31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Nonlinear Bionetwork Structure Inference Using the Random Sampling-High Dimensional Model Representation (RS-HDMR) Algorithm

Miles Miller[1], Xiaojiang Feng[2], Genyuan Li[2], and Herschel Rabitz[2]*

Abstract— **This work presents the random sampling - high dimensional model representation (RS-HDMR) algorithm for identifying complex bionetwork structures from multivariate data. RS-HDMR describes network interactions through a hierarchy of input-output (IO) functions of increasing dimensionality. Sensitivity analysis based on the calculated RS-HDMR component functions provides a statistically interpretable measure of network interaction strength, and can be used to efficiently infer network structure. Advantages of RS-HDMR include the ability to capture nonlinear and cooperative realtionships among network components, the ability to handle both continuous and discrete relationships, the ability to be used as a high-dimensional IO model for quantitative property prediction, and favorable scalability with respect to the number of variables. To demonstrate, RS-HDMR was applied to experimental data measuring the single-cell response of a protein-protein signaling network to various perturbations. The resultant analysis identified the network structure comparable to that reported in the literature and to the results from a previous Bayesian network (BN) analysis. The IO model also revealed several nonlinear feedback and cooperative mechanisms that were unidentified through BN analysis.**

I. Introduction

The development of myriad high-throughput biological measurement techniques has led to the availability of increasingly rich datasets describing the behavior of underlying biological networks. Experimental methods ranging from particle-based and multiplex flow cytometric assays to high-throughput kinase activity assays [1] not only allow for the simultaneous observation of multiple (> 10) network species, but are of high enough resolution to capture complex interactions characteristic of many biological systems. Appropriately designed computational methods are necessary to reliably identify the bionetwork structure from such methods and the correspondingly rich datasets they produce.

Several advances have been made in the development of network inference protocols for analyzing multivariate datasets taken from high-throughput biological measurements. Network identification algorithms based on linearized steady-state models and regression analysis have been developed to identify bionetwork connectivity [2] . Although effective in some situations, particularly in conditions of sparse sampling and noisy data, such algorithms discount nonlinear interactions which can become significant in understanding complex biological networks. Bayesian networks (BNs) [3], clustering algorithms [4], and information-theoretic approaches [5] have been employed to capture both linear and nonlinear relationships. However, probabilistic methods like BN analysis can scale poorly with network complexity, especially when higher-order interactions are considered. As a result, network connections are often simplified as discretized functions. Information loss through discretization amplifies, however, as high-throughput measurement technologies improve and quantitative measurements become more precise.

This article introduces the random sampling – high dimensional model representation (RS-HDMR) algorithm for a quantitative, predictive characterization of nonlinear bionetwork interactions directly from laboratory data [6], [7]. RS-HDMR decomposes interactions among network species into a hierarchy of (usually nonlinear) continuous input-output (IO) component functions, describing both independent and coopeartive interactons among the network components. These component functions can then be utilized to (1) effectively describe the network structure as well as (2) provide quantitative understanding of the network behavior under previously unsampled conditions.

The RS-HDMR algorithm has previously been applied to a broad range of input-output modeling problems [6]; it is employed in this work to quantitatively capture the complex interactions among protein species in the human T-cell signaling network. A map of network structure was generated based on the relative strength of decomposed RS-HDMR functions. Results were then compared to those obtained previously through BN analysis and to descriptions of the signaling network by previous literature [8]. In this illustration, the ability of RS-HDMR to quantitatively capture nonlinear and higher-order relationships was shown to be a significant factor in characterizing several of the target protein network's feedback and cooperative mechanisms.

II. The RS-HDMR Algorithm

RS-HDMR is a tool to deduce nonlinear and cooperative interactions among a set of inputs and outputs. The independent and cooperative effects of n input variables $\mathbf{x} = (x_1, x_2, ...x_n)$ on an output, $y = f(\mathbf{x})$, can be described in terms of a hierarchy of RS-HDMR component functions [7]:

[1]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 (milesm@mit.edu)
[2]Department of Chemistry, Princeton University, Princeton, NJ 08544 (xfeng@princeton.edu, genyuan@princeton.edu, hrabitz@princeton.edu)

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{1 \le i < j \le n} f_{ij}(x_i, x_j) + \dots \quad (1)$$

Here f_0 represents the mean value of $f(\mathbf{x})$ over the sample space, the first-order component function $f_i(x_i)$ describes the generally nonlinear independent contribution of the i^{th} input variable to the output, the second-order component function $f_{ij}(x_i, x_j)$ describes the pairwise cooperative contribution of x_i and x_j, and further terms describe higher order cooperative contributions. The RS-HDMR component functions can be represented and determined in a variety of ways. In this article, they are approximated as weighted orthonormal basis functions in order to reduce the sampling effort and take the following form:

$$f_i(x_i) \approx \sum_{r=1}^{k} \alpha_r^i \varphi_r^i(x_i) \quad (2)$$

$$f_{ij}(x_i, x_j) \approx \sum_{p=1}^{l} \sum_{q=1}^{l'} \beta_{pq}^{ij} \varphi_p^i(x_i) \varphi_q^j(x_j) \quad (3)$$

Where k,l, and l' are integers (generally ≤ 3 for most applications), α_r^i and β_{pq}^{ij} are constant weighting coefficients to be determined, and the basis functions $\{\varphi\}$ are optimized from the distribution of sample datapoints to follow conditions of orthonormality [6]. Basis functions are approximated in this work as non-linear polynomials, where

$$\varphi_1^i(x_i) = a_1 x_i + a_0 \quad (4)$$

$$\varphi_2^i(x_i) = b_2 x_i^2 + b_1 x_i + b_0 \quad (5)$$

$$\varphi_3^i(x_i) = c_3 x_i^3 + c_2 x_i^2 + c_1 x_i + c_0 \quad (6)$$

The coefficients a_0,a_1,b_0,...c_3 are calculated through monte carlo integration (hence the term "Random Sampling" HDMR) under constraints of orthonormality, such that when integrated over all datapoints,

$$\int \varphi_r(x) dx \approx 0 \ \ \forall\, r \quad (7)$$

$$\int \varphi_r^2(x) dx \approx 1 \ \ \forall\, r \quad (8)$$

$$\int \varphi_p(x) \varphi_q(x) dx \approx 0 \ \ (p \neq q) \quad (9)$$

In effect, the orthonormality of the basis functions is optimized for a given set of observed data. Optimal basis functions are then weighted by coefficients (α_r^i and β_{pq}^{ij} for first and second order component functions, respectively), which are calculated from least-squares regression. Biased inference and overfitting is often a problem when analyzing noisy, sparsely sampled, highly correlated data. To address this, only inputs and their respective component functions validated as significant by the statistical *F*-test were included in RS-HDMR expansions as a method of variable selection. The resultant expansion in Eq. (1) serves both as a predictive model of network response due to its input variable interactions and as a statistical representation of the underlying biological system.

The decomposition in Eq. (1) applies to both continuous and discrete data sets. In most cases, the zeroth, first, and second order functions are sufficient to describe the high-dimensional IO relationships of physically realistic systems [6]. When all the component functions are identified from laboratory data, they can be directly used as an equivalent IO model to describe the quantitative network behavior.

In network structure inference, the relative strength of interactions among network components can be quantitatively determined through a global sensitivity analysis based on the respective RS-HDMR component functions. Here the total variance σ^2 of an output $f(x)$ is decomposed into hierarchical contributions from the individual RS-HDMR component functions:

$$\sigma^2 = \int [f(\mathbf{x}) - f_0]^2 \, dx = \sum_{i=1}^{n} \sigma_i^2 + \sum_{1 \le i < j \le n} \sigma_{ij}^2 + \dots \quad (10)$$

The sensitivity indices, $S_l (l = 1, 2, ..., n_p)$, are then defined as the portion of the total variance σ^2 represented by the variance of the l^{th} component function. In addition, the sensitivity indices S_l describing first, second (and higher, if warranted) order component functions of an input variable x_i can then be summed into a measure $S_i^T (i = 1, 2...n)$, describing both independent and higher-order effects of x_i on the output. Under the assumption that relatively more significant and direct network interactions are described by high sensitivity indices (an assumption similar to the Markov condition used in Bayesian analysis), the probability for the existence of a network connection can be ranked by its corresponding S_i^T value.

In many bionetworks, such as cyclic systems, a network component can act as both an input (upstream species) and an output (downstream species). Similar to Bayesian network analysis, RS-HDMR can be used to infer causal interactions using time-series data. However, this work focuses on applying RS-HDMR to biological network inference with no explicit causal relationships. In this case, a separate RS-HDMR IO expansion is formulated using each measured species as the output, $f(\mathbf{x})$, and the remaining species as inputs. As a result, n RS-HDMR IO mappings are determined for a system of n network species. The agglomeration of the RS-HDMR expansions then constitutes a complete predictive model of network behavior.

III. Application to a Cell-Signaling Network

RS-HDMR network inference was applied to infer the structure of a well studied cell signaling network. Data used in this work was taken from high-dimensional cytometry measurements of human primary naive CD4+ T-cells, where individual cells observed in a given cell population describe network behavior under statistically sampled microenviroments [8]. Observed proteins and phoshpolipids include protein kinase C (PKC), Raf, mitogen-activated protein kinases

(MAPKs) Erk1 and Erk2, p38 MAPK, Jnk, AKT, Mek1 and Mek2, protein kinase A (PKA) substrates, phospholipase C-γ (PLC-γ), phosphatidylinositol 4,5-bisphosphate (PIP2), and phosphatidylinositol 3,4,5-triphosphate (PIP3). Nine datasets were first analyzed individually using RS-HDMR, where each dataset corresponds to a different perturbative experimental condition. 11 RS-HDMR IO mappings were computed from each dataset to identify all significant component functions relating the 11 network species, with each function using a measured species as the dependent variable (the output, $f(\mathbf{x})$) and the remaining ten species as the input variables. Each RS-HDMR mapping (99 total for this application) then provides a quantitative description of the relationships between the output variable and its respective inputs.

In addition to the above analysis, 13 sets of laboratory data from activation or inhibition of specific protein species were also paired with data taken from general stimulatory conditions (the control) in order to examine the population-wide effects of exogenous perturbative conditions. Because specific perturbations were not quantified in cytometry measurements, the measured levels of the perturbed species were represented discretely as *high* ($f(\mathbf{x}) = 1$) or *low* ($f(\mathbf{x}) = 0$). RS-HDMR analysis was then conducted on these discrete data sets, using the perturbed species as the output.

Sensitivity analysis results from the 99 RS-HDMR IO mappings from individual datasets and the 13 RS-HDMR IO mappings from the pairwise experiments were aggregated, and the sensitivity measures S_i^T describing network connections were ranked by magnitude. A threshold sensitivity value S_{min} was defined empirically as the lowest value whose corresponding input-output relationship can be considered as a real connection. Fig. 1 shows the 21 high-confidence connections identified using this theshold. Of the 21 connections, all were described to some extent through previous empirical studies in a variety of systems. Three lower confidence ($S_l^T < S_{min}$) network connections were also identified which correspond with three well known connections reported previously (PKA/p38, PKA/Raf, PKC/PIP2). RS-HDMR analysis successfully identified all of the connections revealed through previous BN analysis results [8] using the same data sets, as well as two additional connections (PIP3/Akt and PKC/Plcγ) well-established in the literature but not identified by BN analysis.

In addition to the first-order connections that relate two network components, three statistically significant second-order connections were also revealed by RS-HDMR analysis: (PLCγ, PIP2, PIP3), (PKA, Akt, Erk), and (PKC, Jnk, p38). The most significant one is the connection among PIP2, PIP3, and PLCγ. These three proteins are unique from other measured species in that they have significant negative feedback interaction. Activated PLCγ catalyzes the destructive cleavage of PIP2. The product of PIP2 phosphorylation, PIP3, serves as a docking site for PLCγ and ultimately catalyzes PLCγ phosphorylation and activation. In several RS-HDMR expansions, this feedback interaction was characterized through second-order RS-HDMR component functions.

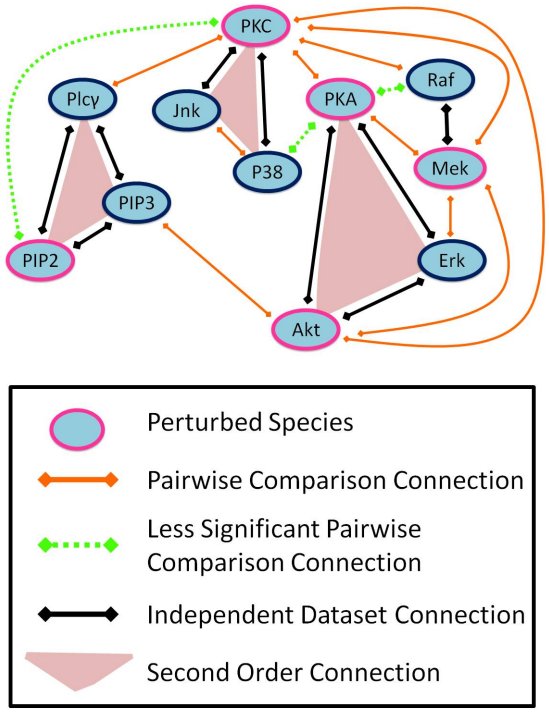

Fig. 1. **HDMR Identified Significant Network Connections.** Significant network interactions ($S_{min} = 0.15$) from individual and pairwise RS-HDMR analysis are graphically represented. Graph edges represent network interactions with total sensitivity measures S_i^T, which account for total first, second, and third order interactions, that are above the threshold S_{min}. Orange lines represent connections identified only through pairwise comparison. Dashed green lines indicate connections well defined in previous literature, but identified to a less significant degree by RS-HDMR analysis ($S_i^T < S_{min}$). Causality is not inferred by RS-HDMR analysis in this application, thus connections are not directional.

Fig. 2 shows two significant first-order component functions and a significant second-order function with Akt as the output. Inspection of these component functions can provide meaningful physical interpretation. For example, the functional dependence of Akt on Erk is monotonically increasing, nearly linear, and consistent across several experimental conditions; this suggests a relatively direct network interaction. However, PKA is neither monotonic nor consistent across experimental conditions at low levels of PKA. Thus while PKA may play a more consistent (and direct) role in positively affecting Akt at high levels, its relationship to Akt at lower levels of PKA may be considered less significant, or possibly more indirect.

Fig. 2 also describes the second-order function with the highest sensitivity index of the nine RS-HDMR expansions with Akt as the output, capturing the cooperative influence of PKA and Erk on Akt. Evidence in the literature supports the presence of complex feedback and cooperative interactions among Erk, Akt, and PKA [9]. Akt may interact with Erk through the Raf/Mek/Erk pathway and with PKA independently of Erk through a CaMKK-mediated pathway. However, PKA has been reported to negatively regulate

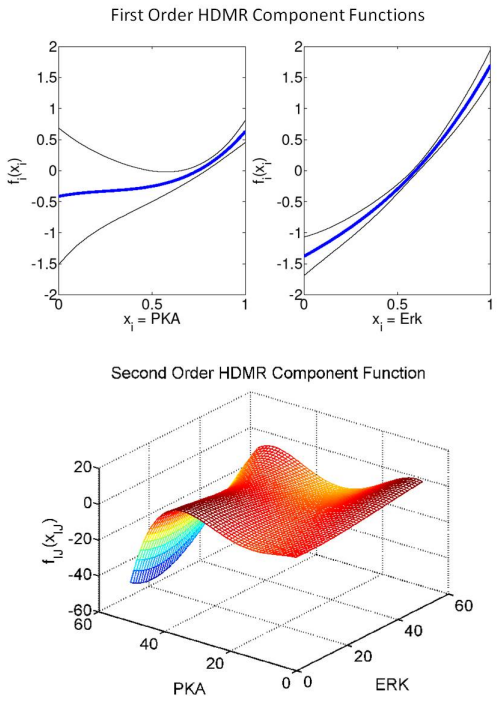

Fig. 2. **First and second order RS-HDMR Component Functions Describing Akt.** First-order RS-HDMR component functions describing interactions between Akt and other network species were averaged over corresponding RS-HDMR functions describing the same network connections under all experimental conditions. Of the ten total first-order component functions, two were determined by F-test to be significant. The first and second most significant functions correspond with the input species Erk and PKA, respectively. The thick blue line describes the mean function, and thin black lines are one standard deviation above and below the mean function. The most significant second order RS-HDMR-component function of the nine RS-HDMR expansions describes the cooperative relationship between Akt, Erk and PKA.

Erk activity by phosphorylating Raf [10]. This cooperative interaction may be one physical explanation for this significant second-order component function observed through RS-HDMR analysis.

The collection of RS-HDMR component functions can be directly used as a quantitative IO model to predict how the network behavior is affected by variations in one or more network components. Importantly, this IO model is obtained without any detailed mechanistic information of the network, hence it can be a useful tool for understanding network behavior and guiding network engineering when construction of mechanistic models is impractical.

IV. Discussion and Conclusion

The RS-HDMR algorithm was empolyed in this work to quantitatively characterize the structure of a cell signaling network, as well as to construct a quantitative IO model of the network behavior. The RS-HDMR predictions are obtained without any mechanistic modeling of the network; information is extracted directly from both continuously and discretely distributed high-dimensional laboratory data. Characterization of nonlinear IO relationships is made computationally manageable without losing significant information by approximating the interactions through a hierarchy of orthonormal basis functions. The restricted approximation of RS-HDMR component functions as orthonormal basis functions allows for a clear physical/statistical interpretation while maintaining robustness to outlier datapoints. The higher-order interactions can be a significant feature of complex bionetworks. In this work, incorporation of such cooperative IO functions significantly improved the quantitative predictive and data-fitting accuracy of RS-HDMR IO mapping, revealing network interactions unobserved through both first-order RS-HDMR and BN analysis. Lastly, RS-HDMR calculation scales well with respect to the number of network species [7] (in comparison, BN optimization often requires heuristic approaches), making it a favorable method for studying complex bionetworks.

V. Acknowledgments

The authors acknowledge funding from NSF, EPA, and DOE.

References

[1] K.A. Janes, J.G. Albeck, L.X. Pend, P.K. Sorger, D.A. Lauffenburger, and M.B. Yaffe, A High-throughput quantitative multiplex kinase assay for monitoring information flow in signaling networks application to sepsis-apoptosis, *Molecular & Cellular Proteomics*, vol. 2, 2003, pp 463-473.

[2] D. di Bernardo, M.J. Thompson, T.S. Gardner, S.E. Chobot, E.L. Eastwood, A.P. Wojtovich, S.J. Elliott, S.E. Schaus, and J.J. Collins, Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks, *Nat. Biotech.*, vol. 23, 2005, 377-383.

[3] P.J. Woolf, W. Prudhomme, L. Daheron, G.Q. Daley, and D.A. Lauffenburger, Bayesian analysis of signaling networks governing embryonic stem cell fate decisions, *Bioinformatics*, vol. 21, 2005, pp 741-753.

[4] P. D'haeseleer, S. Liang, and R. Somogyi, Genetic network inference: from co-expression clustering to reverse engineering, *Bioinformatics*, vol. 16, 2000, pp 707-726.

[5] A.A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, F. Dalla, and A. Califano, ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context, *BMC Bioinformatics*, vol. 7, 2006, pp S1-S7.

[6] G. Li, J. Hu, S.-W. Wang, P.G. Georgopoulos, J. Schoendorf, and H. Rabitz, Random sampling-high dimensional model representation (RS-HDMR) and orthogonality of its different order component functions, *J. Phys. Chem. A*, vol. 110, 2006, pp 2474-2485.

[7] G. Li, C. Rosenthal, and H. Rabitz, High dimensional model representations, *J. Phys. Chem. A*, vol. 105, 2001, pp 7765-7777.

[8] K. Sachs, O. Perez, D. Pe'er, D.A. Lauffenburger, and G.P. Nolan, Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data, *Science*, vol. 308, 2005, pp 523-529.

[9] S. Greco, C. Storelli, and S. Marsigliante, Protein kinase C (PKC)-δ/-ε mediate the PKC/Akt-dependent phosphorylation of extracellular signal-regulated kinases 1 and 2 in MCF-7 cells stimulated by bradykinin, *J. Endocrinology*, vol. 188, 2006, pp 79-89.

[10] B.M.T. Burgering and J.L. Bos, Regulation of Ras-mediated signalling: more than one way to skin a cat, *Trends Biochem. Sci.*, vol. 20, 1995, pp 18-22.