Statistical Inference Problems with Applications to Computational Structural Biology

Parthan Kasarapu

Supervisors:

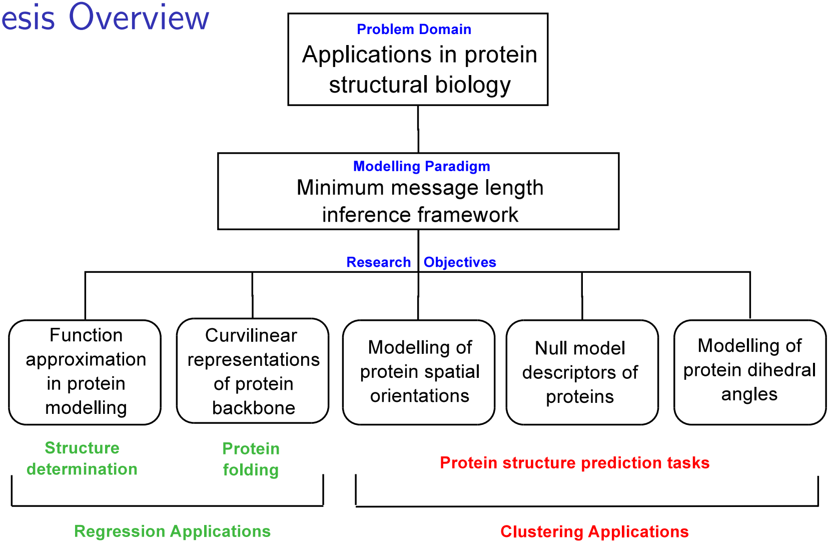
Arun Konagurthu & Maria Garcia de la Banda

6 Aug 2015

Presentation Outline

- Thesis Overview
- Motivation
- Research Summary
  - Statistical modelling
  - Applications to protein structural biology
- Thesis contributions
- Conclusion

Thesis Overview



Problem Domain

Applications in protein structural biology

Modelling Paradigm

Minimum message length inference framework

Research Objectives

Function approximation in protein modelling

Curvilinear representations of protein backbone

Modelling of protein spatial orientations

Null model descriptors of proteins

Modelling of protein dihedral angles

Structure determination

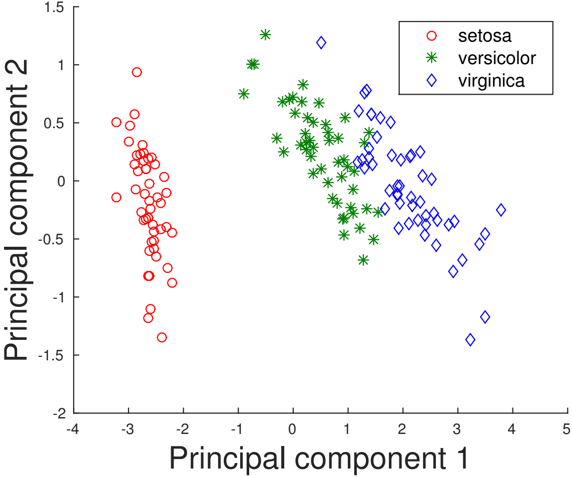
Protein folding

Protein structure prediction tasks

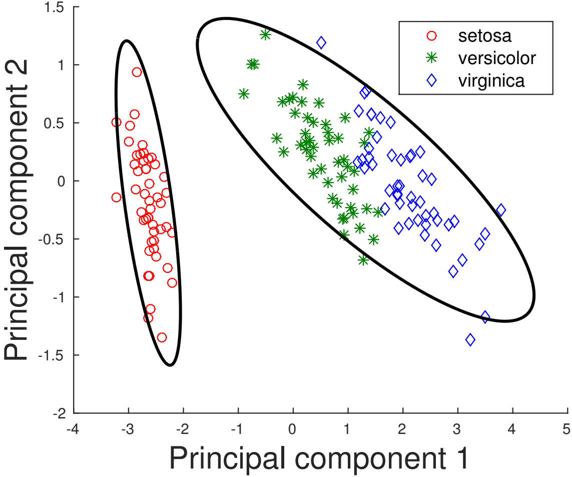
Regression Applications

Clustering Applications

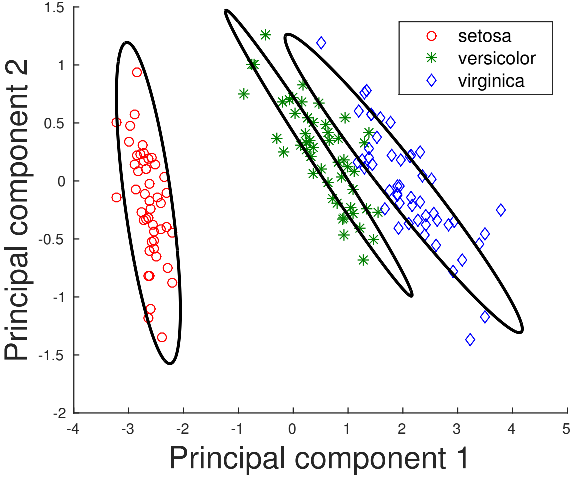
Motivation: How many clusters?



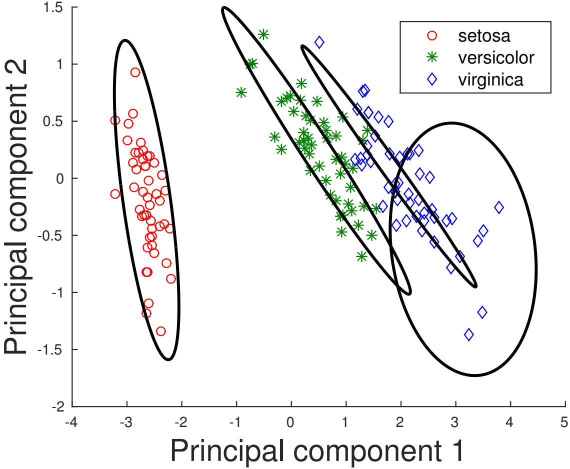
Motivation: How many clusters?



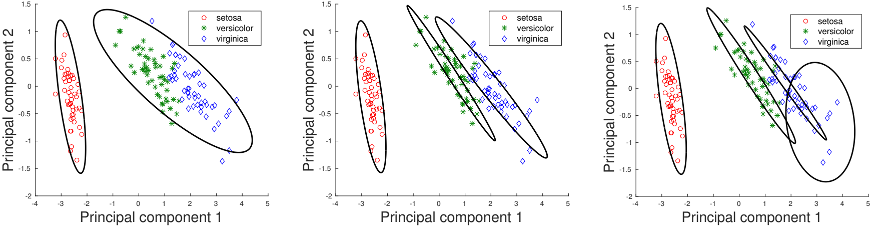
Motivation: How many clusters?



Motivation: How many clusters?



Motivation: How many clusters?



(a) 2-cluster model

(b) 3-cluster model

(c) 4-cluster model

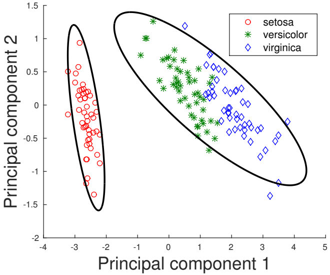
Statistical model selection is important.

Model selection and inference

- Several candidate models: which one to choose?
  - A criterion to compare models ...
  - Based on the model's complexity and the goodness-of-fit

Model selection and inference

- Several candidate models: which one to choose?
  - A criterion to compare models ...
  - Based on the model's complexity and the goodness-of-fit



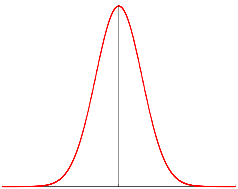
complexity: 2 means + 2 covariance matrices + cluster weights

The typical model selection criteria ...

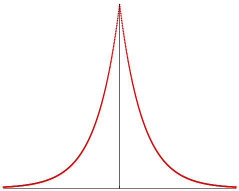
- Various model selection criteria are commonly used ...
  - AIC, BIC, MDL, ...

The typical model selection criteria ...

- Various model selection criteria are commonly used ...
  - AIC, BIC, MDL, ...
- An example of model selection ...



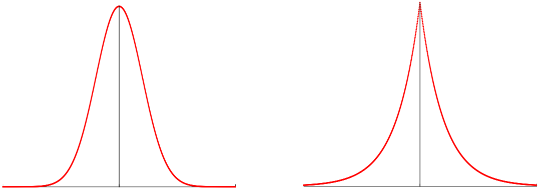
(a) Normal model



(b) Laplace model

The typical model selection criteria ...

- Various model selection criteria are commonly used ...
  - AIC, BIC, MDL, ...
- An example of model selection ...



(a) Normal model (b) Laplace model

- Two parameters for each model (μ & σ)
- Considered to have the same model complexity (limitation)

Minimum Message Length (MML) Framework

Encoding of model (hypothesis) $\mathcal{H}$ **and** data $\mathcal{D}$

$$I(\mathcal{H}\&\mathcal{D}) = \underbrace{I(\mathcal{H})}_{\text{First part}} + \underbrace{I(\mathcal{D}|\mathcal{H})}_{\text{Second part}}$$

Minimum Message Length (MML) Framework

Encoding of model (hypothesis) $\mathcal{H}$ **and** data $\mathcal{D}$

$$I(\mathcal{H}\&\mathcal{D}) = \underbrace{I(\mathcal{H})}_{\text{First part}} + \underbrace{I(\mathcal{D}|\mathcal{H})}_{\text{Second part}}$$

- Two-part message:
  - $I(\mathcal{H})$: model complexity
  - $I(\mathcal{D}|\mathcal{H})$: goodness-of-fit
- Total message length $I(\mathcal{H}\&\mathcal{D})$ is used to compare models.

Minimum Message Length (MML) Framework

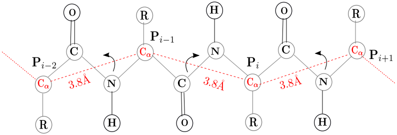
Encoding of model (hypothesis) $\mathcal{H}$ **and** data $\mathcal{D}$

$$I(\mathcal{H}\&\mathcal{D}) = \underbrace{I(\mathcal{H})}_{\text{First part}} + \underbrace{I(\mathcal{D}|\mathcal{H})}_{\text{Second part}}$$

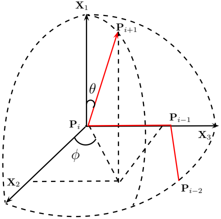
- Two-part message:
  - $I(\mathcal{H})$: model complexity
  - $I(\mathcal{D}|\mathcal{H})$: goodness-of-fit
- Total message length $I(\mathcal{H}\&\mathcal{D})$ is used to compare models.

Model with the least message length is optimal

Problem: Modelling the protein main chain

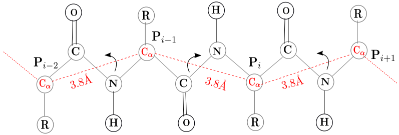


(a) True structure ($C_\alpha - C_\alpha$ data)

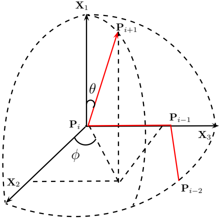


(b) C_α orientations

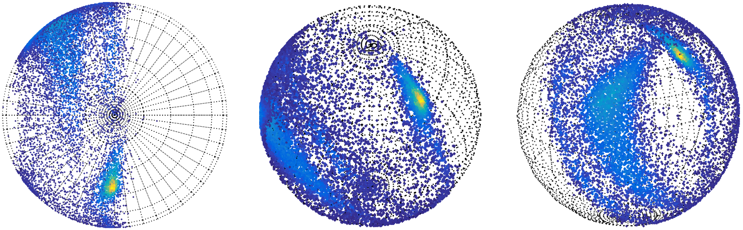
Problem: Modelling the protein main chain



(a) True structure ($C_\alpha - C_\alpha$ data)

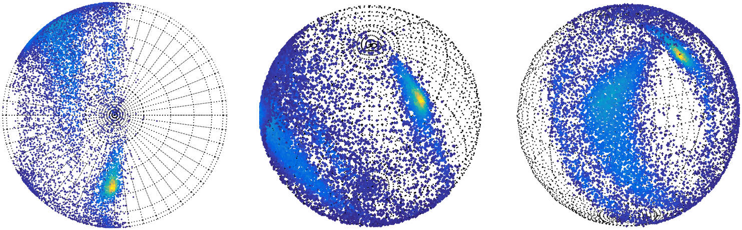


(b) C_α orientations



Empirical distribution of (θ, ϕ)

Modelling of empirical distribution of directional data



- Mixture modelling (Clustering)
  - Data is multi-modal
  - Ideal to find data clusters ...
  - Modelling using *directional* probability distributions

Mixture modelling (Clustering)

Challenges:

- Determination of the number of components
  - Proposed a search method
- Ability to generalize to any probability distribution
  - No assumptions in terms of the nature of data or distribution

Mixture modelling (Clustering)

Challenges:

- Determination of the number of components
  - Proposed a search method
- Ability to generalize to any probability distribution
  - No assumptions in terms of the nature of data or distribution

P. Kasarapu, L. Allison, Minimum message length estimation of mixtures of multivariate Gaussian and von Mises–Fisher distributions, *Machine Learning* (2015) Vol. 100, No. 2-3, Pages 333-378

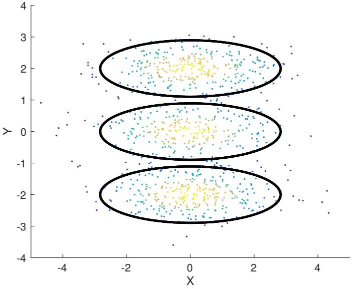
Proposed method to determine clusters of data

Basic idea to determine **number** of clusters

Perturb a K-component mixture through a series of operations so that the mixture escapes a sub-optimal state to reach an improved state.

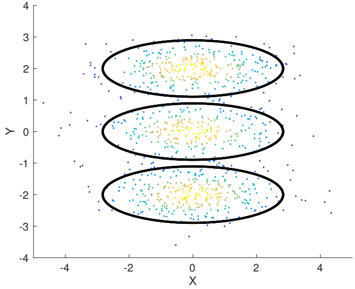
- Operations include ...
  - *Split*
  - *Delete*
  - *Merge*

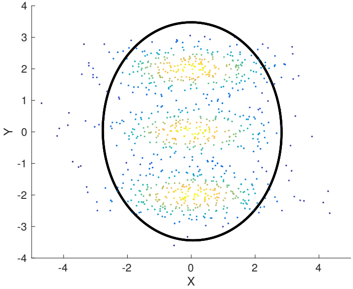
Illustrative example of the search method



Original mixture with three components.

Illustrative example of the search method

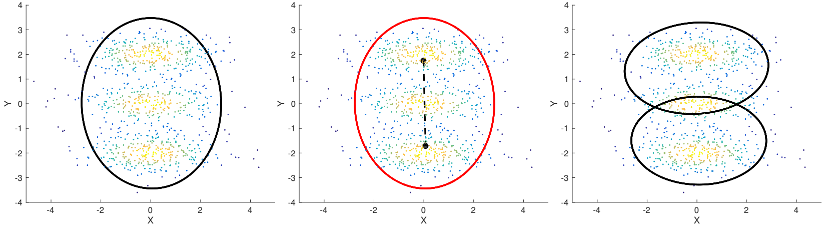




Original mixture with three components.

Begin with a one-component mixture.

Illustrative example of the search method

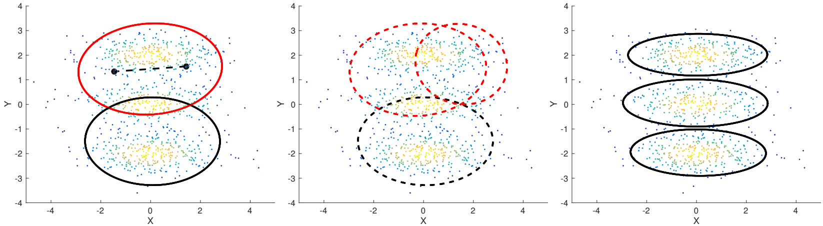


(a) $I = 22793$ bits (b) Splitting (c) $\mathbf{I} = \mathbf{22673}$ bits

Split operation

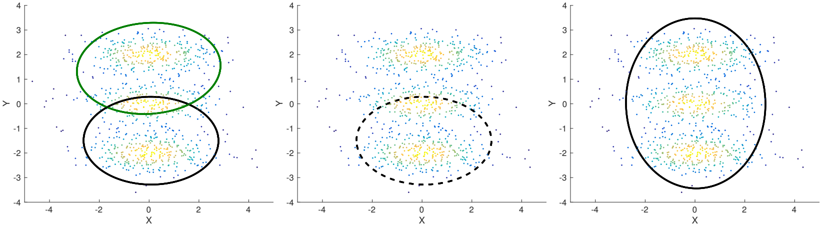
A parent component is split to find locally optimal children leading to a $(K + 1)$-component mixture.

Illustrative example of the search method



(a) Initial means (b) $I = 22691$ bits (c) $\mathbf{I = 22460}$ bits

Illustrative example of the search method

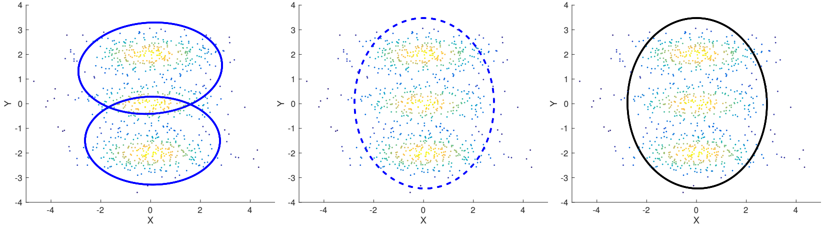


(a) Deleting (b) $I = 25599$ bits (c) $I = 22793$ bits

Delete operation

A component is deleted to find an optimal $(K - 1)$-component mixture.

Illustrative example of the search method

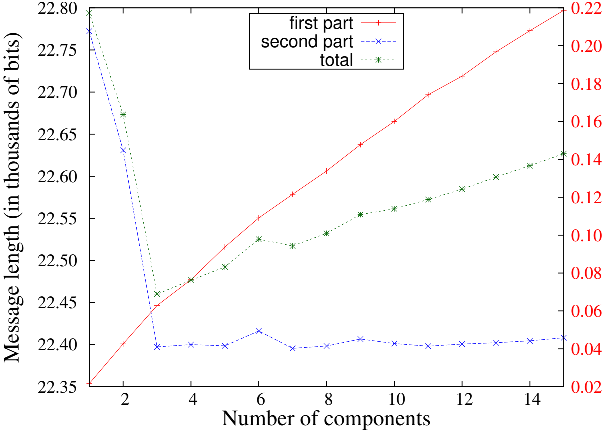


(a) Merging (b) Initialization (c) $I = 22793$ bits

Merge operation

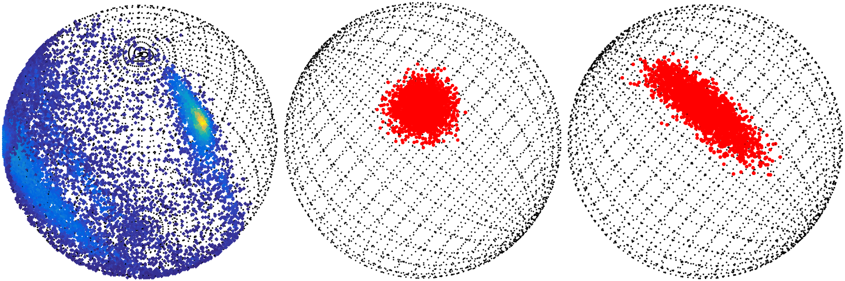
A pair of *close* components are merged to find an optimal $(K-1)$-component mixture.

Evolution of the mixture model



Variation of the individual parts of the total message length with increasing number of components (clusters).

Models of protein data

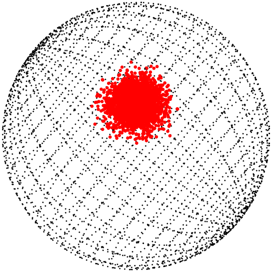


(a) Data

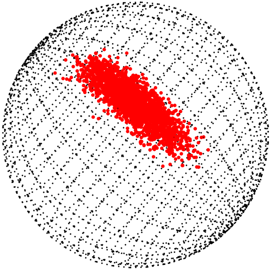
(b) von Mises-Fisher

(c) Kent

Models of protein data



(a) vMF: $\beta = 0$



(b) Kent: $\beta > 0$

Kent probability density function

$$\propto \exp\{ \underbrace{\kappa \boldsymbol{\gamma}_1^T \mathbf{x}}_{\text{linear term}} + \underbrace{\beta(\boldsymbol{\gamma}_2^T \mathbf{x})^2 - \beta(\boldsymbol{\gamma}_3^T \mathbf{x})^2}_{\text{non-linear term}} \}$$

Modelling using Kent distributions

Challenges:

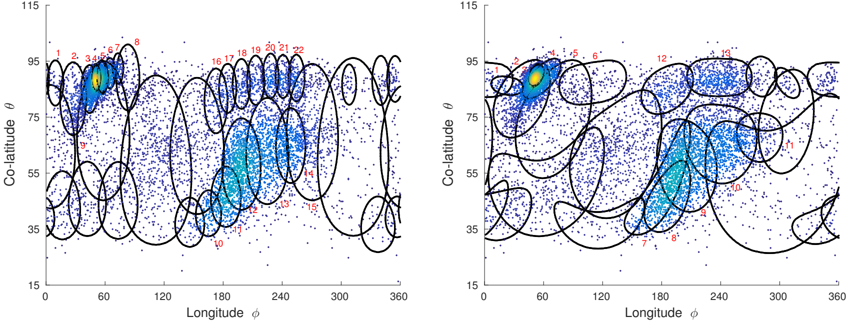
- Complex mathematical form
  - Parameter estimation is a difficult task

Modelling using Kent distributions

Challenges:

- Complex mathematical form
  - Parameter estimation is a difficult task
- Mixture modelling
  - Cluster data on the spherical surface

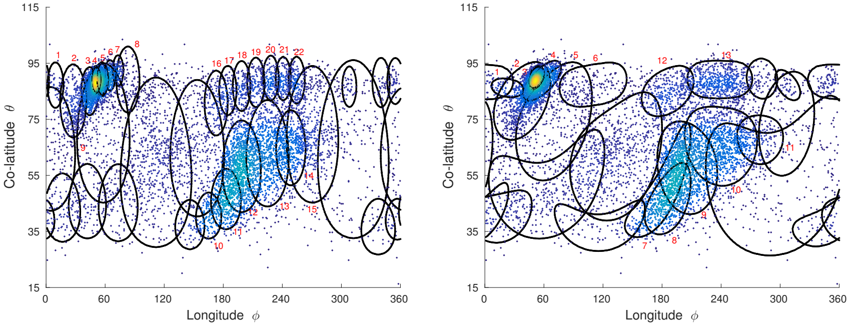
vMF and Kent mixtures of protein directional data



(a) Uncorrelated (35 vMF clusters)

(b) Correlated (23 Kent clusters)

vMF and Kent mixtures of protein directional data



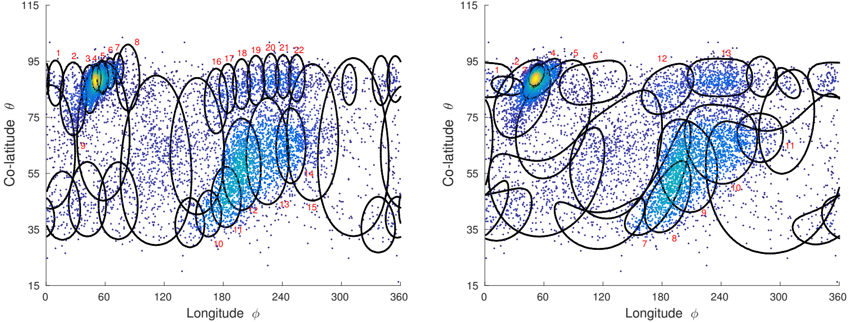
(a) Uncorrelated (35 vMF clusters)

(b) Correlated (23 Kent clusters)

How are these models useful?

- Discovery of frequently occuring patterns
  - Dedicated clusters for helices, strands, etc.
- Clustering profile can be related to protein function
  - Structurally similar proteins will have similar clusters
- *Ab initio* protein structure prediction
  - Random protein generation, homology modelling, template structures, etc.

vMF and Kent mixtures of protein directional data

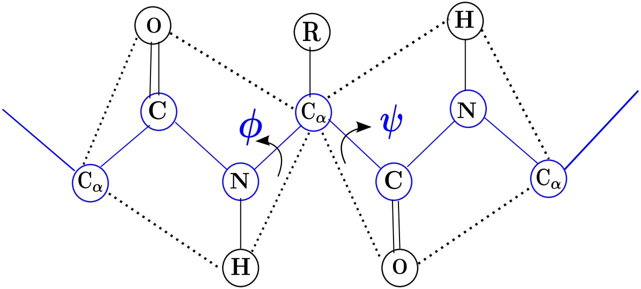


(a) Uncorrelated (35 vMF clusters)

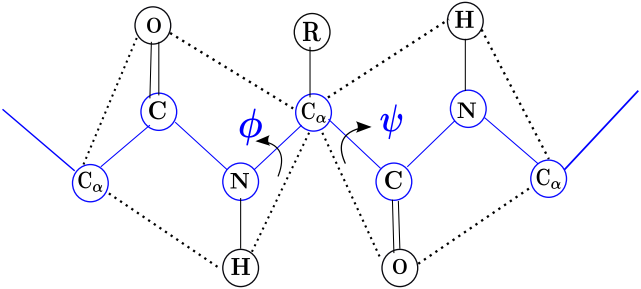
(b) Correlated (23 Kent clusters) - optimal!

Model	Total message length (millions of bits)	Bits per residue
Uniform	6.895	27.434
vMF mixture	6.449	25.656
Kent mixture	**6.442**	**25.630**

Problem: Modelling of protein dihedral angles

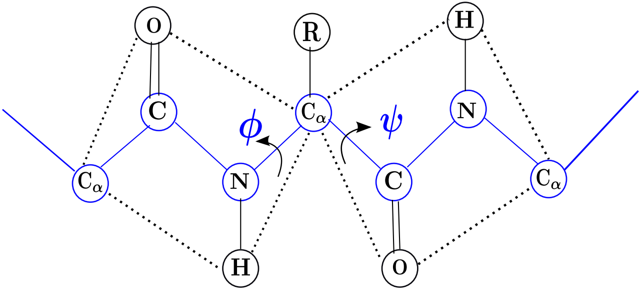


Problem: Modelling of protein dihedral angles



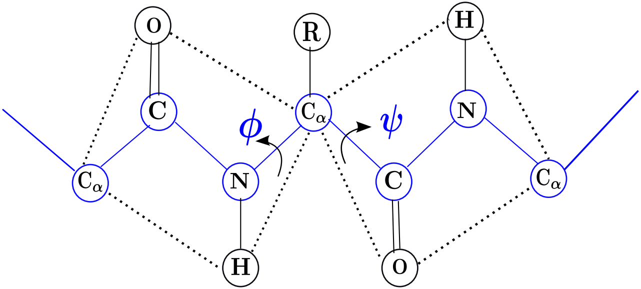
- Modelling protein dihedral angles (ϕ, ψ)
  - $\phi, \psi \in [0, 2\pi)$

Problem: Modelling of protein dihedral angles



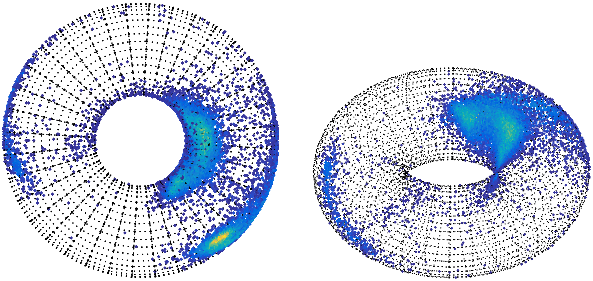
- Modelling protein dihedral angles (ϕ, ψ)
  - $\phi, \psi \in [0, 2\pi)$ represents a point on the torus

Problem: Modelling of protein dihedral angles

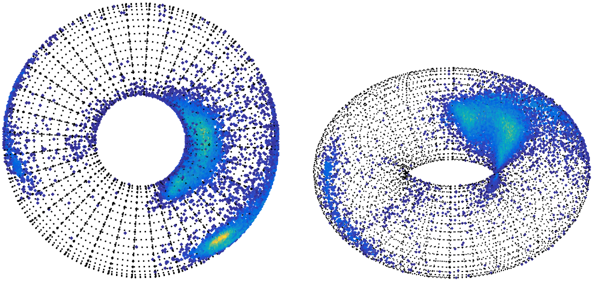


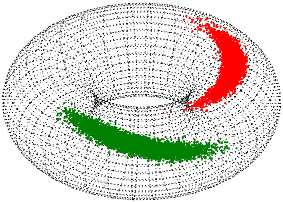
- Modelling protein dihedral angles (ϕ, ψ)
  - $\phi, \psi \in [0, 2\pi)$ represents a point on the torus
  - Cannot be modelled using vMF or Kent
  - Modelled using mixtures of bivariate von Mises (BVM) distributions

Distribution of protein dihedral angles



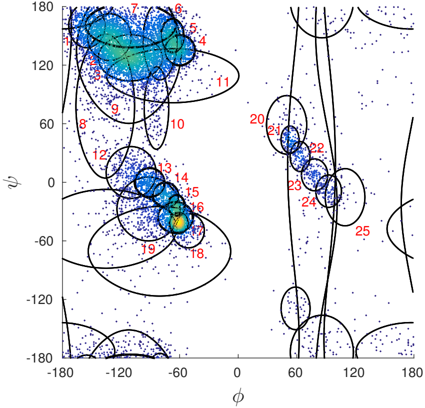
Distribution of protein dihedral angles



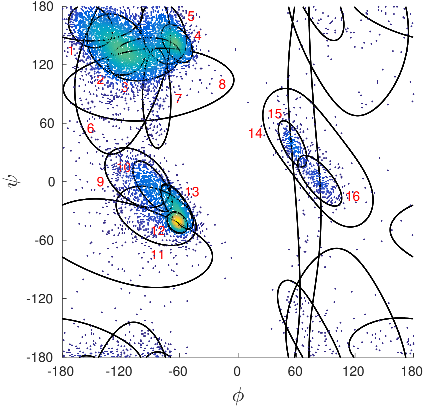


Example BVM distributions

Bivariate von Mises (BVM) clusters of dihedral angle data



No correlation (32 clusters)



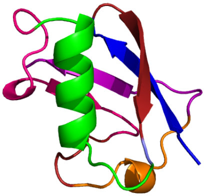
BVM (21 clusters) - optimal!

Problem: Abstraction of protein folding patterns

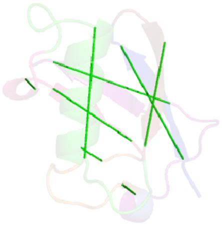
Motivation

- Rapid protein structure comparison
  - Achieved by effective summarization of folding patterns
- Determine functionally similar proteins
  - Achieved by unique representations

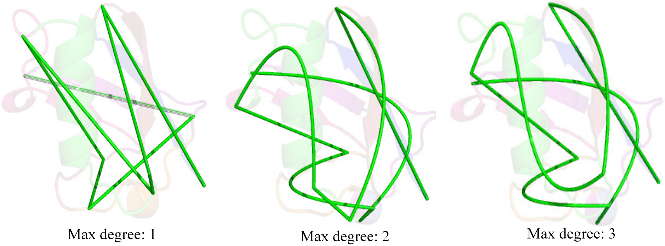
A novel method to abstract protein folding patterns



(a) True structure

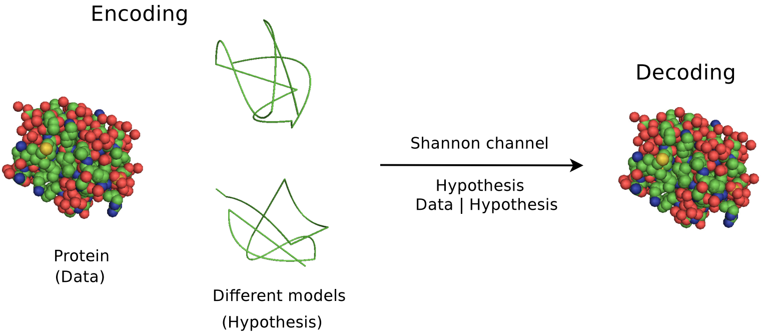


(b) Commonly used



(c) Illustrative non-linear representations (which is optimal?)

Optimal representation



Encoding

Protein
(Data)

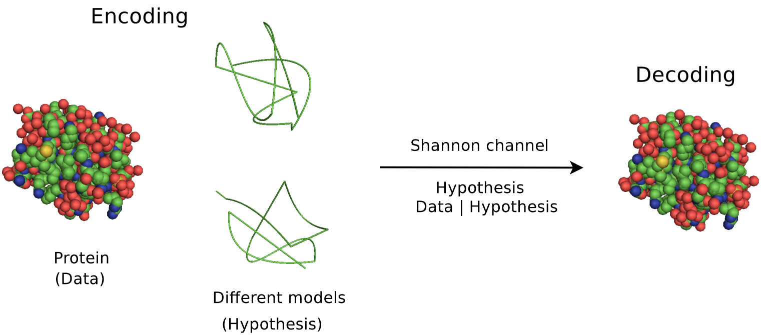
Different models
(Hypothesis)

Shannon channel

Hypothesis
Data | Hypothesis

Decoding

Optimal representation



Encoding

Protein
(Data)

Different models
(Hypothesis)

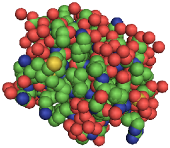
Shannon channel

Hypothesis
Data | Hypothesis

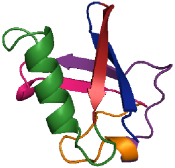
Decoding

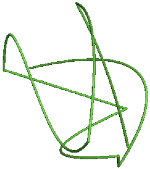
- MML balances the trade-off between
  - Maximize economy of description (compression)
  - Minimize loss of structural information (preservation of geometry)

Merits of this abstraction



Protein

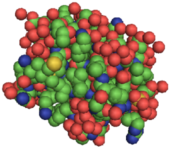




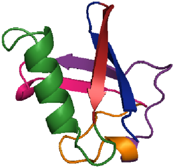
Bezier segmentation

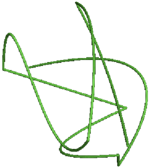
- Does not rely on secondary structure assignment

Merits of this abstraction



Protein

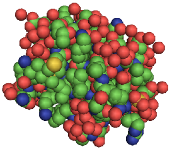




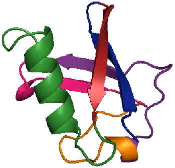
Bezier segmentation

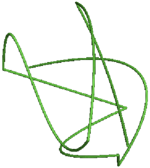
- Does not rely on secondary structure assignment
- Applications in protein structure comparison
  - Database search
  - Comparing the representations

Merits of this abstraction



Protein

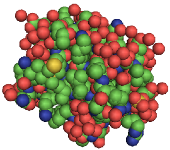




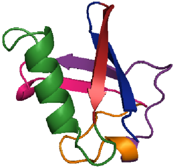
Bezier segmentation

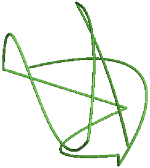
- Does not rely on secondary structure assignment
- Applications in protein structure comparison
  - Database search
  - Comparing the representations - fast

Merits of this abstraction



Protein





Bezier segmentation

- Does not rely on secondary structure assignment
- Applications in protein structure comparison
  - Database search
  - Comparing the representations - fast

P. Kasarapu, M. G. de la Banda, A. S. Konagurthu, On representing protein folding patterns using non-linear parametric curves, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(6):1218-1228 (2014)

Main contributions of my thesis

Theoretical:

- MML-based statistical inference
  - Multivariate von Mises-Fisher (hypersphere)
  - Kent (3D-sphere)
  - Bivariate von Mises (3D-torus)
- Mixture modelling (clustering)
- Non-linear abstractions (regression)

Main contributions of my thesis

Theoretical:

- MML-based statistical inference
  - Multivariate von Mises-Fisher (hypersphere)
  - Kent (3D-sphere)
  - Bivariate von Mises (3D-torus)
- Mixture modelling (clustering)
- Non-linear abstractions (regression)

Applications:

- Structural bioinformatics
- High-dimensional text clustering using vMF mixtures
- Analytical tools for biologists and statisticians

Conclusion

- Data analysis and statistical modelling go hand-in-hand
  - Rigorous models are useful

Conclusion

- Data analysis and statistical modelling go hand-in-hand
    - Rigorous models are useful
- Scope for improving the existing methodologies
    - Extend the current machine learning algorithms

Conclusion

- Data analysis and statistical modelling go hand-in-hand
  - Rigorous models are useful
- Scope for improving the existing methodologies
  - Extend the current machine learning algorithms
- My research has practical implications in data mining, structural biology, etc.

Thank you.