

Multiscale factor models for molecular networks

Justin Guinney^{1,2}, Philip Febbo^{1,3,4}, Mauro Maggioni^{5,6}, and Sayan Mukherjee^{5,6,7}
Institute for Genome Sciences & Policy¹, Department of Medicine²,
Department of Molecular Genetics and Microbiology³, Department of Mathematics⁴,
Department of Computer Science⁵, Department of Statistical Science⁶,
Duke University Durham, NC 27708 U.S.A.

A factor modeling framework is developed that is both predictive of phenotypic or response variation and the inferred factors offer insight with respect to underlying physical or biological processes. The method is general and can be applied to a variety of scientific problems. We focus on modeling complex disease phenotypes (etiology of cancer) as a motivating example. In this setting, the factors capture gene or protein interaction networks at different scales – breadth of the interaction network. The method integrates multiscale analysis on graphs and manifolds developed in applied harmonic analysis with sparse factor models, a mainstay of applied statistics. Specific findings include the association of the TGF- β pathway with prostate cancer recurrence mediated by cell-cycle control and the implication of the p27 pathway in cancer progression. In silico perturbation analyses of the inferred multiscale model suggest that the TGF- β pathway is a dominant pathway in control of cell-cycle deregulation in prostate cancer.

Key Words: diffusion geometry, sparse regression, molecular networks, factor models

1 Introduction

Methods for analyzing high-dimensional data have received much attention in the last decade, driven by increasingly high-throughput techniques for data generation in the social, physical, and biological sciences. An important challenge in all these settings is the inference of models that are both predictive and interpretable – offering insight into the underlying biological, social, or physical phenomena. This can often be restated as understanding the structure and statistical dependencies of variables relevant to prediction of a response, category, or phenotype given high-dimensional data. One modeling principle common across biological and physical sciences is the idea of scale – the phenomena under study is composed of interactions or processes that vary depending on the level of resolution at which they are examined. The other modeling principle is that a sparse or low-dimensional representation captures relevant information of the high-dimensional data – factor modeling and manifold learning are two such examples. We couple these two principles in the framework of multiscale factor models to produce a low-dimensional representation comprised of factors at various scales.

The scientific problem of modeling complex phenotypes or traits serves as a motivating example to highlight the efficacy of this approach. However, the method itself is general and can be applied to a variety of scientific and engineering applications. We consider complex phenotypes as those controlled by many genes and gene products with complex interactions. A common property of complex phenotypes is heterogeneity of both the phenotype and its genetic and molecular basis. Cancer is a complex phenotype where the heterogeneity is derived from two main sources: variability across time or stage of disease and genetic and environmental variability across individuals. The idea of scale is central in oncogenesis since the set of steps by which interactions of genetic, biochemical, and cellular mechanisms with environmental factors driving tumor development vary in complexity of the underlying networks as well as the timescale of the interactions. In this paper scale will refer to the granularity or specificity of molecular and cellular interactions, and spans the range from physical binding of proteins to other molecules, to loosely coupled interactions of molecular pathways and networks.

We address the biological problem of mapping the genetic or expression variation giving rise to phenotypic variation onto sparse multiscale subsets of a putative direct gene (product) interaction network. We formalize this problem by denoting the genes and gene products as a set of nodes \mathcal{V} in a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ where the edges between gene products quantify the direct interaction. This graph can be represented as an association matrix W where the elements W_{ij} encode dependence between two nodes. Gene expression or genetic variation is assayed on the nodes or gene products in the interaction network. A set of n observations of the expression measurements and phenotypes is denoted as $\{(X_i, Y_i)\}_{i=1}^n$ where Y is the phenotype and $X \in \mathbb{R}^m$ are the measurements over the m gene products in the graph. The *multiscale* factor framework specifies the following model for phenotypic variation

$$Y_i = \sum_{l=1}^p \alpha_l \phi_l(X_i) + \varepsilon_i, \quad (1)$$

where ε_i characterizes the noise, $\phi_l(X_i) \equiv \langle X_i, \phi_l \rangle$ is the projection of the observation X_i onto the multiscale factor ϕ_l , and α_l models the relevance of factor l in predicting phenotypic variation. The multiscale bases or factors ϕ_l are constructed from the association matrix W . The index l has two components: $l \equiv (j, k)$, where j parametrizes the scale of the factor – related to number of nodes comprising the factor – and k indexes factors at each scale j . At the finest scale the factors are single genes, and at coarser scales (j increasing) they are linear combinations of highly independent genes. At the coarsest scale these factors correspond to eigengenes or metagenes used in singular value decomposition analysis [8] and sparse factor modeling [2], respectively.

The main innovation of our approach is that inferred factors can be interpreted as subnetworks at various scales of molecular and cellular processes relevant to explaining phenotypic variation.

2 Sparse multiscale factor models

The *multiscale* factor model (1) is specified as

$$Y_i = \sum_{l=1}^p \alpha_l \phi_l(X_i) + \varepsilon_i = \sum_{j=1}^J \sum_{k=0}^{K_j} \alpha_{j,k} \phi_{j,k}(X_i) + \varepsilon_i, \quad i = 1, \dots, n$$

where i indexes the observations and l is a double index. In the second expression scale is made explicit by decomposing l into a scale index j and indexing factors at scale j by index k . *Sparsity* implies that only a few factors are required to explain variation in the response and is helpful in interpreting the inferred model. We use the following generalization of the Lasso estimator [1] to infer a sparse model

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{R}^p} \left[\sum_{i=1}^n |y_i - \sum_{l=1}^p \alpha_l \phi_l(x_i)|^2 + \lambda \sum_{l=1}^p |\alpha_l| \right]. \quad (2)$$

The regularization parameter λ controls the trade-off between fitting the data and sparsity of the solution and is estimated using bootstrapping, see Materials and methods. In the classical Lasso formulation the features are the m coordinates of the data, $\phi_l(x) = \langle x, e_l \rangle$, where e_l is the l -th coordinate basis.

The method of diffusion wavelets is used to construct the multiscale factors, $\{\phi_1, \dots, \dots, \phi_p\}$. A putative interaction graph of the variables (e.g. genes) is either given or inferred, vertices correspond to variables and the edges represent dependence between variables. This graph is represented by an association matrix W with element W_{ij} encoding the dependence between the i -th and j -th variables. This matrix may be given a priori knowledge such as a protein-protein interaction network or defined by local interactions in the data, for example $W_\sigma(x_i, x_j) \equiv \exp(-\|x_i - x_j\|^2/\sigma)$ with $\sigma > 0$ [15, 12]. Given the association matrix the diffusion operator [12] on the graph is defined as

$$T \equiv D^{-1/2} W D^{-1/2}, \quad \text{with } D_{ii} = \sum_j W_{ij}.$$

This operator is related to the graph Laplacian L [15, 12], $L = I - T$. In the case of a Gaussian graphical model, T corresponds to the partial correlation matrix [6, 5]. In manifold learning, the eigenfunctions of this operator are used as global basis factors to capture information on the geometry and local interactions underlying the data. This can be thought of as a nonlinear version of principal components analysis (PCA), the nonlinearity is a function of the eigenfunctions of T [35]. A drawback of this approach is that the eigenfunctions tend to be global, a linear combination of all variables, and hence difficult to interpret.

In biological applications where interpretability may be as important as prediction performance factors with a few genes may be preferred even if they have lower predictive accuracy than these eigenfunctions or eigengenes. In most applications the number of variables is far greater than the number of observations $m \gg n$ resulting in a very large number of few gene factors that are equally predictive, complicating the biological interpretability of factors with few genes. This ambiguity arises from the complex relationships between genes and we address it by constructing a multiscale family of factors, from single genes to eigengenes. The sparsity constraint in the regression model is then used to select predictive factors. Typically we obtain significant factors at different scales.

In order to generate the different scales, we observe that T only encodes local relationships between the variables, the partial correlation matrix is sparse in the terminology of graphical models. Powers of T integrate or propagate local dependencies to more global dependencies. In a graphical model $T^t(i, j)$ corresponds to integrating the partial correlation across all paths of length t from variable i to variable j . To recover the dependencies across all the variables we sum across all path-lengths t

$$(I - T)^{-1} = \sum_{t=1}^{+\infty} T^t,$$

which converges in the complement of the eigenspace of T corresponding to the eigenvalue 1. Similar expansions were used in path analysis [3, 4] to decompose genetic and phenotypic variation and in graphical models [6, 5] for both computation and inference. This expansion can be represented in a truly multiscale fashion as a product of multiscale models by the limit of the following expansion ($J \rightarrow +\infty$)

$$\sum_{t=1}^{2^J} T^t = \left(\sum_{t=1}^{2^{J-1}} T^t \right) (I + T^{2^{J-1}}) = \prod_{t=0}^J (I + T^{2^t}). \quad (3)$$

The factors at scale j , $\{\phi_{j,k}\}_{k=1}^{K_j}$, are constructed by analyzing $T^{2^{j-1}}$ and constructing a basis of sparse vectors spanning its range, which is low-dimensional. $T^{2^{j-1}}$ can be represented by a small matrix acting on the range of $T^{2^{j-2}}$. This decomposition is the key idea in diffusion wavelets, which given the graph \mathcal{G} yields a set of multiscale features or bases $\{\{\phi_{j,k}\}_{k=0}^{K_j}\}_{j=1}^J$. The bases $\phi_{j,k}$ at different scales are related to each other: each $\phi_{j,k}$ is a linear combination of a small number of $\phi_{j-1,k}$'s at the previous finer scale, yielding a hierarchical structure among these factors. As j grows or the

factors become more global and approximate the top eigenvectors of T . For more details on diffusion wavelets see [13].

Given p factors and n observations where $p \gg n$ we would like a sparse set of factors to capture variation in the phenotype in equation (1). The sparsity of the model has a strong dependence on the set of factors used. A sparse representation given one set of factors may be dense with respect to another set. For example, sparsity with respect to a few single gene factors is very different than sparsity with respect to principal components or eigengenes. The multiscale factors capture both extremes as well as scales in between, providing a rich set of bases.

The advantages of the multiscale representations with respect to sparsity and interpretability are due to three aspects of the method. A priori or data adapted knowledge of dependence between the variables is embedded in the bases by the association matrix W . The multiscale bases interpolate between models comprised of single genes and models comprised of global factors models based on the spectral decompositions of the diffusion operator T , (kernel) PCA sparse regression. There is a hierarchical relation between scales, each basis at a coarser scale is a sparse combination of bases at a finer scale

2.1 Comparison with other decomposition and clustering methods

In this section we develop the relation between other matrix factorization approaches and the multiscale factors and relate the multiscale factors with hierarchical clustering.

An important step in many analyses is that of modeling or factorizing the data matrix $X \in \mathbb{R}^{m \times n}$ with the goal of a reduced or compressed representation. The most common representation or factorization is principal component analysis (PCA). The data matrix is represented by the following eigendecomposition $X = U\Sigma V^T$, where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal and $\Sigma \in \mathbb{R}^{m \times n}$ is diagonal with entries $\sigma_1 \geq \dots \sigma_n \geq 0$. The best k rank approximation of X consist of the first k columns of U and V , $X_k = U_k \Sigma_k V_k^T$. When the data matrix is low rank this representation is natural. However interpretation of this representation is challenging since the columns of U and V are typically vectors with all entries non-zero. In this case the factors are the columns of U which corresponds to a mixture of positive and negative coefficients for all the genes, all the factors are global.

An alternative decomposition is a CUR-type decomposition [14] where the data matrix X takes the form $X \approx CUR$, where $C \in \mathbb{R}^{m \times k}$, $U \in \mathbb{R}^{k \times k}$, $R \in \mathbb{R}^{k \times n}$. C and R are subsets of k columns and rows of the data matrix, respectively. U , notwithstanding the name, is not unitary in general. The interpretation of this decomposition is picking k genes and data points from the data for the matrices C and R and the constructing U to well approximate X . U is in general a full matrix and is complicated to interpret. A potential disadvantage of CUR decompositions is they may be poor rank k approximations of the data matrix as compared to PCA. This can be obviated by careful and algorithmically efficient choices of C, U, R , as shown in [14].

In light of these approaches, we interpret the multiscale approach as “interpolating” between the “full” features of PCA and the single features of CUR , by creating a multiscale dictionary of features representing the data. These features are nonlinear functions on the data, capturing nonlinear subspaces characterizing the data. While in

this paper we focus on the predictive aspects of the model, our construction may be used to construct a multiscale representation of the rows and the columns of the data matrix X with dictionaries $\Phi^{\text{rows}} \equiv \{\phi_{j,k}^{\text{rows}}\}$ and $\Phi^{\text{cols}} \equiv \{\phi_{j,k}^{\text{cols}}\}$, respectively. One representation of the data matrix is $X = \Phi_J X \Phi_J^T + \Phi_{J-1} (X - \Phi_J X \Phi_J^T) \Phi_{J-1}^T + \dots$, the data matrix is a superposition of coarse-to-fine approximations.

A related task to factorizing the data is hierarchically grouping the data. Arguably one of the most popular methods used for the unsupervised analysis and identification of biologically meaningful clusters is agglomerative hierarchical clustering [7]. A problem with hierarchical clustering is its sensitivity to the choice of linkage function and a lack of coherent methodology to select link functions. This problem is extenuated when one agglomerates clusters at larger scales, since the decision to agglomerate is based on very basic cluster statistics such as correlation of cluster centers that may not be robust. Another approach to hierarchical grouping is based on recursive partitioning using spectral methods [11, 10]. Unfortunately, these “top-down” approaches do not seem to perform well on complex biological networks which are often characterized by slow spectral decays. While useful for global partitioning of data, spectral methods can magnify errors made early on, making it difficult to infer local partitions of the data. The same problem is inherent in factor models that use principal components to generate factors. We will show that the multiscale factors provide robust hierarchical clusters.

The final two aspects of the multiscale factor model that other factorization or clustering approaches fail to address is that the clusters or low-rank subspaces may not be the most relevant with respect to predicting the response or phenotype. We find predictive factors by searching the redundant dictionary of multiscale factors. In addition none of the other methods can impose a priori or data adapted information of the gene interaction graph \mathcal{G} in factorization or grouping.

3 Results

An intuition of the modeling principles and the flexibility of multiscale factor models is developed over a series of applications. The first two applications highlight the information content in the factors. The next three applications focus on both the predictive and interpretive aspects of the model and illustrate how the model can be used to study the effect of perturbations of a gene network.

3.1 Subphenotype discovery with multiscale factors

We begin with a study of hierarchical clustering as applied to the analysis of gene expression data. In genetically heterogeneous diseases such as cancer, clustering methods have been applied to genome-wide expression data to identify molecularly distinct subphenotypes of patients [26, 28, 30]. Such discoveries are of immense interest as they offer the potential for personalized medicine through targeted therapies and improved prognosis. It is often unclear if inferred clusters are spurious artifacts of the clustering methodology or reflect the intrinsic structure of the data. We explore this problem and contrast hierarchical clustering with clustering using diffusion wavelets.

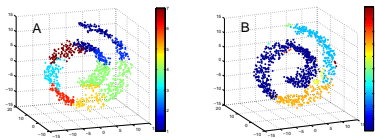


Figure 1: Clustering on a swiss roll. [A] Hierarchical clustering (complete linkage) [B] Clustering with diffusion wavelets.

Fig. 13.1 illustrates a toy example illustrating the advantage of diffusion wavelets over agglomerative hierarchical clustering. The data consists of three classes or groups – a three component mixture model – on a spiral manifold. These three classes are captured by diffusion wavelets, Fig. 13.1b, but not by hierarchical clustering with complete linkage, Fig. 13.1a.

This problem is also seen in real data. In [31] 292 high-risk breast cancer patients were hierarchically clustered into patient subgroups based on 11 scores for each patient computed from signatures of pathway deregulation, a 11×292 data matrix. A few of these clusters were able to stratify patients with respect to recurrence rates. A natural question to ask is how robust are these clusters and do they represent well-defined risk populations? When we apply hierarchical clustering with complete linkage on this same data set, we are able to recapitulate the risk stratifications previously reported [31] (log-rank, $p < .001$). However, clustering with average linkage produces clusters without any significant difference in recurrence rates. This raises some concerns about the stability of the clusters.

Diffusion wavelets can be used to cluster this data, see materials and methods. The bases or factors $\{\phi_{j,k}\}$ can be used to define hierarchical clusters since each factor assigns to each patient a probability of belonging to the factor or cluster. Our objective was to examine which clusters were able to stratify patients by risk. Adding the constraint that we were only interested in moderate to large clusters allowed us to ignore clusters at local scales. We conducted pair-wise tests between clusters at the same scale to see if the two clusters stratified patients according to recurrence rates, see Materials and Methods.

Results from this analysis suggest the presence of global risk patterns. Three significant cluster pairs were found at the coarser scales: one significant cluster pair at the 6th or global scale (log-rank, $p = .006$) and two modestly significant cluster pairs at the 5th scale (log-rank, $p = .06$, $p = .08$). This suggests that there may not be multiple distinct molecular sub- phenotypes as previously reported.

3.2 Gene Ontologies

The Gene Ontology (GO) project [9] has defined a multiscale or hierarchical organization of function-based associations annotating biological processes, molecular functions, and cellular localization. This has become the gold-standard for describing functional relationships between sets of genes.

The Gene Ontology (GO) database was used in [18] to construct a metric or dis-

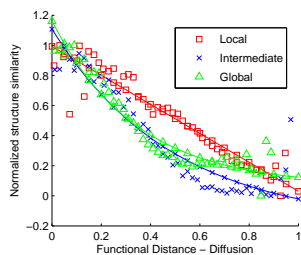


Figure 2: Functional distance (measured on a gene ontology graph) versus structural similarity of protein domains at different scales.

tance functional between protein structure domains that captures variation in the function of the protein. The authors in this study used various kernel functions $K(\cdot)$ or similarity measures to embed the gene ontology graph in a Euclidean space. Distances in this Euclidean space reflect functional distances and it was shown that distance between functional domain subgraphs using these kernel embeddings was highly correlated with protein structure domains, using the DALI Z score [19]. The geodesic distance, computed as the shortest distance on the GO graph, was unable to recover any meaningful structural correlation. This problem is related to the previous example in the sense that geodesic distance corresponds to hierarchical clustering with single linkage. The kernel functions used in [18] were local linear embeddings [16], graph diffusions [12], and the inverse Laplacian to respectively capture local, medium, and global distances. However, the scale between these different kernels is not explicit, making it difficult to intuit how the kernels correspond to different scalings. Diffusion wavelets, on the other hand, provide an explicit and optimally finite number of scales to consider.

We repeated the analysis in [18] using diffusion wavelets for generating kernels, where $K_j(\cdot) = \Phi_j$ at scale j , and found significant correlation between functional distance (on the GO graph) and protein domain structure, Fig. 22. The fit of an exponential decay model to the correlation values shows increasing decay rates with $\tau = \{\sim 0, 2, 3\}$ corresponding to local, intermediate, and global scales respectively, see Materials and Methods. The correspondence between increasing τ and global scales reflects the impact of global topology in the evaluation of functional distances. From Fig. 22, we also observe greater noise at smaller functional distances for local scales. This illustrates that diffusion at higher scales smooths or denoises the evaluation of functional distances.

3.3 Science documents classification

Classification of scientific documents is used to highlight the accuracy of the multi-scale model and the interpretability of the factors. Given a document-word matrix M composed of 1153 words from 1161 articles gathered from Science News, define M_{ij} as the frequency of the j^{th} word in the i^{th} article. Each article is also annotated according to one of eight scientific subjects - Anthropology, Astronomy, Social Sci-

ences, Earth Sciences, Biology, Mathematics, Medicine, and Physics. A preliminary multiscale analysis of the document graph using articles as nodes reveals a complex, hierarchical structure on the data, see Supp. materials. For the supervised analysis, we use M to construct a weighted word graph W with words as nodes and edge weights computed by pair-wise word co-occurrence across documents.

The objective is inference of a discriminative function capable of distinguishing Earth Sciences documents (class A) from all other documents (class B) using the multiscale factor model. Factors were generated using diffusion wavelets on the weighted word graph W and the factor weights were inferred based on equation (2). The data was randomly split into a training set of 80 documents from class A and B with the remaining documents comprising a test set.

We compared the performance to the multiscale factor model with the saturated model (Lasso on the original features), SVD regression, and regression using the eigenvectors of the Laplacian. Bootstrapping was used for fitting all hyper-parameters (see Materials and Methods). The performance metric used was the area under the receiver operating curve (AUC) on the test data. The experiment is run 20 times and results are tabulated (Tbl. 3.3) - the multiscale classification method outperforms the other methods as measured by AUC.

Method	Sat	SVD	Lap	MSF
Avg AUC	.886	.885	.914	.934
Std dev	.028	.038	.037	.022

Table 1: Performance comparison for discrimination of science documents. Lasso is applied to factors generated from following methods: (Sat) original features, (SVD) singular valued components, (Lap) eigenvectors of the Laplacian, (MSF) multiscale factors (diffusion wavelets).

The factors with significant factor loadings are also interpretable. The scale of the factors positively correlated with the Earth Science category reflect levels of specialization within the topic of Earth Science. The fine scale factors contain specific words such as 'nitrogen', words such as 'pest', 'crops', and 'plants' appear at an intermediary scale, and words such as 'weather' and 'forecast' appear at the global scale. This assignment reflects the correspondence between scale and generality with respect to subject matter. It stands to reason that that the highly weighted features of global scaling functions will be those features with the strongest or largest number of dependencies. In Earth Science, the word 'weather' appropriately fits this description.

3.4 Predicting Prostate Cancer Recurrence

The prostate-specific antigen (PSA) test is the current standard for monitoring and assessing prostate cancer (PC) risk in men. While PSA-based screening has likely contributed to the decline of PC mortality in the last decade, its high false positive rate has resulted in overtreatment and increased morbidity. Consequently, there is much interest in finding alternative methods for distinguishing between aggressive and indolent forms of PC.

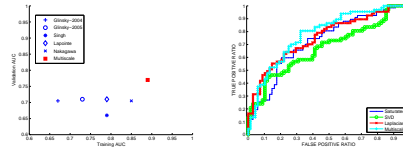


Figure 3: Predicting Prostate Cancer Recurrence, comparison of the multiscale factor models with biomarker and global factor models.

Gene expression profiling has led to several biomarker based models to predict PC recurrence [21, 26, 24, 25, 27]. These models are based on assaying the expression of a few genes. Unfortunately, the models offer only modest improvement over standard clinical models for predicting outcome. Moreover, there is little overlap in the genes used in each of these models. This highlights the perils of gene-based predictor models for complex disease: technical variation, strong genotypic and phenotypic heterogeneity, and complex molecular interactions limit the ability of these models to generalize.

We compared the multiscale factor model to these biomarker based models in predicting PC recurrence, see Fig. 43a. We also compared our model to other regression methods: SVD regression, factor regression using eigenvectors of the graph Laplacian as factors, and Lasso on the original features, see Fig. 43b. The data used was collected from a large prostate cancer data bank composed of 596 patients treated with radical prostatectomies and monitored for up to 10 years [21]. Based on clinical outcome, each patient is grouped into one of three categories: PC with no evidence of disease recurrence (NED, $n = 195$), PC with biochemical recurrence measured by PSA (PSA, $n = 201$), and PC with metastatic recurrence (SYS, $n = 200$). Gene expression data is available for each patient measured on Illumina’s generic cancer array and a custom prostate cancer array, resulting in a total of 1024 mRNA probe measurements for each patient. A dependency graph with each node corresponding to a probe was computed using a thresholded correlation statistic. A multiscale factor model to discriminate PSA from SYS was inferred. The inferred model had eight factors ranging from local to global. AUC on a validation set was used to measure the predictive performance of the models. The multiscale factor model outperforms all other methods.

3.5 PC Multiscale Pathway Analysis

The limitation of the predictive accuracy of single gene models for PC recurrence was observed in the previous section. Inference and interpretation of molecular mechanism giving rise to PC recurrence at the level of single genes suffers the same problem of lack of generalization and robustness. This difficulty has motivated analysis based on *gene sets*, which are often more robust and interpretable than analyses of single genes. Methods for gene set analysis [22, 20, 23] provide statistical evaluation of co-expression of genes in a priori defined sets. The sets capture prior biological knowledge such as the KEGG and BioCarta pathways or are based on experiments using known molecular perturbations.

We infer and interpret multiscale factor models in the space of gene sets that are predictive of PC recurrence. The nodes or variables in this model will be gene sets. Specifically, we use 639 curated genesets defined by the Molecular Signature Database [22]. An association graph reflecting similarity between gene sets is constructed, see Materials and Methods. Gene expression data for 197 patients was compiled from two independent prostate tumor banks ($n = 79$, Memorial; $n = 118$, Catalonia). Each patient was given a clinical annotation of biochemical recurrence measured by PSA elevation. The expression data X is a $m \times n$ matrix with $m = 22,000$ probes assayed and $n = 197$ observations. This data is mapped to gene set enrichment summary statistics Z which is a $g \times n$ matrix with $g = 639$ gene sets. Z_{ij} provides an estimate of the constitutive differential enrichment of the genes in the i -th gene set in the j -th patient, see materials and methods for details. The multiscale factor model was applied using the summary statistics Z as the explanatory variables, recurrence (predicted from the previous multiscale model) as the response categorical response variable, and an association graph reflecting gene set similarity, see Materials and Methods. See Supp. materials Table S1 for details of the inferred model, and gene set membership probabilities for significant factors.

We focus on a few interesting observations resulting from the multiscale analysis. Fig. 4 displays a coarse scale factor that suggests the involvement of TGF- β with cell-cycle mediated pathways such as the G1, G2, and p53 since these pathways are up-regulated in samples with high predicted probability of PC recurrence. Further evidence for functional association with cell-cycle is given by the direct connection between the up-regulated TGF- β cell-cycle gene sets. At a finer scale we observe two TGF- β related gene sets with enrichment in opposite directions, see Fig. 4a. This may reflect differential activation of subpathways as TGF- β has been shown to have both repressive and proliferative roles [29]. Also of interest is the presence and up-regulation of the p27 pathway and its gene- set neighbors involved with ubiquitin mediated proteolysis. The p27 protein is a cell-cycle inhibitor and tumor suppressor. p27 itself is known to be controlled by post-translational degradation via ubiquitination, captured by our model in Fig. 4B.

Using our model we would like to examine the effect of a drug that would disable TGF- β pathways. Of particular interest is whether compensating pathways appear, thereby negating the effect of a drug targeting TGF- β . We simulate the disabling of TGF- β by removing the two TGF- β related pathways in the association graph and then applying the multiscale factor model on the modified graph. Results of this analysis show that no new pathways appear among the selected factors, and most of the pathways involved in cell-cycle deregulation in Fig. 4 do not appear. We do continue to see the ubiquitin proteolysis pathways in Fig. 4B. This indicates that the dependency between p27 ubiquitination and TGF- β is weak and therefore likely unaffected by the disabling of TGF- β pathways.

4 Discussion

The multiscale factor model based on diffusion wavelets is an intuitive and robust framework for analyzing high-dimensional biological data with complex dependen-

cies. The method as presented is applicable to a variety of scientific settings involving high-dimensional data. We show that it is particularly well-suited for the analysis of molecular networks. This addresses a major challenge in computational biology – the development of models that are simultaneously robust and allow for the interpretation of underlying biological mechanism.

We note that the sparsity penalty used in inference of the factor loads in (2) is scale agnostic, there is no scale bias in the optimization. We can imagine situations where prior knowledge suggests preference for certain scales or alternatively we wish to infer which scale or scales should be given preference. The regularization penalty in (2) can be modified to weigh scales differently with a hyper-parameter that controls the emphasis of factors as a function of scale and can be set a priori or adapted based on data.

Finally, we did not utilize direct protein-protein or protein-DNA information from curated databases, such as HPRD [34], in the construction of the interaction graphs. We believe direct binding information is useful in well-annotated model systems such as budding yeast, but can be misleading in highly context-sensitive systems such as human cancer. However, we do believe our modeling framework can be appropriately applied to well characterized binding networks.

5 Materials

5.1 Gene expression data

The 573 breast cancer samples can be downloaded from the Gene Expression Omnibus [GEO] microarray data repository at <http://www.ncbi.nlm.nih.gov/geo> under the GEO accession numbers GSE2034, GSE4922, GSE3143, and GSE7849. These data sets were RMA normalized and corrected for batch effect; see [31] for details.

Microarray data used in the training of the prostate cancer recurrence model is available in GEO with accession number GSE10645. The prostate data used for validation and pathways analysis come from two independent radical prostatectomy series from Washington University (Catalona) and Memorial Sloan Kettering Cancer Center (Memorial). These series are not publicly available.

5.2 Local Similarity and Graph Construction

We studied the following distance metrics for defining local similarity across points or variables - Pearson correlation, Spearman rank correlation, and mutual information. We did not observe any significant differences in results between these metrics, and therefore used Pearson correlation for the results presented in this paper unless otherwise noted. To emphasize local geometry and to induce some sparseness in the graphs, all graphs were thresholded at the 60% percentile.

5.3 Phenotype Clustering in Breast Cancer

The 272 patient node graph is deduced from the similarity matrix $W = \exp(-\|x_i - x_j\|^2 / .5)$, where x is an 11-dimensional vector corresponding to 11 genomic signature scores [31]. Multiple hard clusters are generated by sampling from each $\phi_{j,k}$ using the corresponding vector of probabilities $(|\phi_{j,k}(x_i)|^2)_{i=1}^n$. Survival differences between cluster pairs are evaluated using the log-rank statistic. For details concerning the breast cancer data set and generation of signature enrichment values, see [31] and [33].

5.4 Comparison of Functional Distance and Structural Similarity At Multiple Scales

We use the method from [18] to compare correlations across scales, that is, we fit a decaying exponential using the function $Y = y_0 + Ae^{-\tau x}$. The value of τ describes the rate of decay where larger values of τ correspond to a faster decay rate. In the context of functional distances, we attribute a faster decay rate as incorporating more distant domains in calculating functional distances. We calculate functional distances at scale j using the diffusion metric $d^j(x, y) = \sqrt{K^j(x, x) + K^j(y, y) - 2K^j(x, y)}$ where K is the kernel corresponding to the diffusion operator T . See [13] for details.

5.5 Hyper-parameter fitting

Selection of λ in (2) uses .632 bootstrapping [32]. This method uses sampling with replacement, where each sample has a $1 - 1/n$ probability of not being sampled, tending to produce sparse solutions and good generalizability.

5.6 Pathway Analysis in Prostate Cancer

We merge Illumina and Affymetrix data by matching Entrez Ids, and standardize each row of genes to have mean 0 and variance 1.

The graph of pathways (gene sets) is deduced from the matrix $W_{ij} = -\log(p_{ij})$ where p_{ij} is the probability of overlap between gene sets i and j based on Fisher's Exact statistic. We use the 639 genesets from MsigDB [22] and ignore all genesets comprised of 10 genes or less.

We need to transform the data matrix X_G , a $m \times n$ matrix where m is the number of probes, to Z - a $g \times n$ matrix g is the number of gene sets. This is done using the ASSESS [20] software package for the transformation $X_G \rightarrow Z$, which measures gene set enrichment variation for each sample.

acknowledgments

M. Maggioni is grateful for partial support from NSF (DMS 0650413, CCF 0808847, IIS 0803293), ONR N00014-07-1-0625, and the Sloan Foundation. J. Guinney is grateful for partial support from PC081324, Department of Defense, Prostate Cancer Training Award. S. Mukherjee is grateful for partial support from NSF (DMS 0732260) and NIH Systems biology center grant.

References

- [1] R. Tibshirani (1996) Regression shrinkage and selection via the lasso *J Roy Stat Soc B*, 58 1:267-288.
- [2] West M (2003) Bayesian factor regression models in the “large p, small n” paradigm *Bayesian Statistics 7* 723-732.
- [3] Wright S (1921) Correlation and causation *J Agric Res* 20:557-585.
- [4] Turner ME, Stevens CD (1959) The regression analysis and causal paths *Biometrics*, 15:236-258.
- [5] Malioutov D, Johnson JK, Willsky AS (2006) Walk-sums and belief propagation in Gaussian graphical models *J Mach Learn Res* 9:660-668.
- [6] Jones B, West M (2005) Covariance decomposition of undirected Gaussian graphical models *Biometrika* 92:779-786.
- [7] Eisen M, Spellman P, Brown P, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns *Proc Natl Acad Sci USA* 95:14863-68.
- [8] Alter O, Brown P, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling *Proc Natl Acad Sci USA* 97:10101-06
- [9] Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology *Nature Genet.* 25:25-29.
- [10] Shi J, Malik J (1997) Normalized cuts and image segmentation. *Proc IEEE Comp Soc on Comp Vis and Pat Recog* 731-737.
- [11] Wu Z, Leahy R (1993) An optimal graph theoretic approach to data clustering: theory and application. *IEEE Trans on Pat Anal and Mach Int* 15:1101-3.
- [12] Coifman R, Lafon S, Lee A, Maggioni M, Nadler B, et al (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps *Proc Natl Acad Sci USA* 102:7426-7431.
- [13] Coifman R, Maggioni M (2006) Diffusion wavelets *Appl Comput Harmon Anal* 21:53-94.
- [14] Mahoney, M W, Drineas, P (2009) CUR Matrix Decompositions for Improved Data Analysis *Proc. Natl. Acad. Sci. USA* 106:697-702
- [15] Belkin M, Niyogi P (2004) Semi-supervised learning on Riemannian manifolds *Machine Learning* 56:209-39.
- [16] Roweis S, Saul L (2000) Nonlinear dimensionality reduction by Locally Linear Embedding *Science* 290:2323-26.
- [17] Chapelle O, Schölkopf B, Zien A (2006) Semi-Supervised Learning.

- [18] Lerman G, Shakhnovich B (2007) Defining functional distance using manifold embeddings of gene ontology annotations *Proc Natl Acad Sci USA* 104:11334-39.
- [19] Dietmann S, Park J, Notredame C, et al (2001) A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary *Nucleic Acids Res* 29:55-57.
- [20] Edelman E, Porrello A, Guinney J, Balakumaran B, Bild A, et al. (2006) Analysis of sample set enrichment scores: Assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics* 22:e108-e116.
- [21] Nakagawa T, Kollmeyer T, Morlan B, Anderson, S, Bergstralh E, et al, (2008) A tissue biomarker panel predicting systemic progression after PSA recurrence post-definitive prostate cancer therapy, *Plos One* 3:e2318.
- [22] Subramanian A, Tamayo P, Mootha V, Mukherjee S, Ebert B, et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide profiles. *Proc Natl Acad Sci USA* 102:15545-15550.
- [23] Barry W, Nobel A, Wright F (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* 21:1943-9.
- [24] Singh D, Febbo P, Ros K, Jackson D, Manola J, et al. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1:203-09.
- [25] Glinsky G, Glinskii A, Stephenson A, Hoffman R, Gerald W (2004) Gene expression profiling predicts clinical outcome of prostate cancer. *J Clin Invest* 113:913-23.
- [26] Lapointe J, Li C, Higgins J, van de Rijn M, Bair E, et al. (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci USA* 101:811-6.
- [27] Glinsky G, Berezovska O, Glinskii A (2005) Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer *J Clin Invest* 115:1503-21.
- [28] Loi S, Haibe-Kains B, Desmedit C (2007) Definition of Clinically Distinct Subtypes in Estrogen Receptor-Positive Breast Carcinomas Through Genomic Grade *J Clin Oncol* 25:1239-46.
- [29] Moses H, Yang E, Pietenpol J (1990) TGF- β Stimulation and Inhibition of Cell Proliferation: New Mechanistic Insights *Cell* 63:245-247.
- [30] Sorlie T, Tibshirani R, Parker J (2003) Repeated observation of breast tumour subtypes in independent gene expression data sets. *Proc Natl Acad Sci* 100:8418-23.

- [31] Acharya C, Hsu D, Anders C, Anguiano A, Walters K (2008) Gene expression signatures, clinicopathological features, and individualized therapy in breast cancer *JAMA* 299:1574:87.
- [32] Efron, B, Estimating the error rate of a prediction rule: improvement on cross-validation (1983) *J Am Stat Assoc* 78:316-331.
- [33] Bil A, Yao G, Chang J, et al. (2005) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439:353-57.
- [34] Prasad, T. S. K., Goel, R., Kandasamy, K., Keerthikumar, S., et al. (2009) Human Protein Reference Database *Nucleic Acids Research* 37:D767-D772.
- [35] Jones P, Maggioni M, Schul R (2008) Manifold parametrizations by eigenfunctions of the Laplacian and heat kernels, *Proc. Nat. Acad. Sci.*, 105(6).

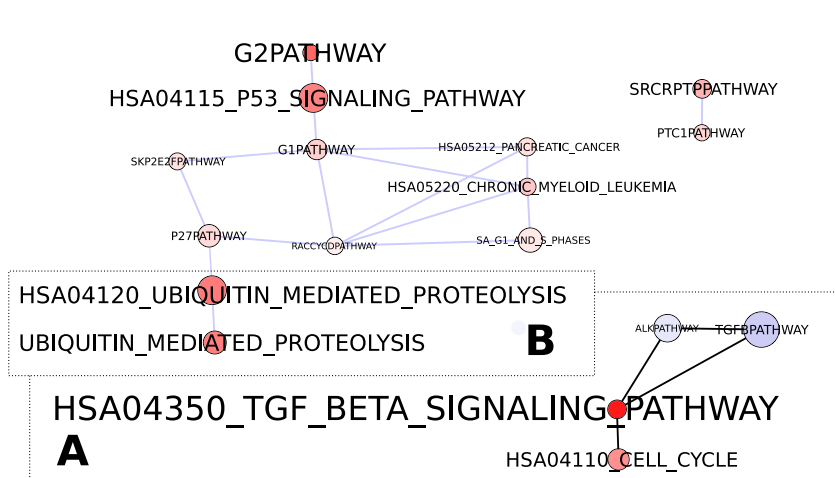


Figure 4: TGF- β scaling functions. The intensity of the red/blue nodes corresponds to the amount of up/down regulation of the pathways between the two classes. The size of the node corresponds to the weight of the pathway in the scaling function. Edges denote connections from the original dependency graph. **A**. Local TGF- β related scaling function, also observed as part of a global scaling function (entire figure). **B**. The only pathways selected after network damage, i.e. removal of the TGF- β pathways.