'omics scales (Figure 1, small size shapes) also precedes integration of more scales (Figure 1, large size shapes). The y-axis of Figure 1 represents the difficulty of the prediction: diagnosis, followed by prognosis and response to therapy. Evidently, seminal studies pertaining to simpler problems (y axis, near origin) precede those addressing more difficult ones (y axis, away from the origin). Further, predicting the outcome using a continuous clinical variable (regression-type problem; e.g. survival; Figure 1, green) is harder than with ordinal variables (light green**;** e.g. cancer stage) or a qualitative one (Figure 1, yellow; cancer subtypes). Two forms of *data integration* are observed in these studies of clinical outcome prediction: (i) the earlier studies *corroborate* a single scale predictor with another scale (Figure 1, triangles), then followed by (ii) more complex *data fusion* (Figure 1, circles), defined as "the process of integration of multiple data and knowledge representing the same real-world object into a consistent, accurate, and useful representation"[‡]. Corroborative 'omics studies have preceded deeper mathematical integration and data fusion for each *type of biological problem* (*y axis*).

Predicting clinical outcome of complex disorders

Mendelian disorders can be diagnosed with a single genetic marker that also provides insight in the underpinning biological mechanisms. Predicting clinical outcome of *complex* diseases has conventionally leveraged this principle using single molecular *biomarkers* that are correlated with outcomes. In spite of a plethora of genomic studies, very few biomarkers have been discovered for complex disorders in the last decade. Breaking from the biomarker tradition, classifiers consisting of multiple molecular features have been accurately designed from single 'omics scales to predict clinical outcome. Multiscale genomic predictors have in theory the potential to surpass single scale ones, which have been demonstrated in practice through significant increase of sensitivity while maintaining the same level of specificity [6].

[‡] http://en.wikipedia.org/wiki/Data_fusion

As shown in Figure 1, studies of cross-scale (2-scales) corroborations of classifiers of clinical outcome (Figure 1, triangles) have first been designed for diagnoses (Golub's team; 2005) [8], prognosis (Huang's team; 2002) [9] and recently addressed response to therapy (Nephew's team; 2009) [10]. Three cross-scale studies pioneered data fusion in each category of clinical predictions: (i) Shen et al used joint latent variable for cancer diagnosis (2009; copy number variation, gene expression) [11], Stuart's team imputed pathway scores from copy number variation and gene expression for cancer prognosis (2010) [12] and (iii) Daemen et al fused microarray and proteomics/genomics data into the kernel of support vector machine and predicted response to therapy (2009) [13].

Miska's pioneered multiscale corroborative methods for clinical classification (miRNA, mRNA, CNV) [14], while Johnson's team was first to utilize them across four genomic scales (copy number variation, mRNA, microRNA and methylation) [15]. Sander's team pioneered the multiscale data fusion to predict time-to-recurrence of serous ovarian tumors (2011) [16]. They integrated four scales with Cox Lasso models (mRNA, microRNA, methylation and copy number variation). They were followed by several other multiscale data fusion models that built from The Cancer Genome Atlas (TCGA) datasets [17] (e.g. matrix factorization techniques [18], integrated pathway scores [19, 20], and mutual exclusivity [20, 21]).

Kim et al classifies clinical prognosis by fusing four 'omics scales: mRNA expression, microRNA expression, copy number variation (CNV), and DNA methylation [7]. While graph-based semi-supervised learning (SSL) [22] was previously used for clinical outcome analysis with a single genomic scale [23], Kim's team initiate its application using multiple scales. They employ a minimum objective function integrating two factors from each 'omics dataset: i) regression errors for all patients (loss function), ii) concordance between class similarity and their underlying feature similarity for every pair of patients (smoothness). The clinical classifier Kim et al. developed leverages multiscale 'omics data fusion and demonstrates more accuracy than using each of the four single 'omics scale separately or any combinations (10 combinations taking 2 or 3 scales at a time). Previous studies were not as systematic and overlooked the lower order combinations. The algorithm also advances computational efficiency due to the impressive use of several straightforward matrix operations, an approach not required when used in single scales by previous authors. Additional predictive power can probably be obtained by modeling, in more detail, some of the 'omics interactions rather than using a single approach to the integration. Further, these approaches should be confirmed in prospective independent studies. SSL should be applied to therapeutic response prediction problems that remain a greater challenge than prognostic.

While the commercial application of omics classifiers will require cost-efficient solutions, a tradeoff is likely to occur between the number of omics scales and the accuracy of clinical decision making. With their comprehensive analysis of the accuracy of every combination of scales, Kim et al. provide a

systematic approach to the optimization of accuracy vs costs. Whether multiscale classifiers can provide sufficient increase in clinical utility to justify their costs remains to be established and a domain of active research (e.g. NIH/NCI PAR-11-151 SPEC II grants funding clinical trials in multi-analytes signatures).

Acknowledgements

The authors appreciate valuable comments by Dr. Ikbel Achour and Ms. Colleen Kenost.

Funding

This work was supported in part by the following NIH Grants: 5UL1RR024999-04, the University of Illinois Cancer Center, and the University of Illinois Center for Clinical and Translational Science (UL1RR029879).

References

[1] Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. Nat Rev Genet. 2010;11:476-86.
[2] Duan S, Huang RS, Zhang W, Bleibel WK, Roe CA, Clark TA, et al. Genetic Architecture of Transcript-Level Variation in Humans. The American Journal of Human Genetics. 2008;82:1101-13.
[3] Aerts S, Lambrechts D, Maity S, Loo PV, Coessens B, Smet FD, et al. Gene prioritization through genomic data fusion. Nature Biotechnology. 2006;24:537-44.
[4] Wang Y-P, Li K-B. Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data. BMC Genomics. 2009;10:218.
[5] Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification. Journal of the National Cancer Institute. 2003;95:14-8.
[6] Kulasingam V, Pavlou MP, Diamandis EP. Integrating high-throughput technologies in the quest for effective biomarkers for ovarian cancer. Nat Rev Cancer. 2010;10:371-8.
[7] Kim D, Shin H, Song YS, Kim JH. Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. Journal of Biomedical Informatics. 2012;45.
[8] Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, et al. MicroRNA expression profiles classify human cancers. Nature. 2005;435:834-8.
[9] Wei SH, Chen C-M, Strathdee G, Jaturon Harnsomburana, Shyu C-R, Farahnaz Rahmatpanah, et al. Methylation microarray analysis of late-stage ovarian carcinomas distinguishes progression-free survival in patients and identifies candidate epigenetic markers. Clinical Cancer Research. 2002;8:2246-52.
[10] Li M, Balch C, Montgomery J, Jeong M, Chung J, Yan P, et al. Integrated analysis of DNA methylation and gene expression reveals specific signaling pathways associated with platinum resistance in ovarian cancer. BMC Medical Genomics. 2009;2:34.
[11] Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25:2906-12.

[12] Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010;26:i237-i45.
[13] Daemen A, Gevaert O, Ojeda F, Debucquoy A, Suykens J, Sempoux C, et al. A kernel-based integration of genome-wide data for clinical decision support. Genome Medicine. 2009;1:39.
[14] Blenkiron C, Goldstein L, Thorne N, Spiteri I, Chin S-F, Dunning M, et al. MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. Genome Biology. 2007;8:R214.
[15] Kim H, Huang W, Jiang X, Pennicooke B, Park PJ, Johnson MD. Integrative genome analysis reveals an oncomir/oncogene cluster regulating glioblastoma survivorship. Proceedings of the National Academy of Sciences. 2010;107:2183-8.
[16] Mankoo PK, Shen R, Schultz N, Levine DA, Sander C. Time to Recurrence and Survival in Serous Ovarian Tumors Predicted from Integrated Genomic Profiles. PLoS ONE. 2011;6:e24709.
[17] Consortium TICG. International network of cancer genome projects. Nature. 2010;464:993-8.
[18] Zhang S, Liu C-C, Li W, Shen H, Laird PW, Zhou XJ. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Research. 2012;40:9379-91.
[19] Kristensen VN, Vaske CJ, Ursini-Siegel J, Van Loo P, Nordgard SH, Sachidanandam R, et al. Integrated molecular profiles of invasive breast tumors and ductal carcinoma in situ (DCIS) reveal differential vascular and interleukin signaling. Proceedings of the National Academy of Sciences. 2012;109:2802-7.
[20] Network TCGA. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61-70.
[21] Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. Genome Research. 2012;22:398-406.
[22] Zhou D BO, Lal TN,Weston J, Scholkopf B. Learning with local and global consistency. Adv Neur Inform Process Syst (NIPS). 2004;16:321-8.
[23] Gui J, Wang S-L, Lei Y-K. Multi-step dimensionality reduction and semi-supervised graph-based tumor classification using gene expression data. Artificial Intelligence in Medicine. 2010;50:181-91.