2. **Large-Scale SVM Regression** [35] Bottou demonstrated that a stochastic gradient descent solver for a variety of learning problems (including support vector machine optimization) is able to scale with extremely large datasets, while converging to the predictive performance of traditional optimization algorithms.

3. **Large-Scale L-Regularized Learning** [36] Stochastic coordinate descent (a method related to stochastic gradient descent, but with a slightly different update rule), can be used to learn sparse regression models, with small training-times, even for data sets where both the dimensionality and the number of training-points is large. This is the approach used in our preliminary work.

 The net effect of this analysis will be to find the weighting of different factors that will lead us to conclude that two genes in some species are correlated. Then, using available Arabidopsis time-series data [22], and other datasets that are currently being generated in our lab and others, we will combine correlation with time-series [22,26-29] and perturbation approaches using Graphical Lasso [37] to form causal networks.

**REFERENCES CITED**

1. Katari, MS, Nowicki, SD, Aceituno, FF, Nero, D, Kelfer, J, Thompson, LP, Cabello, JM, Davidson, RS, Goldberg, AP, Shasha, DE, Coruzzi, GM, and Gutierrez, RA, *VirtualPlant: a software platform to support systems biology research.* Plant Physiol, 2010. **152**(2): p. 500-515.

2. Shannon, PT, Reiss, DJ, Bonneau, R, and Baliga, NS, *The Gaggle: an open-source software system for integrating bioinformatics software and data sources.* BMC Bioinformatics, 2006. **7**: p. 176.

3. Waltman, P, Kacmarczyk, T, Bate, AR, Kearns, DB, Reiss, DJ, Eichenberger, P, and Bonneau, R, *Multi-species integrative biclustering.* Genome Biology, 2010. **11**(9): p. R96.

4. Chiu, JC, Lee, EK, Egan, MG, Sarkar, IN, Coruzzi, GM, and DeSalle, R, *OrthologID: automation of genome-scale ortholog identification within a parsimony framework.* Bioinformatics, 2006. **22**(6): p. 699-707.

5. Goloboff, PA, Farris, JS, and Nixon, KC, *TNT, a free program for phylogenomic analysis.* Cladistics, 2008. **24**(5): p. 774-786.

6. Cibrián-Jaramillo, A, De la Torre-Bárcena, J, Lee, E, Katari, M, Little, D, Stevenson, D, Martienssen, R, Coruzzi, G, and DeSalle, R, *Using Phylogenomic Patterns and Gene Ontology to Identify Proteins of Importance in Plant Evolution.* Genome Biol and Evol, 2010. **2**(0): p. 225.

7. de la Torre-Barcena, JE, Kolokotronis, SO, Lee, EK, Stevenson, DW, Brenner, ED, Katari, MS, Coruzzi, GM, and DeSalle, R, *The impact of outgroup choice and missing data on major seed plant phylogenomics using genome-wide EST data.* PLoS One, 2009. **4**(6): p. e5764.

8. Lee EK, C-JA, Kolokotronis SO, Katari MS, Stamatakis A, Ott M, Chiu JC, Little DP, Stevenson DW, Brenner ED, McCombie RW, Martienssen RA, Coruzzi GM, DeSalle R, *Functional Phylogenomics of the Seed Plants.* PloS Genetics, 2011. **Conditionally accepted**.

9. Gifford, ML, Dean, A, Gutierrez, RA, Coruzzi, GM, and Birnbaum, KD, *Cell-specific nitrogen responses mediate developmental plasticity.* Proc Natl Acad Sci (USA), 2008. **105**(2): p. 803-808.

10. Gutierrez, RA, Gifford, ML, Poultney, C, Wang, R, Shasha, DE, Coruzzi, GM, and Crawford, NM, *Insights into the genomic nitrate response using genetics and the Sungear Software System.* J Exp Bot, 2007. **58**(9): p. 2359-2367.

11. Gutierrez, RA, Lejay, LV, Dean, A, Chiaromonte, F, Shasha, DE, and Coruzzi, GM, *Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in Arabidopsis.* Genome Biol, 2007. **8**(1): p. R7.

12. Nero, D, Katari, MS, Kelfer, J, Tranchina, D, and Coruzzi, GM, *In silico evaluation of predicted regulatory interactions in Arabidopsis thaliana.* BMC Bioinformatics, 2009. **10**: p. 435.

13. Thum, KE, Shin, MJ, Gutierrez, RA, Mukherjee, I, Katari, MS, Nero, D, Shasha, D, and Coruzzi, GM, *An integrated genetic, genomic and systems approach defines gene networks regulated by the interaction of light and carbon signaling pathways in Arabidopsis.* BMC Syst Biol, 2008. **2**: p. 31.

14. Wang, R, Tischner, R, Gutierrez, RA, Hoffman, M, Xing, X, Chen, M, Coruzzi, G, and Crawford, NM, *Genomic analysis of the nitrate response using a nitrate reductase-null mutant of Arabidopsis.* Plant Physiol, 2004. **136**(1): p. 2512-2522.

15. Gutierrez, RA, Stokes, TL, Thum, K, Xu, X, Obertello, M, Katari, MS, Tanurdzic, M, Dean, A, Nero, DC, McClung, CR, and Coruzzi, GM, *Systems approach identifies an organic nitrogen-responsive gene network that is regulated by the master clock control gene CCA1.* Proc Natl Acad Sci (USA), 2008. **105**(12): p. 4939-4944.

16. Warde-Farley, D, Donaldson, SL, Comes, O, Zuberi, K, Badrawi, R, Chao, P, Franz, M, Grouios, C, Kazi, F, Lopes, CT, Maitland, A, Mostafavi, S, Montojo, J, Shao, Q, Wright, G, Bader, GD, and Morris, Q, *The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function.* Nucleic Acids Res, 2010. **38 Suppl**: p. W214-220.

17. Moreno-Risueno, MA, Busch, W, and Benfey, PN, *Omics meet networks-using systems approaches to infer regulatory networks in plants.* Curr Opin Plant Biol, 2009.

18. Ashburner, M, Ball, CA, Blake, JA, Botstein, D, Butler, H, Cherry, JM, Davis, AP, Dolinski, K, Dwight, SS, Eppig, JT, Harris, MA, Hill, DP, Issel-Tarver, L, Kasarskis, A, Lewis, S, Matese, JC, Richardson, JE, Ringwald, M, Rubin, GM, and Sherlock, G, *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nature Genetics, 2000. **25**(1): p. 25-29.

19. Mewes, HW, Amid, C, Arnold, R, Frishman, D, Guldener, U, Mannhaupt, G, Munsterkotter, M, Pagel, P, Strack, N, Stumpflen, V, Warfsmann, J, and Ruepp, A, *MIPS: analysis and annotation of proteins from whole genomes.* Nucleic Acids Res, 2004. **32**(Database issue): p. D41-44.

20. Poultney, CS, Gutierrez, RA, Katari, MS, Gifford, ML, Paley, WB, Coruzzi, GM, and Shasha, DE, *Sungear: interactive visualization and functional analysis of genomic datasets.* Bioinformatics, 2007. **23**(2): p. 259-261.

21. Ferro, A, Giugno, R, Pigola, G, Pulvirenti, A, Skripin, D, Bader, GD, and Shasha, D, *NetMatch: a Cytoscape plugin for searching biological networks.* Bioinformatics, 2007. **23**(7): p. 910-912.

22. Krouk, G, Mirowski, P, LeCun, Y, Shasha, DE, and Coruzzi, GM, *Predictive network modeling of the high-resolution dynamic plant transcriptome in response to nitrate.* Genome Biology, 2010. **11**(12): p. R123.

23. Mirowski, P, Madhavan, D, Lecun, Y, and Kuzniecky, R, *Classification of patterns of EEG synchronization for seizure prediction.* Clin Neurophysiol, 2009. **120**(11): p. 1927-1940.

24. Craigon, DJ, James, N, Okyere, J, Higgins, J, Jotham, J, and May, S, *NASCArrays: a repository for microarray data generated by NASC's transcriptomics service.* Nucleic Acids Res, 2004. **32**(Database issue): p. D575-577.

25. Margolin, AA, Nemenman, I, Basso, K, Wiggins, C, Stolovitzky, G, Dalla Favera, R, and Califano, A, *ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.* BMC bioinformatics, 2006. **7 Suppl 1**: p. S7.

26. Greenfield, A, Madar, A, Ostrer, H, and Bonneau, R, *DREAM4: Combining genetic and dynamic information to identify biological networks and dynamical models.* PLoS One, 2010. **5**(10): p. e13397.

27. Kim, S, Kim, J, and Cho, KH, *Inferring gene regulatory networks from temporal expression profiles under time-delay and noise.* Comput Biol Chem, 2007. **31**(4): p. 239-245.

28. Madar, A, Greenfield, A, Vanden-Eijnden, E, and Bonneau, R, *DREAM3: network inference using dynamic context likelihood of relatedness and the inferelator.* PLoS One, 2010. **5**(3): p. e9803.

29. Yu, J, Smith, VA, Wang, PP, Hartemink, AJ, and Jarvis, ED, *Advances to Bayesian network inference for generating causal networks from observational biological data.* Bioinformatics, 2004. **20**(18): p. 3594-3603.

30. Altschul, SF, Madden, TL, Schaffer, AA, Zhang, J, Zhang, Z, Miller, W, and Lipman, DJ, *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucl.Acids Res., 1997. **25**(17): p. 3389-3402.

31. Li, L, Stoeckert, CJ, Jr., and Roos, DS, *OrthoMCL: identification of ortholog groups for eukaryotic genomes.* Genome Res, 2003. **13**(9): p. 2178-2189.

32. O'Brien, KP, Remm, M, and Sonnhammer, EL, *Inparanoid: a comprehensive database of eukaryotic orthologs.* Nucleic Acids Res, 2005. **33**(Database issue): p. D476-480.

33. Breiman, L, *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.

34. Huynh-Thu, VA, Irrthum, A, Wehenkel, L, and Geurts, P, *Inferring regulatory networks from expression data using tree-based methods.* PLoS One, 2010. **5**(9).

35. Bottou, L. *Large-scale machine learning with stochastic gradient descent*. in *Proceedings of the 19th International Conference on Computational Statistics*. 2010: Springer:Paris.

36. Shalev-Shwartz, S and Tewari, A. *Stochastic methods for l 1 regularized loss minimization*. in *Proceedings of the 26th Annual International Conference on Machine Learning*. 2009: ACM:Montreal, Quebec, Canada.

37. Menendez, P, Kourmpetis, YA, ter Braak, CJ, and van Eeuwijk, FA, *Gene regulatory networks from multifactorial perturbations using Graphical Lasso: application to the DREAM4 challenge.* PLoS One, 2010. **5**(12): p. e14147.

38. Kanehisa, M, Goto, S, Kawashima, S, Okuno, Y, and Hattori, M, *The KEGG resource for deciphering the genome.* Nucleic Acids Res, 2004. **32 Database issue**: p. D277-280.

39. de Folter, S, Immink, RG, Kieffer, M, Parenicova, L, Henz, SR, Weigel, D, Busscher, M, Kooiker, M, Colombo, L, Kater, MM, Davies, B, and Angenent, GC, *Comprehensive interaction map of the Arabidopsis MADS Box transcription factors.* Plant Cell, 2005. **17**(5): p. 1424-1433.

40. Ding, X, Richter, T, Chen, M, Fujii, H, Seo, YS, Xie, M, Zheng, X, Kanrar, S, Stevenson, RA, Dardick, C, Li, Y, Jiang, H, Zhang, Y, Yu, F, Bartley, LE, Chern, M, Bart, R, Chen, X, Zhu, L, Farmerie, WG, Gribskov, M, Zhu, JK, Fromm, ME, Ronald, PC, and Song, WY, *A rice kinase-protein interaction map.* Plant physiology, 2009. **149**(3): p. 1478-1492.

41. Popescu, SC, Popescu, GV, Bachan, S, Zhang, Z, Gerstein, M, Snyder, M, and Dinesh-Kumar, SP, *MAPK target networks in Arabidopsis thaliana revealed using functional protein microarrays.* Genes Dev, 2009. **23**(1): p. 80-92.

42. Popescu, SC, Popescu, GV, Bachan, S, Zhang, Z, Seay, M, Gerstein, M, Snyder, M, and Dinesh-Kumar, SP, *Differential binding of calmodulin-related proteins to their targets revealed through high-density Arabidopsis protein microarrays.* Proc Natl Acad Sci U S A, 2007. **104**(11): p. 4730-4735.

43. Rohila, JS, Chen, M, Chen, S, Chen, J, Cerny, RL, Dardick, C, Canlas, P, Fujii, H, Gribskov, M, Kanrar, S, Knoflicek, L, Stevenson, B, Xie, M, Xu, X, Zheng, X, Zhu, JK, Ronald, P, and Fromm, ME, *Protein-protein interactions of tandem affinity purified protein kinases from rice.* PLoS One, 2009. **4**(8): p. e6685.

44. Gholami, AM and Fellenberg, K, *Cross-species common regulatory network inference without requirement for prior gene affiliation.* Bioinformatics, 2010. **26**(8): p. 1082-1090.

45. Mutwil, M, Obro, J, Willats, WG, and Persson, S, *GeneCAT--novel webtools that combine BLAST and co-expression analyses.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W320-326.

46. Yu, H, Luscombe, NM, Lu, HX, Zhu, X, Xia, Y, Han, JD, Bertin, N, Chung, S, Vidal, M, and Gerstein, M, *Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs.* Genome Res, 2004. **14**(6): p. 1107-1118.

47. Geisler-Lee, J, O'Toole, N, Ammar, R, Provart, NJ, Millar, AH, and Geisler, M, *A predicted interactome for Arabidopsis.* Plant Physiol, 2007. **145**(2): p. 317-329.

48. Bader, G, Betel, D, and Hogue, C, *BIND: the Biomolecular Interaction Network Database.* Nucleic Acids Res., 2002. **31**: p. 248.

49. Masoudi-Nejad, A, Goto, S, Endo, TR, and Kanehisa, M, *KEGG bioinformatics resource for plant genomics research.* Methods Mol Biol, 2007. **406**: p. 437-458.

50. Zhang, P, Foerster, H, Tissier, CP, Mueller, L, and Paley, S, *MetaCyc and AraCyc. Metabolic pathway databases for plant research.* Plant Physiol., 2005. **138**: p. 27.

51. Tatusov, RL, Galperin, MY, Natale, DA, and Koonin, EV, *The COG database: a tool for genome-scale analysis of protein functions and evolution.* Nucleic Acids Res, 2000. **28**(1): p. 33-36.

52. Zhang, J, *Evolution by gene duplication: an update.* TRENDS in Ecology and Evolution, 2003. **18**(6): p. 292-298.

53. Barrett, T, Troup, DB, Wilhite, SE, Ledoux, P, Rudnev, D, Evangelista, C, Kim, IF, Soboleva, A, Tomashevsky, M, and Edgar, R, *NCBI GEO: mining tens of millions of expression profiles--database and tools update.* Nucleic Acids Res, 2007. **35**(Database issue): p. D760-765.

54. Griffiths-Jones, S, Grocock, RJ, van Dongen, S, Bateman, A, and Enright, AJ, *miRBase: microRNA sequences, targets and gene nomenclature.* Nucleic Acids Res, 2006. **34**(Database issue): p. D140-144.

55. Gustafson, AM, Allen, E, Givan, S, Smith, D, Carrington, JC, and Kasschau, KD, *ASRP: the Arabidopsis Small RNA Project Database.* Nucleic Acids Res., 2005. **33**: p. D637.

56. Lu, S, Sun, YH, Shi, R, Clark, C, Li, L, and Chiang, VL, *Novel and mechanical stress-responsive microRNAs in Populus trichocarpa that are absent from Arabidopsis.* Plant Cell, 2005. **17**: p. 2186.

57. Wang, L, Xie, W, Chen, Y, Tang, W, Yang, J, Ye, R, Liu, L, Lin, Y, Xu, C, Xiao, J, and Zhang, Q, *A dynamic gene expression atlas covering the entire life cycle of rice.* Plant J, 2009.

58. Settles, B, *Active learning literature survey.* Computer Sciences Technical Report, 2010. **1648**.

59. King, RD, Rowland, J, Oliver, SG, Young, M, Aubrey, W, Byrne, E, Liakata, M, Markham, M, Pir, P, Soldatova, LN, Sparkes, A, Whelan, KE, and Clare, A, *The automation of science.* Science, 2009. **324**(5923): p. 85-89.

60. King, RD, Whelan, KE, Jones, FM, Reiser, PG, Bryant, CH, Muggleton, SH, Kell, DB, and Oliver, SG, *Functional genomic hypothesis generation and experimentation by a robot scientist.* Nature, 2004. **427**(6971): p. 247-252.

61. Lewis, D and Gale, W, *Training text classifiers by uncertainty sampling.* Research and Development in Information Retrival, 1994.

62. Settles, B and Craven, M. *An analysis of active learning strategies for sequence labeling tasks*. in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2008: ACL Press.

63. Settles, B, Craven, M, and Ray, S. *Multiple-instance active learning*. in *Advances in Neural Information Processing Systems (NIPS)*. 2008: MIT Press.

64. Cui, J, Li, P, Li, G, Xu, F, Zhao, C, Li, Y, Yang, Z, Wang, G, Yu, Q, and Shi, T, *AtPID: Arabidopsis thaliana protein interactome database--an integrative platform for plant systems biology.* Nucleic Acids Res, 2008. **36**(Database issue): p. D999-1008.

65. Davuluri, R, Sun, H, Palaniswamy, S, Matthews, N, Molina, C, Kurtz, M, and Grotewold, E, *AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors.* BMC Bioinformatics, 2003. **4**(1): p. 25.

66. Lu, C, Tej, SS, Luo, S, Haudenschild, CD, Meyers, BC, and Green, PJ, *Elucidation of the small RNA component of the transcriptome.* Science, 2005. **309**: p. 1525.

67. Mueller, LA, Zhang, P, and Rhee, SY, *AraCyc: a biochemical pathway database for Arabidopsis.* Plant Physiol, 2003. **132**(2): p. 453-460.

68. Rzhetsky, A, Iossifov, I, Koike, T, Krauthammer, M, Kra, P, Morris, M, Yu, H, Duboue, PA, Weng, W, Wilbur, WJ, Hatzivassiloglou, V, and Friedman, C, *GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data.* J Biomed Inform, 2004. **37**(1): p. 43-53.