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. In particular, I have worked on the analysis and classification of time series within biological systems, including gene-expression data and the ultra-high frequency calls of mice.

Doctoral Dissertation: Alignment and Classification of High-Dimensional Gene-Expression Data

The subject of my doctoral research was 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 asked 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? This is illustrated in Figure 1. In pursuing answers to these questions, I 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 was the need for faster, more cost-efficient protocols for characterizing the potential toxicity of industrial chemicals. I worked 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 helped 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. My algorithms have been applied to systems such as developing stem cells and to animals with important regulatory genes knocked out, in order to elucidate the function of the underlying networks.

They are also broadly applicable in domains that involve classifying data that have multiple dimensions over time. For example, I 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 probabilistically classified new signs by matching them to those already in a database. I 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.

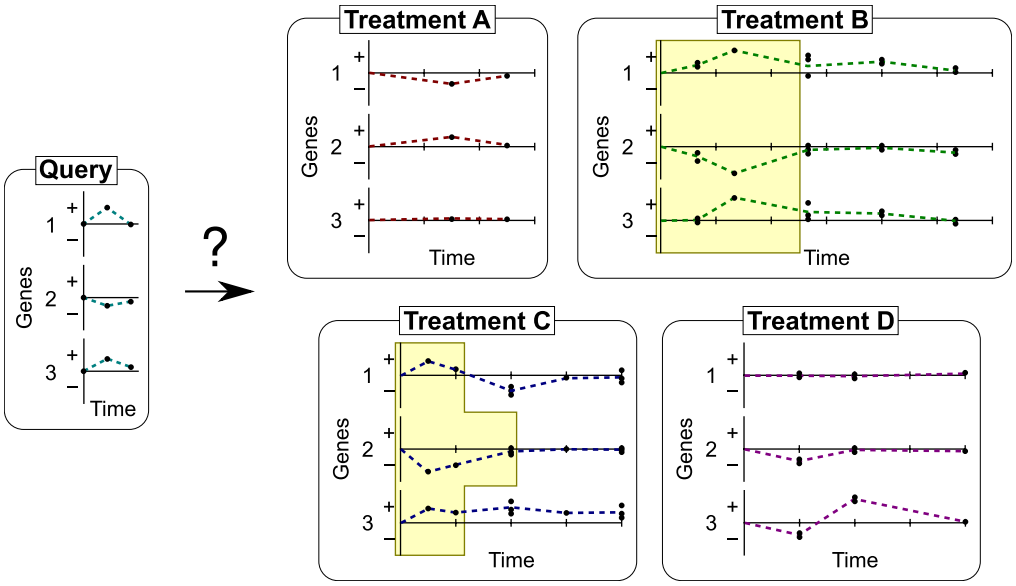

Figure 1: A toy example showing the expression (relative to baseline) of several genes over time, after some treatment has been applied. Here we are trying to find which of the known treatments the query most resembles. Expression levels at unobserved times must be reconstructed, and the query must be temporally aligned to possible treatments. Here, either Treatment B or Treatment C might be a good answer.

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 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 showed that this method more accurately predicts missing data than the more traditional use of B-splines.

In order to compare and classify time series (as in Figure 1), 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 dynamic-programming method commonly 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.

I then focused on finding clusters of genes that are warped together. Up to this point, I assumed that all the genes should be aligned the same way, even though this is obviously a simplification. I improved our alignments and classifications by allowing each cluster of genes to be aligned separately (as in Treatment C in Figure 1). Note that I did not cluster the expression profiles directly, as other research groups had done before. These methods only group together genes that exhibit similar expression levels. I clustered the alignments themselves, which allowed me to group together genes that vary in the same way between treatments, even though their expression levels may be very different. I found my clusters through a variation of the Expectation Maximization algorithm, alternately reassigning genes to different clusters and recalculating clusters until the method converges. Thus this method aligns and finds 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.

Postdoctoral Work: Automated Extraction of Mouse Ultrasonic Vocalizations

As a postdoctoral fellow, I have applied my bioinformatics knowledge to understanding the ultrasonic (i.e. ultra-high frequency) vocalizations made by laboratory mice. Again, I asked two main questions. First, given an audio recording, what is the best way to identify and extract the vocalizations within it? Second, what is the meaning of these vocalizations? Do any of them seem to correlate with observed behaviors of the animals?

Motivation

Mental illness and brain damage in people is often accompanied by difficulties in properly expressing emotions and responding to emotions in others. These conditions—including autistic spectrum disorder, schizophrenia, drug addiction, and traumatic brain injury—have a high social cost. Understanding them better has the potential to improve quality of life for millions of people, and save billions of dollars. Because these conditions express themselves in the expression and emotional state of those afflicted, it would be a logical step to quantify such attributes in any model organism used to study them. However, our understanding of individual mouse vocalizations and their connection to the animal's affective state remains primitive.

The task is thus to establish a baseline of the normal, expected vocalizations emitted by mice under certain stimuli. Mouse vocalizations appear as high-frequency whistles, and (with a few exceptions, such as mating calls and the distress calls emitted by an isolated mouse pup) no one has yet established a firm link between a mouse's state and the sounds it emits. Once the sounds are decoded and a baseline has been established, we may study mice we believe to model a specific mental disease or injury (e.g. those with certain genes knocked out, or with particular lesions on their brains) to assess how they differ. Further, we may then apply experimental treatments to these animals, to ascertain if they help mitigate the damage.

However, biologists are impeded in this task by the large amount of labor involved. Parsing a minute-long recording to identify all the microseconds-long calls, while leaving out background noise, can easily take a day's work.

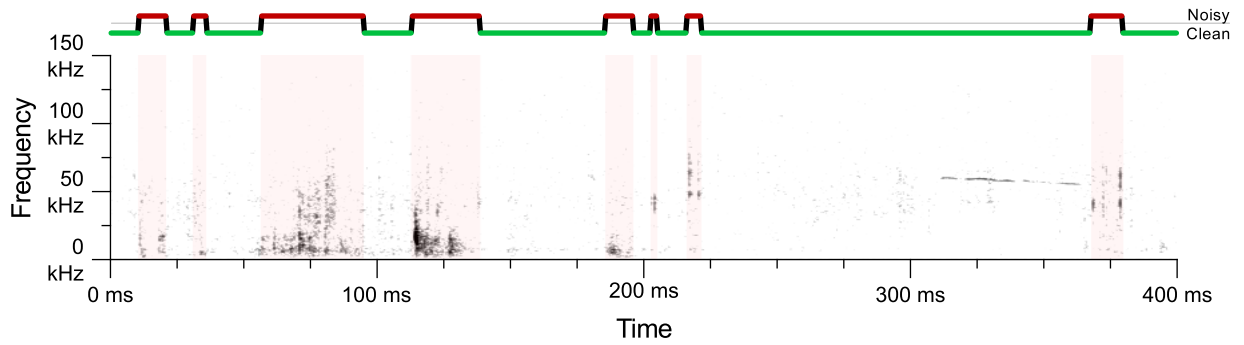

Figure 2: Finding noise in a recording of a mouse using a hidden Markov model. The model has two states: "noisy" and "clean," illustrated by the thick jagged path on top which is high during noisy times, and low during clean ones. Note the bona fide whistle-like vocalization between about 310 ms and 365 ms, which the HMM does not interpret as noise.

Complex bioacoustics programs can help, but they present a new problem. Such programs require a great deal of training to use, and even then they require a great deal of human supervision and setting of parameters to do the job well. There is a great opportunity here to develop an intelligent call detection and classification program that learns from human-labeled data, and applies that knowledge to classify new data.

Technical Contributions

I developed a filter based on a hidden Markov model, which is illustrated in Figure 2. First, a human labels the noise within several spectrograms by hand. (A spectrogram is a graph obtained by a sliding fast Fourier transform, that shows the frequencies of the pure tones emitted over time.) Certain frequencies are not commonly emitted by mice, and so are likely indicative of background noise. The algorithm finds all of these by looking at which areas the human designated as noisy, and which as clean. Thus, given a new time point with all of its associated frequencies, the model can calculate a probability that it was emitted at a noisy time or at a clean time. Further, by using a hidden Markov model, we can efficiently take into account the classifications of the times immediately before and after the given time. We have shown that the resulting filter functions just as well or better on mouse vocalization data than noise-reduction filters currently used. Not only is it efficient (employing the Viterbi algorithm, and thus running in $O(n)$ time where n is the length of the recording), but it is based on human-labeled input rather than explicit mathematical parameters. This will make it much more intuitive for biologists to use, since they can train it themselves.

I also developed a hidden Markov model technique to determine the single frequency that was most likely emitted by an animal at any given time. Ultrasonic mouse calls often resemble whistles, in that they usually consist of a single frequency that is modulated over time (see for example the call in Figure 2, starting at about 310 ms). By using an HMM to trace the whistle, we can easily classify the call as being flat, upward-modulated, chevron-shaped, etc. Some studies (including my own) have suggested that these different shapes may be associated with different states of the animal. By extracting all of this data quickly and accurately, we will be able to quickly ascertain if this is true.

Future Work

The unifying theme through my research has been the use of time-series analysis algorithms to differentiate among data within biological domains. I have developed expertise with many methods, adapted several to new problems, and written my own. These algorithms have been used by myself and others to answer important biological questions.

My future work will likely be geared toward analyzing the calls of mice. I have spent the time during my postdoc developing contacts in the field in Portland and beyond, and I believe there are several interesting questions that I can help answer. Further, several aspects of these questions are, when taken independently, within the scope of what an undergraduate student may do in the course of a summer. I already have experience supervising undergraduate research, and I look forward to introducing new students to my work. A talented student could even use one of these problems as the basis for a senior thesis.

Most analyses done by the bioacoustics community treat differences in frequency the same, regardless of what frequency they are. For example, two calls modulated upward by 10 kHz will be treated identically, regardless of whether they started at 40 kHz or at 80 kHz. However, terrestrial vertebrates perceive pitch on a logarithmic scale, not a linear one. Thus it is likely that a 10-kHz modulation is perceived to be twice as great when the call started at 40 kHz. My experiments have shown that as the base frequency of a call increases, the expected modulation does as well. That the bioacoustics community uses the linear scale is likely due to limitations of the fast Fourier transform itself. However, wavelets offer a fascinating opportunity to use a frequency transform that is inherently log-scaled. I would especially like to investigate the use of Morelet wavelets to create spectrograms. This could easily improve the quality of our analyses.

When comparing two similar calls (e.g. two upward-modulated calls), one would not expect them to necessarily change frequency in lockstep. As I mentioned above, dynamic time warping is the algorithm that is most commonly used to address this problem. However, dynamic time warping assumes that each series being compared has the same dimensionality: either two numbers are compared as they change over time, or two constant-sized vectors of numbers are compared. In the case of mouse vocalizations, this is not necessarily true. Multiple calls start and stop as time progresses, and it may be that the message is found in the interplay between these juxtaposed elements. Thus not every call has a valid frequency at all times, and sometimes two calls might overlap in time. Thus, a time-warping algorithm is needed that can successfully align this sort of data. I believe I can develop such an algorithm, and bring it to bear on vocalization data.

Finally, the automated extraction of salient information from vocalization data opens up many new areas of study. By observing differences in mouse vocalizations between experimental conditions, we will open up a new window into the behavior of these animals. With the connections I have developed during my postdoctoral tenure, I expect in the near future to focus my algorithms on many novel behavioral issues studied by my colleagues.