TEXT ENCODING FOR PROTEIN STRUCTURE REPRESENTATION

Jun Tan
Donald Adjeroj
• Biological background
• The Problem and our goal
• Related research
• General idea
• Implementations
• Summary
BIOLOGICAL BACKGROUND [1]
PROTEIN

- One of the four of life's basic building blocks
- DNA -> RNA -> Protein
- Peptide bound: the link between two amino acids
- Polypeptide: chain of amino acide
- Once the chain of amino acide is in its final shape, it is called protein
- Twenty types of amino acids
- Three group: COOH, NH2 and R
Protein can have very complex shapes, and the final form is essential to its intended function.

- **Primary structure**: the chain of amino acids reaches its final form.
- **Secondary structure** describes common folding patterns.
- **Tertiary structure** describes the overall three-dimensional structure of a single folded amino acid chain.
- **Quaternary structure** for protein with multiple chains describes all subunits consist of the protein.
PROTEIN

Primary structure determines all other structures
However...
**PROTEIN**

- The shape of the 3D protein structure has a direct impact on its function.
- Secondary structure is much more conserved than sequence (primary structure), over evolution. [2]
THE PROBLEM AND OUR GOAL
THE ORIGINAL FORMAT

- Saved as x, y, z coordinates for each atom along the chain
- Complicated operations for even simple tasks
OUR GOAL

- Simplify the representation
- keep important information such that it retains the biological meanings
- Demo the performance: search similar structures in a protein domain database
- 80 query domains, database size: 23500
STRUCTURAL CLASSIFICATION OF PROTEIN (SCOP) DATABASE

- Protein Domain: part of a protein that can evolve, function and exist independently of the rest of the protein chain
- Manually classified
- Hierarchical structure: Class, Fold, Super-family, family
RELATED RESEARCH
DIRECTLY ALIGN 3D SHAPES

- High accuracy
- Involve complex operation. Time consuming
- Example: DALI [3, 4](distance alignment matrix method) algorithm
CONVERT 3D SHAPE INTO 2D TEXTURES
CONVERT 3D SHAPE INTO STRING

- Not as accurate as 3D method but close
- Much faster
- Example: Ramachandran codes [5]
FRAGMENT APPROACH [6]

- Library of fragments/short structure motifs (hand picked?)
- Represent protein structure as the frequency of the fragments
- Bag of words method
GENERAL IDEA

- Decompose a shape into a sequence of segments
- Represent the segments with basic primitives: segment type, segment length and transition angle between segments
- Encoded into shape string
- Answer biological question by applying string/text algorithms on the shape strings
- N-Gram, TF/IDF and cosine similarity are used when compare similarity between shape strings
IMPLEMENTATIONS
DIHEDRAL ANGLES (RAMACHANDRAN ANGLES)

- One of the most important local parameters that control protein folding
- Three angles:
  1. $\phi$ involves atoms $C'-N-C^\alpha-C'$
  2. $\psi$ involves atoms $N-C^\alpha-C'-N$
  3. $\omega$ involves atoms $C^\alpha-C'-N-C^\alpha$ (usually 0 or 180 due to peptide bond)
RAMACHANDRAN PLOT
CLUSTERING DIHEDRAL ANGLES

Graph 1: Within groups sum of squares vs. Number of Clusters

Graph 2: Optic Order (angles)
PRECISION VS RECALL

4 Clusters

6 Clusters
FOLD VS RECALL

4 Clusters

6 Clusters
CLASS VS RECALL

4 Clusters

6 Clusters
TRIPLES

- Dihedral angles involve three consecutive residues only.
- Pick three residues/points that can best represent a segment of a given length
- The three residues is selected as following:
  1. Select the first and last residue A and B
  2. Select residue C such that the distance $d$ from C to straight line segment AB is maximized
  3. Using three distances to represent the triple: $d$, $|AB|$, $\max(|AC|, |BC|)$
- Another predefined parameter determines how much two adjacent fragments overlap.
DISTRIBUTION (SEGMENT SIZE = 5)
DISTRIBUTION (SEGMENT SIZE = 10)
PRECISION VS RECALL

Triples, 6 Clusters

Dihedral angles, 6 Clusters
FOLD VS RECALL

Triples, 6 Clusters

Dihedral angles, 6 Clusters
CLASS VS RECALL

Triples, 6 Clusters

Dihedral angles, 6 Clusters
SUMMARY
SIGNIFICANCE

- Avoid alignment. Runs fast: $O(n)$ complexity.
- Automatically learn important patterns. No predefined fragment libraries are needed.
WEAKNESS AND FUTURE WORK

• Performance is not as good as alignment based methods
• Possible improvement one: Using multiple strings
QUESTIONS?
REFERENCE

1. An Introduction to Proteins