Differential Flux Balance Analysis of Quantitative Proteomic Data on Protein Interaction Networks

Biaobin Jiang
Department of Biological Sciences
Purdue University
West Lafayette, IN 47907
Email: bjiang@purdue.edu

David F. Gleich
Department of Computer Science
Purdue University
West Lafayette, IN 47907
Email: dgleich@purdue.edu

Michael Gribskov
Department of Biological Sciences
Department of Computer Science
Purdue University
West Lafayette, IN 47907
Email: gribskov@purdue.edu

***Abstract*—Protein fluxes provide a more refined notion of protein abundance than raw counts alone by considering potential channels based on protein interaction networks. We propose a novel method to estimate protein fluxes in a protein interaction network using a linear programming model based on the framework of flux balance analysis. When we combine this estimate of protein fluxes with a protein-centric network measure, inspired by egocentric network analysis in sociology, we discover that the fluxes of proteins encoded by hypermutated genes in colon cancer have substantially higher alterations in cancer cells than the protein quantities alone. These alterations remain statistically significant under different network perturbations. We conclude that the importance of a change in the quantity of a protein is determined not only by the protein itself, but also by its network neighbors.**

Index Terms*—biological interactions, linear programming, cancer, biological system modeling, mass spectroscopy.

I. Introduction

Systems biology is the interdisciplinary study of the cooperative behavior of biological molecules through complex interactions in a biological system. A fundamental task in systems biology is to uncover the rules governing how molecules select their interacting partners in a complex interaction network. Whether and how, for example, a protein changes its *friendship* under different physiological conditions given a protein physical interaction network is unclear.

High throughput technologies enable comprehensive measurements of various molecular profiles that are useful for the study of complex diseases, such as cancers [1, 2, 3]. By comparing these profiles in different conditions, one can identify both qualitative and quantitative molecular alterations, such as genetic mutations and differential protein abundance in signaling pathways, respectively. However, identical genetic mutations are rarely identified in different patients, but rather are often found in common signaling pathways [4, 5]. Attempts have been made to investigate how genetic variants disrupt protein interactions [6, 7]. But these methods did not incorporate quantitative protein abundance data, and therefore cannot be used to interpret how structurally abnormal proteins caused by genetic mutations mediate interaction dynamics in signaling pathways.

Quantitative changes in protein interactions can be experimentally measured by AP-SWATH (Affinity Purification combined with Sequential Window Acquisition of all THeoretical spectra) mass spectrometry [8, 9]. However, currently the AP-SWATH technique is limited to small-scale studies due to the insufficient precision of statistical estimation for interacting protein abundances. And large-scale proteome-wide studies of quantitative changes in protein-protein interaction networks still depend on computational modeling. From a computational perspective, thermodynamic or kinetic modeling has been used to offer a precise quantitative map of transcriptional regulatory pathways [10]. However, the application scale of this method is usually limited to less than 10 transcription factors due to its high computational cost and the difficulty of obtaining the required kinetic parameters. In sum, both AP-SWATH and thermodynamic or kinetic modeling only work on small-scale studies. Extending the both methods to large-scale studies is an active research topic in systems biology community.

Linear modeling is able to model high-throughput large-scale data sets, and is widely used to study biological networks. Li *et al.* constructed a bipartite network between exon fragments and transcripts to estimate transcript abundance from mRNA sequencing data using a modified regularized least squares model [11]. Wang *et al.* reconstructed a transcriptional regulatory network from multiple microarray data sets by linear programming [12]. Duarte *et al.* utilized Flux Balance Analysis (FBA), a model based on linear programming, to reconstruct a human metabolic network [13]. However, to our knowledge, there are few studies using linear models to analyze proteome-wide quantitative data in a large-scale protein interaction network. In fact, FBA can be extended from metabolic networks to protein interaction networks under reasonable assumptions (see Methods).

To this end, we propose a linear programming model based on the FBA framework, to estimate *protein flux* (for definition, see Methods) in a protein interaction network, and demonstrate its use on proteome-wide quantitative data in colon cancer. In the Methods section, we make three basic assumptions to adapt the network-based proteomic model to the framework of FBA, and then mathematically describe the linear programming model and the egocentric network metric used in evaluation. In the Results section, we describe the quantitative proteomic data sets; illustrate the distribution of protein fluxes; and finally examine the predictive performance of the estimated protein fluxes within the egocentric networks of hypermutated genes, and also the performance robustness under different network perturbations.

978-1-4799-7591-4/15/$31.00 ©2015 IEEE

II. METHODS

Flux Balance Analysis (FBA) is widely used in metabolic networks [14]. It calculates the fluxes of metabolites through the network of biochemical reactions based on reaction stoichiometry. Similarly, given one protein with multiple binding partners in a protein interaction network, we would like to estimate the proportions of the protein binding to each of its partners. This binding fraction is termed *protein flux* in this study.

FBA can be viewed as a linear programming model [14]. Given a set of stoichiometric constraints, FBA aims to optimize a predefined objective function, e.g., to maximize a set of fluxes. Similarly, the goal of the proposed model in this study is to maximize the sum of all protein fluxes in the interaction network. The rationale for this objective function lies in two facts. On one hand, many proteins cannot function alone in a living cell. Instead, they bind to their network partners in a functional group to fulfill biological functions *in vivo*. On the other hand, proteins are intrinsically expensive to produce, and it is inefficient to produce proteins in excess of their binding partners.

A. Model Assumption

The proposed model is subject to the following two assumptions:

- No Stoichiometry: each protein copy can only bind one single copy of its neighboring proteins in the network. And the ratio of each binding pair of protein copies is 1:1, since currently no large-scale stoichiometric data are available. Similarly to the application of FBA in biochemical reaction networks, proteome-wide stoichiometric data can be naturally incorporated into our FBA-based model once they can be measured in high throughput manner.
- Independence: protein binding depends only on the abundance of the two proteins. Other complicated factors like protein locations, binding affinity and regulatory mechanism are not considered in this study, since these factors cannot be simplified into the linear structure of FBA. Instead, modeling protein locations, binding affinity and regulatory mechanism requires a spatial, high-order and time-varying system model. Solving this complicated model is computationally expensive, and cannot be applied to large scale proteome-wide data so far. In fact, the independence assumption is similar to assuming complete and rapid mixing of protein copies.

B. Model Construction

The model construction starts with a protein interaction network and a list of protein quantity data measured by quantitative proteomic techniques. We denote the protein interaction network as an undirected graph with a symmetric adjacency matrix $\boldsymbol{G} \in \mathbb{R}^{m\times m}$ where m is the number of proteins, and $G_{ij} = 1$ $(i, j = 1, \cdots, m)$ if protein i physically interacts with protein j, and 0 otherwise. Then the adjacency matrix is converted into an incidence matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ where n is the number of edges in the graph (normally $m \ll n$), and $A_{ik} = A_{jk} = 1$ $(k = 1, \cdots, n)$ if $G_{ij} = 1$ and 0 otherwise, where $i < j$. In fact, the incidence matrix shows the relationship between nodes and edges in a graph. We denote the protein quantity data as a vector $\mathbf{b} \in \mathbb{R}^m$ and the protein flux of each interaction as $\mathbf{x} \in \mathbb{R}^n$. The model is designed to maximize the total interaction fluxes, i.e., $\mathbf{c}^T\mathbf{x}$ where $\mathbf{c}^T$ is an all-one vector. The portion of bound proteins in the flux is calculated as $\boldsymbol{A}\mathbf{x}$; this portion of any protein cannot exceed its total quantity, i.e., $\boldsymbol{A}\mathbf{x} \leq \mathbf{b}$. The estimated fluxes cannot be negative, i.e., $\mathbf{x} \geq \mathbf{0}$. In sum, we derive an FBA-like model based on linear programming as

$$\begin{aligned} \underset{\mathbf{x}}{\text{maximize}} \quad & \mathbf{c}^T\mathbf{x} \\ \text{subject to} \quad & \boldsymbol{A}\mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0}. \end{aligned} \tag{1}$$

We empirically set the lower bound of $\mathbf{x}$ as 0.001 other than exact 0 for two reasons. First, to further compare the fold change of protein fluxes in two conditions, we need to calculate $\log_2(x^{c_1}/x^{c_2})$, and it has no meaning when x^{c_2} exactly equals to 0. Second, in practice we found that, given different protein abundance $\mathbf{b}$, the lower bound of $\mathbf{x}$ set to a value less than 0.001 yields different boundary values in the solutions when we use the interior point method to solve the linear programming problem.

C. Evaluation Metric

To test if differential protein flux prioritizes disease-associated genes better than differential protein quantity, we have devised a novel protein-wise metric based on an Egocentric Network (or *EgoNet*). The EgoNet of one node in a graph is defined as a local subnetwork comprising that node, its direct neighbors and the edges among them. In the literature of social and information networks, EgoNet analysis is frequently used to identify important structural and anomalous types of nodes [15, 16]. In this study, similarly, the flux changes in the EgoNet of one protein help identify how altered quantities affect a local network region centered at that protein. For a flux network, the EgoNet matrix of one protein t is defined as

$$\boldsymbol{Z}_{(t)}(i^*, j^*) = \begin{cases} x_{k^*} & \text{if } (i^*, j^*) \in EgoNet(t); \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where k^* is the corresponding index of edge (i^*, j^*) in flux vector $\mathbf{x}$. Under two different conditions c_1 and c_2, the total flux change of a protein t within its EgoNet can be quantified using the Frobenius norm as $\mathbf{s}_E(t)$ (Equation 3). In contrast, we define two baseline scores for protein t between conditions c_1 and c_2 as the total flux change to neighbors $\mathbf{s}_N(t)$ (Equation 4) and the quantity change $\mathbf{s}_B(t)$ (Equation 5), respectively as,

$$\mathbf{s}_E(t) = \|\boldsymbol{Z}^{c_1}_{(t)} - \boldsymbol{Z}^{c_2}_{(t)}\|_F \tag{3}$$

$$\mathbf{s}_N(t) = \mathbf{a}_t^T|\mathbf{x}^{c_1} - \mathbf{x}^{c_2}| \tag{4}$$

$$\mathbf{s}_B(t) = |b_t^{c_1} - b_t^{c_2}| \tag{5}$$

where $\mathbf{a}_t^T$ is the t-th row of matrix $\boldsymbol{A}$.

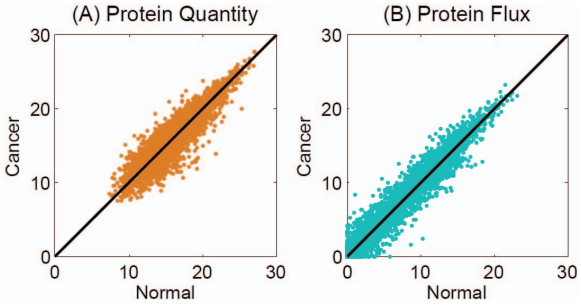

Fig. 1. Scatter plot of protein quantities (A) and fluxes (B) in normal (x-axis) vs. cancer (y-axis) conditions.

III. Results

A. Data Sets

There are two data sets needed in differential FBA. One is the protein-protein physical interaction network, which can be downloaded from BioGRID [17]. The other is the protein quantity data (absolute copy numbers), which is obtained from an extensive quantitative proteome study of colon normal tissue and adenocarcinoma [18]. After ID mapping across these two data sets using BioMart [19], we identified $6,334$ proteins with measured quantities in both conditions (normal and cancerous) and $49,337$ physical interactions among them. Due to the large range of measured protein quantities (10^2 to 10^8), we performed a log-scaling, as shown in Figure 1(A).

B. Distribution of Differential Fluxes

Given the protein quantities $\mathbf{b}_n$ in normal colon tissue and $\mathbf{b}_c$ in colon cancer, respectively, the linear programming model (Equations 1) was solved to estimate the protein fluxes $\mathbf{x}_n$ and $\mathbf{x}_c$, respectively (Figure 1(B)). Comparing Figure 1 (A) and (B), we find that majority of protein quantities and fluxes show no change between normal and cancer conditions. However, a portion of the fluxes are close to zero, even though their linked proteins are abundant, indicating that some of the flux channels (protein interactions) are shut down under specific pathological conditions.

To highlight significant changes in protein quantities and fluxes, we illustrate the distribution of $\log_2$ fold changes of the ratios of cancerous to normal conditions in Figure 2. A subset of interactions show significant $\log_2$ fold changes ($5+$ folds) compared to the overall $\log_2$ fold changes in protein quantities ($0.2+$ folds). This suggests that the proposed model is able to correctly combine the changes in protein quantities and interactions. In this case, one can find an associated set of interaction fluxes that explain the change in protein quantities. For instance, given the up-regulation of one protein, the proposed model is able to inform us which fluxes are concurrently up-regulated and which are not responsive or down-regulated.

Our FBA-based linear model is scalable for larger data sets. Using the solver, *linprog* in MATLAB, it normally takes around one minute to solve the model with our data set. We used the default algorithm in the solver, the interior point method, which has proven to be a polynomial-time algorithm in solving linear programming problems [20].

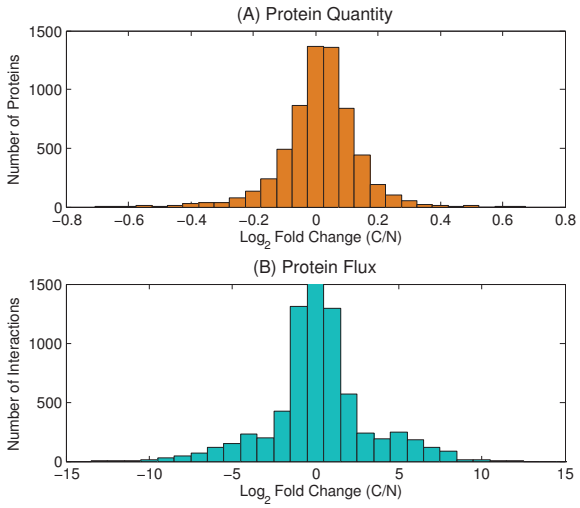

Fig. 2. Histogram of $\log_2$ fold changes (C/N, Cancer over Normal conditions) in protein quantities (A) and protein fluxes (B). The most abundant fold change bin in (B), located within $[-0.5, 0.5]$, is truncated at $1,500$ for visualization convenience. The actual value is $43,798$ interactions.

C. Identification of Known Cancer Genes

To evaluate whether significant flux changes are associated with proteins related to colon cancer, we first collected 18 hypermutated genes from a comprehensive genomic study of colon cancer reported in The Cancer Genome Atlas (TCGA) [3]. We first tested the null hypothesis that the cancer-related proteins with increasing (decreasing) quantities up-regulate (down-regulate) all the fluxes to their network neighbors. For each hypermutated gene/protein, we used a scatter plot to examine the relationship between its quantity fold-change and flux fold-changes (Figure 3). Generally, we can see that there is no positive relationship between the fold changes of protein quantity and protein flux. This rejects the null hypothesis and suggests that an up-regulated (or down-regulated) protein does not necessarily up-regulate (or down-regulate) all of the fluxes to its neighbors. For example, TP53, a well-known oncogene [21], is up-regulated by around 0.3 folds in quantity, whereas its flux fold changes span a wide range (± 8 folds) in cancer cells. Using our model, one can narrow down a large number of fluxes into a small set, and perform further precise modeling, or experimental validation using AP-SWATH, for example.

To further test whether flux changes in EgoNet can be used to predict these mutated genes in colon cancer, we scored each protein using the three Equations (3), (4) and (5), and examined the score ranks of these mutated genes using Receiver Operating Characteristic (ROC) curves (Figure 4). The Area Under the Curve (AUC) indicates the predictive performance of the three metrics. As shown in Figure 4, we find that the EgoNet-based metric achieves the AUC of 0.7327, whereas the other two baseline scores based on the difference only of protein quantity and flux changes to neighbors have the AUCs of 0.4759 and 0.7169, respectively. In particular, at a 0.2 false positive rate, the EgoNet-based metric achieves a true positive rate of around 0.55, whereas the protein quantity change and flux change to neighbors achieve only about 0.2 and 0.45, respectively. This suggests that protein quantity changes influence not only the fluxes

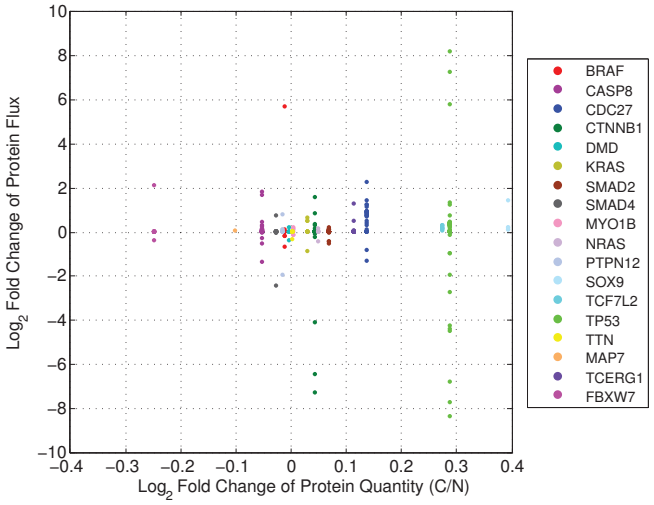

Fig. 3. Fold Change of Protein Quantity (x-axis) and Fluxes (y-axis) in genes that are hypermutated in colon cancer.

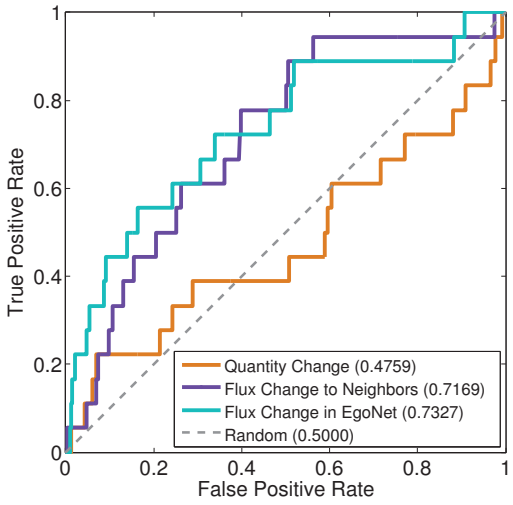

Fig. 4. Receiver Operating Characteristic (ROC) curves in the evaluation of hypermutated gene prediction. The Area Under the Curves (AUCs) are shown in the brackets.

flowing out to their network neighbors, but also the fluxes between their neighbors. In addition, it reveals that the proteins with cancer-related mutations have no significant changes in quantities. Nevertheless, using the proposed differential FBA combined with the egocentric network analysis, we discovered that genetic alterations in fact have much stronger impacts on protein fluxes within the EgoNet than protein quantities alone.

To examine the robustness of cancer-associated protein identification, we altered the protein interaction network and examined whether the prediction performance is robust to network perturbation. We first randomly reassigned the protein abundance data to different nodes in the same network, and found that the prediction performance (Area Under the ROC curve, AUROC) dramatically drops to a random level (Figure 5). Next, we tested whether our method is robust against network topology noise by randomly removing a proportion of edges (while ensuring that every protein has at least one edge). We find that the performance of our method drops slowly until

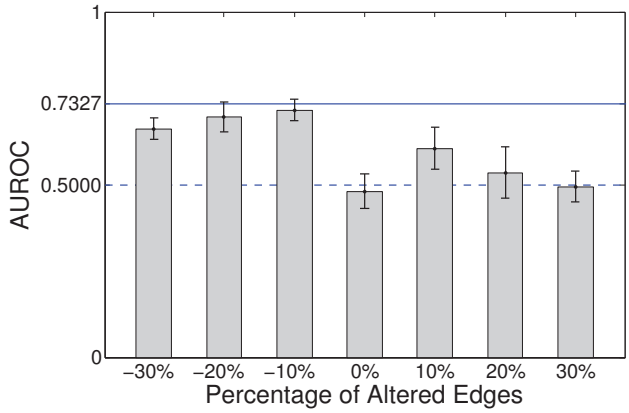

Fig. 5. Area Under the ROC curves (AUROC) in robustness test with randomly perturbed networks. In x-axis, negative percentages denote the proportion of edges randomly removed; positive percentages denote random addition of edges; and 0% denotes random shuffle of protein abundance data. In y-axis, the bars and error bars indicate the means and standard deviations of AUROCs from 10 repeated experiments under each type of network perturbations. AUROC $= 0.5000$ (blue dashed line) indicates the performance of random prediction; and AUROC $= 0.7327$ (blue solid line) indicates the original performance of our method without network perturbation, as shown in Figure 4.

30% of edges are removed (Figure 5). In contrast, randomly adding 10% extra edges results in a significant decline of the performance from 0.7327 to around 0.6, and even worse when 30% extra edges are added in (Figure 5). In sum, this perturbation test suggests that the network topology and the protein abundance data have strong associations with each other. Also, it demonstrates that our method is robust to the network data even in the presence of a relatively high false positive rate.

IV. Conclusion

In this paper, we have presented a computational method based on flux balance analysis to estimate protein fluxes throughout the protein interaction network subject to a balance assumption. We show that the difference in protein quantities can be combined with the protein interactions assuming one-hop balanced diffusion in the network. We also show that the protein flux changes within egocentric networks have a stronger association with the genetic mutational status of the corresponding protein-coding genes than the protein quantity changes. To our knowledge, this is the first attempt to extend flux balance analysis, which is widely used to study metabolic networks, to network-based analysis of quantitative proteomic data.

In future work, we would like to incorporate multiple *omic* data sets into our framework. And so far, we have assumed the stoichiometric ratio between two binding proteins is 1:1. As stoichiometric data *in vivo* become more available, they can be integrated with higher-level network information about functional modules to refine the estimation of protein fluxes.

Acknowledgment

Principal investigator Dr. David F. Gleich would like to acknowledge support from the NSF through CAREER CCF-1149756.

REFERENCES

[1] Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.

[2] Cancer Genome Atlas Research Network, "Comprehensive genomic characterization of squamous cell lung cancers," *Nature*, vol. 489, no. 7417, pp. 519–525, 2012.

[3] Cancer Genome Atlas Network, "Comprehensive molecular characterization of human colon and rectal cancer," *Nature*, vol. 487, no. 7407, pp. 330–337, 2012.

[4] L. Ding, G. Getz, D. A. Wheeler, E. R. Mardis, M. D. McLellan, K. Cibulskis, C. Sougnez, H. Greulich, D. M. Muzny, M. B. Morgan *et al.*, "Somatic mutations affect key pathways in lung adenocarcinoma," *Nature*, vol. 455, no. 7216, pp. 1069–1075, 2008.

[5] N. J. Krogan, S. Lippman, D. A. Agard, A. Ashworth, and T. Ideker, "The cancer cell map initiative: Defining the hallmark networks of cancer," *Molecular Cell*, vol. 58, no. 4, pp. 690–698, 2015.

[6] Q. Zhong, N. Simonis, Q.-R. Li, B. Charloteaux, F. Heuze, N. Klitgord, S. Tam, H. Yu, K. Venkatesan, D. Mou *et al.*, "Edgetic perturbation models of human inherited disorders," *Molecular Systems Biology*, vol. 5, no. 1, 2009.

[7] N. Sahni, S. Yi, M. Taipale, J. I. F. Bass, J. Coulombe-Huntington, F. Yang, J. Peng, J. Weile, G. I. Karras, Y. Wang *et al.*, "Widespread macromolecular interaction perturbations in human genetic disorders," *Cell*, vol. 161, no. 3, pp. 647–660, 2015.

[8] B. C. Collins, L. C. Gillet, G. Rosenberger, H. L. Röst, A. Vichalkovski, M. Gstaiger, and R. Aebersold, "Quantifying protein interaction dynamics by swath mass spectrometry: application to the 14-3-3 system," *Nature Methods*, vol. 10, no. 12, pp. 1246–1253, 2013.

[9] J.-P. Lambert, G. Ivosev, A. L. Couzens, B. Larsen, M. Taipale, Z.-Y. Lin, Q. Zhong, S. Lindquist, M. Vidal, R. Aebersold *et al.*, "Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition," *Nature Methods*, vol. 10, no. 12, pp. 1239–1245, 2013.

[10] Y. Setty, A. E. Mayo, M. G. Surette, and U. Alon, "Detailed map of a cis-regulatory input function," *Proceedings of the National Academy of Sciences*, vol. 100, no. 13, pp. 7702–7707, 2003.

[11] J. J. Li, C.-R. Jiang, J. B. Brown, H. Huang, and P. J. Bickel, "Sparse linear modeling of next-generation mRNA sequencing (RNA-seq) data for isoform discovery and abundance estimation," *Proceedings of the National Academy of Sciences*, vol. 108, no. 50, pp. 19 867–19 872, 2011.

[12] Y. Wang, T. Joshi, X.-S. Zhang, D. Xu, and L. Chen, "Inferring gene regulatory networks from multiple microarray datasets," *Bioinformatics*, vol. 22, no. 19, pp. 2413–2420, 2006.

[13] N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, and B. Ø. Palsson, "Global reconstruction of the human metabolic network based on genomic and bibliomic data," *Proceedings of the National Academy of Sciences*, vol. 104, no. 6, pp. 1777–1782, 2007.

[14] J. D. Orth, I. Thiele, and B. Ø. Palsson, "What is flux balance analysis?" *Nature Biotechnology*, vol. 28, no. 3, pp. 245–248, 2010.

[15] L. Akoglu, M. McGlohon, and C. Faloutsos, "Oddball: Spotting anomalies in weighted graphs," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2010, pp. 410–421.

[16] D. F. Gleich and C. Seshadhri, "Vertex neighborhoods, low conductance cuts, and good seeds for local community methods," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD)*, Aug. 2012, pp. 597–605.

[17] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Research*, vol. 34, no. suppl 1, pp. D535–D539, 2006.

[18] J. R. Wiśniewski, P. Ostasiewicz, K. Duś, D. F. Zielińska, F. Gnad, and M. Mann, "Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma," *Molecular Systems Biology*, vol. 8, no. 1, 2012.

[19] J. Zhang, S. Haider, J. Baran, A. Cros, J. M. Guberman, J. Hsu, Y. Liang, L. Yao, and A. Kasprzyk, "BioMart: a data federation framework for large collaborative projects," *Database*, vol. 2011, p. bar038, 2011.

[20] J. Renegar, "A polynomial-time algorithm, based on newton's method, for linear programming," *Mathematical Programming*, vol. 40, no. 1-3, pp. 59–93, 1988.

[21] A. Petitjean, M. Achatz, A. Borresen-Dale, P. Hainaut, and M. Olivier, "TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes," *Oncogene*, vol. 26, no. 15, pp. 2157–2165, 2007.