My research interests are in *machine learning* and its applications in *computational biology*. Machine learning is the branch of computer science that is concerned with recognizing patterns in data sets and using this new knowledge to characterize new data. Computational biology is the study of developing algorithms to solve important problems in biological domains. The subject of my doctoral research has been the study of high-dimensional time series data, such as the *expression levels* of genes as they vary over time. (These expression values represent the "activity levels" of the genes in a cell which vary due to external stimuli, circadian rhythms, developmental stages, etc.) I ask two questions. First, given two such high-dimensional time series, what is the best way to align them so that similarities are made apparent? Second, given a database of labeled examples and an unlabeled query, which label should be assigned to the query? In pursuing answers to these questions, I have developed methods to interpolate missing data, to align accurately and quickly, and to simultaneously cluster and calculate independent alignments for each cluster.

Motivation

One immediate motivation for my work is the need for faster, more cost-efficient protocols for characterizing the potential toxicity of industrial chemicals. I have been working with Prof. Christopher Bradfield's lab (Univ. of Wisconsin, Department of Oncology) to develop such an assay. More than 80,000 chemicals are used commercially, and approximately 2,000 new ones are added each year. This number makes it impossible to properly assess the toxicity of each compound in a timely manner using conventional methods. However the effects of toxic chemicals may often be predicted by how they influence global gene expression over time. For example, certain genes are good indicators of an inflammatory reaction, because they either ramp up or shut down their activity levels in response to exposure. If an uncharacterized chemical is seen to affect these genes in a similar way, there is a good chance that it is an inflammatory agent as well. Using technologies such as microarrays or RNA sequencing, it is possible to measure the expression of thousands of genes simultaneously following exposure to an uncharacterized chemical. We can thus create a profile for it, and classify it by comparing it to the profiles of several well-characterized treatments. It is likely that these gene-expression profiles will soon become a standard component of toxicology assessment and government regulation of drugs and other chemicals.

Of course, analyzing time series in such a way is applicable to many domains beside toxicogenomics. An overarching challenge in modern biology is to model the complete metabolic, signaling, and regulatory networks (i.e. the "circuit diagram") of a cell or organism. My research helps toward this ultimate goal by finding regularities within the gene-expression profiles that are a crucial part of this network. Regularities may include genes that are co-expressed, or genes with different expression levels but that respond in a similar way to a particular treatment. For example, I have worked with Prof. James Thomson's lab (Univ. of Wisconsin, Department of Anatomy), applying my algorithms to stem cell gene-expression data. The stem cells in question were exposed to different treatments which cause them to differentiate. We are trying to identify dependencies among the expression levels of various genes involved in the differentiation. My algorithms aid this task by pooling the various time series together, quantifying the differences observed. This helps elucidate the network of genes at play. I have also applied my algorithms on another data set from the Bradfield lab, on mice that have had MOP3—a crucial circadian rhythm gene—knocked out. By aligning these profiles with those from wildtype mice, we can identify sets of genes that are regulated in a similar manner by the MOP3 genes.

My methods are also of use in tasks unrelated to biology or gene expression. They are broadly applicable in domains that involve classifying data that have multiple dimensions over time. For example, I have successfully used them to align and classify sign language data. In this case, each dimension of the data is not a gene's expression level, but the position of a hand or finger as it varies over time. In this way I can probabilistically classify new signs by matching them to those already in a database. I have also applied my algorithms to speech-recognition and electroencephalogram data sets. All of these domains have a time component, but this is not necessary in order to use my algorithms. For example, my methods can be used in order to align chromatography data, in which features vary over distance instead of time.

Technical Contributions

Gene-expression time series tend to be sampled irregularly, and very sparsely in time. Time series that one wishes to compare may not even have been sampled at the same times. One of the first challenges is to perform some kind of interpolation to mitigate these problems. Previously, other research groups had successfully interpolated gene-expression data using *B-splines*, a kind of piecewise polynomial. However they had all assumed that the points to fit were evenly spaced, which often leads to undefined interpolations when dealing with our sparse data. I refined the spline fitting to allow for more unevenly sampled data. I also noticed that B-splines as traditionally used tend to overfit the data, resulting in interpolations that intercept the observed points but vary wildly between them. I developed a method using *"smoothing splines"* which relaxes the intercepting requirement in order to prevent this overfitting from happening. I have shown that this method more accurately predicts missing data than the more traditional use of B-splines.

In order to compare and classify time series, we need to determine which parts of a pair of given series are most similar. The alignment task involves calculating which parts of the series should be aligned with one another in order to maximize the similarity between the two. This is often called *time warping*, because the times compared are subtly shifted for this purpose. In addition, it is possible when comparing gene-expression data that one series has not advanced as much as the other, appearing to be truncated. It is thus important to *"short"* the warp, allowing the end of one of the series to remain unaligned. Most work in alignment has used a method called *dynamic time warping*. This is a fast algorithm, but often aligns suboptimally on expression data. It must make all comparisons between specific times, without considering those times immediately before or after. I developed a segment-based method, that partitions the series into an equal number of segments and then compares corresponding segments. All times within a pair of segments can be compared as a unit, eliminating the locality problem of dynamic time warping. My method also allows the user to program in penalty factors that can vary depending on the domain. I showed that this novel algorithm classifies and aligns more accurately than the previous methods, when working with our toxicology data.

This algorithm must search for the best partitioning of each series into a fixed number of segments, and doing so is computationally expensive. Even with dynamic programming optimization, its time complexity remains $O(n^5)$ (where n is the length of the interpolated series). By contrast, dynamic time warping has a complexity of $O(n^2)$. I developed several heuristics, however, to speed up the search. The first heuristic disallows segment partitionings in which the segment boundaries found in each series are very different from each other. This speeds the search up by a constant factor, and actually can improve accuracy as it serves as a regularization method. Another heuristic I developed is to do a first pass using a method similar to dynamic time warping, and then restrict the segment search so that the segments found closely match that first alignment. This can speed the time complexity up to $O(n^3)$ without significantly hurting alignment and classification accuracy. I also developed a third heuristic that searches for segment boundaries in one series at a time, and goes back and forth between the two until it converges on a local maximum score. This also works in $O(n^3)$ time.

My most recent focus has been on finding *clusters* of genes that are warped together. Up to this point, I have assumed that all the genes should be aligned the same way, even though this is obviously a simplification. I have further improved our alignments and classifications by allowing each cluster of genes to be aligned separately. Note that I am not clustering the expression profiles directly, as other research groups have done before. These methods only group together genes that exhibit similar expression levels. I am clustering the alignments themselves, which allows me to group together genes that vary in the same way between treatments, even though their expression levels may be very different. I find my clusters through a variation of the *Expectation Maximization* algorithm, alternately reassigning genes to different clusters and recalculating clusters until the method converges. Thus I align and find clusters simultaneously. When two genes react in a similar way to a different treatment or condition, it may be that they are both responding to the same (possibly hidden) stimulus. My clustering method associates the two of them together, thus indicating that there may be a connection.

Future Work

There are several directions in which I plan to take my work. One weakness in most of it is that I assume that there exists an implicit zero-point that we know is already aligned. For example, in my toxicology work I assume that series are aligned at the time of exposure. It should be possible to design a previous step that searches both series for closely aligned seed points from which to anchor the alignment. The alignment could then proceed both forwards and backwards from such a point. This is similar to a *BLAST* search, which performs a similar step when aligning DNA or protein sequence data. Doing this would greatly expand the utility of my algorithms to other domains.

Another unsolved problem is how to extrapolate unseen dosages of known treatments. For example, what if an unknown treatment appears to have similar effects to a drug like acetaminophen, but much more potent? A robust algorithm should be able to tell a researcher that it has similar properties to a known drug, but at strengths heretofore unseen. Here, it might be possible to use knowledge extracted about the underlying gene network. Knowing which genes are linked will give us a better idea about how they should be expressed in unseen strengths, and allow us to better classify these treatments when we see them.

It might also be possible to adapt my methods for use in comparing different species. This would require a looser definition of a "gene," since those often vary between species. We would also need some way to take into account genes that do not have an ortholog in one of the species being compared. But in principle, we can align the expression values of related genes across species and uncover differences in their interrelations. This could provide valuable insights about how the genetic networks evolved.

I could also compare similar species, one of which has an important or valuable ability. For example, many bacteria are capable of fixing ultraviolet damage to their genomes. Some amphibians can regrow lost limbs and organs. Both of these abilities have obvious medical applications. By comparing these species with related ones that lack the feature in question, we can better map out the genetic processes that are taking place.

I anticipate that these lines of research will continue to yield interesting results for years to come.