**Summary of large scale identification and molecular characterization of phenotypes using state-space models**

A fundamental goal in computational biology is to characterize which and how genes give rise to phenotypes. At high-level, genes displaying similar phenotypes correspond to small regulatory sub-systems and thus are expected to exhibit similar expression trajectories. Therefore, we developed a state-of-the-art method to identify the mechanisms that govern gene activity by exploiting expression data using state-space models, dimensionality reduction techniques and biological networks. Our method focuses on a given groups of genes, and evaluates their expression dynamics patterns, thus uncovering the regulatory effects that govern them. This is done by decomposing the gene expression into *internal* contributions, from genes within the group, and *external* contributions, from genes outside the group. Specifically, we use a state space model to represent the temporal gene expression dynamics and identify principal temporal dynamic patterns. In order to handle limited time samples, we use dimensionality reduction approaches that will identify canonical temporal expression trajectories (e.g. degradation, growth, damped oscillation). In this way, we are able to untangle the regulatory effects from various contributors.

***Background on State-Space Models***

The state-space model has been widely used in engineering [28]. It models the dynamical system output as a function of both the current **internal** system state, and the **external** input signal. A commonly used example in engineering is the vehicle cruise control system where the internal system state is the vehicle’s speed. Based on the road conditions (external input signal), the cruise control will output the required fuel amounts in order to keep the desired speed. In biology state-space models have been used in the study of regulatory networks and in particular in the analysis of temporal gene expression data [29, 30, 31]. Compared to other methods that calculate the expression correlation between individual genes, the state-space model has the advantage that it predicts the temporal-causal relationships at the system level i.e., the state at a time is determined by the state and external input at the prior time point.

One of the early adopters, Wu et al (2004) used state-space equations to model the gene expression from microarray data. The authors describe the gene expression profiles as observation variables whose values are modelled by a linear combination of the current internal state variables (e.g. regulatory elements expression levels). Factor analysis was used to identify the internal state variables and calculate their expression values. The results suggest that it is possible to unambiguously determine a gene expression dynamic pattern from a limited time-course dataset. More recently Mar and Quackenbush (2009) [32] have used state-space models to study cell differentiation. They decompose the state-space gene expression trajectories into components one representing the changes inherent to the biological process (cellular phenotypic changes) and another component that captures the cell response to the perturbation (variation in gene expression levels).

However, these models have been able to account for only a limited number of genes. In fact, it is not feasible to use state space models for describing systems composed of thousands of genes due to the limited amount of data available, which would not allow us to learn all the required model parameters.

We combined state-space models with dimensionality reduction techniques in order to model the gene expression data on a genome-wide scale.

***Previous Experience in Networks***

The Gerstein lab has carried out projects in biological networks for over a decade. We have made extensive contributions in the analysis of genomic data using network frameworks [37]. In particular, we have integrated regulatory networks with gene expression to uncover different kinds of dynamic sub-networks [38]. We also developed methods to analyse the regulatory networks of a variety of species from yeast to human, using a wide range of data [39, 40, 41, 42]. In the following we give details on our results in network biology.

**Networks and Function.** Biological networks, normally large in scale, are organized with topological structures in the form of interacting modules. We have previously collaborated in developing various methods to identify the functional modules of biological networks. We developed a method to extract metabolic modules from metagenomic data, enabling the identification of pathways that are expressed under different environmental conditions [43]. We have also developed a way to identify nearly complete, fully connected modules (cliques) present in network interactions [44], and we have been using networks to map various kinds of functional genomics data [41]. For example, by mapping gene-expression data onto yeast regulatory network, we identified different sub-networks that are active in different conditions [38].

More recently we have developed a computational approach OrthoClust [45], to extract meaningful new information from gene expression data using biological networks. OrthoClust is a universal computational framework that integrates co-association networks of individual species using gene orthology relationships to enable the identification of functional modules formed by species-specific or conserved gene. Leveraging on the modENCODE RNA-seq data for *C. elegans* and *D. melanogaster* we used OrthoClust functional module predictions to infer putative functions of uncharacterized elements (e.g. non-coding RNAs) based on the guilt-by-association principle.

**Integrating Networks with Biological Data.** Acentral problem in computational biology is how to integrate different types of biological data in order to obtain more reliable predictions. We have developed several techniques in this area. One method, which is particularly relevant to this project, combined a widely heterogeneous set of biological networks, ranging from co-expression relationships to similar phylogenetic profiles, to predict genome-wide protein-protein interactions. Moreover, we explored the limits of genomic data integration, assessing the degree to which the predictive power increases with the addition of more features [46].

***Previous Experience in Phenotype Predictions***

Gerstein lab has developed a correlation-based method [33] that was able to discover genotype-phenotype associations combining phenotypic information from a biomedical informatics database, GIDEON, with the molecular information contained in COGs [34]. Evolutionary relationships between species are also used in methods that predict phenotypes by exploiting the concept of phenolog. The idea behind these methods is to associate a given gene with the phenotype which is most common among its orthologues [35].

Recently, much research has also been carried out to characterize and predict disease phenotypes. Inherited diseases that are phenotypically similar share disease-associated cellular components: they are linked by common molecular machinery whose normal functioning is somehow perturbed [36]. In other words, the disease modules of phenotypically similar diseases should be located closely on the interactome.

***References***

28. Li, Z., Shaw, S. M., Yedwabnick, M. J., & Chan, C. Using a state-space model with hidden variables to infer transcription factor activities. *Bioinformatics* **22**, 747-54 (2006).29. Wu, F. X., Poirier, G. G., & Zhang, W. J. Inferring gene regulatory networks with time delays using a genetic algorithm. *Syst Biol (Stevenage)* **152**, 67-74 (2005).30. Rangel, C., Angus, J., Ghahramani, Z., Lioumi, M., Sotheran, E., Gaiba, A., Wild, D. L., & Falciani, F. Modeling T-cell activation using gene expression profiling and state-space models. *Bioinformatics* **20**, 1361-72 (2004).31. Mar, J. C. & Quackenbush, J. Decomposition of gene expression state space trajectories. *PLoS Comput Biol* **5**, e1000626 (2009).32. Goh, C.-S., Gianoulis, T. A., Liu, Y., Li, J., Paccanaro, A., Lussier, Y. A., & Gerstein, M. Integration of curated databases to identify genotype-phenotype associations.. *BMC Genomics* **7**, 257 (2006).33. Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., & Natale, D. A. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).34. Robinson, P. N. & Webber, C. Phenotype ontologies and cross-species analysis for translational research. *PLoS Genet* **10**, e1004268 (2014).35. Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabási, A.-L. The human disease network. *Proc Natl Acad Sci U S A* **104**, 8685-90 (2007).36. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., & Gerstein, M. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449-53 (2003).37. Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., & Gerstein, M. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431**, 308-12 (2004).38. Cheng, C., Yan, K.-K., Hwang, W., Qian, J., Bhardwaj, N., Rozowsky, J., Lu, Z. J., Niu, W., Alves, P., Kato, M., Snyder, M., & Gerstein, M. Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS Comput Biol* **7**, e1002190 (2011).39. Yan, K.-K., Fang, G., Bhardwaj, N., Alexander, R. P., & Gerstein, M. Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks.. *Proc Natl Acad Sci U S A* **107**, 9186-9191 (2010).40. Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R., Min, R., Alves, P., Abyzov, A., Addleman, N., Bhardwaj, N., Boyle, A. P., Cayting, P., Charos, A., Chen, D. Z., Cheng, Y., Clarke, D., Eastman, C., Euskirchen, G., Frietze, S., Fu, Y., Gertz, J., Grubert, F., Harmanci, A., Jain, P., Kasowski, M., Lacroute, P., Leng, J., Lian, J., Monahan, H., O'Geen, H., Ouyang, Z., Partridge, E. C., Patacsil, D., Pauli, F., Raha, D., Ramirez, L., Reddy, T. E., Reed, B., Shi, M., Slifer, T., Wang, J., Wu, L., Yang, X., Yip, K. Y., Zilberman-Schapira, G., Batzoglou, S., Sidow, A., Farnham, P. J., Myers, R. M., Weissman, S. M., & Snyder, M. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100 (2012).41. Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., Yip, K. Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., Alves, P., Chateigner, A., Perry, M., Morris, M., Auerbach, R. K., Feng, X., Leng, J., Vielle, A., Niu, W., Rhrissorrakrai, K., Agarwal, A., Alexander, R. P., Barber, G., Brdlik, C. M., Brennan, J., Brouillet, J. J., Carr, A., Cheung, M.-S., Clawson, H., Contrino, S., Dannenberg, L. O., Dernburg, A. F., Desai, A., Dick, L., Dose, A. C., Du, J., Egelhofer, T., Ercan, S., Euskirchen, G., Ewing, B., Feingold, E. A., Gassmann, R., Good, P. J., Green, P., Gullier, F., Gutwein, M., Guyer, M. S., Habegger, L., Han, T., Henikoff, J. G., Henz, S. R., Hinrichs, A., Holster, H., Hyman, T., Iniguez, A. L., Janette, J., Jensen, M., Kato, M., Kent, W. J., Kephart, E., Khivansara, V., Khurana, E., Kim, J. K., Kolasinska-Zwierz, P., Lai, E. C., Latorre, I., Leahey, A., Lewis, S., Lloyd, P., Lochovsky, L., Lowdon, R. F., Lubling, Y., Lyne, R., MacCoss, M., Mackowiak, S. D., Mangone, M., McKay, S., Mecenas, D., Merrihew, G., Miller, D. M. r., Muroyama, A., Murray, J. I., Ooi, S.-L., Pham, H., Phippen, T., Preston, E. A., Rajewsky, N., Ratsch, G., Rosenbaum, H., Rozowsky, J., Rutherford, K., Ruzanov, P., Sarov, M., Sasidharan, R., Sboner, A., Scheid, P., Segal, E., Shin, H., Shou, C., Slack, F. J., Slightam, C., Smith, R., Spencer, W. C., Stinson, E. O., Taing, S., Takasaki, T., Vafeados, D., Voronina, K., Wang, G., Washington, N. L., Whittle, C. M., Wu, B., Yan, K.-K., Zeller, G., Zha, Z., Zhong, M., Zhou, X., Ahringer, J., Strome, S., Gunsalus, K. C., Micklem, G., Liu, X. S., Reinke, V., Kim, S. K., Hillier, L. W., Henikoff, S., Piano, F., Snyder, M., Stein, L., Lieb, J. D., & Waterston, R. H. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.. *Science* **330**, 1775-1787 (2010).42. Gianoulis, T. A., Raes, J., Patel, P. V., Bjornson, R., Korbel, J. O., Letunic, I., Yamada, T., Paccanaro, A., Jensen, L. J., Snyder, M., Bork, P., & Gerstein, M. B. Quantifying environmental adaptation of metabolic pathways in metagenomics.. *Proc Natl Acad Sci U S A* **106**, 1374-1379 (2009).43. Yu, H., Paccanaro, A., Trifonov, V., & Gerstein, M. Predicting interactions in protein networks by completing defective cliques.. *Bioinformatics* **22**, 823-829 (2006).44. Yan, K.-K., Wang, D., Rozowsky, J., Zheng, H., Cheng, C., & Gerstein, M. OrthoClust: an orthology-based network framework for clustering data across multiple species. *Genome Biol* **15**, R100 (2014).45. Lu, L. J., Xia, Y., Paccanaro, A., Yu, H., & Gerstein, M. Assessing the limits of genomic data integration for predicting protein networks.. *Genome Res* **15**, 945-953 (2005).46. Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrín-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H. Y., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A., & Greenblatt, J. F. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**, 637-43 (2006).