Consensus Spectral Techniques and Machine Learning

Debra Knisley and Jeff Knisley

Institute for Quantitative Biology
East Tennessee State University

AMS Special Session on Big Data: Mathematical and
Statistical Modeling, Tools, Services, and Training

Outline

1. **Tools and Training:** Initial comments, and then some software tools
2. **Modeling:** Where Consensus Spectral Techniques fit in.

Big Data is complex

- A very large – and *complex* – data set or network
  - Complex (roughly) means High dimensional + multi-scale (think “fractals in ^{n}” for large n)
  - Complexity is important in part due to *emergence* (roughly)
    - Smaller scales influence what happens on larger scales
    - Phenological (large scale) as a consequence of genomic or proteomic (small scale)
    - Epidemiological (large scale) as a consequence of local interactions (small scale)
- Large data sets focused on small scale activities (e.g., microarrays) used to explain large scale behaviors
- *Another Implication: Every Big Data Problem is Unique!!*

Big Data Requires Different types of Models

- The earth ($\approx 6 \times 10^{24}$ kg) is to a human ($\approx 10^{2}$ kg) as a human is to a protein ($\approx 5.5 \times 10^{-23}$ kg)
  - Earth-Moon gravitational system (two body problem):
    - Ignores the billions of humans running around on the earth
    - is a *descriptive model* (i.e., single scale)
  - Proteomics considers both the human and protein scales simultaneously
    - Requires *predictive models* that allow multi-scale
    - Most predictive models are *algorithms*
- Machine learning is a type of predictive modeling

Big Data and Software/Analysis

- A fairly standard process
  1. Exploration (Plots, histograms, basic stats, etc)
  2. Pre-processing (rescaling, centering, normalizing, missing data, etc)
  3. Predictive Modeling Algorithm (clustering/classification/regression)
  4. Metrics (How did we do??)
- Software tools implement this process
  - Togaware Rattle (http://rattle.togaware.com/) based on R
  - sklearn (http://scikit-learn.org/stable/) based on Python

Togaware Rattle (http://rattle.togaware.com/)

- Great, Great tool – Easy to use, easy to customize (generates R code available in Log pane)
- Only for relatively small data sets with fairly limited complexity

SKlearn (http://scikit-learn.org/stable/)

- Another great tool, but requires knowledge of Python and various libraries such as Numpy, Scipy, Pandas
- How SKLearn is used in general:
  1. Data as Numpy Matrices (Features/Factors, Class)
  2. from sklearn import preprocessing # pre-process the data
  3. from sklearn import cross_validation # train/validation/test partitioning
  4. Model = MachineLearningObject ()
     Model.fit(TrainingFactors, TrainingClass)
     TestingClassPredictions = Model.predict(TestingFactors)
  5. from sklearn import metrics

Preprocessing is very Important!!

- **Principal component analysis** (PCA) is a statistical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.
- **Independent component analysis** (ICA) is a computational method for separating a multivariate signal into additive subcomponents by assuming that the subcomponents are non-Gaussian signals and that they are all statistically independent from each other.
- Goal: Representation of the Data that reduces dimensionality and redundancy
- Consensus Model: A representation that quantifies what a collection of observations have in common
  - Across an entire collection of data (unsupervised)
  - Within individual classes of a set of labelled data (supervised)

Consensus Spectrum

- Some History: Protein coding regions of DNA predicted from 3-base periodicity of coding regions via Fourier Transforms
  - Key concept: Homologous families = Collections with similar characteristics
  - Various methods used to construct a single representative genome from a homologous family
- Similar methods for residue chains
  - Transform sequences into the frequency domain (via Fourier Transform)
  - Convolution in the time domain = multiplication in the frequency domain
  - Product of two spectral representations retains only the information the sequences have in common

Protein Example

Example: Cytochrome C functional group plays (more or less) the same role in all living cells.

- Human: GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATN
- Horse: GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNKGITWKEETLMEYLENPKKYIPGTKMIFAGIKKKTEREDLIAYLKKATN
- Dog: GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFSYTDANKNKGITWGEETLMEYLENPKKYIPGTKMIFAGIKKTGERADLIAYLKKATK
- Wheat: GNPDAGAKIFKTKCAQCHTVDAGAGHKQGPNLHGLFGRQSGTTAGYSYSAANKNKAVEWEENTLYDYLLNPKKYIPGTKMVFPGLKKPQDRADLIAYLKKATS
- Rice: GNPKAGEKIFKTKCAQCHTVDKGAGHKQGPNLNGLFGRQSGTTPGYSYSTANKNMAVIWEENTLYDYLLNPKKYIPGTKMVFPGLKKPQERADLISYLKEATS

The General Approach

- **Signal Processing:** Amino acid sequence GDVEKGKK... is converted to a sequence $\{x_k\}_{k=1}^N$, where each x_j is a (numerical) amino acid *descriptor* .
- **Z-transform:** The Z-transform of a sequence $\{x_n\}_{n=1}^N$ is a function of the form
$$X(z) = \sum_{k=1}^{N} x_k z^{-k}$$
- **Frequency Domain:** For $z = e^{i\omega t}$, the quantity
$$P(\omega) = |X(z)|^2 = 2\sum_{k=1}^{N}\sum_{k=1}^{N} x_k x_j \cos((k-j)\omega)$$
is the *power spectrum* of the signal.
- Consensus is $P_1(\omega) P_2(\omega) \cdot \ldots \cdot P_m(\omega)$, for $j = 1, \ldots, m$ sequences in the Homologous family
- Consensus is a generalization of cross-correlation of sequences

Example: Electron Ion Interaction Potentials

- Electron Ion Interaction Potentials (EIIP) can be used to find protein “hot spots” – i.e., regions of proteins that are most likely to bind to other proteins
  - Each Amino Acid (AA) has a specific EIIP number
  - For example, GDVEKGKK... is converted into
    $$0.005, 0.1263, \; 0.0057, \; 0.07610, \; 0.0371, \; 0.005, \; 0.0371, \; 0.0371, \ldots$$
  - Method: For a large family of functionally related proteins (such as Cytochrome C), we do the following:
    - Use *cross-correlation* to find a *consensus model* for the entire family
    - Cross-Correlation converts two sequences $\{x_k\}$, $\{y_k\}$ into a sequence $\{w_k\}$ via
      $$w_k = \sum_{j=1}^{n} x_j \cdot y_{j+k}$$
    - Consensus model is a model for the entire functional family that is formed from cross-correlations
    - Power spectrum of consensus model then reveals the “fundamental frequency” of the protein family

Example: Electron Ion Interaction Potententials

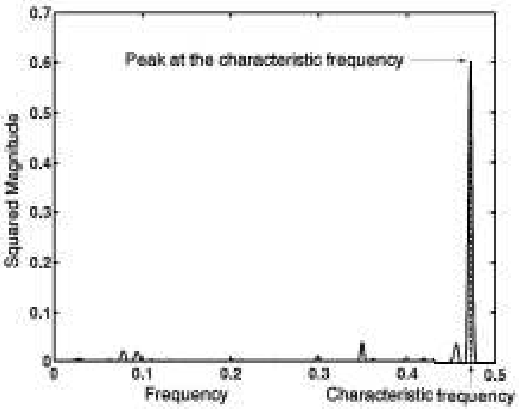


Figure : Consensus spectrum of the Cytochrome C functional group (EIIP) (from Ramachandran, 2008)

But are we loosing too much information?

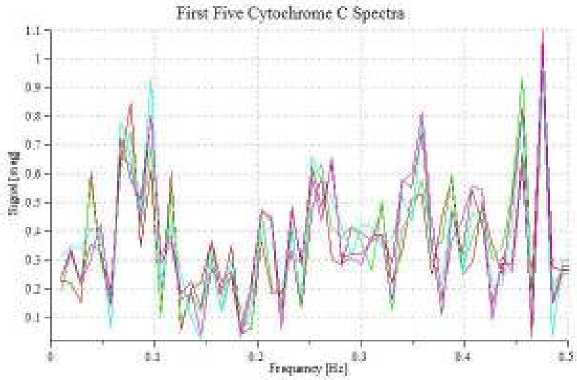


Figure : Spectra of the First Five Cytochrome C Sequences

But are we loosing too much information?

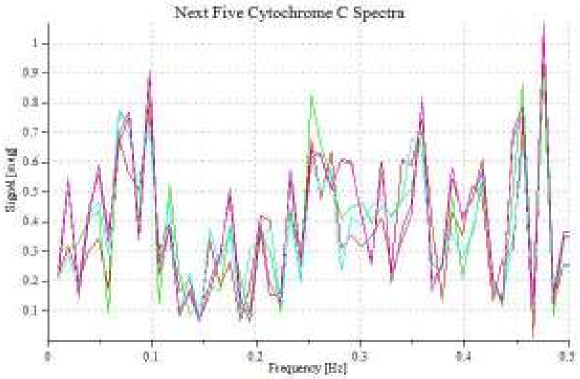


Figure : Spectra of the Next Five Cytochrome C Sequences

But are we loosing too much information?

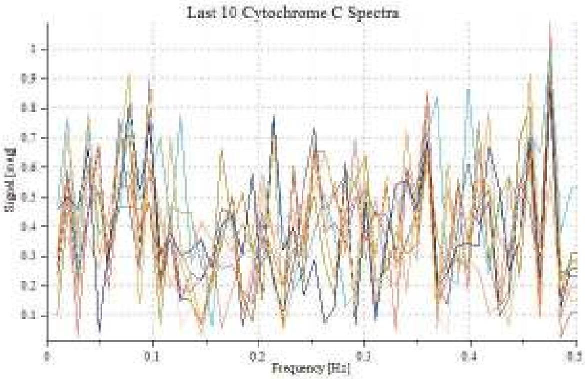


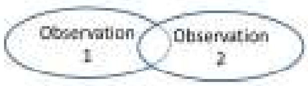
Figure : Spectra of the Last 10 Cytochrome C Sequences

Generalizing the Method

- Use a resampling approach
  - Randomly choose a subset of sequences
  - Compute consensus spectrum (via $P_1(\omega) P_2(\omega) \cdot \ldots \cdot P_m(\omega)$)
  - The average over resamples is consensus spectrum
- Issue: No way to map the consensus spectrum back to the sequence domain
  - Use something other than the product
  - That vanishes if any datum vanishes
  - In particular, use some type of mean

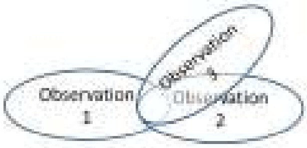
The Arithmetic or Zero Mean (AZM)

- (Spectral) average over the intersection of the (frequency) support of the observations.
- That is, we only want averages to reflect processes common throughout the observations (the homologous family)



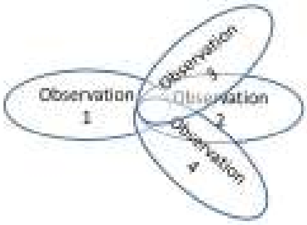
The Arithmetic or Zero Mean (AZM)

- (Spectral) average over the intersection of the (frequency) support of the observations.
- That is, we only want averages to reflect processes common throughout the observations (the homologous family)



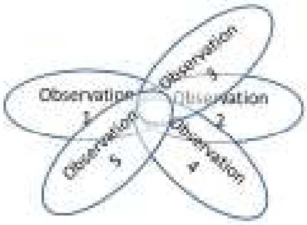
The Arithmetic or Zero Mean (AZM)

- (Spectral) average over the intersection of the (frequency) support of the observations.
- That is, we only want averages to reflect processes common throughout the observations (the homologous family)



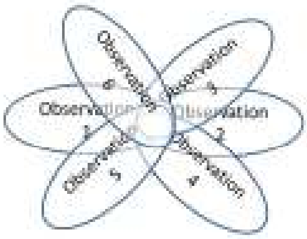
The Arithmetic or Zero Mean (AZM)

- (Spectral) average over the intersection of the (frequency) support of the observations.
- That is, we only want averages to reflect processes common throughout the observations (the homologous family)



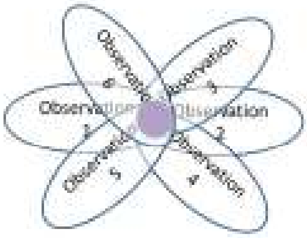
The Arithmetic or Zero Mean (AZM)

- (Spectral) average over the intersection of the (frequency) support of the observations.
- That is, we only want averages to reflect processes common throughout the observations (the homologous family)



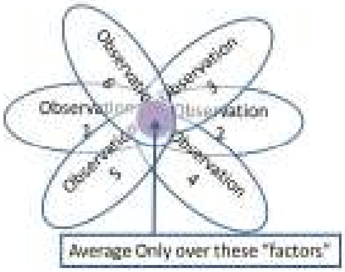
The Arithmetic or Zero Mean (AZM)

- (Spectral) average over the intersection of the (frequency) support of the observations.
- That is, we only want averages to reflect processes common throughout the observations (the homologous family)



The Arithmetic or Zero Mean (AZM)

- (Spectral) average over the intersection of the (frequency) support of the observations.
- That is, we only want averages to reflect processes common throughout the observations (the homologous family)



Average Only over these "factors."

The AZM

- Desired Properties of an AZM:
  - If $|x_i| \approx |x_j|$, $i, j = 1, \ldots, n$, and $|x_i| \not\approx 0$, then
    $$AZM(x_1, \ldots x_n) \approx \frac{x_1 + \ldots + x_n}{n}$$
  - If $x_i = 0$ for some $i = 1, \ldots n$, and $p = 1$, then
    $$AZM(x_1, \ldots x_n) = 0$$
  - If $x_r \approx 0$ for some r in $\{1, \ldots, n\}$, then
    $$AZM(x_1, \ldots x_n) \approx 0$$
  - The function $AZM(x_1, \ldots, x_n)$ is smooth

General Structure

- Let $w : \quad \to \quad$ such that
  $$w(0) = 0, \ w'(0) \neq 0,$$
  with $w(x) \quad 0$ if $x \neq 0$ and
  $$\lim_{x \to \infty} w(x) = 1$$
- For such a w, define
  $$AZM_w(x_1, \ldots x_n) = \begin{cases} 0 & \textit{if} \quad x_j = 0 \textit{ for some } j \\ \dfrac{\sum_{j=1}^{n} x_j \prod_{k \neq j} w(x_k)}{\sum_{j=1}^{n} \prod_{k \neq j} w(x_k)} & \textit{otherwise} \end{cases}$$

Expected Value

- Simplifies to
  $$AZM_w(\mathbf{x}) = \begin{cases} 0 & \textit{if} \quad x_j = 0 \textit{ for some } j \\ \dfrac{\sum_{j=1}^{n} \frac{x_j}{w(x_j)}}{\sum_{j=1}^{n} \frac{1}{w(x_j)}} & \textit{otherwise} \end{cases}$$
- Natural to define the expected value of a function f by
  $$E(f(\mathbf{x})) = \begin{cases} 0 & \textit{if} \quad x_j = 0 \textit{ for some } j \\ \dfrac{\sum_{j=1}^{n} \frac{f(x_j)}{w(x_j)}}{\sum_{j=1}^{n} \frac{1}{w(x_j)}} & \textit{otherwise} \end{cases}$$

Generalized Consensus Modeling

- Given a homologous family $\{x_i\}_{i=1}^{m}$ (residue chain or DNA or networks or...)
  - **Spectral:** The family w.r.t. to a new basis $\{e_j\}_{j=1}^{n}$as
    $$x_i = \sum a_{i,j} e_j$$
  - **BootStrapping:** For each j, bootstrap over $AZM(a_{i_1,j}, \ldots, a_{i_s,j})$, where s is the sample size for each bootstrap resample
  - **Spectrum:** If c_j is the bootstrap mean, then the sequence $\{c_j\}_{j=1}^{n}$ is the *consensus spectrum* of the homologous family
  - **Model:** The consensus model is then defined to be
    $$C = \sum c_j e_j$$
- Repeat for classes, use projections, etc, to produce an ICA-like representation of the data

AZM + Resampling = Consensus Spectrum

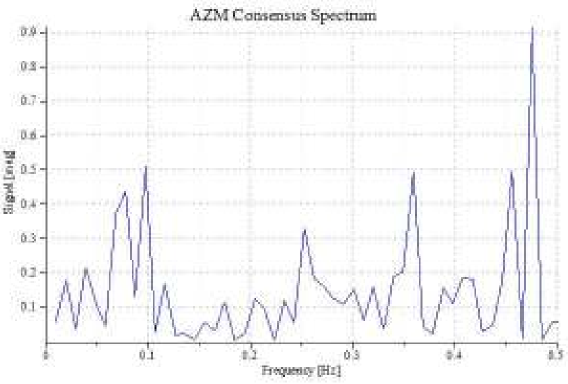


Figure : Consensus spectrum of the Cytochrome C functional group (EIIP)

Consensus Spectrum for a different w(x)

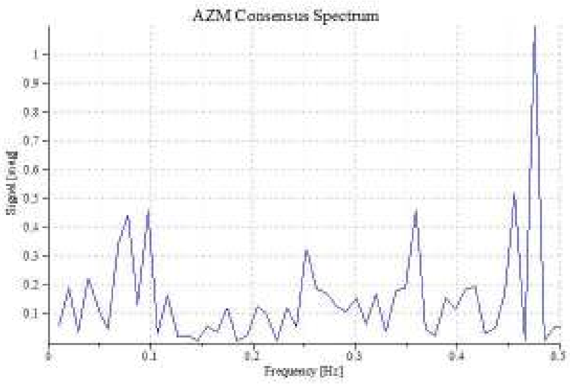


Figure : Another Consensus spectrum of the Cytochrome C functional group (EIIP)

Another Approach

- For fixed $p \quad 0, p \approx 0$ define

$$AZM\left(x_1,\ldots x_n\right)=\begin{cases} 0 & \text{if} \quad x_j = 0 \text{ for some } j \\ \dfrac{\sum_{j=1}^{n} x_j \left|\prod_{k\neq j} x_k\right|^p}{\sum_{j=1}^{n}\left|\prod_{k\neq j} x_k\right|^p} & \text{otherwise}\end{cases}$$

  Useful, but it is not smooth.
- Examples: ($a \neq 0,\ b \neq 0,\ c \neq 0$)

$$\begin{aligned} AZM\left(a,b\right) &= \frac{a\left|b\right|^p+\left|a\right|^p b}{\left|b\right|^p+\left|a\right|^p} = \frac{a/\left|a\right|^p+b/\left|b\right|^p}{1/\left|a\right|^p+1/\left|b\right|^p} \\ AZM\left(a,b,c\right) &= \frac{a\left|bc\right|^p+b\left|ac\right|^p+c\left|ab\right|^p}{\left|bc\right|^p+\left|ac\right|^p+\left|ab\right|^p}\end{aligned}$$

Another Approach

- In general, we can write

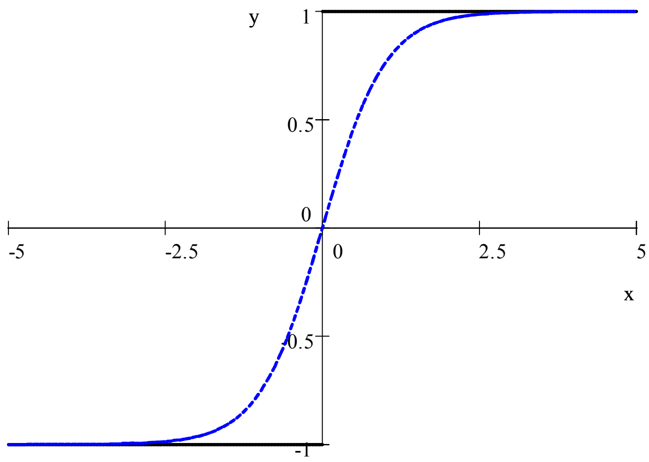
$$AZM\left(x_1,\ldots x_n\right)=\begin{cases} 0 & \text{if} \quad x_j = 0 \text{ for some } j \\ \left(\sum_{j=1}^{n}\frac{1}{\left|x_j\right|^p}\right)^{-1}\sum_{j=1}^{n}\frac{x_j}{\left|x_j\right|^p} & \text{otherwise}\end{cases}$$

- If $p = 1$ and $x_j \neq 0$ for all $j = 1,\ldots,n$, then

$$AZM\left(x_1,\ldots x_n\right) = H\left(\left|x_1\right|^p,\ldots,\left|x_n\right|^p\right)\left(\frac{1}{n}\sum_{j=1}^{n}\frac{x_j}{\left|x_k\right|^p}\right)$$

  where $H\left(x_1,\ldots x_n\right)$ is the *harmonic mean*.

Notice that $\frac{x}{|x|} \approx \tanh\left(x\right)$ for $x \not\approx 0$



The AZM

- Consequently, we define (one variation – there are many)

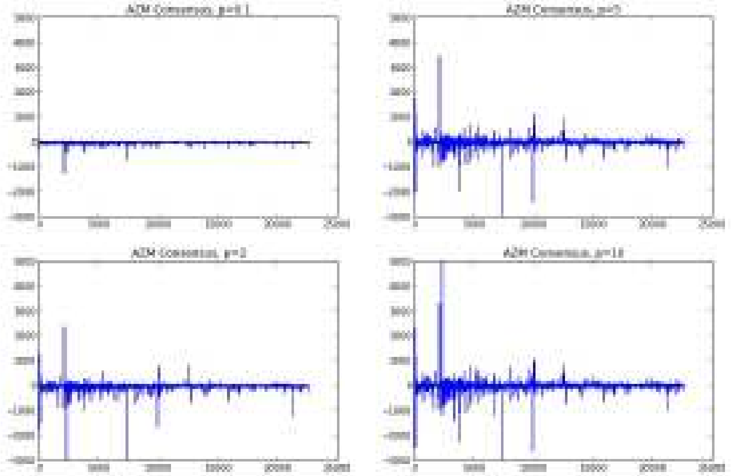
$$AZM\left(x_1,\ldots x_n\right) = H\left(\left|x_1\right|^p,\ldots,\left|x_n\right|^p\right)\sum_{j=1}^{n}\left|x_k\right|^{1-p}\tanh\left(\alpha x_j\right) \qquad (1)$$

  - This is a smooth function at 0
  - The parameter α can be used to *tune* the AZM
- Equation (1) looks like a simple *neural network*

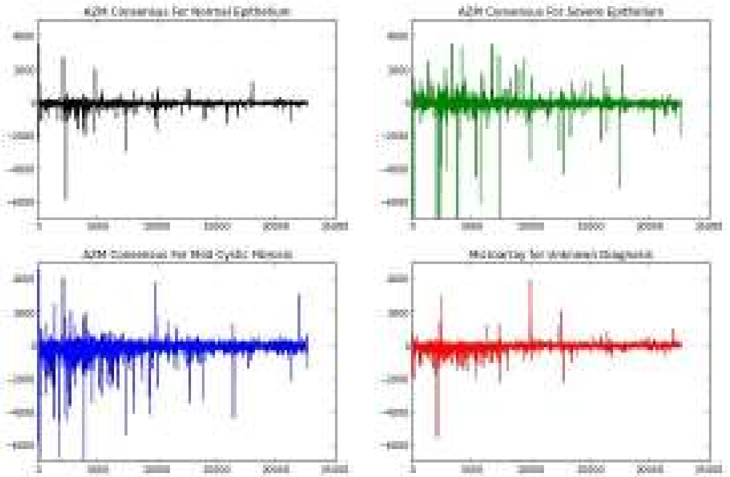
Example: Microarray Data

From (Wright, 2006):
This investigation used high-density oligonucleotide microarray analysis of nasal respiratory epithelium to investigate the molecular basis of phenotypic differences in CF by (1) identifying differences in gene expression between DeltaF508 homozygotes in the most severe 20th percentile of lung disease by forced expiratory volume in 1 s and those in the most mild 20th percentile of lung disease and (2) identifying differences in gene expression between DeltaF508 homozygotes and age-matched non-CF control subjects.
20 subjects – Microarray Data from Epithelium
3 classes – Normal (11), Mild (4), Severe(5)

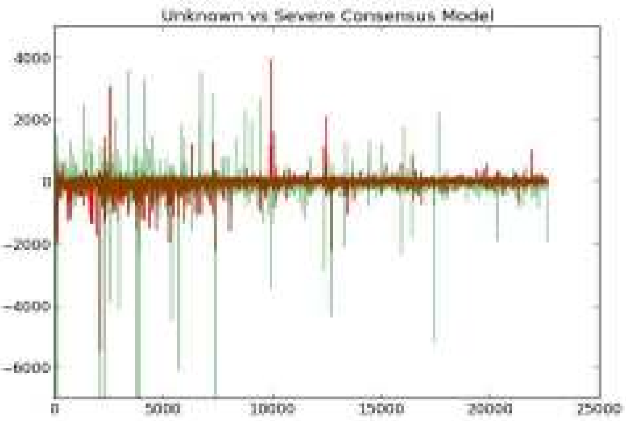
Effects of Parameter p – All Subjects



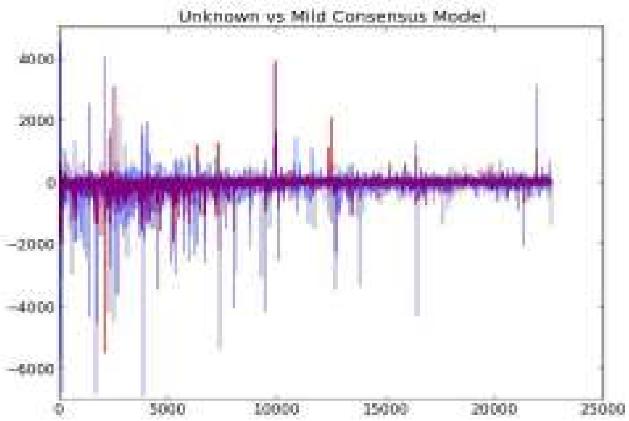
The 3 diagnoses



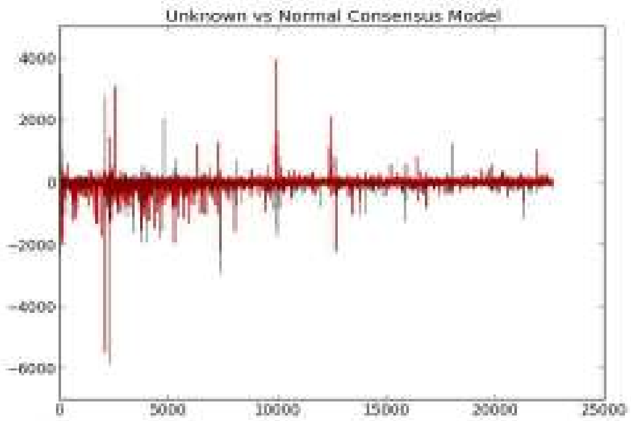
Unknown versus Severe Cystic Fibrosis



Unknown Versus Mild Cystic Fibrosis



Unknown Versus Normal



Conclusions

- Consensus Spectrum is based on Signal Processing and Bio-electrical Parameters
  - Can be generalized to a Linear Algebra Context (change of basis)
  - Can be based on (some) non-physical measurements
- Consensus Models + Spectrum extend the Original by
  - Incorporating resampling with a non-standard mean (AZM)
  - Using vertex-weighted graph invariants instead of biophysical parameters
- Future Directions: Characterizing what a consensus model tells us mathematically about a homologous family of sequences

Any Questions

Thank you!

- Rushdi A, Tuqan J. *The role of the symbolic-to-numerical mapping in the detection of DNA periodicities.* Proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics, GENSIPS '08 2008, 1-4.
- Datta S, Asif A, Wang H. *Prediction of protein coding regions in DNA sequences using Fourier spectral characteristics.* Proceedings of the IEEE Sixth International Symposium on Multimedia Software, 2004, 160-63.
- P. Ramachandran, A. Antoniou, *Identification of Hot Spot Locations in Proteins Using Digital Filters*, *IEEE Journal of Selected Topics in Signal Processing*, Volume 2, issue 3, Pages 378-389, 2008.
- A. L. Rockwood, D. K. Crocket, J. R. Oliphant, and K. S. J. Elenitoba-Johnson, *Sequence Alignment by Cross-Correlation*, *Journal of Biomolecular Techniques*, Volume 16, Pages 453-458, 2005.
- A. Sabarish and T. Thomas, *A Frequency Domain Approach to Protein Sequence Similarity Analysis and Functional Classification*,

Signal and Image Processing: An International Journal(SIPIJ), Vol.2, No.1, March 2011.

- C. H. Trad, Q. Fang, and I. Cosic, *Protein Sequence Comparison Based on the Wavelet Transform Approach*, *Protein Engineering*, vol. 15, no.3, Pages 193-203, 2002.
- Wright JM, Merlo CA, Reynolds JB, Zeitlin PL, Garcia JG, Guggino WB, Boyle MP. *Respiratory epithelial gene expression in patients with mild and severe cystic fibrosis lung disease.* Am J Respir Cell Mol Biol. 2006 Sep;35(3):327-36. Epub 2006 Apr 13.
- P. P. Vaidyanathan and B. Yoon, *The Role of Signal-Processing Concepts in Genomics and Proteomics*, *Journal of the Franklin Institute*, 341 (2004), Pages 111-135.
- Y. Yadav and S. Wadhwani, *Determination of Characteristic Frequency for Identification of Hot Spots in Proteins*, *International Journal of Electrical and Electronics Engineering* , Volume 1, Issue 1, 2011.