Predicting Protein Structure and Examining Similarities of Protein Structure by Spectral Analysis Techniques

Hong Gu, Melanie Abeysundera, Krista Collins, Chris Field

Department of Mathematics and Statistics
Dalhousie University

ANU BioinfoSummer2006, Dec.4-8, 2006

Outline

Spectral envelope method

- The spectral envelope method (Stoffer et al., 2000) is a frequency domain analysis of categorical time series.
- Categorical time series, X_t, $t = 0, \pm 1, \pm 2, \ldots$ with the finite state space $C = \{c_1, c_2, \ldots, c_k\}$ can be represented by the k-dimensional time series $\mathbf{Y}_t$ through a set of index variables $e_1, e_2, \ldots, e_k$.
- $\mathbf{Y}_t$ can be converted back to a univariate real valued time series through $X_t(\beta) = \beta' \mathbf{Y}_t$.
- The spectral envelope chooses the scalings at each frequency to maximize the periodicity of $X_t(\beta)$:

$$\lambda(\omega) = \sup_{\beta} \left\{ \frac{\beta' f_Y^{re}(\omega) \beta}{\beta' V \beta} \right\},$$

where $f_Y^{re}(\omega)$ is the real portion of the spectral density of $\mathbf{Y}_t$.

Spectral envelope method

- The spectral envelope method (Stoffer et al., 2000) is a frequency domain analysis of categorical time series.
- Categorical time series, X_t, $t = 0, \pm 1, \pm 2, \ldots$ with the finite state space $C = \{c_1, c_2, \ldots, c_k\}$ can be represented by the k-dimensional time series $\mathbf{Y}_t$ through a set of index variables $e_1, e_2, \ldots, e_k$.
- $\mathbf{Y}_t$ can be converted back to a univariate real valued time series through $X_t(\beta) = \beta' \mathbf{Y}_t$.
- The spectral envelope chooses the scalings at each frequency to maximize the periodicity of $X_t(\beta)$:

$$\lambda(\omega) = \sup_{\beta} \left\{ \frac{\beta' f_Y^{re}(\omega) \beta}{\beta' V \beta} \right\},$$

where $f_Y^{re}(\omega)$ is the real portion of the spectral density of $\mathbf{Y}_t$.

Spectral envelope method

- The spectral envelope method (Stoffer et al., 2000) is a frequency domain analysis of categorical time series.
- Categorical time series, X_t, $t = 0, \pm 1, \pm 2, \ldots$ with the finite state space $C = \{c_1, c_2, \ldots, c_k\}$ can be represented by the k-dimensional time series $\mathbf{Y}_t$ through a set of index variables $e_1, e_2, \ldots, e_k$.
- $\mathbf{Y}_t$ can be converted back to a univariate real valued time series through $X_t(\beta) = \beta' \mathbf{Y}_t$.
- The spectral envelope chooses the scalings at each frequency to maximize the periodicity of $X_t(\beta)$:

$$\lambda(\omega) = \sup_{\beta} \left\{ \frac{\beta' f_Y^{re}(\omega) \beta}{\beta' V \beta} \right\},$$

where $f_Y^{re}(\omega)$ is the real portion of the spectral density of $\mathbf{Y}_t$.

Spectral envelope method

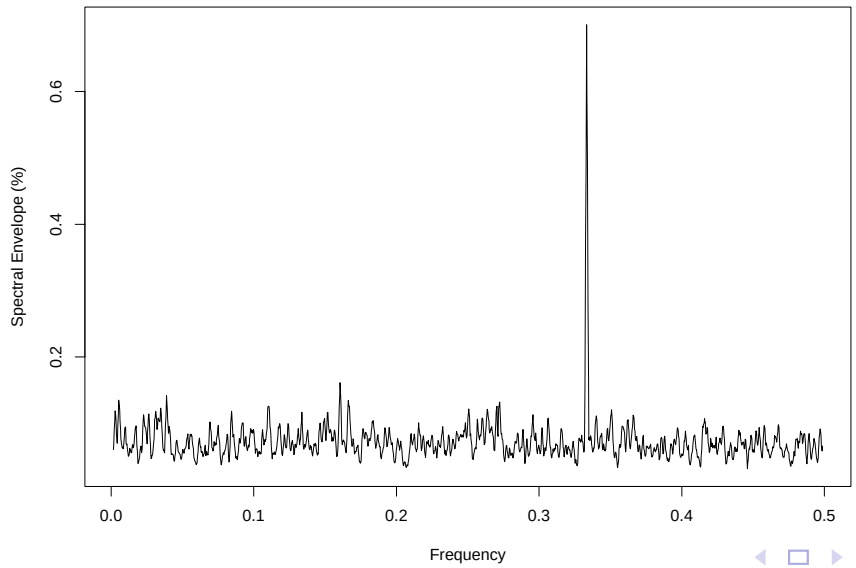
- The spectral envelope method (Stoffer et al., 2000) is a frequency domain analysis of categorical time series.
- Categorical time series, X_t, $t = 0, \pm 1, \pm 2, \ldots$ with the finite state space $C = \{c_1, c_2, \ldots, c_k\}$ can be represented by the k-dimensional time series $\mathbf{Y}_t$ through a set of index variables $e_1, e_2, \ldots, e_k$.
- $\mathbf{Y}_t$ can be converted back to a univariate real valued time series through $X_t(\beta) = \beta' \mathbf{Y}_t$.
- The spectral envelope chooses the scalings at each frequency to maximize the periodicity of $X_t(\beta)$:

$$\lambda(\omega) = \sup_{\beta} \left\{ \frac{\beta' f_Y^{re}(\omega) \beta}{\beta' V \beta} \right\},$$

where $f_Y^{re}(\omega)$ is the real portion of the spectral density of $\mathbf{Y_t}$.

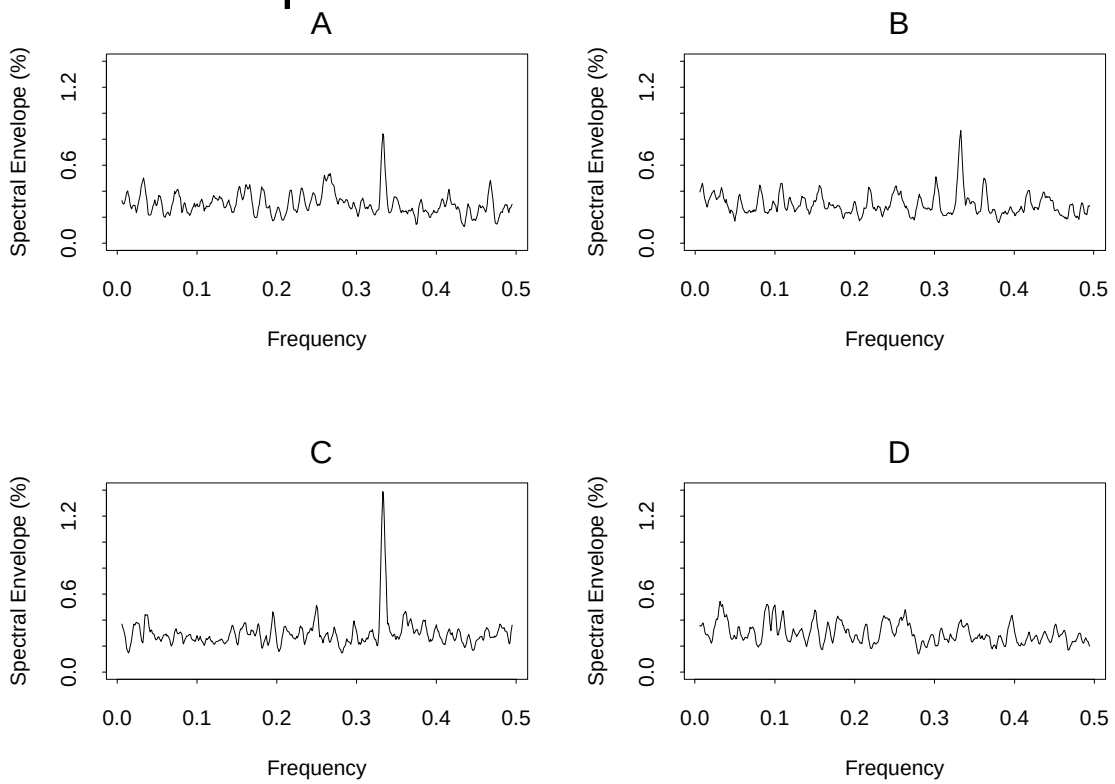
Spectral envelope application in DNA sequences

- The spectral envelopes of protein coding DNA sequences indicate that each gene contains a strong signal at a cycle of approximately 3 (Stoffer et al.,1993, 1998, 2000), which corresponds to the codons.
- DNA sequence example: Spectral Envelope of the Epstein-Barr virus gene BNFR1 human herpesvirus 4 (bp 1736 to 3957).



Spectral envelope application in DNA sequences

- Use spectral envelope to search for coding and non-coding regions in DNA sequences.
- Spectral Envelope of the Epstein-Barr virus gene BNRF1 partitioned into 4 quarters.



Protein structure

- The basic principal in protein folding is to pack the hydrophobic side chains into the interior of the molecule with the hydrophilic side chains on the surface.
- The structure hierarchy of proteins
  - Primary structure: amino acid sequence.
  - Secondary structure: local folding determined by regions of the protein's amino acid sequence.
  - Tertiary structure: arrangement of secondary structures into domains or motifs.
  - Quaternary structure: interaction of two or more tertiary structures to form a functional region.

Protein structure

- The basic principal in protein folding is to pack the hydrophobic side chains into the interior of the molecule with the hydrophilic side chains on the surface.
- The structure hierarchy of proteins
  - Primary structure: amino acid sequence.
  - Secondary structure: local folding determined by regions of the protein's amino acid sequence.
  - Tertiary structure: arrangement of secondary structures into domains or motifs.
  - Quaternary structure: interaction of two or more tertiary structures to form a functional region.

Protein structure

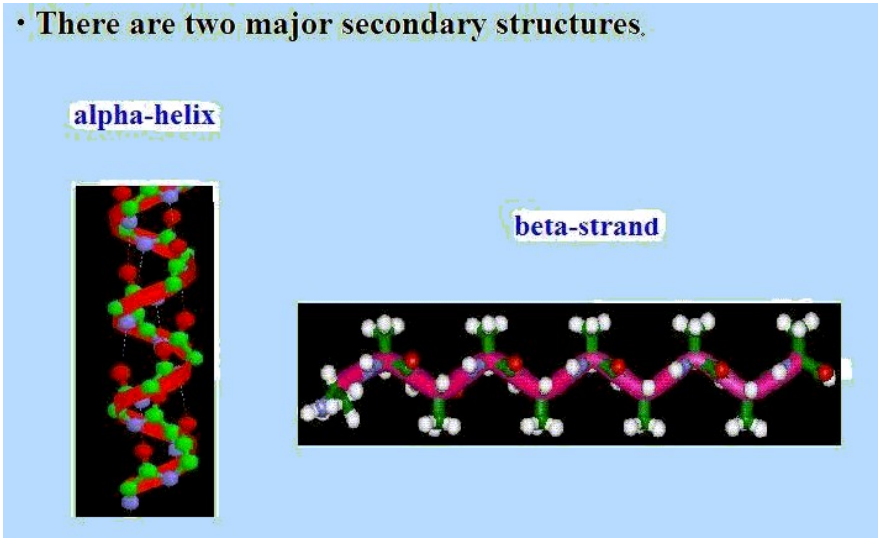
- The basic principal in protein folding is to pack the hydrophobic side chains into the interior of the molecule with the hydrophilic side chains on the surface.
- The structure hierarchy of proteins
  - Primary structure: amino acid sequence.
  - Secondary structure: local folding determined by regions of the protein's amino acid sequence.
  - Tertiary structure: arrangement of secondary structures into domains or motifs.
  - Quaternary structure: interaction of two or more tertiary structures to form a functional region.

Protein structure

- The basic principal in protein folding is to pack the hydrophobic side chains into the interior of the molecule with the hydrophilic side chains on the surface.
- The structure hierarchy of proteins
  - Primary structure: amino acid sequence.
  - Secondary structure: local folding determined by regions of the protein's amino acid sequence.
  - Tertiary structure: arrangement of secondary structures into domains or motifs.
  - Quaternary structure: interaction of two or more tertiary structures to form a functional region.

Protein secondary structure

The secondary structure of most proteins is composed of combinations of α-helices and β-sheets which are connected by loop regions or turns.



Spectral envelope and protein secondary structure

The secondary structural elements of proteins are a natural application of the spectral envelope:

- α-helices-3.6 residues, spectral envelope peak at $\omega = 0.277$.
- β-sheets- 2.3 residues, a peak at $\omega = 0.435$.
- Others: turns have a period of 4 residues and 3_{10}-helices have an ideal periodicity of 3.

Spectral envelope and protein secondary structure

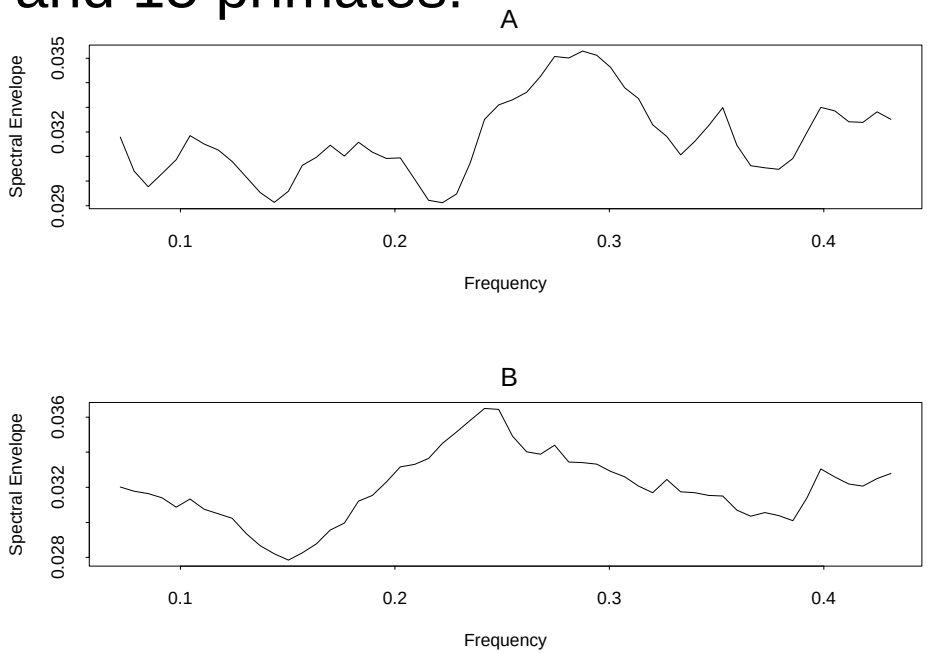
- Secondary structure motifs: combinations of α-helices, β-sheets and loops. e.g. helix-loop-helix motif and the β-α-β motif.
- The combination of motifs with other secondary structural elements makes the domains of protein tertiary structure.
- The secondary structure within a domain is the repetition of structural elements and motifs.
- The secondary structure of an entire protein should illustrate periodicity not only in the structural elements but also in the repeated motifs.
- The spectral envelope should reveal multiple peaks corresponding to different periodicities.

Spectral envelope and protein secondary structure

- Secondary structure motifs: combinations of α-helices, β-sheets and loops. e.g. helix-loop-helix motif and the β-α-β motif.
- The combination of motifs with other secondary structural elements makes the domains of protein tertiary structure.
- The secondary structure within a domain is the repetition of structural elements and motifs.
- The secondary structure of an entire protein should illustrate periodicity not only in the structural elements but also in the repeated motifs.
- The spectral envelope should reveal multiple peaks corresponding to different periodicities.

Example: the myoglobin sequences

The average spectral envelope of myoglobin sequences from 13 cetaceans and 15 primates:



- Cetaceans peak at $\omega = 0.288$. The primates peak at $\omega = 0.242$. Both are indicative of the α-helices.
- The difference may reflect small changes in the structure of the proteins.

Protein structure and its prediction

- Among vast amounts of protein sequence data, only a small number of protein structures have been determined by
  - X-ray chrystallography (for the majority of structures available in protein databanks).
  - NMR (Protein nuclear magnetic resonance spectroscopy)
- The prediction of protein structure remains an unsolved problem. Existing methods are mainly comparative:
  - Homology modelling (sequence alignment)
  - Protein threading

Some issues with comparative models

- A protein sequence may assume multiple conformations depending on location on the genome – Primary structure may not totally decide higher order structure.
- Some proteins are structurally similar but sequentially unrelated – comparative methods often invalid.

Some issues with comparative models

- A protein sequence may assume multiple conformations depending on location on the genome – Primary structure may not totally decide higher order structure.
- Some proteins are structurally similar but sequentially unrelated – comparative methods often invalid.

Predicting protein structure by spectral techniques

- The method requires no comparison between sequences.
- Predictors extracted from the spectral envelope:
  - An 80% bootstrap confidence threshold was obtained for each frequency and the median taken to get a single threshold value.
  - The spectra of the proteins were divided into p (p=100) frequency bands and the mean values above the threshold in each band was calculated as predictors.
- The problem is formulated as a standard supervised learning problem.
- CART (classification and regression tree) is used to demonstrate the results of such formulation.

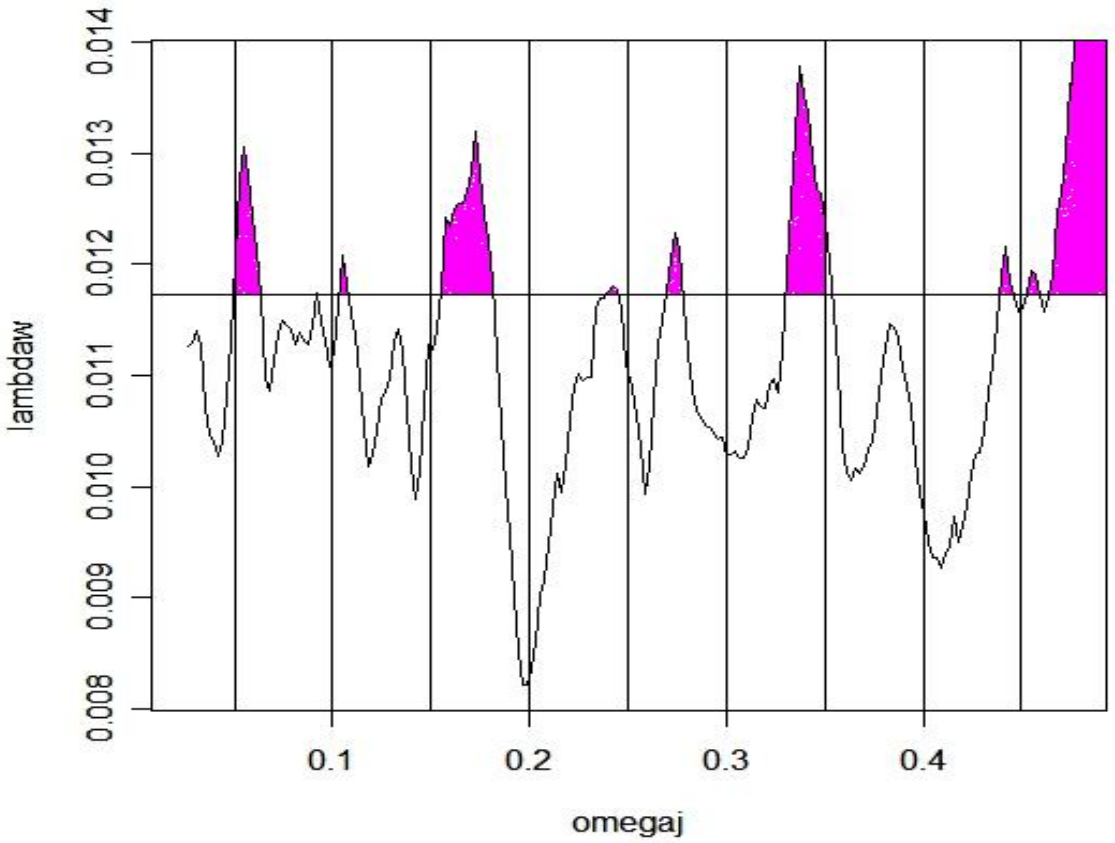
Predicting protein structure by spectral techniques

- The method requires no comparison between sequences.
- Predictors extracted from the spectral envelope:
  - An 80% bootstrap confidence threshold was obtained for each frequency and the median taken to get a single threshold value.
  - The spectra of the proteins were divided into p (p=100) frequency bands and the mean values above the threshold in each band was calculated as predictors.
- The problem is formulated as a standard supervised learning problem.
- CART (classification and regression tree) is used to demonstrate the results of such formulation.

Predicting protein structure by spectral techniques

- The method requires no comparison between sequences.
- Predictors extracted from the spectral envelope:
  - An 80% bootstrap confidence threshold was obtained for each frequency and the median taken to get a single threshold value.
  - The spectra of the proteins were divided into p (p=100) frequency bands and the mean values above the threshold in each band was calculated as predictors.
- The problem is formulated as a standard supervised learning problem.
- CART (classification and regression tree) is used to demonstrate the results of such formulation.

Predicting protein structure by spectral techniques



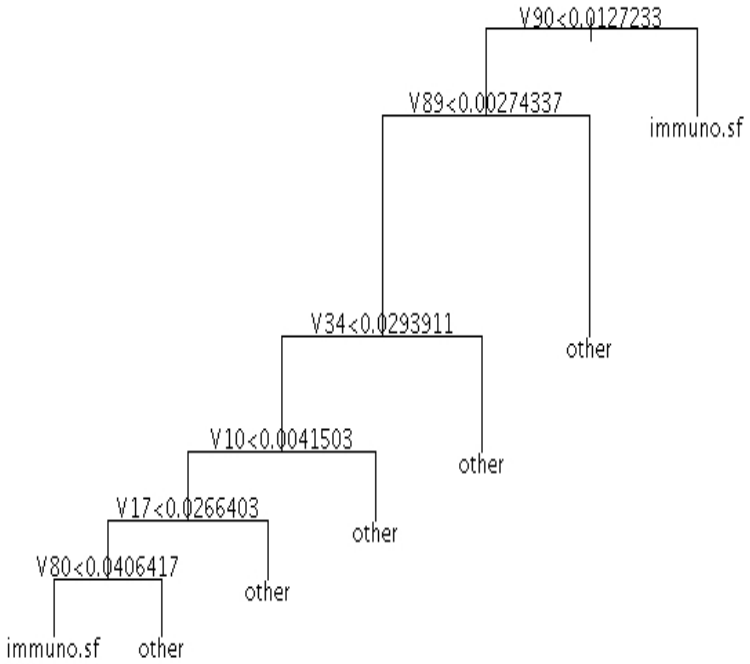
The Structural Hierarchy in SCOP

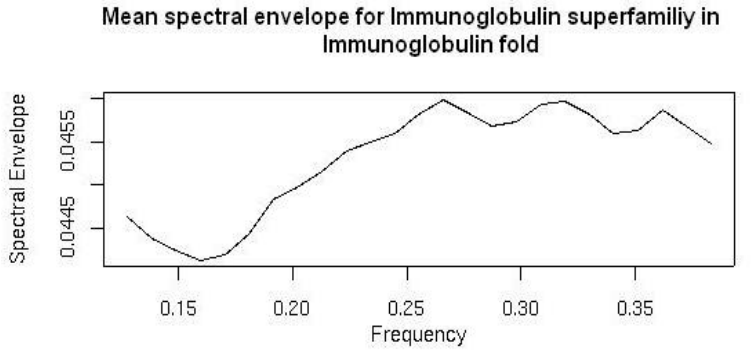
- Class [all alpha proteins (1390), all beta proteins (1578)]

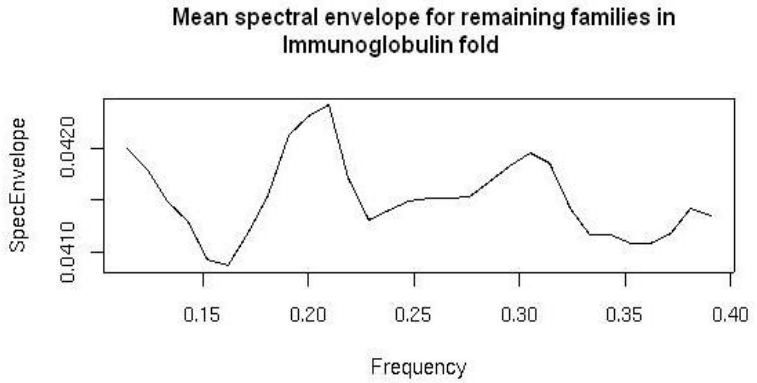
  (Errors: 11.12%, 18.60%, 19.54%, nodes: 43)
- Fold
  - all alpha proteins [Globin-like (516), EF hand-like (334), cytochrome (540)] (11.22%, 22.73%, 22.01%; 38)
  - all beta proteins [Immunoglobulin beta sandwhich (530), SH3-like barrel (322), OB-fold (726)] (18.89%, 29.40%, 30.04%; 43)
- Superfamily
  - Immunoglobulin beta sandwhich [Immunoglobulin (308),Other (222)] (7.93%, 15.09%, 14.84%; 7)
  - OB-fold [Bacterial enterotoxins (498), Other superfamilies (228)] (4.96%, 6.06%, 8.82%; 6)
- Family
  - Immunoglobulin [C1 (270), Other (38)] (7.79%, 11.04%, 16.23%; 3)
  - Bacterial enterotoxins [Bacterial AB5 toxins (318),Superantigen toxins (180)](0.80%, 2.41%, 5.22%; 6)

Classification within Immunoglobulin beta sandwhich

Superfamily: Immunoglobulin (308),Other (222) (6.79%, 7.93%, 16.69%; 7)

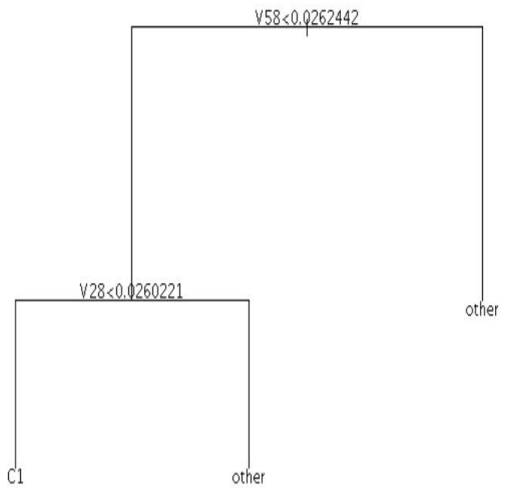




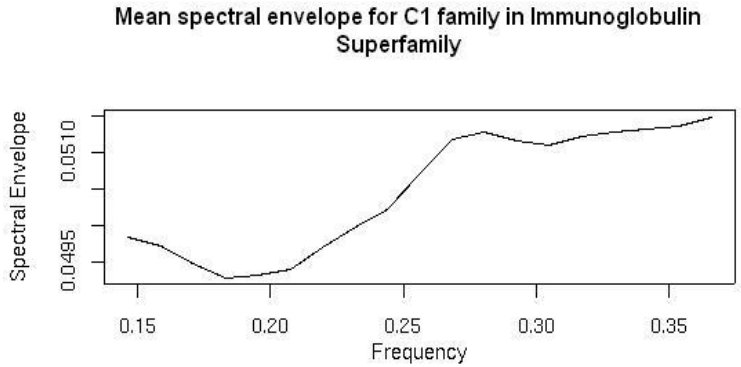


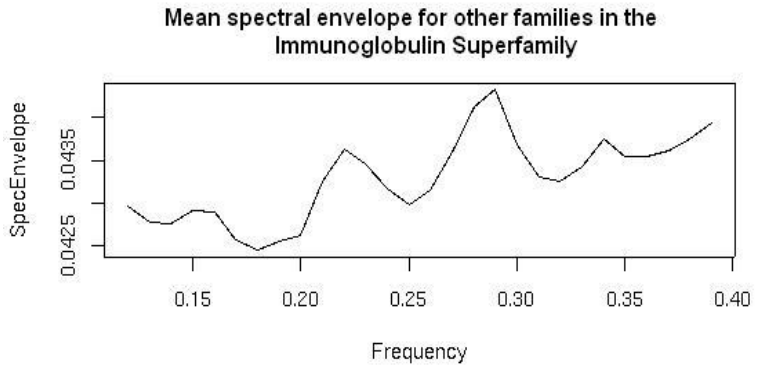
Classification within Immunoglobulin Superfamily

Family: C1 (270), Other (38) (7.14%, 7.14%, 18.83%; 8)



1





Spectral Envelope Covariance: method

- Denote the multivariate time-series of two categorical time sequences X_{1t} and X_{2t} with the same state space as $\mathbf{Y}_{1t}$ and $\mathbf{Y}_{2t}$.
- A common scaling β is assigned to both sequences.
- The squared spectral covariance is

$$\mathrm{Cov}_{12}^2(\omega) = \sup_{\beta:\beta'\beta=1} |\beta' f_{12}(\omega)\beta|^2$$

- f_{12} is the cross-spectral density between $\mathbf{Y}_{1t}$ and $\mathbf{Y}_{2t}$.

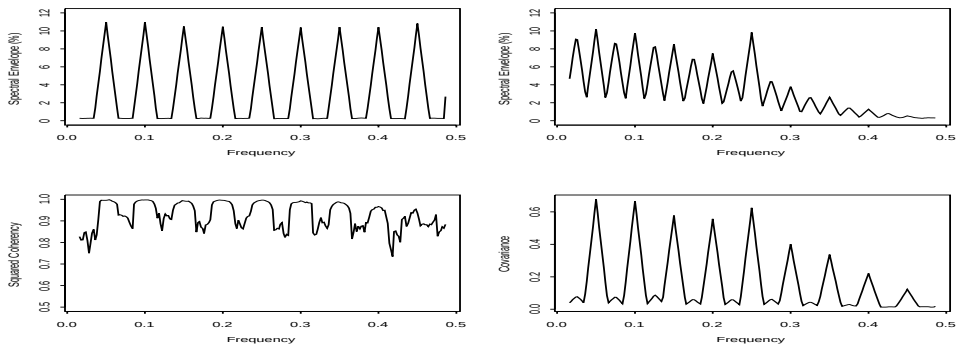
Spectral Envelope Covariance: Computation

- $$\mathrm{Cov}_{12}^2(\omega;\beta) = \sup_{\beta:\beta'\beta=1} \{[\beta' f_{12}^{re}(\omega)\beta]^2 + [\beta' f_{12}^{im}(\omega)\beta]^2\}.$$
- $A_s^{re}(\omega) = [f_{12}^{re}(\omega) + f_{12}^{re}(\omega)']/2$
  $A_s^{im}(\omega) = [f_{12}^{im}(\omega) + f_{12}^{im}(\omega)']/2$
- $$\mathrm{Cov}_{12}^2(\omega;\beta) = \sup_{\beta:\beta'\beta=1} \{[\beta' A_s^{re}(\omega)\beta]^2 + [\beta' A_s^{im}(\omega)\beta]^2\}.$$
- Algorithm:
  1. Initialization: set $\beta_0 \leftarrow \varepsilon_1(A_s^{re})$ or $\varepsilon_1(A_s^{im})$.
  2. Iterate until convergence:

$$\beta_j = \varepsilon_1[A_s^{re}\beta_{j-1}\beta_{j-1}'A_s^{re} + A_s^{im}\beta_{j-1}\beta_{j-1}'A_s^{im}].$$

Spectral covariance and coherency

- The spectral covariance is a non-standardized version of the coherency.
- It maximizes the product of the coherency and the spectral envelopes.
- Avoid the peaks of high coherency generated by random variation.
- A good measure of sequence similarity.



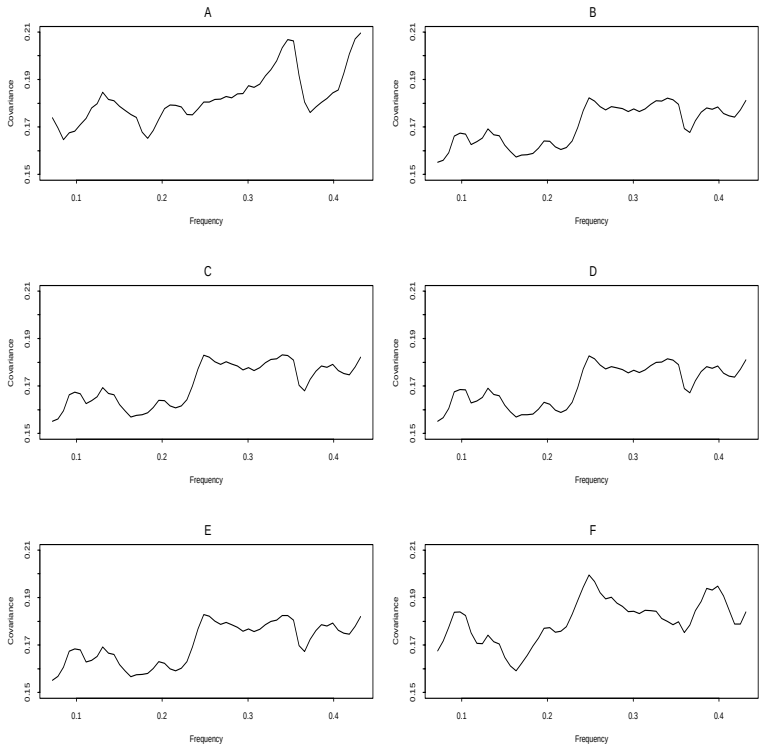
A similarity measure based on spectral covariance

- Threshold: the mean of the upper confidence bounds of 95% bootstrap confidence intervals over all different frequencies between two randomly generated sequences with equal probability of selecting any base.
- The total spectral covariance (after threshold) is calculated for a pair of sequences as a measure of their similarity.
- A dissimilarity measure is defined as:
  $diss(x_i, x_j) = 1 - \frac{sim(x_i, x_j)}{max(sim(x_i, x_j))}$.
- A clustering algorithm (e.g. neighbor-joining tree) can be applied to examine the difference between sequences.

Mammalian myoglobin sequences pairwise spectral covariance

A) minke whale and fin whale, B) minke whale and gorilla, C) minke whale and chimp, D) fin whale and gorilla, E) fin whale and chimp, F) gorilla and chimp.

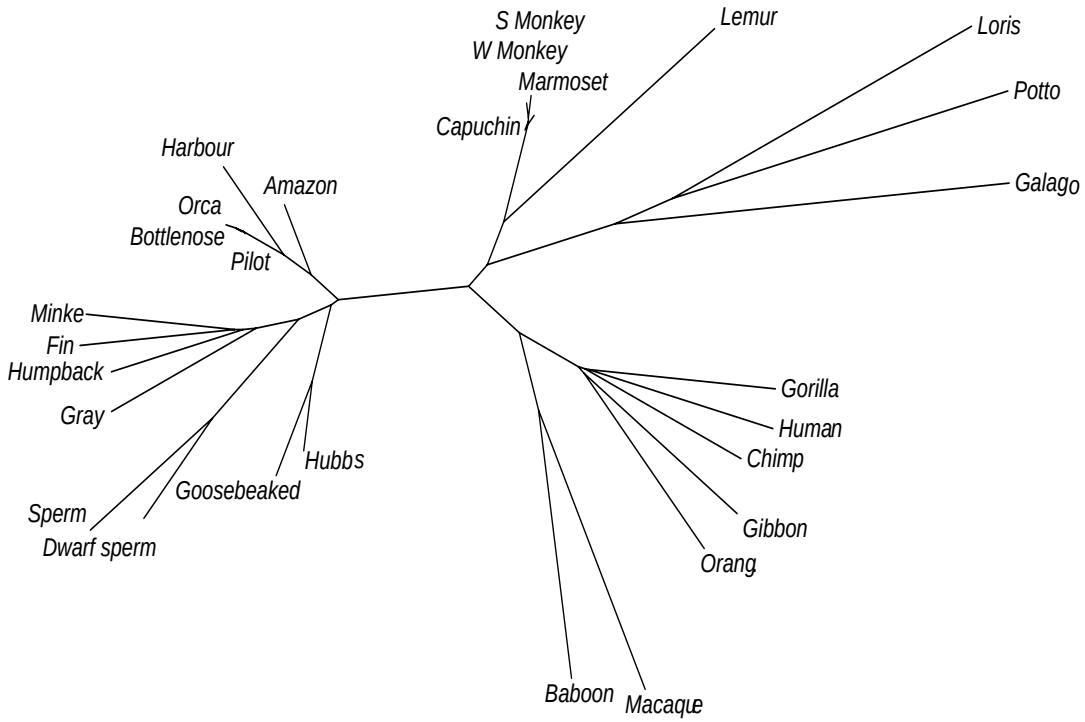
- Two cetacean peak near $\omega = 0.34$.
- Two primate peak near $\omega = 0.24$.
- Between a cetacean and primate: two smaller peaks at $\omega = 0.24$ and 0.34.



Neighbor-joining tree based on spectral covariance

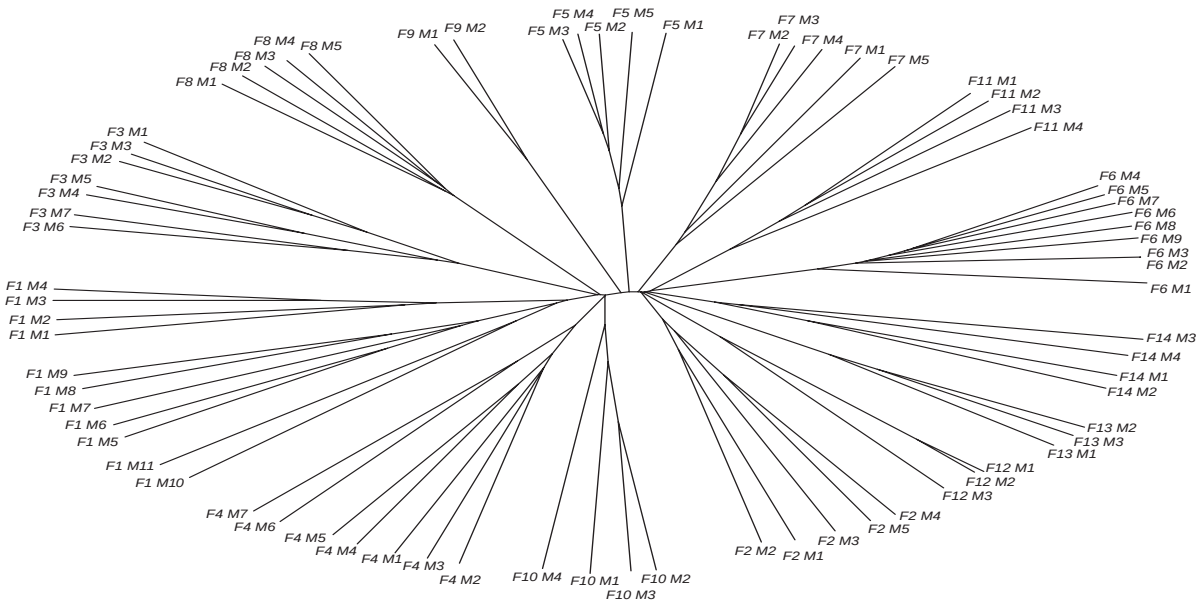
Excellent agreement with the phylogenetic tree (Naylor and Gerstein, 2000) except minor difference (human is closer to the gorilla than to the chimpanzee).

- Phylogenetic tree needs alignment. Spectral covariance tree alignment free.
- phylogenetic tree: taxonomic relationships.
- Spectral covariance tree: structure similarities.



Separate the protein families within a superfamily

- NJ tree of 74 species from different (14) protein families (PANDIT database) clearly separates all families and illustrates the relationships between the families.
- Within the family: comparable to ML phylogenetic trees.
- ML tree among the families not available.



Summary

- Spectral envelope can relate periodicity to protein structure, provides good method in protein structure prediction.
- Spectral covariance can be used to distinguish proteins within and between protein families.
- ML tree and spectral covariance tree not always consistent
  - ML tree is based on the site patterns from the alignment and the model assumes the independence of all sites and a specific Markov process involved in the evolution.
  - The spectral method is alignment free, not assuming independence, but makes use of the correlations between the sites. It provides a way to examine sequences with a lot of similarity in the structure but are not well aligned.

Summary

- Spectral envelope can relate periodicity to protein structure, provides good method in protein structure prediction.
- Spectral covariance can be used to distinguish proteins within and between protein families.
- ML tree and spectral covariance tree not always consistent
  - ML tree is based on the site patterns from the alignment and the model assumes the independence of all sites and a specific Markov process involved in the evolution.
  - The spectral method is alignment free, not assuming independence, but makes use of the correlations between the sites. It provides a way to examine sequences with a lot of similarity in the structure but are not well aligned.