Bioinformatics, 31(8), 2015, 1250–1257
doi: 10.1093/bioinformatics/btu800
Advance Access Publication Date: 4 December 2014
Original Paper

OXFORD

Systems biology

SMARTS: reconstructing disease response networks from multiple individuals using time series gene expression data

Aaron Wise[1] and Ziv Bar-Joseph[1,2,*]

[1]Lane Center for Computational Biology and [2]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA

*To whom correspondence should be addressed.

Associate Editor: Igor Jurisica

Received on September 6, 2014; revised on October 22, 2014; accepted on November 26, 2014

Abstract

Motivation: Current methods for reconstructing dynamic regulatory networks are focused on modeling a single response network using model organisms or cell lines. Unlike these models or cell lines, humans differ in their background expression profiles due to age, genetics and life factors. In addition, there are often differences in start and end times for time series human data and in the *rate* of progress based on the specific individual. Thus, new methods are required to integrate time series data from multiple individuals when modeling and constructing disease response networks.

Results: We developed Scalable Models for the Analysis of Regulation from Time Series (SMARTS), a method integrating static and time series data from multiple individuals to reconstruct condition-specific response networks in an unsupervised way. Using probabilistic graphical models, SMARTS iterates between reconstructing different regulatory networks and assigning individuals to these networks, taking into account varying individual start times and response rates. These models can be used to group different sets of patients and to identify transcription factors that differentiate the observed responses between these groups. We applied SMARTS to analyze human response to influenza and mouse brain development. In both cases, it was able to greatly improve baseline groupings while identifying key relevant TFs that differ between the groups. Several of these groupings and TFs are known to regulate the relevant processes while others represent novel hypotheses regarding immune response and development.

Availability and implementation: Software and Supplementary information are available at http://sb.cs.cmu.edu/smarts/.

Contact: zivbj@cs.cmu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Several methods have been developed for modeling regulatory networks (Hecker *et al.*, 2009). While most of these methods focus on static networks, recent methods have also been developed for reconstructing dynamic regulatory networks (Bar-Joseph *et al.*, 2012). These methods, which usually integrate time series gene expression data with other types of (often static) genomic data, are specifically appropriate for modeling response and developmental processes that involve several factors that interact with their targets at different time points.

To date, such methods have been focused on modeling-specific responses or developmental processes, usually in a model organism

© The Author 2014. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

or a cell line. Examples include modeling mouse stem cell development (Mendoza-Parra *et al.*, 2011), yeast response to stress (Ernst *et al.*, 2007), plant hormone response (Chang *et al.*, 2013) and human t-cell development (Rangel *et al.*, 2004). In theory, modeling may also be beneficial for studying human development and disease response using time series data collected from individuals. Dynamic models provide important information regarding TFs that differentiate good from bad responders for a specific treatment (which may suggest new directions for clinical interventions), identify pathways that are differentially regulated and suggest groupings of patients for further treatment and/or analysis based not only on the observed expression changes but also on the underlying networks that lead to these expression profiles.

While modeling dynamic human response networks is an important goal, current methods for reconstructing regulatory networks are not appropriate for this task. Unlike model organisms or cell lines, humans differ in their background expression profiles due to age, genetics and life factors. Thus, while repeat data from animal studies are usually very helpful, time series data from different human subjects are usually not easy to integrate. In addition, there are often differences in start and end times for time series human data (e.g. the first time a patient sees a doctor for a specific infection may be very different between individuals when considering the actual time an individual has been infected). There are also differences in the *rate* of progress so that one day for one individual may represent a longer or shorter period for another (Kaminski and Bar-Joseph, 2007). Finally, there are also several differences in regulatory relationships between individuals which often result from small differences in their genomes (Kasowski *et al.*, 2010). Thus, obtaining dynamic, condition and response-specific models from human data is still a major challenge.

In this article, we present a new method for inferring dynamic, discriminatory, regulatory networks using static data and time series from multiple individuals. In contrast to previous techniques, our method, Scalable Models for the Analysis of Regulation from Time Series (SMARTS), uses multiple time series expression experiments (TSEEs) to identify the underlying regulatory dynamics. SMARTS aligns the different datasets, determines appropriate modeling time and resolution and uses an iterative procedure to learn groupings and models in an unsupervised manner. Following the assignment and modeling phases, SMARTS proposes new hypotheses (in the form of transcription factor activity) that aim to explain the differences between the groupings it identifies.

We applied SMARTS to human influenza infection and mouse brain development data. In both cases, by relying on regulatory information, SMARTS was able to greatly improve the assignment of datasets to models compared with a baseline method that only used gene expression data. In addition, SMARTS identified several TFs as discriminatory between the different groups, some of which are known to be related to the conditions studied and others which represent novel hypotheses.

2 Methods

SMARTS uses an iterative procedure to build models from sets of TSEEs (see Fig. 1). We first align our TSEEs so that all the datasets are on a common biological time scale. Next, we use clustering to learn an initial grouping for the TSEEs using our aligned time series. Finally, we iterate between creating regulatory models for sets of TSEEs and assigning TSEEs to models until convergence. In addition to the regulatory models derived by SMARTS, which can be used to identify specific activated pathways and genes, we also use statistical analysis to identify TFs that *differ* in activity between models. Such TFs are important for understanding differences between different individuals or populations.

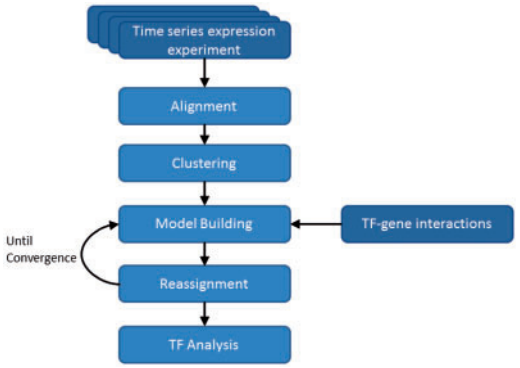

Fig. 1. Flowchart of the SMARTS algorithm. SMARTS uses two types of data (dark blue), a set of TSEEs and a list of TF–gene interactions. It performs an iterative process (lighter blue) to learn dynamic regulatory models and to assign individuals to these models. Analyzing the models leads to the identification of transcription factors with activity that differs between individual time series (Color version of this figure is available at *Bioinformatics online*.)

SMARTS can be applied on any type of data which produces time series expression measurements, such as microarrays, RNA-seq and *in situ* hybridization (ISH). Static TF-gene interaction data can be used based on the specific domain at hand, or some general data are included with SMARTS for multiple organisms.

2.1 Synchronizing individual time series

The first phase of the algorithm involves the establishment of a common biological time scale among the various TSEEs. Due to differences across organisms (including humans) in metabolism, age or genetic factors, the rates of biological processes may differ (Aach and Church, 2001; Bar-Joseph *et al.*, 2008; Lin *et al.*, 2008). For example, the incubation period of a flu infection may vary across individuals based on their prior exposures and other immunological factors. Thus, we first perform pairwise alignments for all pairs in our dataset using a method described in Bar-Joseph *et al.* (2003).

As part of the iterative alignment process, we compute a gene-wise penalty term based on the residual error after alignment with the current set of parameters. This term is used to update the global alignment parameters as discussed in Bar-Joseph *et al.* (2003). Such weighted alignment allows us to focus on genes that are key participants in the response being studied (which will likely agree on the correct alignment, and have low residual error) while minimizing the impact of background genes that may differ among individuals for other reasons (age, life style, gender, etc.). (See the Supplementary text for more details.)

As part of the alignment process, we calculate the pairwise alignment error between datasets (i.e. after alignment, the amount of average residual difference between genes in the two datasets). We then use this error (or distance) matrix to perform an initial clustering of datasets. Any of several clustering methods can be used for this initial dataset assignment. Here, we use spectral clustering (Shi and Malik, 2000) for cases where we are interested in two clusters. For clusterings with more than two clusters, we use affinity propagation (Frey and Dueck, 2007), as it tended to produce clusters that were more balanced than spectral clustering.

2.2 Model building

Clustering allows us to obtain an initial grouping of the different individuals. However, this clustering is only based on the observed expression data (which is often noisy) and furthermore uses all genes to compute the distance between the datasets. In contrast, the condition of interest (which is the unifying factor for all individuals/datasets in an experiment) is likely only affecting a small percentage of the genes, and these genes are regulated by an even smaller number of TFs and pathways. Thus, to improve our understanding of the condition and our ability to determine the different groups of individuals in our input set, we build regulatory models for groups of our input time series datasets.

2.2.1 The Dynamic Regulatory Events Miner (DREM)

To construct a regulatory model for each group, SMARTS extends DREM (Ernst *et al.*, 2007). DREM was developed and applied to model the dynamics of a *single* expression experiment at a time. In contrast, SMARTS attempts to jointly model sets of TSEEs which require us to modify DREM in a number of ways discussed below.

DREM uses an input–output Hidden Markov model (IOHMM) to construct dynamic regulatory models. An IOHMM is an extension of a Hidden Markov model (HMM) that allows output variables to be explicitly conditioned on a set of input variables. In the case of DREM, the IOHMM framework allows us to model the path of a gene through a time series experiment using regulatory models conditional on the transcription factors known to regulate each gene. These 'regulatory' models combine a single TSEE with static protein–DNA interaction data (from DNA-binding motifs, ChIP–chip or ChIP–seq data). Thus, DREM allows us to create models where states represent the emission of a specific gene value at a given time point, and transitions represent how the expression of the gene evolves to the next time point (conditional on any regulatory activity that may have occurred). Ultimately, the model is used to identify regulatory events, points in the time series where a set of genes that were previously coexpressed diverge. These splits correspond to states in the HMM and are annotated with the TFs that are predicted to regulate genes in the outgoing paths. Thus, we can identify the time of regulatory events, and the TFs that cause them. See Figure 2 for an example of what these regulatory models look like.

To determine the set of TFs associated with each split, DREM learns a L1-regularized logistic regression classifier. The classifier uses the binding profile of a gene (the set of TFs that regulate it) to predict its next state going out of the split.

DREM has been successful at modeling biological processes and systems in a variety of organisms, such as stress response in yeast and *Escherichia coli* (Ernst *et al.*, 2007, 2008; Gitter *et al.*, 2013), and mouse development (Mendoza-Parra *et al.*, 2011; Roy *et al.*, 2010; Schulz *et al.*, 2013).

2.2.2 Building models on multiple datasets

Unlike the original DREM method, which is aimed at a single dataset from one individual, in SMARTS we wish to reconstruct a regulatory network model from multiple individuals and/or datasets. There are four important ways in which we modify the IOHMM framework in order to support the reconstruction of such models:

- **Aligning and sampling constituent time series:** As discussed above, we must align TSEEs to be on the same biological time scale. This means that even if all datasets were sampled at the same time points, after alignment we may not have the same set of points across individuals. We thus need to select a set of points to which we apply the (discrete) HMM model and recover the values for these points from the different datasets. To select a set of points for the model, we calculate the cumulative distribution of measured time points across all datasets, and choose the n time points which evenly split the density of these measurements. A density-based method is used since the rate of measurement is typically informative about when regulatory activity occurs, and it ensures that the points chosen maximize the amount of information we use from the data.

  Once a time scale and time points are chosen, we use cubic splines to obtain a continuous representation for genes in all datasets and sample these splines at the specific time points selected correcting for the alignment parameters. Since all alignments are pairwise, we must choose a baseline dataset for our canonical time scale. We choose the class medioid—the dataset with the lowest alignment error to other datasets in the class. After choosing the baseline dataset, the splines from each dataset are sampled at the points aligned to the canonical points in the baseline.
- **Extending the emission/transition model:** To allow for multiple readings from a gene across the different individuals we change the way we compute the maximum likelihood for the emission and transition parameters. For the emission, we simply use all copies of a gene to learn a Gaussian model.

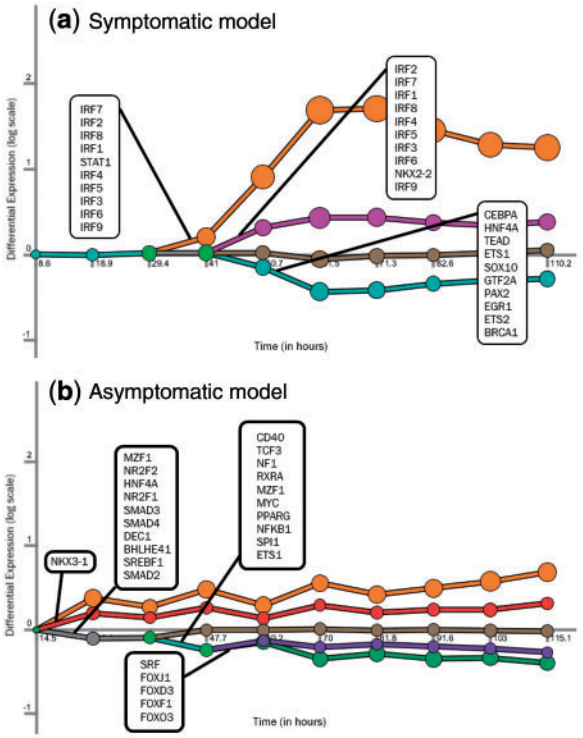

Fig. 2. Regulatory models for human influenza patients. Each model represents the regulatory program of a set of flu patients. Each path represents the activity (in terms of differential expression) of a set of genes over time. Split nodes occur when a regulatory event is predicted to cause a set of genes which were previously regulated similarly to diverge. Split paths are annotated with the set of TFs predicted to regulate the split event; only statistically significant TFs are shown. In (**a**), the orange and magenta paths show increased differential expression, as well as regulatory activity by known immune response factors in the IRF family. The top path contains many immune–response-related genes, such as TLR7 and IL1RA (Color version of this figure is available at *Bioinformatics online*.)

Transition probabilities are calculated based on the TFs which are predicted to regulate a given gene. At each split node, a logistic regression classifier is learned based on the static binding data input. We use a multiplicative model to account for the presence of multiple copies of a gene, where

$$p_{\text{trans}}(g^t) = \prod_{\{d \in D_m \text{ and } g_d^t \text{ exists}\}} p_{\text{trans}}(g_d^t),$$

where D_m is the set of all datasets belonging to model m and g_d^t is the measurement of gene g at time t in dataset d. Thus, the transition probability is a product of all of the observed examples of time series making the transition; this has the effect of emphasizing transition probabilities observed more frequently.

- **Allowing different start and end points for individual time series:** We cannot guarantee that all time series will be of the same length and, furthermore, after alignment even time series of the same length may be aligned such that the first or last time points of one time series extend beyond those of another. Thus, we must be able to model time series beginning and ending. We do so by adding implicit 'begin' and 'end' states to the IOHMM. Individuals 'skip' all states before and after their alignment-defined start and end times. This is done by connecting all states to the new start and end states.
- **Enforcing that all copies of the same gene in different individuals follow the same path in the model:** Since we assume that all individuals assigned to the same model are regulated by the same TFs and pathways, we learn a single regulatory network for these individuals. To achieve this we use the joint likelihood over all genes and their time points when determining the maximum likelihood path. Each gene is assigned its individual maximum likelihood path:

$$L(D, g; m) = \max_p \prod_{d \in D_m} \prod_t p^p_{\text{trans}}(g_d^t) p^p_{\text{em}}(g_d^t)$$

for each path p, gene g, dataset d in the set of all datasets for model m (D_m) and time t. g_d^t is the measurement of gene g in dataset d at time t. These likelihoods are calculated using the forward–backward algorithm and used to compute the model parameters.

2.3 Model assignment and reassignment

Once models have been built for each of our classes, we reassign each dataset to the model which best describes it. We do this in two ways. First, we do a number of iterations using a subset of TFs with highly differentially expressed targets. After this process has converged (or a predefined number of iterations have been performed) we perform a set of 'refinement' iterations using more of the genes in the datasets.

2.3.1 TF-based assignment

While assignment of datasets (or individuals) to models can be performed using every gene modeled in the IOHMM, we found that in many cases using only the most discriminating TFs leads to better performance. This approach provides two major advantages: First, it allows us to look at only a few, most important factors to explain the observed expression patterns making the models easier to interpret. Second, such a process allows us to overcome the dimensionality problem (tens of thousands of genes and a few individuals) by reducing the model dimensionality to a few factors that explain most of the changes.

See the Supplementary text for details in how these TFs are chosen.

2.3.2 Refinement assignment

Once our TF-based iterations have converged to a fixed set of TFs (and thus class assignments), we perform a set of 'refinement' iterations, which takes into account many more genes in the model. We choose the top n genes using a likelihood ratio score: We first calculate for each gene and each TSEE the log likelihood ratio between the TSEE's assigned model and the best other model. We then take the average of this (log) score over all TSEEs (which is equivalent to multiplication in probability space). We rank all genes based on this score, where the highest scoring genes are the most discriminating. We take the genes with scores in the top 50% of all scores, and use them to build the final model.

2.4 Cross-model analysis

Finally, to identify outcome-related transcription factors, we define a 'differential P-value' that measures the difference in a TF's activity between models using a randomization test. We perform this test for each TF and model. The test is based on all genes predicted to be regulated by the TF, and represents the average log likelihood ratio of these genes in the given model compared with the best other model:

$$s(m,t) = \sum_{d \in D_m} \left(\left[\frac{\sum_{g \in G_t} \ell(g_d; m)}{||G_t||} \right] - \left[\max_{n \in M \setminus m} \frac{\sum_{g \in G_t} \ell(g_d; n)}{||G_t||} \right] \right).$$

Here we compute the score for model m and transcription factor t, where d is a TSEE from the set assigned to model m (D_m), G_t is the set of genes regulated by TF t, $\ell(g_d; m)$ is the log likelihood of gene g measured in dataset d given model m, and M is the set of all models.

We then perform a randomization test on gene sets t^* with 1000 draws of randomly selected differential genes, where $||G_{t^*}|| = ||G_t||$, and obtain the score $s(m,t^*)$. Our P-value is the percentage of randomization scores better than our TF of interest's score.

3 Results

While the major goal of SMARTS is to learn unsupervised models of gene regulation for groups of individuals or datasets, we initially tested it on a human flu supervised dataset for which we know the correct groupings so that results can be compared with the ground truth. Next, we used it to analyze time series data from mouse brain development for which much less is known about the correct groupings.

3.1 Applying SMART to data from human flu patients

We applied SMARTS to a set of TSEEs from Huang *et al.* (2011) containing 17 patients who were (voluntarily) infected with influenza. Nine of the 17 patients developed a symptomatic influenza infection; the other eight patients were asymptomatic. Each patient was measured for 16 time points over 132 h.

We used static TF–gene interaction data consisting of predicted interactions using the method and data from Ernst *et al.* (2010). For each TF, the top 100 predicted gene targets were chosen. For more information on the TF–gene interaction data, see the Supplementary text. We performed SMARTS analysis on the 17 time series using $k = 2$ as the number of requested models. The resulting models are presented in Figure 2.

Table 1 presents a comparison of SMARTS assignment of asymptomatic and symptomatic TSEEs with two other methods that have been used to analyze and model dynamic biological

Table 1. Clustering of symptomatic and asymptomatic flu datasets

	Asymptomatic	Symptomatic
(a) Clustering after residual error clustering		
Class 1	5	7
Class 2	2	2
(b) Clustering after SMARTS without alignment or network-based gene selection		
Class 1	5	9
Class 2	2	0
(c) Clustering after SMARTS algorithm		
Class 1	0	8
Class 2	7	1

Each panel shows the clustering of the flu patient time series into two classes; the values in the table represent the number of time series in the intersection of a given condition and class label. (a) Baseline clustering using only gene expression data. (b) A simplified version of SMARTS clustering using regulatory information, but neither using alignment nor selecting a subset of genes based on their relevance. (c) Full SMARTS clustering shows better separation between symptomatic and asymptomatic time series, with only one patient mislabeled

processes: Clustering based on gene expression, and methods that do not account for differences in response initiation and rates (Ernst *et al.*, 2007). As can be seen, using a clustering method that only utilizes the gene expression data leads to an unbalanced grouping (with 12 datasets assigned to class 1, and 4 assigned to class 2). This clustering uses the average residual error between genes after time series alignment as a distance metric. Thus, it does not use the regulatory network model underlying SMARTS. Using a one-tailed Fisher exact test, we find that the separation between labels (symptomatic or not) for the clustering assignment is not significant ($P = 0.61$). Similarly, a simplified version of SMARTS's regulatory network model that uses neither time series alignment nor network-based gene selection (see Section 2.3) is also unable to correctly separate these two classes of patients. As Table 1, panel b shows, this method separates two asymptomatic patients in one class while the other class contains five asymptomatic patients and nine symptomatic patients ($P = 0.175$). In contrast, the SMARTS result (Table 1 panel c) misclassifies only one dataset (see Section 4 for more on this), and has a one-tailed Fisher exact P-value of 0.0007. These results indicate that both the network-based gene selection and alignment aspects of SMARTS are pivotal in its ability to accurately model different response networks. See the Supplementary Results for additional comparisons of SMARTS with simpler methods.

3.2 Reconstructed flu symptomatic and asymptomatic networks

The differences between the symptomatic and asymptomatic models (Fig. 2) are visually striking, and clearly show that a much stronger immune response occurs in patients that develop flu symptoms. This can be seen, for example, in the symptomatic model which includes a path (colored in orange) containing a large set of highly upregulated genes. The TFs associated with this path are predominantly STAT1 and Interferon Response Factor (IRF) factors, both well known to be involved in immune response (Taniguchi *et al.*, 2001). The genes associated with this path include many well known to be involved in immune response, including TLR7, IL1RA and TRIM22. In the asymptomatic model, the levels of differential expression are in general much lower. This is also indicative of another strength of our models, that by incorporating multiple time series, we smooth out any extraneous noise.

We next used our differential test statistic to identify transcription factors that differed in activity between the two models. The top TFs (all of which have $P < 0.001$ using our differential score) are members of the IRF family, RXRA and STAT1. STAT1 and the IRF factors are well known to be involved in immune response (Taniguchi *et al.*, 2001), and RXRA has also been identified as a regulator of immune response (Du *et al.*, 2005). It is worth noting that several IRFs show up in our model. This may be in part due to a difficulty in differentiating between TFs in a family: if our static TF–gene interaction data provide very similar predictions for the targets of TFs in a family, we will not be able to determine which specific family member(s) are actually active.

A list of all TFs and their differential P-values can be found on the Supplementary website.

3.3 Mouse developing brain analysis

In addition to modeling sets of human individual TSEEs, SMARTS can be applied to other domains where sets of time series represent a single condition. We used SMARTS to analyze time series expression data from the developing mouse brain (the Allen Brain Atlas—Henry and Hohmann, 2012). This data consist of gene expression measurements from ISH studies taken at seven time points from multiple mouse brain tissues during development. In all, about 2000 genes are measured at each time point, for each brain region. We restrict our analysis to the 25 brain regions for which there is ISH data at all seven time points.

The static TF–gene interaction data that we used as input for the brain development models were obtained by integrating sequence and tissue-specific epigenetic data using the PIPES method (Zhong *et al.*, 2013). To perform the analysis using PIPES we used DNase data from developing and adult mouse brain tissues (14 days, 18 days and 8 weeks) (Stamatoyannopoulos *et al.*, 2012). For more information on the TF–gene interaction data, see the Supplementary text.

Since these measurements are already synchronized (using genetically similar mice) we did not perform alignment, but simply used the time points as given. The brain regions are described in an ontology at a number of levels of granularity. At the highest level, 3 of these 25 regions were classified as 'midbrain', 12 were 'hindbrain' and 10 were 'forebrain'. We first performed SMARTS with two groups (models). The resulting models and assignments successfully distinguished forebrain from hindbrain, and improved on the baseline clustering that only relied on the expression data. After our initial clustering (based on average error between the curves of all genes) which only used the time series ISH data, we obtained extremely unbalanced clusters, with no clear relation to brain region (Table 2 panel a). Using a one-tailed Fisher exact test, the separation between labels was not significant ($P = 0.1948$). After our initial iterations of SMARTS using the subset of TFs determined to be differential (see Section 2.3.1), the new models greatly improved upon the initial assignment ($P = 0.0007$) as can be seen in Table 2, panel b. However, the best result was obtained after completing the additional refinement iterations (see Section 2.3.2), which resulted in only one forebrain region mislabeled (Table 2, panel c) ($P < 0.0001$).

The models can be seen in Figure 3. We identified differentially active TFs using our cross-class P-value analysis (see Section 2.4). A number of TFs were identified as differentially active between conditions with $P < 0.001$. Of these TFs, many are known to be

Table 2. Clustering of forebrain and hindbrain datasets before and after the SMARTS algorithm

	Forebrain	Hindbrain
(a) Clustering after residual error clustering		
Class 1	2	0
Class 2	8	12
(b) Clustering after TF-based iterations		
Class 1	6	0
Class 2	2	12
(c) Clustering after SMARTS algorithm		
Class 1	9	0
Class 2	1	12

Each panel shows the clustering of brain region time series into two classes. Each value in the table represents the number of time series in the intersection of a given brain region and class label. (a) Baseline clustering using only gene expression data. (b) Clustering after the TF-based iterations of SMARTS, but before the refinement iterations. Note that two forebrain regions are not assigned to either class. (c) Clustering after SMARTS algorithm only misclassifies a single brain region. The three midbrain regions are classified, but omitted from the table

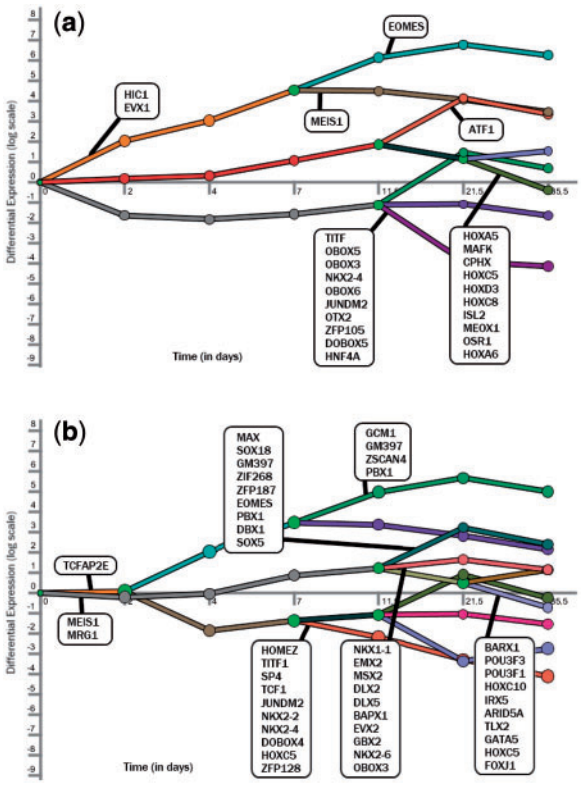

Fig. 3. Regulatory models for 2-class developing brain. (a) Forebrain model. (b) Hindbrain model (Color version of this figure is available at *Bioinformatics online*.)

differentially active in the developing brain. For example, SIX6 activity is largely centered in the forebrain (diencephalon) (Conte *et al.*, 2005), EMX2 is centered in the telencephalon (also forebrain) (Yoshida *et al.*, 1997), and DLX5 is predominantly expressed in the forebrain (Ruest *et al.*, 2003). See the Supplementary website for a full list of significant TFs and their *P*-values.

3.4 Four-way classification of brain development

Unlike our binary classification of healthy versus diseased individuals, brain structure is hierarchical, and so more than two developmental networks may be active across the different tissues. To further analyze this process, we ran SMARTS with $k = 4$ to capture a more detailed level of regulatory control during brain development. The classification resulting from this analysis is presented in Table 3. While the results substantially recapitulate topographical regions, they also deviate slightly from the ontological groupings. Class 1 is the hindbrain, excluding the most anterior structures (rhombomere 1 and isthmus). Class 2 represents the midbrain, including the most anterior portion of the hindbrain (rhombomere 1 and isthmus) and the most posterior portion of the forebrain (pretectum) indicating that the difference between these structures (hind, mid and fore) is not fully encompassed by the discrete naming convention. Indeed, there are known developmental processes that extend beyond the borders of these defined regions. For example, Irving and Mason (2000) describe how patterning of the midbrain and rhombomere 1 are both signaled by the isthmus. Class 3 contains the diencephalon, and also groups the hypothalamus (classified as secondary prosencephalon) with the thalamic regions (in the diencephalon). In fact, the hypothalamus is frequently considered part of the diencephalon, and our analysis supports that view. Class 4 contains the remainder of the secondary prosencephalon. We also performed our differential *P*-value analysis on the four-way classification (see the Supplementary website). Some of our differential TF activity is supported by the literature. For example, we predict that IPF1 is differentially active in the diencephalon, which is a result supported by Perez-Villamil *et al.* (1999).

4 Discussion

We have presented SMARTS, a novel algorithm for classifying and modeling TSEEs. SMARTS is able to build regulatory models from sets of TSEEs, and to classify time series into these models. The SMARTS framework integrates data from many individual gene expression time series with TF–gene interaction data, allowing for novel forms of analysis, such as analyzing and classifying human disease time series or modeling the differentiation of tissues during embryonic development.

We applied SMARTS to flu response data and have shown how the modeling framework greatly improves upon baseline clustering. SMARTS can also point to outlier datasets. In the flu data SMARTS identified one of the TSEEs as an outlier indicating that neither of the models is better at explaining this time series when compared with a null model (all genes assumed to have mean 0 and a single standard deviation). Figure 4a presents the model for just the outlier dataset. It appears that the first time point of the dataset is wildly divergent from the later time points. Though labeled asymptomatic by Huang *et al.* (2011), it is clear looking at this individual model that the patient does not resemble either the symptomatic or asymptomatic models in Figure 2. Thus, it appears that our outlier labeling is correct.

Figure 4b presents a DREM model for the only TSEE that SMARTS misclassified in the flu analysis (SMARTS labeled it as asymptomatic whereas it is really a symptomatic dataset). This patient, labeled in Huang *et al.* (2011) as patient 15, is the mildest case labeled as symptomatic (as can be seen in their Fig. 1b). Furthermore, as can be seen in Figure 4b, the patient developed symptoms much later in the time course than the consensus

Table 3. Brain region groupings for four-class developing brain model

Class 1	Class 2	Class 3	Class 4
Rhombomere 10: medullary hindbrain (medulla)	Collicular (rostral) midbrain tectum: mesomere 1	Peduncular (caudal) hypothalamus: secondary prosencephalon	Preoptic telencephalon: secondary prosencephalon
Rhombomere 11: medullary hindbrain (medulla)	Isthmus: prepontine hindbrain	Pretectal tegmentum: diencephalon	Roof plate of evaginated telencephalic vesicle: secondary prosencephalon
Rhombomere 2: prepontine hindbrain	Preisthmic midbrain tectum: mesomere 2 (preisthmus or caudal midbrain)	Prethalamic tegmentum: diencephalon	Subpallium: secondary prosencephalon
Rhombomere 3: pontine hindbrain	Preisthmic tegmentum: mesomere 2 (preisthmus or caudal midbrain)	Prethalamus: diencephalon	
Rhombomere 4: pontine hindbrain	Pretectum: diencephalon	Thalamic tegmentum: diencephalon	
Rhombomere 5: pontomedullary hindbrain	Rhombomere 1: prepontine hindbrain	Thalamus: diencephalon	
Rhombomere 6: pontomedullary hindbrain			
Rhombomere 7: medullary hindbrain (medulla)			
Rhombomere 8: medullary hindbrain (medulla)			
Rhombomere 9: medullary hindbrain (medulla)			

Each brain region is listed, followed by its ontological category

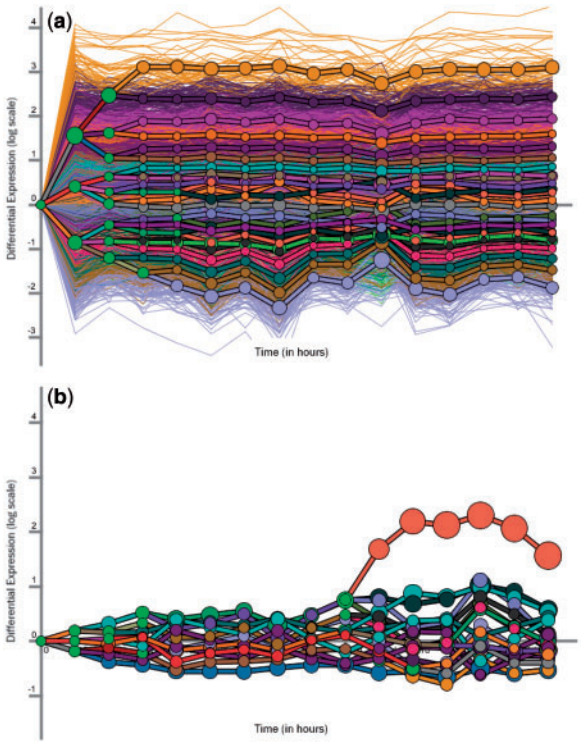

Fig. 4. Individual models of misclassified and unclassified patients. (**a**) Unclassified patient. Gene expression at time point 0 deviates highly from other time points, resulting in a model that does not resemble a 'normal' symptomatic or asymptomatic patient. (**b**) Misclassified patient. The spike in immune response genes can be seen in the reddish brown (topmost) path (Color version of this figure is available at *Bioinformatics online.*)

symptomatic model (Fig. 2a). It seems that a combination of borderline class membership and difficult alignment is the most likely cause of this misclassification.

SMARTS allows the analysis of sets of data at differing levels of granularity. We were able to recover the major fore/hind grouping of brain development when looking at a two-class analysis of developing brain regions. When we performed four-class analysis, we retained this general grouping, but with increasing specificity: two forebrain groups, encompassing different regions of the forebrain, a posterior hindbrain group and an anterior hindbrain/midbrain group.

Beyond merely building models, our technique allows for the identification of regulatory factors that differ between models. This could allow the discovery of transcription factors that are not just active during a disease, but differentially active in different disease states. This advance will allow for the increasingly granular analysis of regulatory activity in disease progression.

SMARTS is designed to scale, and is multithreaded at its core. Though both of the datasets analyzed here contain around 20 individual time series, SMARTS is capable on running on substantially larger experiments—the bottleneck is the availability of data.

While currently only a few patient datasets exist for our application, we expect the amount of clinical data to increase and such data are often temporal in nature. That said, as we showed in Section 3, the method can be applied to non-clinical data (animal models) as well, leading to important insights regarding the grouping of different conditions or tissues based on their regulatory program.

One difficulty in methods involving the grouping or clustering of data is the selection of the number of groups to use (in our case, the parameter k). In both of the examples we present, we choose the number of groups a priori: in the case of the human flu data, the number of groups is known (2); in the case of the mouse brain data, we believe we have showed that looking at the data at multiple granularities is valuable. In the scenario where the number of groups is unknown, since we use a likelihood-based algorithm, penalized likelihood techniques such as the Akaike information criterion can be used.

Future work will involve the application of SMARTS to patient data in a clinical setting, a setting which requires a more sophisticated look at time series alignment. In the clinical setting, the disease onset date is unknown, only known is the date the patient first

presents to the doctor. Thus, new techniques must be developed to better align time series with ambiguous start times.

Funding

The work was supported in part by National Institute of Health [grant number 1 U54 HL127624-01 to Z.B.J.], by the National Science Foundation [grant number DBI- 1356505 to Z.B.J.] and by the James S. McDonnell Foundation Scholars Award in Studying Complex Systems.

Conflict of Interest: none declared.

References

Aach,J. and Church,G.M. (2001) Aligning gene expression time series with time warping algorithms. *Bioinformatics*, **17**, 495–508.

Bar-Joseph,Z. *et al.* (2003) Continuous representations of time-series gene expression data. *J. Comput. Biol.*, **10**, 341–356.

Bar-Joseph,Z. *et al.* (2008) Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells. *Proc. Natl Acad. Sci.*, **105**, 955–960.

Bar-Joseph,Z. *et al.* (2012) Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.*, **13**, 552–564.

Chang,K.N. *et al.* (2013) Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in arabidopsis. *Elife*, **2**, e00675.

Conte,I. *et al.* (2005) Comparative analysis of six3 and six6 distribution in the developing and adult mouse brain. *Dev. Dyn.*, **234**, 718–725.

Du,X. *et al.* (2005) An essential role for rxrα in the development of th2 responses. *Eur. J. Immunol.*, **35**, 3414–3423.

Ernst,J. *et al.* (2007) Reconstructing dynamic regulatory maps. *Mol. Syst. Biol.*, **3**, 74.

Ernst,J. *et al.* (2008) A semi-supervised method for predicting transcription factor–gene interactions in *Escherichia coli*. *PLoS Comput. Biol.*, **4**, e1000044.

Ernst,J. *et al.* (2010) Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res.*, **20**, 526–536.

Frey,B.J. and Dueck,D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–976.

Gitter,A. *et al.* (2013) Linking the signaling cascades and dynamic regulatory networks controlling stress responses. *Genome Res.*, **23**, 365–376.

Hecker,M. *et al.* (2009) Gene regulatory network inference: data integration in dynamic models. A review. *Biosystems*, **96**, 86–103.

Henry,A.M. and Hohmann,J.G. (2012) High-resolution gene expression atlases for adult and developing mouse brain and spinal cord. *Mamm. Genome*, **23**, 539–549.

Huang,Y. *et al.* (2011) Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection. *PLoS Genet.*, **7**, e1002234.

Irving,C. and Mason,I. (2000) Signalling by fgf8 from the isthmus patterns anterior hindbrain and establishes the anterior limit of hox gene expression. *Development*, **127**, 177–186.

Kaminski,N. and Bar-Joseph,Z. (2007) A patient-gene model for temporal expression profiles in clinical studies. *J. Comput. Biol.*, **14**, 324–338.

Kasowski,M. *et al.* (2010) Variation in transcription factor binding among humans. *Science*, **328**, 232–235.

Lin,T.-h. *et al.* (2008) Alignment and classification of time series gene expression in clinical studies. *Bioinformatics*, **24**, i147–i155.

Mendoza-Parra,M.A. *et al.* (2011) Dissecting the retinoid-induced differentiation of f9 embryonal stem cells by integrative genomics. *Mol. Syst. Biol.*, **7**, 538.

Perez-Villamil,B. *et al.* (1999) The pancreatic homeodomain transcription factor idx1/ipf1 is expressed in neural cells during brain development. *Endocrinology*, **140**, 3857–3857.

Rangel,C. *et al.* (2004) Modeling t-cell activation using gene expression profiling and state-space models. *Bioinformatics*, **20**, 1361–1372.

Roy,S. *et al.* (2010) Identification of functional elements and regulatory circuits by drosophila modencode. *Science*, **330**, 1787–1797.

Ruest,L.-B. *et al.* (2003) Dlx5/6-enhancer directed expression of Cre recombinase in the pharyngeal arches and brain. *Genesis*, **37**, 188–194.

Schulz,M.H. *et al.* (2013) Reconstructing dynamic microrna-regulated interaction networks. *Proc. Natl Acad. Sci.*, **110**, 15686–15691.

Shi,J. and Malik,J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **22**, 888–905.

Stamatoyannopoulos,J.A. *et al.* (2012) An encyclopedia of mouse dna elements (mouse encode). *Genome Biol.*, **13**, 418.

Taniguchi,T. *et al.* (2001) Irf family of transcription factors as regulators of host defense. *Annu. Rev. Immunol.*, **19**, 623–655.

Yoshida,M. *et al.* (1997) Emx1 and emx2 functions in development of dorsal telencephalon. *Development*, **124**, 101–111.

Zhong,S. *et al.* (2013) Predicting tissue specific transcription factor binding sites. *BMC Genomics*, **14**, 796.