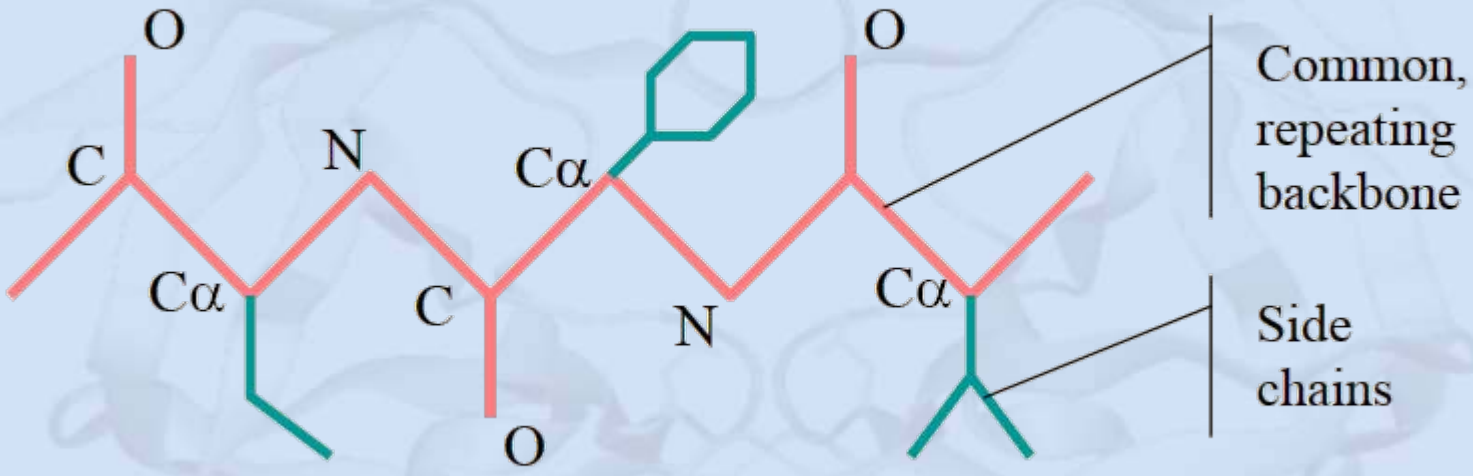
Classification of Protein Secondary Structure

Jay Yagnik and Karthik Raman

Supercomputer Education & Research Centre

Protein Structure



Primary – Sequence of amino acids

Secondary – Local conformation of chain

Tertiary – Global 3-D structure

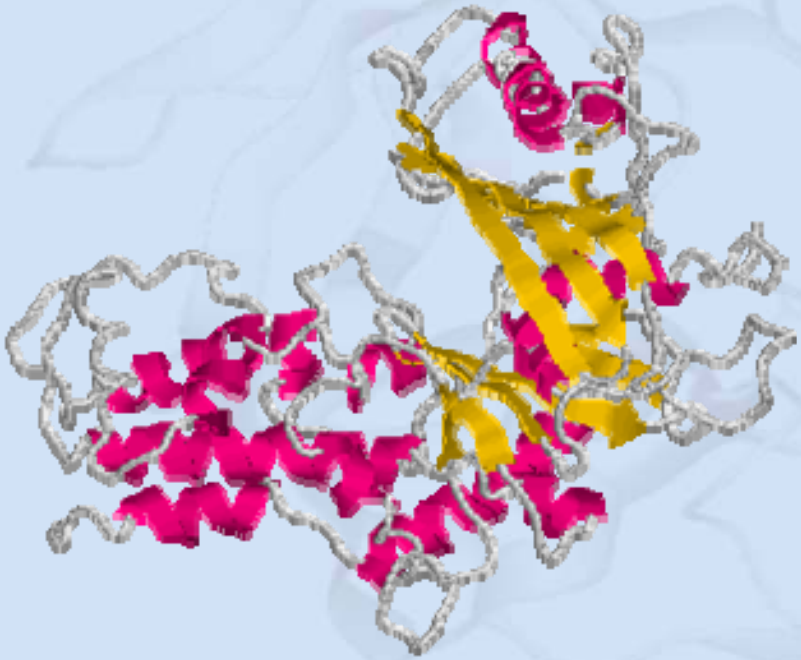
Quaternary – Associations of polypeptide chains

Elements of Secondary Structure

α-helix

β-sheets

Typical proteins



How to classify?

Why classify proteins?

- Structure conserved more than sequence
- Structure is basis for function
- Order vast amounts of structural data
- Common philosophy of studying groups rather than individuals

Strategies for classification

- Structure alignment and comparison
  - DALI, VAST, CE
- Feature Extraction
  - COFE, FastMap, Knot theoretic methods
- Probabilistic Methods
  - PRIDE

Structural Classification Databases

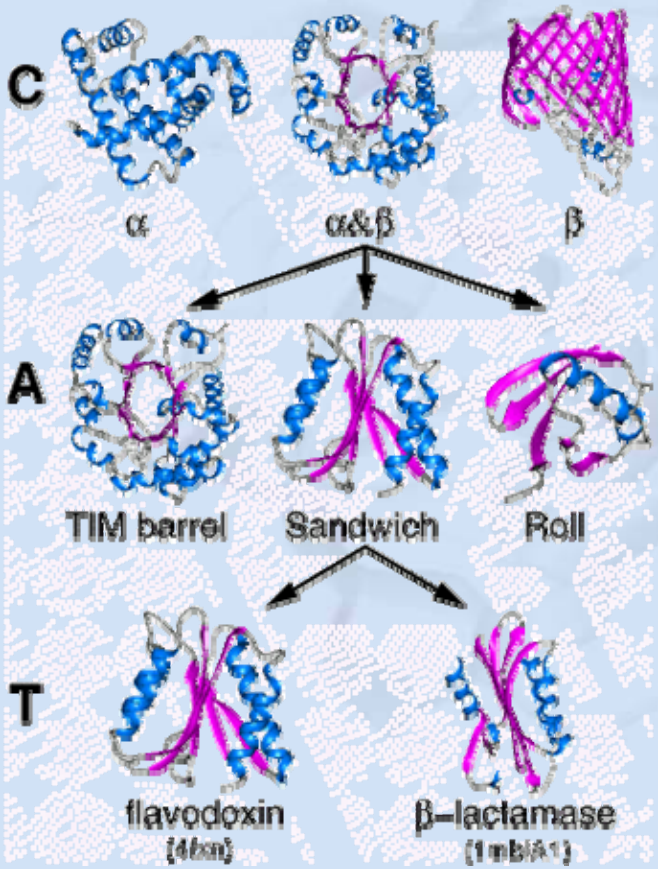
- SCOP (Structural Classification of Proteins)
- CATH (Class Architecture Homologous Superfamilies)

SCOP

Structural Classification of Proteins

- Fully Manually Curated
- Hierarchy
  - Family
  - Superfamily
  - Common fold
  - Class
    - All-α
    - All-β
    - α / β
    - α + β
    - multi-domain

CATH



- Class
- Architecture
- Topology
- Homologous superfamily
- Sequence family

Structure Comparison & Alignment

- Superposition of 3-D structures
- Closely examine SSEs
- Distances between C-α and C-β atoms
- Significance of similarity must be evaluated
- e. g.: FSSP, VAST, CE

FSSP

Families of Structurally Similar Proteins

- Fully automated
- Unlike CATH/SCOP
- *Z*-score > 2.0 is significant
- Uses DALI algorithm
- *Z*-score is useful for automatic classification

Distance Matrix Alignment (DALI)

- Structures represented as 2D arrays of C-α distances
- Similar structures have similar inter-residue distances
- Secondary structure similarity inferred from diagonal overlaps
- Off-diagonal similarities correspond to tertiary structure

Vector Alignment Search Tool

- Represents structures as SSEs
- Topology inferred from type, directionality and connectivity of SSEs
- *p*-value for statistical significance of similarity

Combinatorial Extension

- Comparison of octameric fragments – Aligned Fragment Pairs
- Based on local geometry rather than global orientation of SSEs or topology
- Possible *combinations* of AFPs are *extended* to give an optimal alignment

Demerits of Structural Comparison & Alignment

- Rely on pair-wise comparison
- Expensive and slow
- Triangular inequality not satisfied

Feature Extraction based approaches

- Distance Preserving
  - FastMap
  - COFE
- Absolute Descriptors

Distance Preserving Methods

- Use the distance given by alignment algorithms and try to come up with features which behave the same way in the distance framework
- Distance calculation expensive, so try to optimize on number of distance evaluations

COMDS (Complex Object Multi-Dimensional Scaling)

- Distances between objects are defined by some non-Euclidean method
- Objects are complex (not part of Euclidean space)
- Distance evaluation is expensive

FastMap

- Take points with longest distance
- Project all others on that line and take the distances as the feature value
- Factor the distance out and repeat

COFE
(Complex Object Feature Extraction)

- Chooses sets of points called Reference sets
- Features are the distance of the closest point in the reference set.
- No. of features = No. of reference sets.

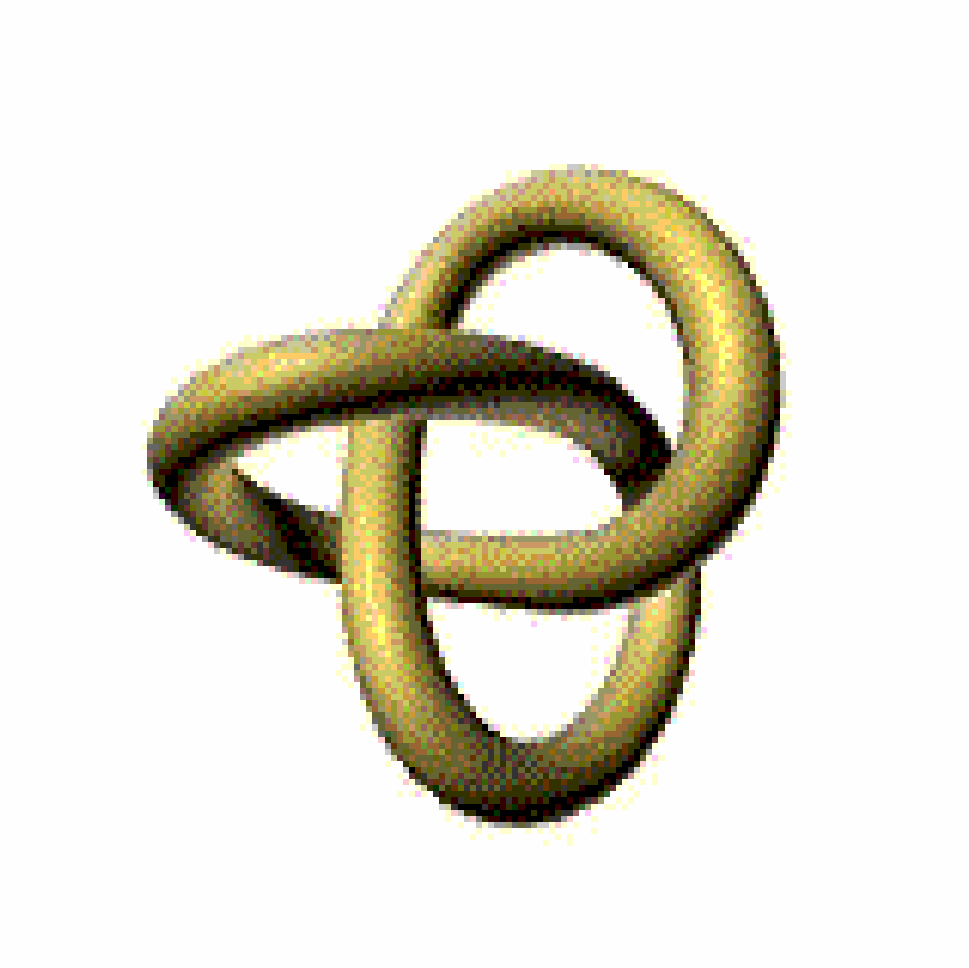
COFE vs FastMap

- Asymptotically similar performance
- But with small number of features COFE performs better
- Number of distance evaluations are significantly less in COFE due to possibility of optimisations

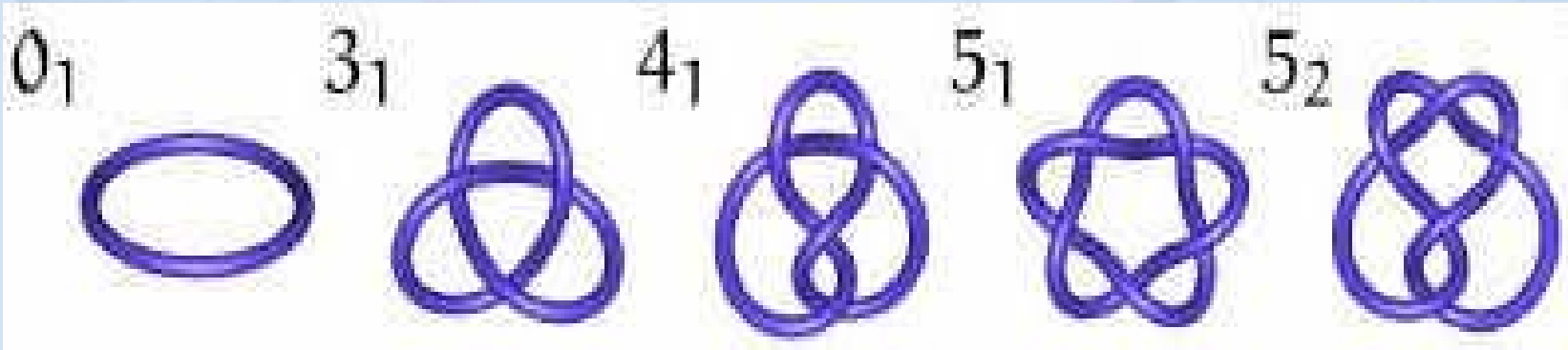
Absolute Descriptors

- Vassiliev Knot Invariants
- Distance Histograms

Knot Theory



Knot Diagrams



Knot Invariants

- Average Crossing Number (First order *Vassiliev* invariant)
- Minimum crossing number
- Unknotting number

Vassiliev Knot Invariants

- Set of mathematical descriptors based on the number of crossings seen in knot diagrams
- Average Crossing Number is a first order Vassiliev Knot Invariant
- Independent of rotation,scale, translation
- Depend only on the geometric shape of the structure

Application to Protein Classification

- Fain and Rogen applied it on CATH.
  - Used Invariants up to order 3.
  - Extracted features from CATH in < 2hrs.
  - Equivalent to 400 million pair wise alignments…….a workstation would take several hundred years to do this.
  - 96% classification accuracy.

Distance Histograms

- Histograms of the inter C-α distances using some window length
- PRIDE (Probability of Identity)
  - Uses histograms from window lengths of 3-30
  - Compares them using average histogram as the expected histogram and applying chi-square goodness of fit test.