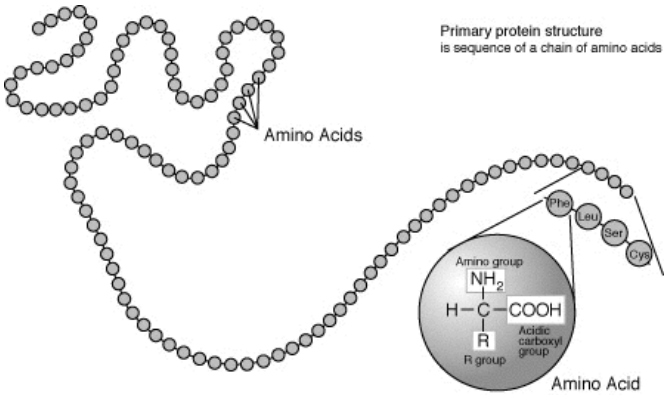
Joint Estimation of Multiple Bivariate Densities of Protein Backbone Angles using an Adaptive Exponential Spline Family

Mehdi Maadooliat
mehdi@mscs.mu.edu

Department of Mathematics, Statistics and Computer Science
Marquette University

November 18, 2013

What is a Protein? (Primary Structure)



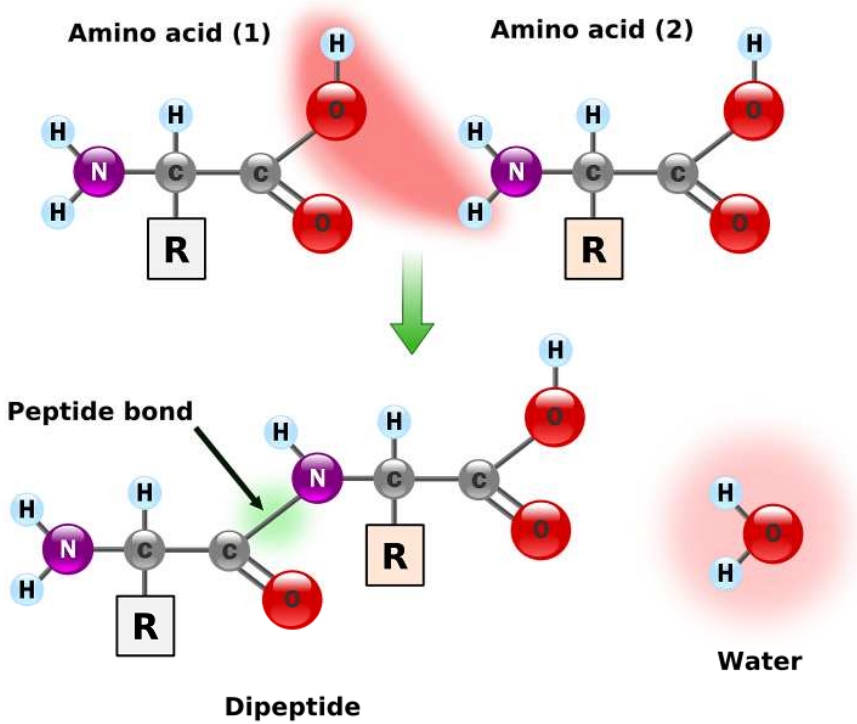
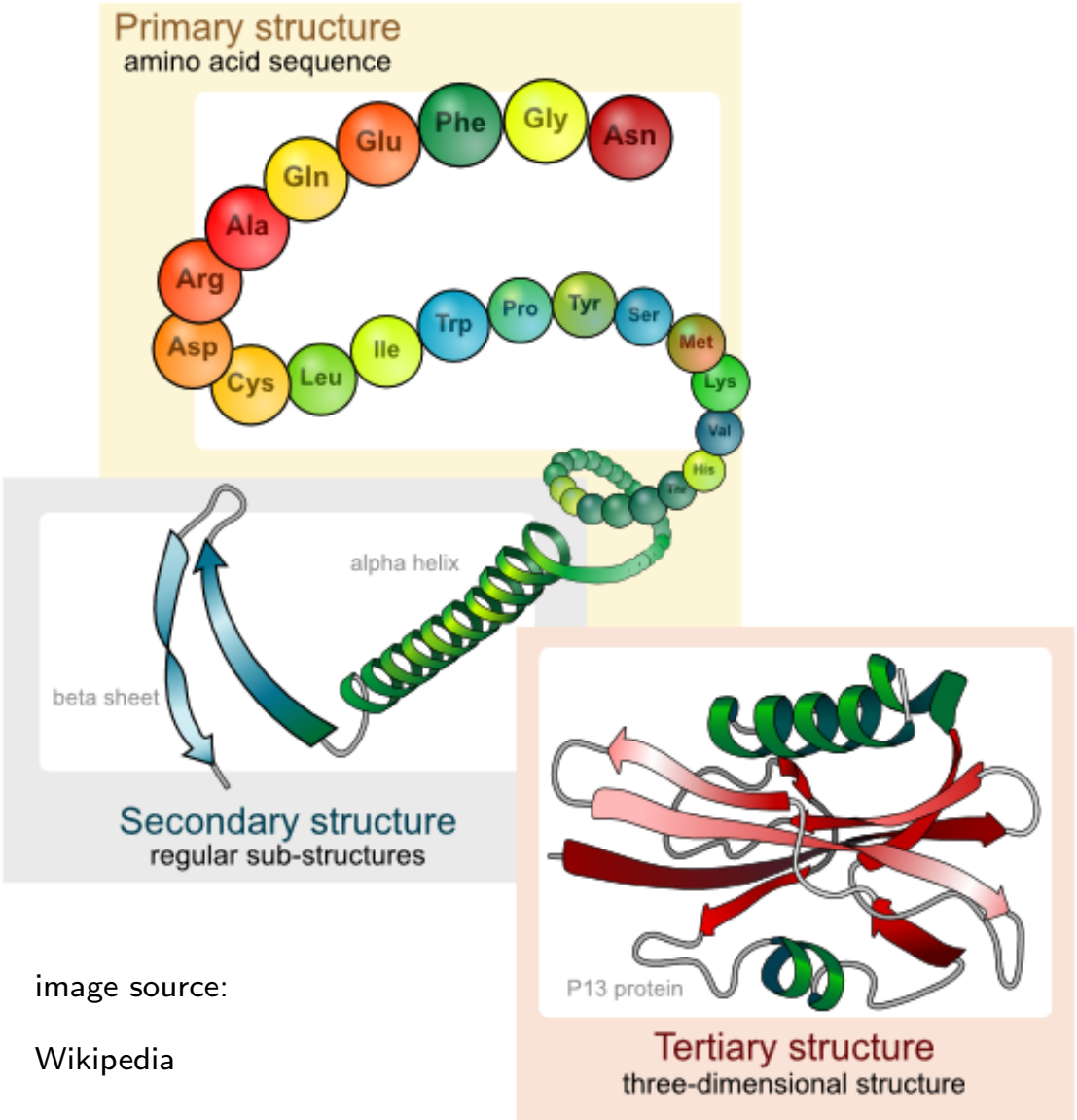


image source: Wikipedia

What is a Protein? (3D Structure)



- The Protein Data Bank (PDB) is a **repository** for the **3D structure** of proteins

image source:

Wikipedia

What is a Protein? (3D Structure)

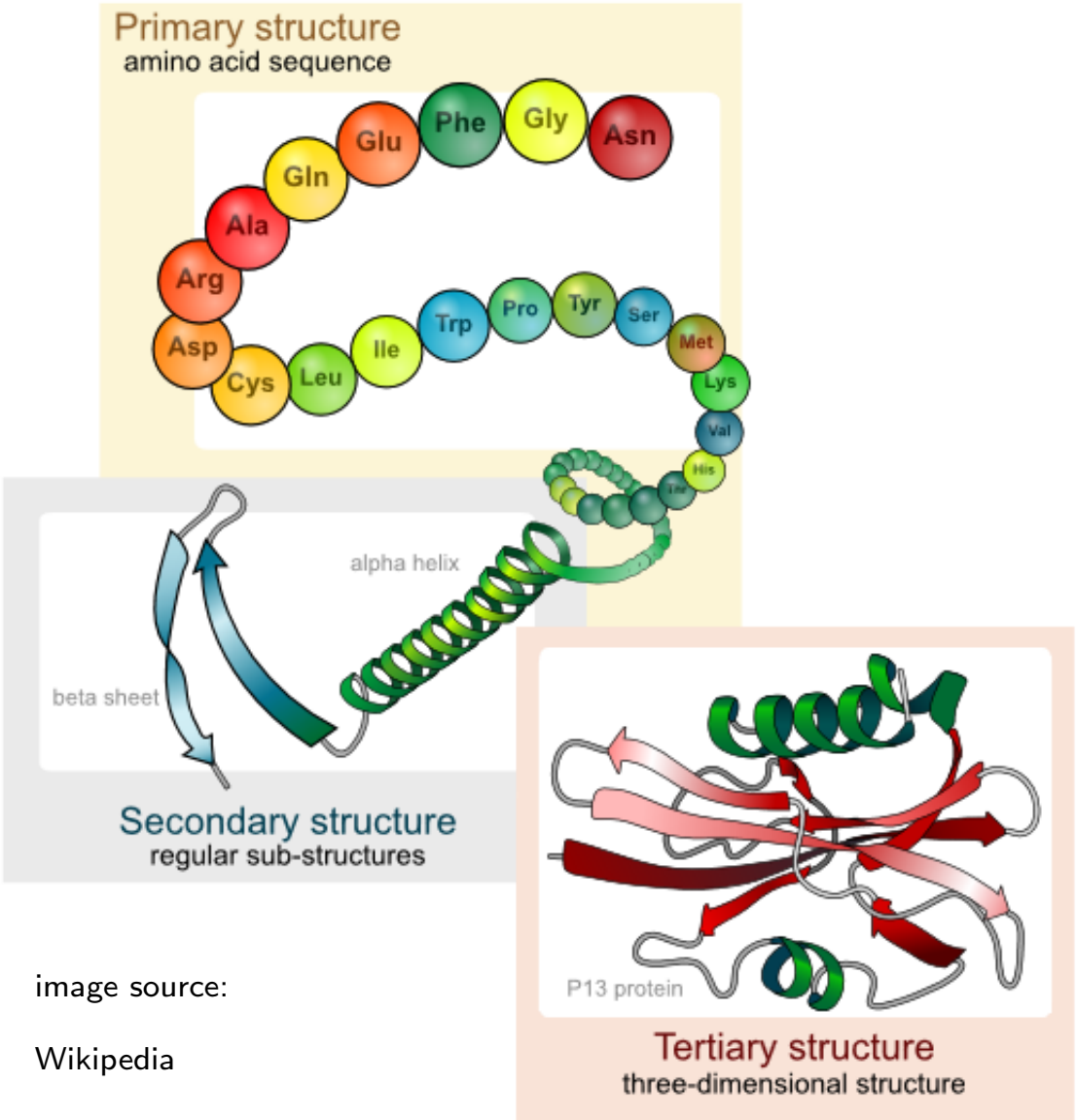


image source:

Wikipedia

- The Protein Data Bank (PDB) is a **repository** for the **3D structure** of proteins

Application 1: Protein Structure Classification

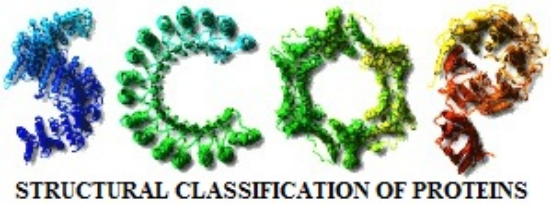
- A good classification method can
  - Reveal the **evolutionary relationship** between the proteins
  - Help to understand how protein **functions**
- Structural Classification of Proteins (SCOP) is a widely used database which has been constructed **manually** by visual inspection
- A reliable automatic protein classification method is not yet available
  - Most of existing methods depend on distance–based similarity measures and are biased by sequence alignments (Rogen and Fain, 2003)
- Classification with **PSJDE** is completely alignment free

Application 1: Protein Structure Classification

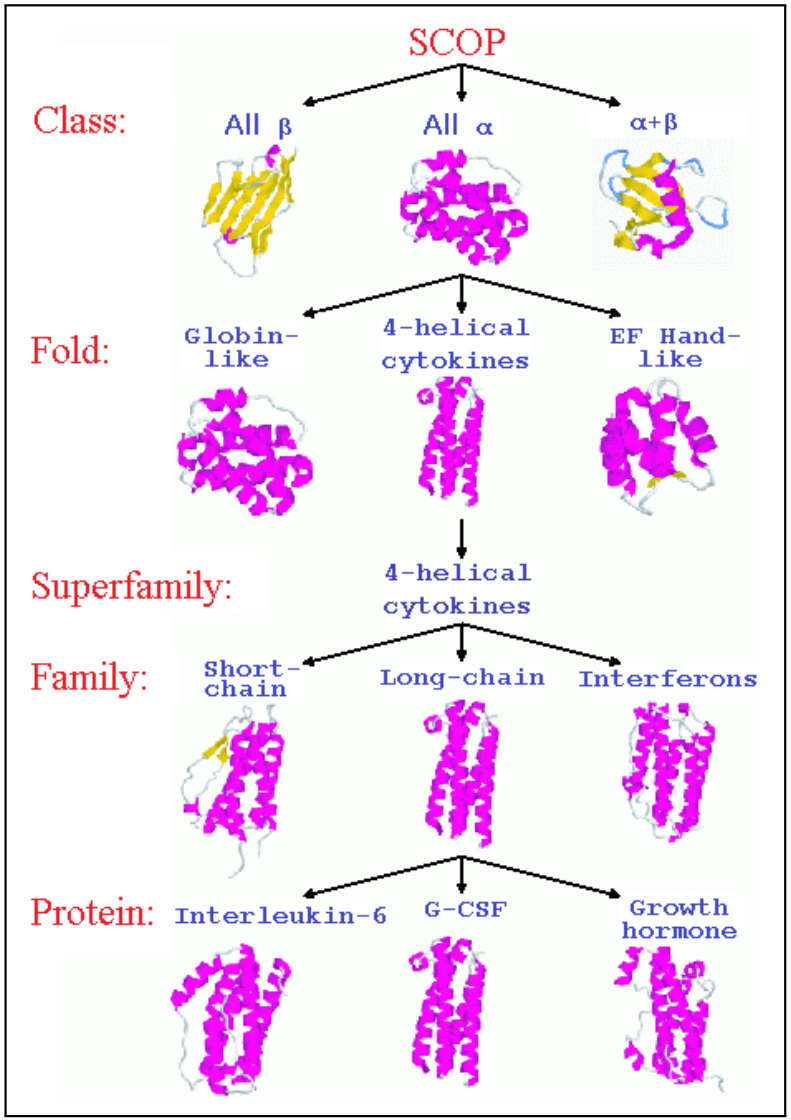
- A good classification method can
  - Reveal the **evolutionary relationship** between the proteins
  - Help to understand how protein **functions**
- Structural Classification of Proteins (SCOP) is a widely used database which has been constructed **manually** by visual inspection
- A reliable automatic protein classification method is not yet available
  - Most of existing methods depend on distance−based similarity measures and are biased by sequence alignments (Rogen and Fain, 2003)
- Classification with **PSJDE** is completely alignment free

Application 1: Protein Structure Classification

- A good classification method can
  - Reveal the **evolutionary relationship** between the proteins
  - Help to understand how protein **functions**
- Structural Classification of Proteins (SCOP) is a widely used database which has been constructed **manually** by visual inspection
- A reliable automatic protein classification method is not yet available
  - Most of existing methods depend on distance−based similarity measures and are biased by sequence alignments (Rogen and Fain, 2003)
- Classification with **PSJDE** is completely alignment free



- A **manual** classification of protein structural domains
  - **Class**: Types of folds, e.g., $\beta-$sheets
  - **Fold**: The different shapes of domains within a class
  - Proteins with the same shapes but having little sequence or functional similarity are placed in different **superfamilies**
  - Proteins having the same shape and some similarity of sequence and/or function are placed in **families**



Application 2: Protein Structure Prediction

- The literature has focused on **variety** of methods:
  - Template−based modeling
  - Template−free modeling
  - Fragment assembly methods
  - Angular−sampling−based methods
  - Consensus servers

- Essential need for **model assessment**
  - 3D coordinate−based measurements:
    - Root mean square distance (RMSD)
    - TM score
    - GDT score
  - PSJDE score

Application 2: Protein Structure Prediction

- The literature has focused on **variety** of methods:
  - Template−based modeling
  - Template−free modeling
  - Fragment assembly methods
  - Angular−sampling−based methods
  - Consensus servers

- Essential need for **model assessment**
  - 3D coordinate−based measurements:
    - ★ Root mean square distance (RMSD)
    - ★ TM score
    - ★ GDT score
  - PSJDE score

Application 2: Protein Structure Prediction

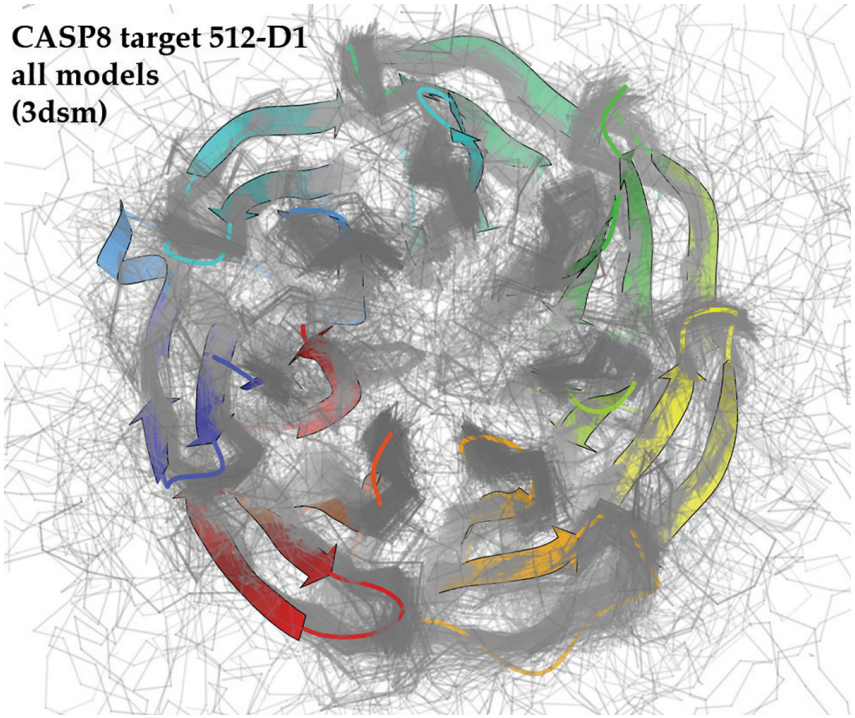
- The literature has focused on **variety** of methods:
  - Template−based modeling
  - Template−free modeling
  - Fragment assembly methods
  - Angular−sampling−based methods
  - Consensus servers

- Essential need for **model assessment**
  - 3D coordinate−based measurements:
    - ★ Root mean square distance (RMSD)
    - ★ TM score
    - ★ GDT score
  - **PSJDE score**

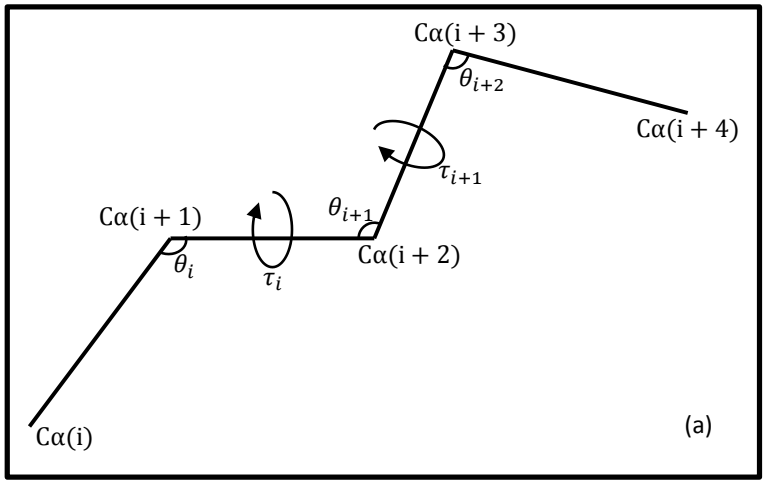
Critical Assessment of Protein Structure Prediction

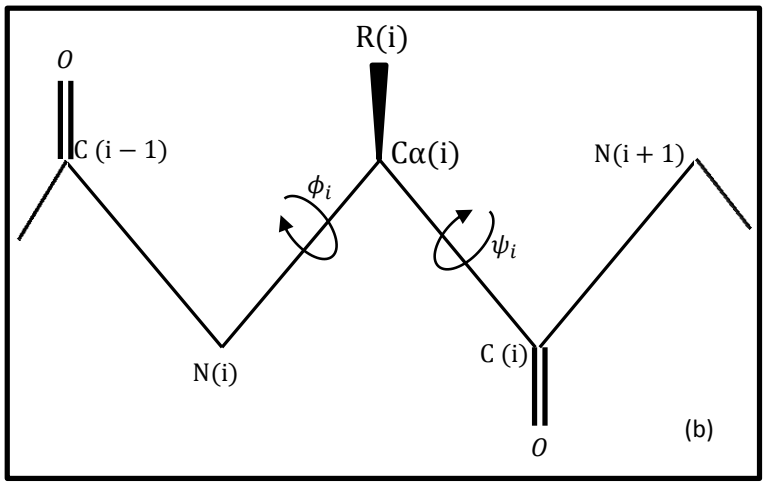
Amino acid sequence (328 AA):

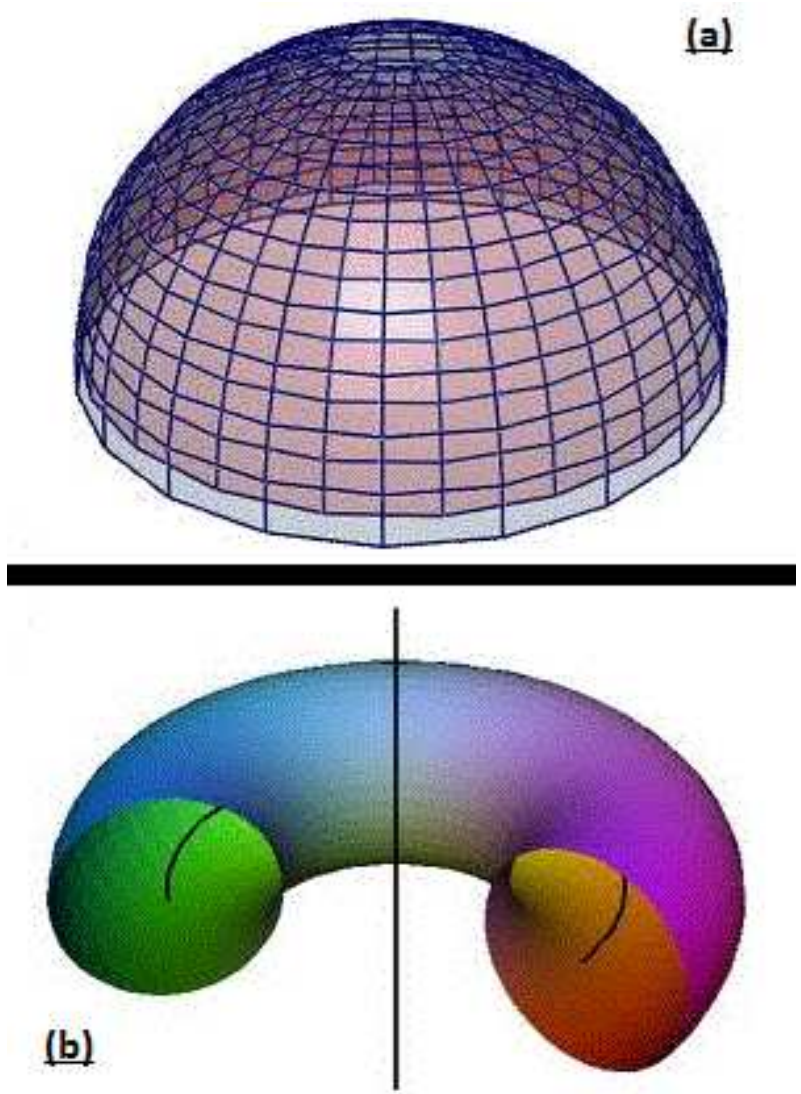
ASGLFITNEGNFQYSNATLSYYDPATCEVENEVFYRANGFKLGDVAQSMVIRDGIGWIVVNNSHVIFAIDINTFKEVGRITGFTSPRYI
HFLSDEKAYVTQIWDYRIFIINPKTYEITGYIECPDMDMESGSTEQMVQYGKYVYVNCWSYQNRILKIDTETDKVVDELTIGIQPTS
LVMDKYNKMWTITDGGYEGSPYGYEAPSLYRIDAETFTVEKQFKFKLGDWPSEVQLNGTRDTLYWINNDIWRMPVEADRVPVR
PFLEFRDTTKYYGLTVNPNNGEVYVADAIDYQQQGIVYRYSPQGKLIDEFYVGIIPGAFCWKLEHHHHHH



Dihedral/Planar Angles







Ramachandran Plot

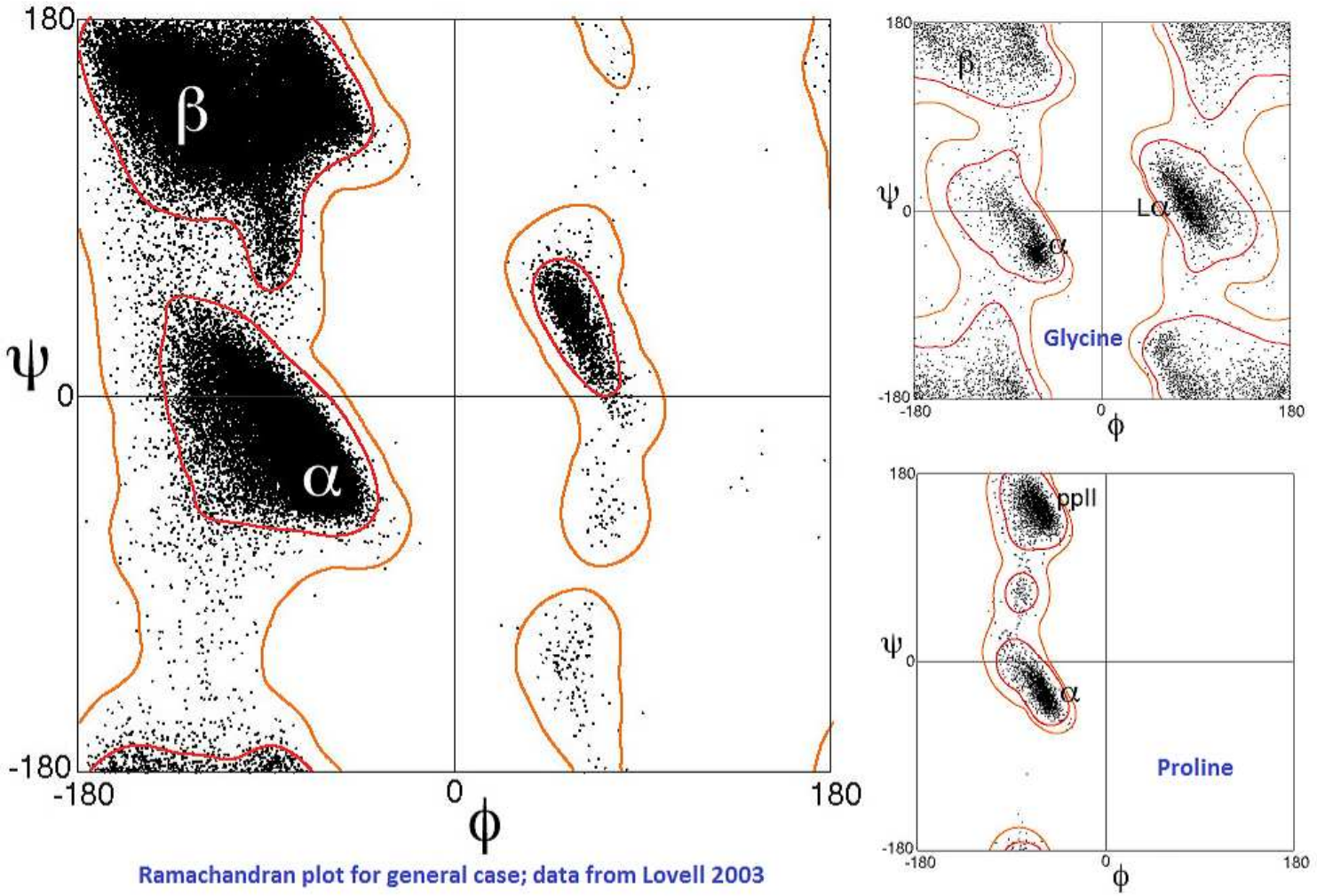


image source: Wikipedia

Existing Models

- **Von Mises−Fisher** distribution for the random "p" −dimensional unit vector (Mardia, 1975)
- **Mixtures** of bivariate von Mises (Mardia et al., 2007)
- Bayesian approach: **Dirichlet process** mixture of bivariate von Mises distributions (Lennox et al., 2009)
- **Modified** kernel density estimator (Maadooliat et al., 2012a)

Existing Models

- **Von Mises−Fisher** distribution for the random "p" −dimensional unit vector (Mardia, 1975)
- **Mixtures** of bivariate von Mises (Mardia et al., 2007)
- Bayesian approach: **Dirichlet process** mixture of bivariate von Mises distributions (Lennox et al., 2009)
- **Modified** kernel density estimator (Maadooliat et al., 2012a)

SCOP – Structural Classification of Proteins

CASP – Critical Assessment of protein Structure Prediction

Penalized Spline Joint Density Estimator

- Consider a collection of densities $f_i, \qquad i = 1, \cdots, m$
- Modeling the densities in an **exponential family** structure:

$$f_i(x) = \frac{\exp \omega_i(x)}{\int \exp \omega_i(x)\, dx} = \exp\left\{ \sum_{k=1}^{K} \phi_k(x)\alpha_{ik} - c_i \right\}$$

- Each log density, up to a constant, can be represented by a **common** set of basis: $\{\phi_k(x), k = 1, \cdots, K\}$. Say $\boxed{\log\{f_i(x)\} := \omega_i(x) - c_i}$,

$$\text{where} \qquad \omega_i(x) = \sum_{k=1}^{K} \phi_k(x)\alpha_{ik}, \tag{1}$$

and $c_i = \log\{\int \exp \omega_i(x)\, dx\}$ is a normalizing constant

Penalized Spline Joint Density Estimator

- Consider a collection of densities $f_i, \qquad i = 1, \cdots, m$
- Modeling the densities in an **exponential family** structure:

$$f_i(x) = \frac{\exp \omega_i(x)}{\int \exp \omega_i(x)\, dx} = \exp\left\{ \sum_{k=1}^{K} \phi_k(x)\alpha_{ik} - c_i \right\}$$

- Each log density, up to a constant, can be represented by a **common** set of basis: $\{\phi_k(x), k = 1, \cdots, K\}$. Say $\boxed{\log\{f_i(x)\} := \omega_i(x) - c_i}$,

$$\text{where} \qquad \omega_i(x) = \sum_{k=1}^{K} \phi_k(x)\alpha_{ik}, \tag{1}$$

and $c_i = \log\{\int \exp \omega_i(x)\, dx\}$ is a normalizing constant

Penalized Spline Joint Density Estimator

- Consider a collection of densities $f_i, \qquad i = 1, \cdots, m$
- Modeling the densities in an **exponential family** structure:

$$f_i(x) = \frac{\exp \omega_i(x)}{\int \exp \omega_i(x)\, dx} = \exp\left\{ \sum_{k=1}^{K} \phi_k(x)\alpha_{ik} - c_i \right\}$$

- Each log density, up to a constant, can be represented by a **common** set of basis: $\{\phi_k(x), k = 1, \cdots, K\}$. Say $\boxed{\log\{f_i(x)\} := \omega_i(x) - c_i}$,

$$\text{where} \qquad \omega_i(x) = \sum_{k=1}^{K} \phi_k(x)\alpha_{ik}, \tag{1}$$

and $c_i = \log\{\int \exp \omega_i(x)\, dx\}$ is a normalizing constant

Some Facts about PSJDE

- Small K means that the exponential family is of **low dimension**
- The basis function ϕ_k's are **data adaptive**
- We call $\phi_k(x)\alpha_{ik}$ in (1) as the kth **component** of the basis expansion and α_{ik} the kth **component score** for the ith density
- The **adaptive** basis functions is a subspace of a rich family ($L \gg K$) of **fixed** basis functions: $\{b_\ell(x), \ell = 1, \cdots, L\}$,

$$\phi_k(x) = \sum_{\ell=1}^{L} b_\ell(x)\theta_{\ell k}, \qquad k = 1 \ldots, K \tag{2}$$

Some Facts about PSJDE

- Small K means that the exponential family is of **low dimension**
- The basis function ϕ_k's are **data adaptive**
- We call $\phi_k(x)\alpha_{ik}$ in (1) as the kth **component** of the basis expansion and α_{ik} the kth **component score** for the ith density
- The **adaptive** basis functions is a subspace of a rich family ($L \gg K$) of **fixed** basis functions: $\{b_\ell(x), \ell = 1, \cdots, L\}$,

$$\phi_k(x) = \sum_{\ell=1}^{L} b_\ell(x)\theta_{\ell k}, \qquad k = 1 \ldots, K \tag{2}$$

Implementing PSJDE

- Using (1) and (2), and rewriting $\omega_i(x)$ in vector−matrix form:

$$\omega_i(x) = \boldsymbol{\phi}(x)^\top \boldsymbol{\alpha}_i = \mathbf{b}(x)^\top \boldsymbol{\Theta} \boldsymbol{\alpha}_i \tag{3}$$

- Now consider x_{ij}'s $(j = 1, \cdots, n_i)$ from the ith distribution $(i = 1 \ldots, m)$. The log−likelihood function has the following form:

$$\ell(\boldsymbol{\Theta}, \mathbf{A}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left\{ \omega_i(x_{ij}) - \log \int \exp \omega_i(x)\, dx \right\} \tag{4}$$

Implementing PSJDE – Derivatives

- Since $\omega_i(x) = \mathbf{b}(x)^\top \boldsymbol{\Theta} \boldsymbol{\alpha}_i$, therefore:

$$\frac{\partial}{\partial \boldsymbol{\alpha}_i} \sum_{j=1}^{n_i} \omega_i(x_{ij}) = \boldsymbol{\Theta}^\top \sum_{j=1}^{n_i} \mathbf{b}(x_{ij}), \qquad \frac{\partial}{\partial \boldsymbol{\theta}_k} \sum_{j=1}^{n_i} \omega_i(x_{ij}) = \alpha_{ik} \sum_{j=1}^{n_i} \mathbf{b}(x_{ij})$$

- Denote $\boldsymbol{\beta}_i = \boldsymbol{\Theta} \boldsymbol{\alpha}_i$ so that $\omega_i(x) = \mathbf{b}(x)^\top \boldsymbol{\beta}_i$:

  Using properties of the exponential family:

$$\frac{\partial}{\partial \boldsymbol{\beta}_i} \log \int \exp \omega_i(x)\, dx = E^{\omega_i}\{\mathbf{b}(X)\},$$

$$\frac{\partial}{\partial \boldsymbol{\beta}_i \boldsymbol{\beta}_i^\top} \log \int \exp \omega_i(x)\, dx = \mathrm{var}^{\omega_i}\{\mathbf{b}(X)\}$$

Implementing PSJDE – Derivatives

- Since $\omega_i(x) = \mathbf{b}(x)^\top \boldsymbol{\Theta} \boldsymbol{\alpha}_i$, therefore:

$$\frac{\partial}{\partial \boldsymbol{\alpha}_i} \sum_{j=1}^{n_i} \omega_i(x_{ij}) = \boldsymbol{\Theta}^\top \sum_{j=1}^{n_i} \mathbf{b}(x_{ij}), \qquad \frac{\partial}{\partial \boldsymbol{\theta}_k} \sum_{j=1}^{n_i} \omega_i(x_{ij}) = \alpha_{ik} \sum_{j=1}^{n_i} \mathbf{b}(x_{ij})$$

- Denote $\boldsymbol{\beta}_i = \boldsymbol{\Theta} \boldsymbol{\alpha}_i$ so that $\omega_i(x) = \mathbf{b}(x)^\top \boldsymbol{\beta}_i$:

  Using properties of the exponential family:

$$\frac{\partial}{\partial \boldsymbol{\beta}_i} \log \int \exp \omega_i(x)\, dx = E^{\omega_i}\{\mathbf{b}(X)\},$$
$$\frac{\partial}{\partial \boldsymbol{\beta}_i \boldsymbol{\beta}_i^\top} \log \int \exp \omega_i(x)\, dx = \mathrm{var}^{\omega_i}\{\mathbf{b}(X)\}$$

Implementing PSJDE – Smoothness

- The roughness penalty approach (Green and Silverman, 1994):

$$\arg \min_{(\boldsymbol{\Theta}, \mathbf{A})} \quad -2\,\ell(\boldsymbol{\Theta}, \mathbf{A}) + \lambda\,\mathsf{PEN}(\phi) \tag{5}$$

where $\mathsf{PEN}(\phi) = \mathrm{tr}\{\boldsymbol{\Theta}^\top \mathbf{R} \boldsymbol{\Theta}\} = \sum_{i=k}^{K} \boldsymbol{\theta}_k^\top \mathbf{R} \boldsymbol{\theta}_k$

- The penalty parameter (λ) is chosen by minimizing the AIC (Akaike, 1973):

$$\mathrm{AIC}(\lambda) = -2\,\ell(\boldsymbol{\Theta}, \mathbf{A}) + 2\,\mathrm{df} \quad \text{where}$$

$$\mathrm{df} = \sum_{k=1}^{K} \mathrm{trace}\left\{ \left[\sum_{i=1}^{m} n_i \alpha_{ik}^2\, \widehat{\mathrm{var}}_i\{\mathbf{b}(X)\} + \lambda \mathbf{R} \right]^{-1} \left[\sum_{i=1}^{m} n_i \alpha_{ik}^2\, \widehat{\mathrm{var}}_i\{\mathbf{b}(X)\} \right] \right\}$$

This is an approximation to the **leave–one–out cross–validation** (Gu, 2002)

Implementing PSJDE – Smoothness

- The roughness penalty approach (Green and Silverman, 1994):

$$\arg \min_{(\boldsymbol{\Theta},\mathbf{A})} \quad -2\,\ell(\boldsymbol{\Theta},\mathbf{A}) + \lambda\,\mathrm{PEN}(\phi) \tag{5}$$

where $\mathrm{PEN}(\phi) = \mathrm{tr}\{\boldsymbol{\Theta}^\top \mathbf{R}\boldsymbol{\Theta}\} = \sum_{i=k}^{K} \boldsymbol{\theta}_k^\top \mathbf{R}\boldsymbol{\theta}_k$

- The penalty parameter (λ) is chosen by minimizing the AIC (Akaike, 1973):

$$\mathrm{AIC}(\lambda) = -2\,\ell(\boldsymbol{\Theta},\mathbf{A}) + 2\,\mathrm{df} \quad \text{where}$$

$$\mathrm{df} = \sum_{k=1}^{K} \mathrm{trace}\left\{ \left[\sum_{i=1}^{m} n_i \alpha_{ik}^2\, \widehat{\mathrm{var}}_i\{\mathbf{b}(X)\} + \lambda \mathbf{R} \right]^{-1} \left[\sum_{i=1}^{m} n_i \alpha_{ik}^2\, \widehat{\mathrm{var}}_i\{\mathbf{b}(X)\} \right] \right\}$$

This is an approximation to the **leave−one−out cross−validation** (Gu, 2002)

Implementing PSJDE – Identifiability

- If $\mathbf{U}$ is a $K \times K$ orthogonal matrix, then $\boldsymbol{\Theta}\boldsymbol{\alpha}_i = \overbrace{(\boldsymbol{\Theta}\mathbf{U})}^{\tilde{\boldsymbol{\Theta}}}\overbrace{(\mathbf{U}^\top\boldsymbol{\alpha}_i)}^{\tilde{\boldsymbol{\alpha}}_i}$
- To gain identifiability, we require that:
  - $\boldsymbol{\Theta}^\top\boldsymbol{\Theta} = \mathbf{I}$
  - $\mathbf{A}^\top\mathbf{A} = \mathbf{D}^2$ be a diagonal matrix
  - Reorder the columns of $\mathbf{A}$ such that the diagonal of $\mathbf{D}^2$ is decreasing
  - The first non−zero element of each column of $\boldsymbol{\Theta}$ is positive
- If the diagonal elements of $\mathbf{D}$ are all different:
  - Consider the singular value decomposition (SVD): $\tilde{\boldsymbol{\Theta}}\tilde{\mathbf{A}}^\top = \boldsymbol{\Theta}\mathbf{D}\mathbf{A}^\top$
  - Let $\mathbf{A} := \tilde{\mathbf{A}}\mathbf{D}$, so the identifiability is uniquely defined by SVD

Implementing PSJDE – Identifiability

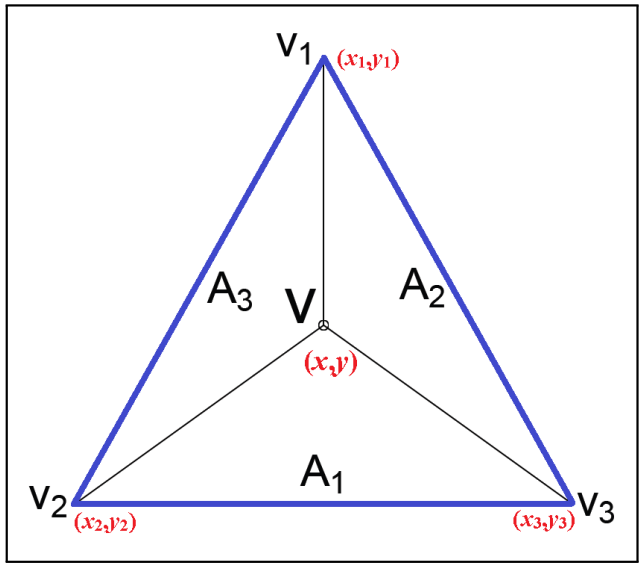
- If $\mathbf{U}$ is a $K \times K$ orthogonal matrix, then $\boldsymbol{\Theta}\boldsymbol{\alpha}_i = \overbrace{(\boldsymbol{\Theta}\mathbf{U})}^{\widetilde{\boldsymbol{\Theta}}}\overbrace{(\mathbf{U}^\top\boldsymbol{\alpha}_i)}^{\widetilde{\boldsymbol{\alpha}}_i}$
- To gain identifiability, we require that:
  - $\boldsymbol{\Theta}^\top\boldsymbol{\Theta} = \mathbf{I}$
  - $\mathbf{A}^\top\mathbf{A} = \mathbf{D}^2$ be a diagonal matrix
  - Reorder the columns of $\mathbf{A}$ such that the diagonal of $\mathbf{D}^2$ is decreasing
  - The first non−zero element of each column of $\boldsymbol{\Theta}$ is positive
- If the diagonal elements of $\mathbf{D}$ are all different:
  - Consider the singular value decomposition (SVD): $\widetilde{\boldsymbol{\Theta}}\widetilde{\mathbf{A}}^\top = \boldsymbol{\Theta}\mathbf{D}\bar{\mathbf{A}}^\top$
  - Let $\mathbf{A} := \bar{\mathbf{A}}\mathbf{D}$, so the identifiability is uniquely defined by SVD

Triangulation

- Barycentric coordinates: Given a triangle A, any point $v \in \mathbb{R}^2$ can be written as

$$v = b_1 v_1 + b_2 v_2 + b_3 v_3,$$

where (b_1, b_2, b_3) is the barycentric coordinates of v with respect to A



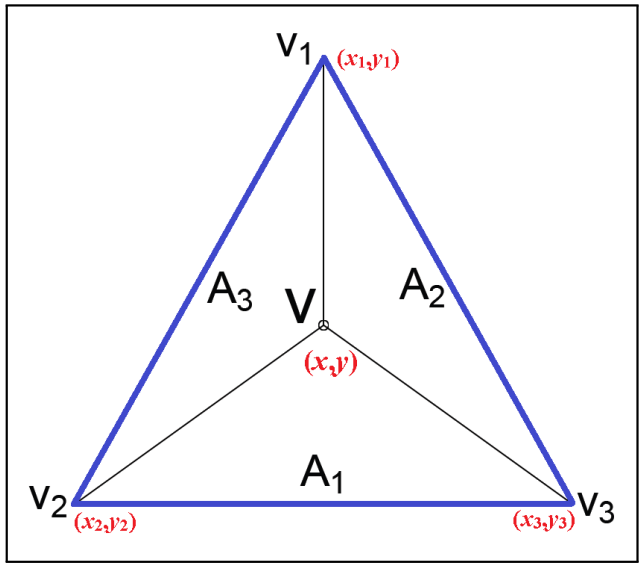
- Some facts:
  - $b_1 + b_2 + b_3 = 1$
  - $b_i = \dfrac{\text{Area of } A_i}{\text{Area of } A}, \quad i = 1, 2, 3$
  - $v \in A \iff b_i \geq 0, \quad i = 1, 2, 3$

Triangulation

- Barycentric coordinates: Given a triangle A, any point $v \in \mathbb{R}^2$ can be written as

$$v = b_1 v_1 + b_2 v_2 + b_3 v_3,$$

where (b_1, b_2, b_3) is the barycentric coordinates of v with respect to A



- Some facts:
  - $b_1 + b_2 + b_3 = 1$
  - $b_i = \dfrac{\text{Area of } A_i}{\text{Area of } A}, \quad i = 1, 2, 3$
  - $v \in A \iff b_i \geq 0, \quad i = 1, 2, 3$

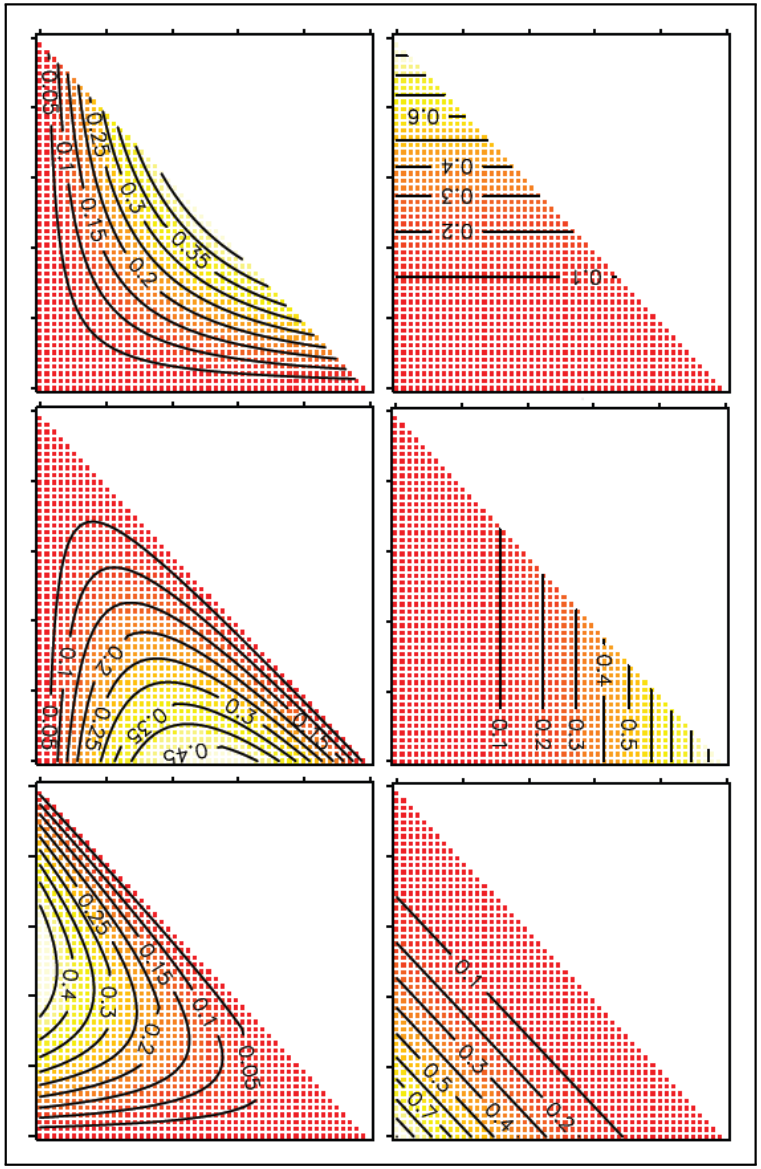
- Bivariate B−splines: Given a triangle A and a point $v \in A$, define:

$$B_{d,ijk}(v) := \frac{d!}{i!j!k!} b_1^i b_2^j b_3^k, \qquad i+j+k=d$$

- For a given triangle A and $d = 2$, B_d contains six functions (see right panel):

- B_d is a basis for $P_d(A)$: $\forall g \in P_d(A)$, $\exists c_{ijk}$:

$$g(v) = \sum_{i+j+k=d} c_{ijk} B_{d,ijk}(v)$$



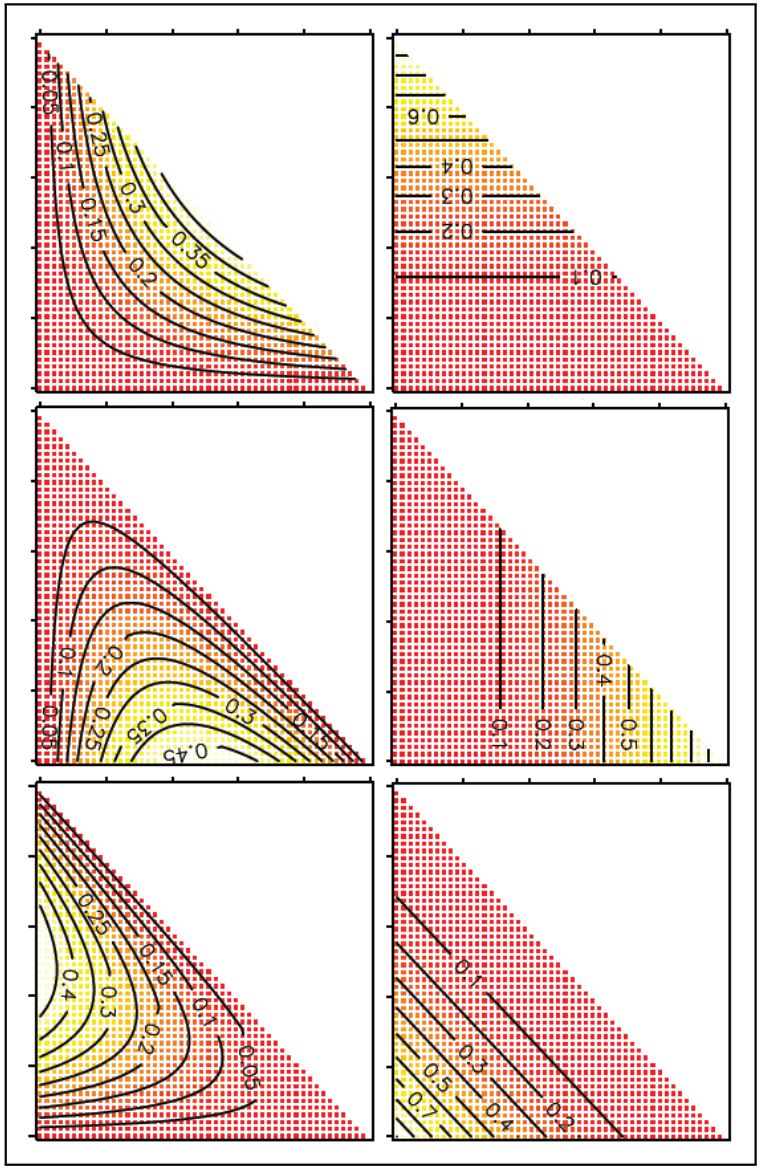
- Bivariate B−splines: Given a triangle A and a point $v \in A$, define:

$$B_{d,ijk}(v) := \frac{d!}{i!j!k!} b_1^i b_2^j b_3^k, \qquad i+j+k=d$$

- For a given triangle A and $d = 2$, B_d contains six functions (see right panel):

- B_d is a basis for $P_d(A)$: $\forall g \in P_d(A)$, $\exists c_{ijk}$:

$$g(v) = \sum_{i+j+k=d} c_{ijk} B_{d,ijk}(v)$$



Theorem

Suppose there are two triangles A_1 and A_2 sharing edge e. ω is any direction not parallel to common edge e and $D^n_\omega p(v)$ is nth order derivative in direction ω at point v. Then

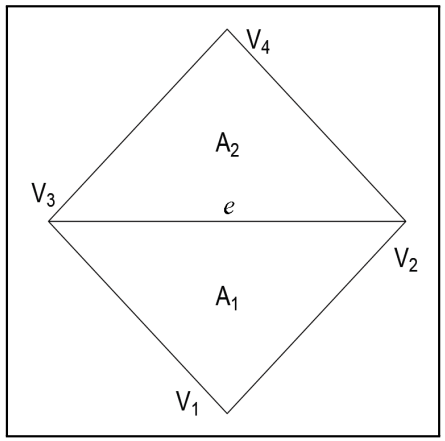
$$D^\ell_\omega p_1(v) = D^\ell_\omega p_2(v), \quad \forall v \in e \text{ and } \ell = 0, \cdots, r \quad \textit{iff}$$

$$c^{(2)}_{ijk} = \sum_{\nu+\mu+\kappa=\ell} c^{(1)}_{\nu,k+\mu,j+\kappa} B^{(1)n}_{\nu\mu\kappa}(v_4), \quad j+k = d-\ell, \ \ell = 0, \cdots, r.$$

- Hint: $\mathbf{Ac} = 0$ is the linear constraint for smoothness of the boundaries

$\Downarrow$

spanning set for the null space of $\mathbf{A}$ is a basis that satisfies the constraint



Theorem

Suppose there are two triangles A_1 and A_2 sharing edge e. ω is any direction not parallel to common edge e and $D^n_\omega p(v)$ is nth order derivative in direction ω at point v. Then

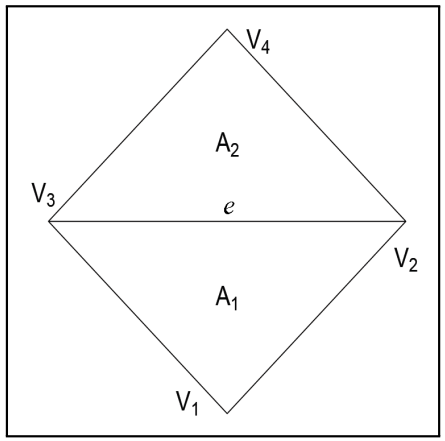
$$D^\ell_\omega p_1(v) = D^\ell_\omega p_2(v), \quad \forall v \in e \text{ and } \ell = 0, \cdots, r \quad \textit{iff}$$

$$c^{(2)}_{ijk} = \sum_{\nu+\mu+\kappa=\ell} c^{(1)}_{\nu,k+\mu,j+\kappa} B^{(1)n}_{\nu\mu\kappa}(v_4), \quad j+k = d-\ell, \ \ell = 0, \cdots, r.$$

- Hint: $\mathbf{Ac} = 0$ is the linear constraint for smoothness of the boundaries

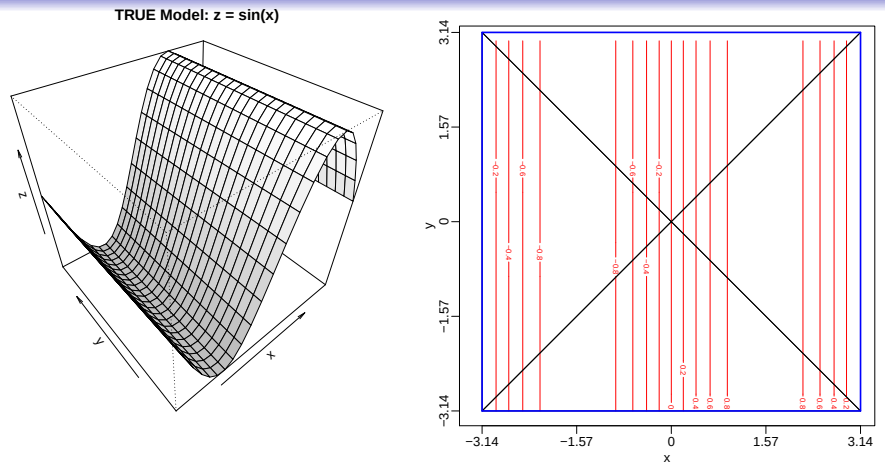
$$\Downarrow$$

spanning set for the null space of $\mathbf{A}$ is a basis that satisfies the constraint



Angular Constraints

Toy Example



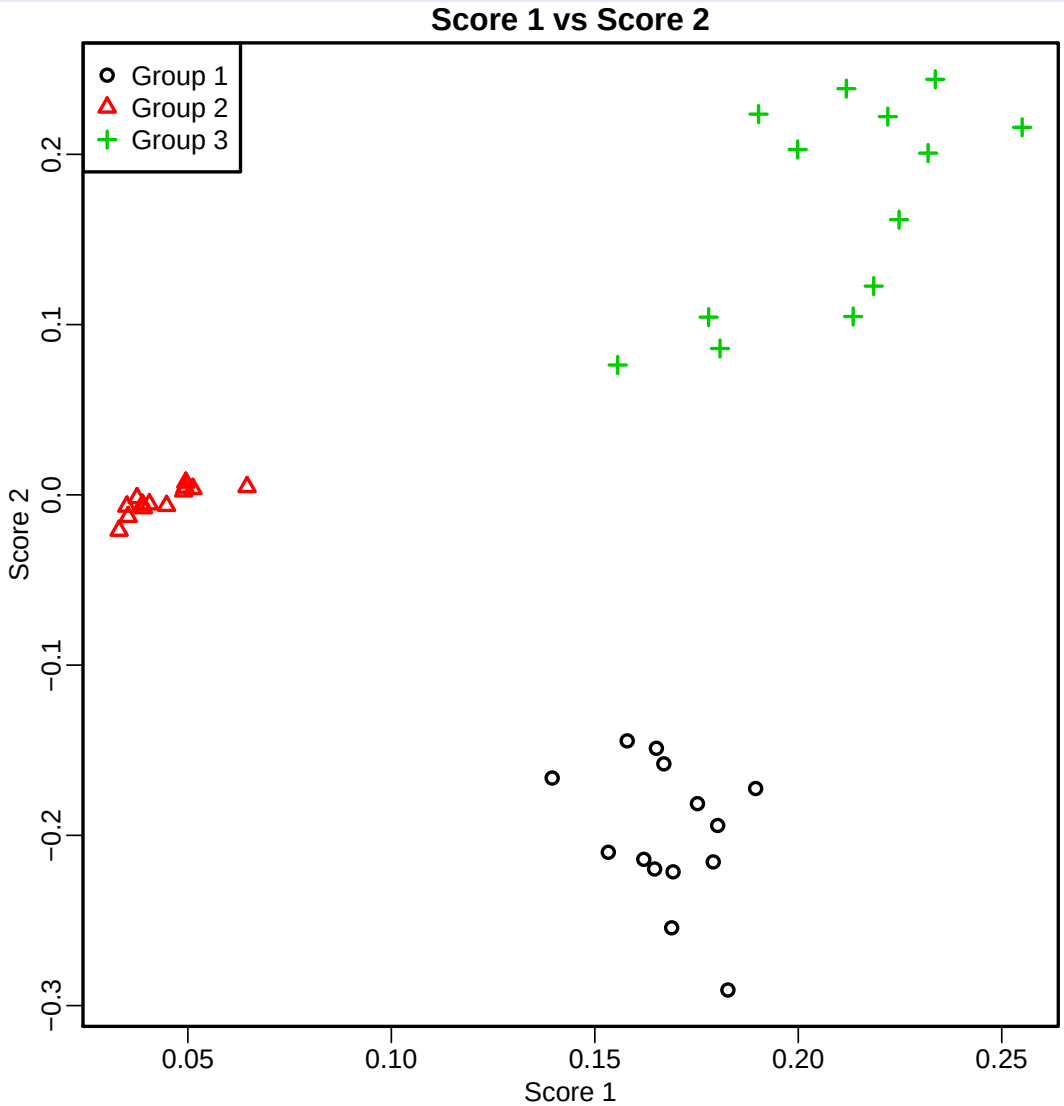


Figure : An arbitrary picked case with $\ell = 42$ and $n = 50$.
Scatterplot of the first two scores from the fitted PSJDE model.

Distances

- For distribution functions F and G with corresponding densities f and g:

$$\mathrm{IAD}(F,G) = \int |f(\mathbf{x}) - g(\mathbf{x})| d\mathbf{x},$$

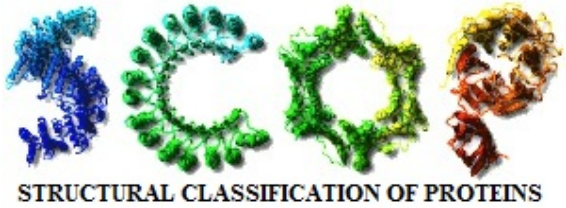
$$\mathrm{HLD}(F,G) = \left[\int \left\{ \sqrt{f(\mathbf{x})} - \sqrt{g(\mathbf{x})} \right\}^2 d\mathbf{x} \right]^{1/2},$$

$$\mathrm{SKLD}(F,G) = \int \{f(\mathbf{x}) - g(\mathbf{x})\} \log \left\{ \frac{f(\mathbf{x})}{g(\mathbf{x})} \right\} d\mathbf{x}$$

Table : Comparison of KDE and PSJDE for three sample sizes ($n = 30, 50, 100$), and three different numbers of distributions ($\ell = 6, 18, 42$) in 100 simulation runs.

# of dist.	Distance Measure	$n = 30$ IAD	$n = 30$ HLD	$n = 30$ SKLD	$n = 50$ IAD	$n = 50$ HLD	$n = 50$ SKLD	$n = 100$ IAD	$n = 100$ HLD	$n = 100$ SKLD
$\ell = 6$	KDE	0.720	0.163	2.983	0.638	0.137	2.417	0.581	0.122	2.085
		(0.042)	(0.016)	(0.363)	(0.041)	(0.015)	(0.319)	(0.042)	(0.014)	(0.273)
	PSJDE	0.632	0.117	1.468	0.543	0.092	1.081	0.495	0.077	0.827
		(0.033)	(0.010)	(0.162)	(0.032)	(0.009)	(0.130)	(0.030)	(0.008)	(0.094)
$\ell = 18$	KDE	0.710	0.159	2.885	0.655	0.143	2.555	0.601	0.128	2.216
		(0.041)	(0.016)	(0.358)	(0.042)	(0.015)	(0.324)	(0.042)	(0.014)	(0.278)
	PSJDE	0.521	0.088	1.052	0.483	0.079	0.895	0.458	0.070	0.739
		(0.033)	(0.009)	(0.127)	(0.034)	(0.009)	(0.112)	(0.032)	(0.008)	(0.088)
$\ell = 42$	KDE	0.691	0.152	2.718	0.642	0.138	2.461	0.579	0.121	2.084
		(0.041)	(0.016)	(0.350)	(0.042)	(0.015)	(0.322)	(0.042)	(0.014)	(0.274)
	PSJDE	0.477	0.079	0.941	0.446	0.073	0.821	0.424	0.063	0.666
		(0.034)	(0.009)	(0.121)	(0.035)	(0.009)	(0.109)	(0.032)	(0.007)	(0.084)

Protein Classification – Hard Task



- **8 domains** from All alpha proteins/Globin–like/Globin–like/Globins/Myoglobin/Sperm whale (Physeter catodon)
- **7 domains** from All alpha proteins/Globin–like/Globin–like/Globins/Myoglobin/Slug sea hare (Aplysia limacina)
- **10 domains** from All alpha proteins/Globin–like/Globin–like/Globins/Myoglobin/Pig (Sus scrofa)
- **8 domains** from All alpha proteins/Globin–like/Globin–like/Globins/Myoglobin/Horse (Equus caballus)

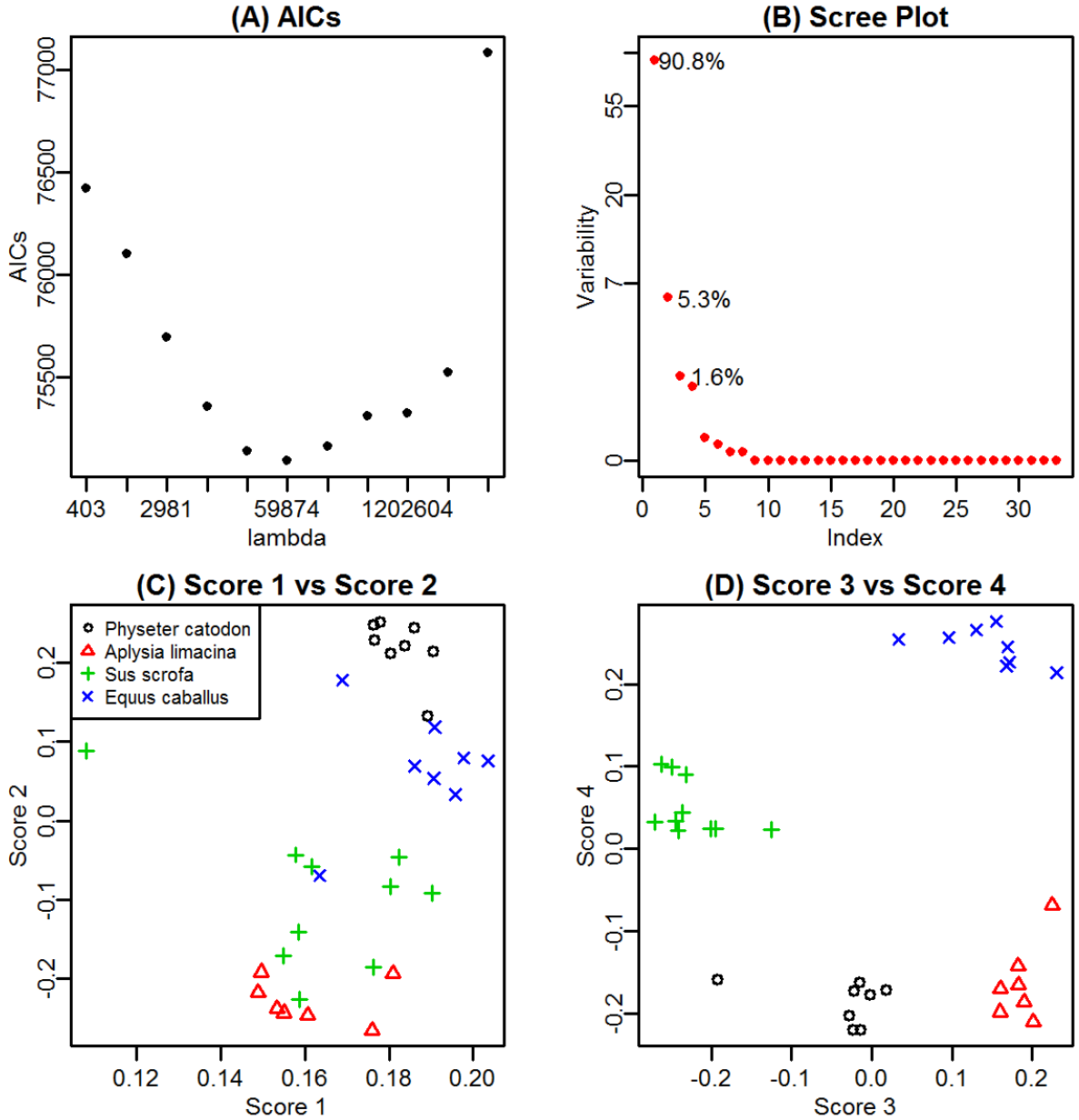


Figure : A **hard classification** task with 33 domains from four protein classes, separated at the bottom of SCOP hierarchy.

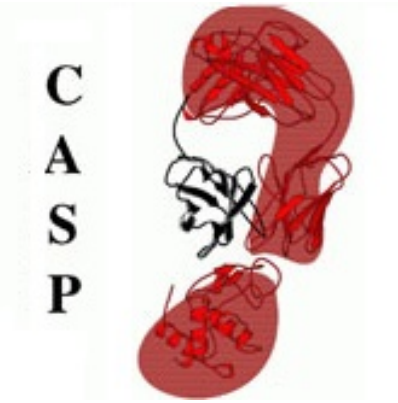
(A) AIC plot;

(B) scree plot;

(C) scatterplot of score 1 vs score 2;

(D) scatterplot of score 3 vs score 4.

Model Assessment – Server Correlation

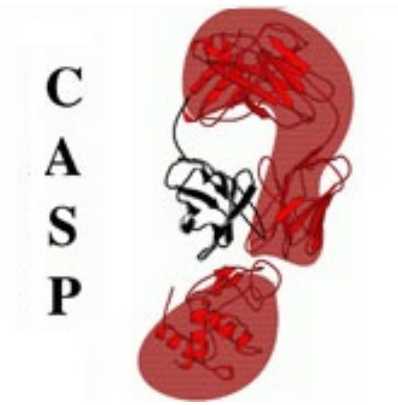


- Choose a target: **T0630** with 132 AA

MRAPIPEPKPGDLIEIFRPFYRHWAIYVGDGYVVHLAPPSEVAGAGAASVMSALTDKAIV
KKELLYDVAGSDKYQVNNKHDDKYSPLPCSKIIQRAEELVGQEVLYKLTSENCEHFVNEL
RYGVARSDQVRD

- Specify a list of **servers**:
  - HHpredA, HHpredB and HHpredC
  - M.COM–CLUSTER, M.COM–CONST., M.COM–NOVEL and M.COM–REFINE
  - PconsD, PconsM, PconsR, ProQ2 and Pcomb
  - RaptorX, RaptorX–MSA and RaptorX–Boost
  - QUARK, Zhang–Server
  - ROSETTA

Model Assessment – Server Correlation



- Choose a target: **T0630** with 132 AA

MRAPIPEPKPGDLIEIFRPFYRHWAIYVGDGYVVHLAPPSEVAGAGAASVMSALTDKAIV
KKELLYDVAGSDKYQVNNKHDDKYSPLPCSKIIQRAEELVGQEVLYKLTSENCEHFVNEL
RYGVARSDQVRD

- Specify a list of **servers**:
  - HHpredA, HHpredB and HHpredC
  - M.COM−CLUSTER, M.COM−CONST., M.COM−NOVEL and M.COM−REFINE
  - PconsD, PconsM, PconsR, ProQ2 and Pcomb
  - RaptorX, RaptorX−MSA and RaptorX−Boost
  - QUARK, Zhang−Server
  - ROSETTA

Table : Comparing the distances between the **18 servers** from **6 groups** in CASP9 with the real protein structures for **T0630**.

Servers	IAD	HLD	SKLD
2 - HHpredA	0.014	0.096	0.056
3 - HHpredB	0.014	0.096	0.056
4 - HHpredC	0.014	0.096	0.056
5 - M.COM–CLUSTER	0.008	0.063	0.031
6 - M.COM–CONST.	0.010	0.066	0.040
7 - M.COM–NOVEL	0.008	0.063	0.031
8 - M.COM–REFINE	**0.006**	**0.052**	**0.025**
9 - PconsD	0.052	0.181	0.211
10 - PconsM	0.029	0.131	0.118
11 - PconsR	0.011	0.077	0.044
12 - ProQ2	0.017	0.103	0.069
13 - Pcomb	0.012	0.091	0.048
14 - RaptorX	0.028	0.123	0.115
15 - RaptorX–MSA	0.028	0.123	0.115
16 - RaptorX–Boost	0.028	0.123	0.115
17 - QUARK	0.069	0.186	0.283
18 - Zhang–Server	0.039	0.134	0.157
19 - ROSETTA	0.021	0.114	0.084

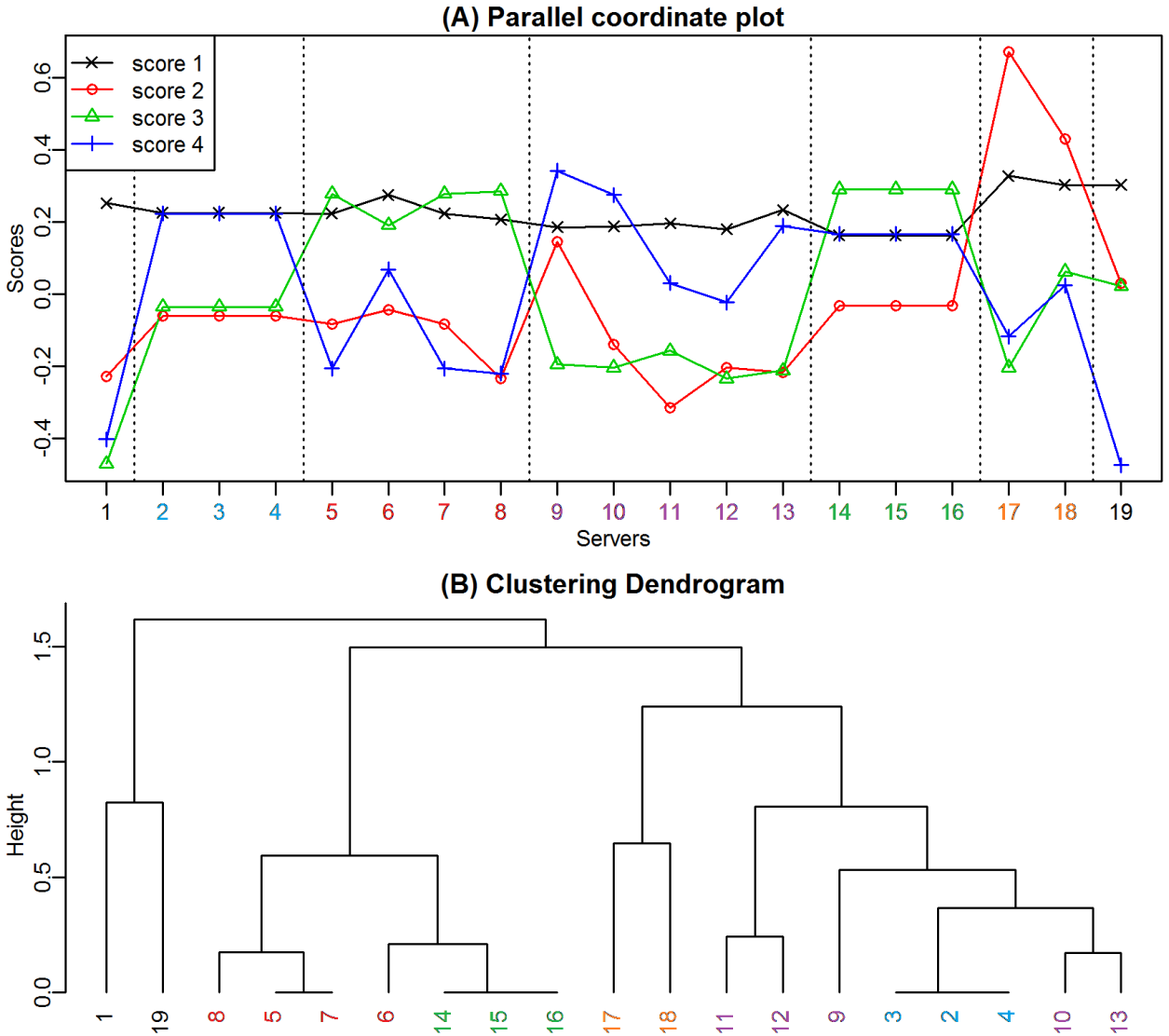


Figure : Model assessment for a CASP9 target (**T0630**).

(A) parallel coordinate plot of top four scores;

(B) hierarchical clustering, based on the top four scores.

Conclusion

- A novel approach for **joint** estimation of multiple bivariate densities
- PSJDE is statistically **more efficient** than separate estimation
- The common adaptive basis, **significantly reduce** the dimensionality
- The scores give a **concise representation** of the corresponding densities
- Useful for protein **clustering**, model **assessment**, and server **correlation detection**
- Future research:
  - Protein Dynamics
  - Further research is needed to improve the computational efficiency for handling larger scale

Conclusion

- A novel approach for **joint** estimation of multiple bivariate densities
- PSJDE is statistically **more efficient** than separate estimation
- The common adaptive basis, **significantly reduce** the dimensionality
- The scores give a **concise representation** of the corresponding densities
- Useful for protein **clustering**, model **assessment**, and server **correlation detection**
- Future research:
  - Protein Dynamics
  - Further research is needed to improve the computational efficiency for handling larger scale

Acknowledgement

- Collaborators:
  - Xin Gao, Assistant Professor at KAUST in Computer Science
  - Lan Zhou, Assistant Professor at Texas A&M University in Statistics
  - Jianhua Huang, Professor at Texas A&M University in Statistics

- Support by King Abdullah University of Science and Technology

Thank You

References

Green, P. and Silverman, B. (1994), *Nonparametric regression and generalized linear models: a roughness penalty approach*, Chapman & Hall/CRC.

Gu, C. (2002), *Smoothing Spline ANOVA Models*, Springer Series in Statistics, Springer.

Lai, M. and Schumaker, L. (2007), *Spline Functions on Triangulations*, Encyclopedia of Mathematics and Its Applications, Cambridge University Press.

Maadooliat, M., Gao, X., and Huang, J. Z. (2012a), "Assessing protein conformational sampling methods based on bivariate lag-distributions of backbone angles," *Briefings in Bioinformatics*.

Maadooliat, M., Zhou, L., Gao, X., and Huang, J. Z. (2012b), "Joint estimation of multiple bivariate densities of protein backbone angles using an adaptive exponential spline family," *Under Revision, JASA*.

Mardia, K. V. (1975), "Statistics of Directional Data (Com: P371-392)," *Journal of the Royal Statistical Society, Series B: Methodological*, 37, 349–371.