Published in final edited form as:

Mt Sinai J Med. 2010 ; 77(4): 345–349. doi:10.1002/msj.20192.

Protein Kinase Target Discovery From Genome-Wide Messenger RNA Expression Profiling

Avi Ma'ayan^{1,2,3} and John C. He^{1,4,5}

¹ Department of Pharmacology and Systems Therapeutics, Mount Sinai School of Medicine, New York, NY

² Systems Biology Center New York, Mount Sinai School of Medicine, New York, NY

³ Experimental Therapeutics Institute, Mount Sinai School of Medicine, New York, NY

⁴ Department of Medicine, Mount Sinai School of Medicine, New York, NY

⁵ James J. Peters VA Medical Center, Bronx, New York

Abstract

Genome-wide messenger RNA profiling provides a snapshot of the global state of the cell under different experimental conditions such as diseased versus normal cellular states. However, because measurements are in the form of quantitative changes in messenger RNA levels, such experimental data does not provide direct understanding of the regulatory molecular mechanisms responsible for the observed changes. Identifying potential cell signaling regulatory mechanisms responsible for changes in gene expression under different experimental conditions or in different tissues has been the focus of many computational systems biology studies. Most popular approaches include promoter analysis, gene ontology, or pathway enrichment analysis, as well as reverse engineering of networks from messenger RNA expression data. Here we present a rational approach for identifying and ranking protein kinases that are likely responsible for observed changes in gene expression. By combining promoter analysis; data from various chromatin immunoprecipitation studies such as chromatin immunoprecipitation sequencing, chromatin immunoprecipitation coupled with paired-end ditag, and chromatin immunoprecipitation-on-chip; protein-protein interactions; and kinase-protein phosphorylation reactions collected from the literature, we can identify and rank candidate protein kinases for knock-down, or other types of functional validations, based on genome-wide changes in gene expression. We describe how protein kinase candidate identification and ranking can be made robust by cross-validation with phosphoproteomics data as well as through a literature-based text-mining approach. In conclusion, data integration can produce robust candidate rankings for understanding cell regulation through identification of protein kinases responsible for gene expression changes, and thus rapidly advancing drug target discovery and unraveling drug mechanisms of action.

Keywords

cell signaling; ChIP-X enrichment analysis; computational systems biology; mRNA profiling; promoter analysis; regulatory mechanisms; transcription factors

Address Correspondence to: Avi Ma'ayan, Department of Pharmacology and Systems Therapeutics Mount Sinai School of Medicine, New York, NY, avi.maayan@mssm.edu.

DISCLOSURES

Potential conflict of interest: Nothing to report.

As progress in biomedical sciences leads to breakthrough in biotechnology, biotechnology is fueling back progress in biomedical sciences. However, as molecular data collection in mammalian cell biology research becomes more extensive and complex, advances in computation are now critical for continual progress.¹ Computational systems biology is playing an increasingly important role in many aspects of biomedical research.² Emerged as the result of the sequencing of the yeast, mouse, and human genomes, computational systems biology and bioinformatics (CSB) aims to develop robust theoretical models that explain how molecular components give rise to cellular, tissue, and organism phenotypes at the genome-wide scale.³ Biotechnological advances in instruments capable of measuring molecular components within cells at a genome-wide level, together with the infusion of ideas from physics and mathematics for data analysis and advances from computer science for data sharing, storage, search, and visualization, are expected to lead to a surge in biomedical breakthroughs in translational research in the near future.

Although genome-wide proteomic approaches improve rapidly, currently and throughout the past decade the most widely available and cost-effective genome-wide expression data collected is at the RNA level. Typical studies examine cells under different experimental conditions such as control versus treated, or disease versus normal intracellular states. Since quantitative changes in mRNA levels do not directly explain how regulatory molecular mechanisms are altered to induce changes in gene expression, and in turn lead to changes in cellular phenotype, identifying such regulatory mechanisms has been the focus of many CSB studies. This is because such understanding will enable us to, among other things, better control cell behavior with small molecules, and in turn translate such ability to therapeutics development. Most popular approaches for data interpretation of changes in gene expression include promoter analysis,^{4,5} gene ontology,⁶ and pathway enrichment analyses,⁷ as well as reverse engineering of networks from mRNA expression data.⁸ The ultimate goal of many of these studies is to identify and rank potential target genes/proteins that if knocked down would explain the observed changes by reversing them. Such protein targets may ultimately become drug targets. In this article, we describe how protein kinases that are likely responsible for observed genome-wide changes in gene expression at the messenger RNA (mRNA) level can be identified and ranked by a rational approach. In the first phase, one combines promoter analysis of microarray results with data from various chromatin immunoprecipitation (ChIP-X) studies, such as chromatin immunoprecipitation sequencing (ChIP-Seq), chromatin immunoprecipitation coupled with paired-end ditag (ChIP-PET), and chromatin immunoprecipitation-on-chip (ChIP-chip). Then, using protein-protein interactions and kinase-protein phosphorylation reactions collected from the public domain, one can potentially identify and rank candidate protein kinases for knock-down and/or other types of functional validations.

We discuss how predictions made by this approach can be cross-validated using other types of high-throughput experimental data, such as phosphoproteomics and protein/DNA arrays, as well as using lists of protein kinases extracted automatically from the literature.

The overall workflow is shown schematically in Figure 1 and discussed below.

DETECTING CHANGES IN MESSENGER RNA EXPRESSION

The first step of our CSB approach is to identify the genes (mRNAs) that are differentially expressed under 2 conditions. This is a standard procedure that can be carried out using statistical tests such as analysis of variance or an adjusted *t* test, and/or unsupervised clustering approaches such as hierarchical clustering or principal component analyses.

PROMOTER ANALYSIS AND CHROMATIN IMMUNOPRECIPITATION ENRICHMENT ANALYSIS

To link changes in gene expression to the molecular mechanisms responsible for the observed changes, we can first apply promoter analysis using binding site matrices obtained from databases such as TRANSFAC4 or JASPAR.5 This method computationally scans the DNA sequence in the proximity of genes' coding regions, seeking enrichment of binding sites for annotated transcription factor binding logo-motifs. Such an approach can identify and rank a list of transcription factor candidates responsible for the observed changes by computing binding-site enrichment for all the genes that changed in expression significantly. Transcription factor binding site enrichment can be computed for all genes that significantly changed in mRNA expression, or by dividing regulation for genes that were differentially increased or decreased in expression compared with the control. Alternatively, we can generate a list of most likely transcriptional regulators by cross-referencing the genes that increased or decreased in expression with previously published ChIP-X studies. Such studies report the binding of specific transcription factors in proximity to gene coding regions. By compiling the results from many ChIP-X studies, we can obtain a global picture of transcriptional activity of many transcription factors. Although such data is collected in many cell types and across different mammalian organisms under different conditions, it has the advantage that it considers the chromatin state of the cell and as such is expected to reduce false positives, a critical limitation of the binding logo-motif promoter scanning approach. Both the promoter analyses and the ChIP-X enrichment analyses produce ranked lists of transcription factors that most likely regulate genes that significantly increased or decreased in mRNA expression.

Such lists can be compared for overlap to assess consistency.

CONNECTING IDENTIFIED TRANSCRIPTION FACTORS

Most analyses stop at this stage; however, our approach next step is to “connect” the transcription factors detected by the ChIP-X enrichment and/or by the promoter scanning approach using known experimentally reported protein-protein interactions. Several tools have been developed for using prior knowledge about protein-protein interaction networks to build subnetworks that connect lists of “seed nodes” given as input.^{9–12} We have developed Genes2Networks¹⁰ and successfully used it for finding pathways responsible for neurite outgrowth¹³ and predicting a novel disease gene, SHOC2, that was found to contain a mutation that can cause a Noonan-like syndrome.¹⁴ Once we have built a subnetwork that connects the transcription factors, we can convert the proteins (nodes) within this subnetwork to a list (of proteins). Such a list of proteins can then be fed into Kinase Enrichment Analysis (KEA), a Web tool we developed for identifying protein kinases that are enriched with substrates in a given input list of proteins¹⁵; or into sequence-based motif discovery tools such as Motif-X¹⁶ or GPS¹⁷ for identifying the protein kinases that most likely regulate the subnetwork anchored with the transcription factors detected to regulate the observed changes in gene expression. Alternatively, programs such as Motif-X or similar sequence-based approaches^{17,18} can be used to associate protein kinases enriched with substrate sequences in the subnetwork. The final output of the workflow is a ranked list of protein kinases as targets for experimental validation through knock-down, overexpression, or dominant negative perturbations.

VARIATIONS IN WORKFLOW

Some of the steps in the process can be skipped. We can, for example, look for kinases that significantly changed in their mRNA level directly. Alternatively, we can look for kinases in

the protein-protein interactions subnetworks created from the transcription factors detected computationally, avoiding the KEA or Motif-X analyses. We also have the option to only use transcription factor binding motif promoter scanning and not ChIP-X, or to use only KEA and not Motif-X. How can we evaluate which approach works best? To answer this question we can bring in additional experimental evidence for cross-validation. For example, if we have data describing changes in protein/DNA interactions,¹³ or proteomics data reporting quantitative changes in protein levels (eg, profiling the entire nuclear proteome¹⁹), or, more directly, stable isotope labeling with amino acids in cell culture (SILAC) phosphoproteomics data describing quantitative changes in phosphoprotein levels,²⁰ we can evaluate and validate the different steps of the analysis. Protein/DNA interaction arrays such as those developed by Panomics/Affymetrix (www.panomics.com/index.php?id=product18) can validate the transcription factors detected by the promoter analysis or the ChIP-X enrichment analysis. SILAC phosphoproteomics followed by a KEA or Motif-X analysis can validate the final output of the protein kinase rankings. In addition, we can see if the candidate protein kinases detected through the analysis were previously implicated with the biological process we aim to better understand. Ultimately, we can employ an RNA interference (RNAi) screen targeting all kinases to see if our rankings reflect our predictions. The different validation approaches are summarized in the schematic provided in Figure 1 under the title Cross-Validation. All the computational analyses presented here can be achieved by meshing the programs Genes2Networks,¹⁰ Lists2Networks,²¹ KEA,¹⁵ and ChEA (unpublished) developed by the Ma'ayan laboratory.

CONCLUSION

Protein kinases are great candidate drug targets because they display global effects through their ability to phosphorylate many targets exhibiting broad changes in cell behavior. Small-molecule inhibitors that target protein kinases can be easily screened for hits through in vitro phosphorylation assays. Protein kinases contain regulatory structural domains as well as catalytic domains amenable for small-molecule targeting. Here we have described several alternatives for inferring and ranking protein kinases, both computationally and experimentally, and for cross-validation of predictions initially made based on gene expression changes. However, there is still a need to validate the approach presented experimentally. With our approach there are many “free parameters” as well as “process routes” for making the ranked kinase predictions. Hence, protein kinase rankings can result in many different output tables depending on the computational and/or experimental routes taken and parameters chosen for cutoffs and algorithmic setup in each step. Although cross-validation could be useful for directing the right route and parameter selection, experimental validation, through functional studies as well as through short hairpin RNA (shRNA) or RNAi screens, is probably the best approach for evaluating the approach presented. CSB is now at a stage where data integration across regulatory layers can be used to infer and unravel regulation at the molecular level of mammalian cells at a degree of detail never possible before. Computational approaches for data integration and creative algorithms that can extract new knowledge from the increasingly complex, multidimensional, diverse datasets are of critical importance in this milieu.

Acknowledgments

The research presented is supported by NIH grants R01-DK088541 to JCH and AM, R01-DK078897 to JCH, and P50-GM071558-01A27398 to AM.

References

1. Ma'ayan A. Insights into the organization of biochemical regulatory networks using graph theory analyses. *J Biol Chem.* 2009; 284:5451–5455. [PubMed: 18940806]
2. Ma'ayan A. Network integration and graph analysis in mammalian molecular systems biology. *IET Syst Biol.* 2008; 2:206–221. [PubMed: 19045817]
3. Ma'ayan A, Blitzer RD, Iyengar R. Toward predictive models of mammalian cells. *Annu Rev Biophys Biomol Struct.* 2005; 34:319–349. [PubMed: 15869393]
4. Matys V, Kel-Margoulis OV, Fricke E, et al. TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* 2006; 34(suppl 1):D108–D110. [PubMed: 16381825]
5. Portales-Casamar E, Thongjuea S, Kwon AT, et al. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 2010; 38(suppl 1):D105–D110. [PubMed: 19906716]
6. The Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.* 2010; 38(suppl 1):D331–D335. [PubMed: 19920128]
7. Kanehisa M, Goto S, Furumichi M, et al. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 2010; 38(suppl 1):D355–D360. [PubMed: 19880382]
8. Margolin A, Nemenman I, Basso K, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics.* 2006; 7(suppl 1):S7. [PubMed: 16723010]
9. Wu J, Vollenius T, Ovaska K, et al. Integrated network analysis platform for protein-protein interactions. *Nat Methods.* 2009; 6:75–77. [PubMed: 19079255]
10. Berger SI, Posner JM, Ma'ayan A. Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases. *BMC Bioinformatics.* 2007; 8:372. [PubMed: 17916244]
11. Lee SA, Chan CH, Chen TC, et al. POINTE: protein interactome with sub-network analysis and hub prioritization. *BMC Bioinformatics.* 2009; 10:114. [PubMed: 19379523]
12. Antonov AV, Dietmann S, Rodchenkov I, Mewes HW. PPI spider: a tool for the interpretation of proteomics data in the context of protein-protein interaction networks. *Proteomics.* 2009; 9:2740–2749. [PubMed: 19405022]
13. Bromberg KD, Ma'ayan A, Neves SR, Iyengar R. Design logic of a cannabinoid receptor signaling network that triggers neurite outgrowth. *Science.* 2008; 320:903–909. [PubMed: 18487186]
14. Cordeddu V, Di Schiavi E, Pennacchio LA, et al. Mutation of SHOC2 promotes aberrant protein N-myristoylation and causes Noonan-like syndrome with loose anagen hair. *Nat Genet.* 2009; 41:1022–1026. [PubMed: 19684605]
15. Lachmann A, Ma'ayan A. KEA: kinase enrichment analysis. *Bioinformatics.* 2009; 25:684–686. [PubMed: 19176546]
16. Schwartz D, Gygi SP. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat Biotechnol.* 2005; 23:1391–1398. [PubMed: 16273072]
17. Xue Y, Zhou F, Zhu M, et al. GPS: a comprehensive www server for phosphorylation sites prediction. *Nucleic Acids Res.* 2005; 33(suppl 2):W184–W187. [PubMed: 15980451]
18. Linding R, Jensen LJ, Pasculescu A, et al. NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Res.* 2008; 36(suppl 1):D695–D699. [PubMed: 17981841]
19. Lu R, Markowitz F, Unwin RD, et al. Systems-level dynamic analyses of fate change in murine embryonic stem cells. *Nature.* 2009; 462:358–362. [PubMed: 19924215]
20. Blagoev B, Ong SE, Kratchmarova I, Mann M. Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nat Biotechnol.* 2004; 22:1139–1145. [PubMed: 15314609]
21. Lachmann A, Ma'ayan A. Lists2Networks: integrated analysis of gene/protein lists. *BMC Bioinformatics.* 2010; 11:87. [PubMed: 20152038]

Fig 1. Workflow process for inferring protein kinases from mRNA gene expression changes with added methods for cross-validation. **Abbreviations:** ChIP-X, chromatin immunoprecipitation studies; mRNA, messenger RNA; TFBS, transcription factor binding site.