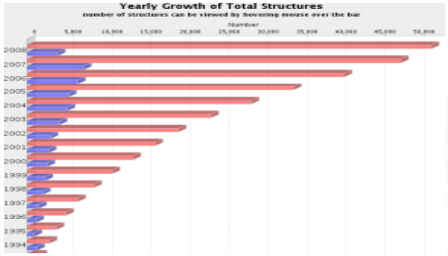
Contacts-based indexing, retrieval and alignment of protein structures

Ahmet Sacan, I. Hakki Toroslu, and Hakan Ferhatosmanoglu
The Ohio State University

Introduction

- Protein structure is more conserved than sequence and can provide valuable information about functional and evolutionary relationships.

- The rapid growth in the protein structure database (PDB) makes finding similar protein structures a real challenge



- Problems to solve:
  - Retrieve structures similar to a query
  - Obtain structural superposition of query with the retrieved structures.

Structure similarity

- The similarity of two protein structures are defined in terms of their alignment.
- The quality of an alignment is measured by the following metrics:
  - Error: Root Mean Square Deviation

$$RMSD = \sqrt{\frac{\sum_i d_i}{N}}$$

  - Coverage:

$$\%N\text{: percentage of residues aligned}$$

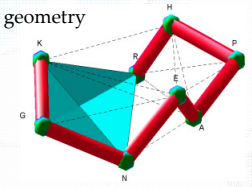
  - Quality: a combined, normalized measure.

$$TM\text{-}score = \frac{1}{L_{target}} \sum_i^N \frac{1}{1 + (\frac{d_i}{d_0(L_{target})})^2}$$

Methods

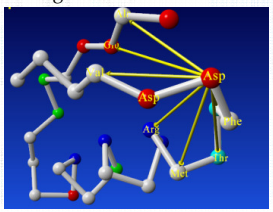
1. Residue Contacts

- **Observation:** Residues in similar structures share similar inter-residue contacts.
- Use Delaunay Tessellation to extract the contacts
  - Well=defined
  - Captures local geometry



2. Contact Strings

- Encode the contacts of a residue based on their sequence order along the backbone.
- Record both the amino acid (AA) type and the secondary structure (SS) information in the contact string.



1jbe Asp13: $V_E\ D_C \parallel D_C^{\#} \parallel F_C\ S_H\ M_H\ -\ E_C\ -$

3. Comparing Contact Strings

- Sequence alignment can be used to obtain the distance between two contact strings.

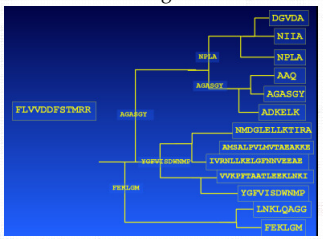
1jbe Asp13: $V_E\ D_C \parallel D_C^{\#} \parallel F_C\ S_H\ M_H\ -\ E_C\ -$
1s8n Asp21: $E_C\ -\ \parallel D_C^{\#} \parallel E_C\ A_H\ R_H\ G_C\ D_C\ G_H$

- Apply alignment in a piece-wise fashion around the central residues to enforce alignment of central residues and to save time.
- We need a residue scoring matrix that captures **provides** both AA and SS information.
- We construct such a matrix from an amino acid substitution matrix and a secondary structure element substitution matrix, using weighted summation:

$$M = w_1 AA + w_2 SS$$

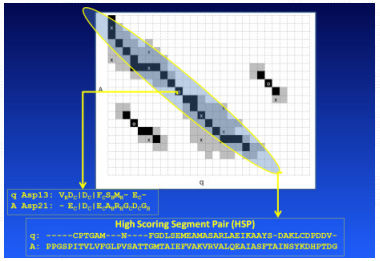
4. Indexing

- Seller's Theorem (1974)
  "If a metric substitution matrix is used, the resulting alignment scores also form a metric."
- We prove that:
  "Weighted summation of two metric functions is also metric."
- The metricity of the distances allows the use **of** distance-based indexing.

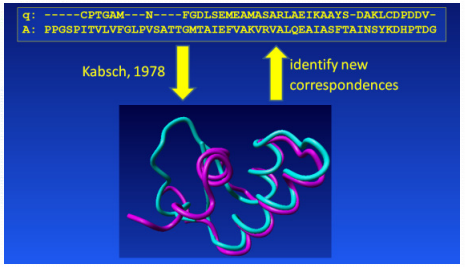


5. Seed Extension

- A contact string hit is extended on both sides. We introduce several improvements over the classical extension idea used in BLAST.



6. Structural superposition



Experiments

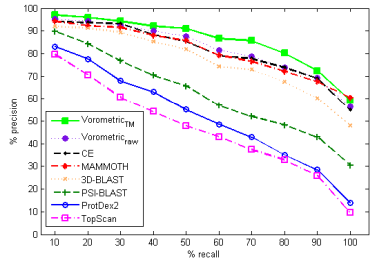
Alignment Quality

- "ten difficult pairs" dataset. Our method is called "Vorometric"

method	RMSD (Å)	%N (query)	quality (TM)
CE	3.17	83.4	0.60
SSAP	4.37	**88.1**	0.59
DaliLite	2.82	80.0	0.61
Vorolign†	**2.28**	51.7	0.56
Vorometric	3.02	84.8	**0.65**

Similarity Search

- ASTRAL-90 database: 34,055 proteins



	avg. precision (%)	time per query	superposition
Vorometric-TM	**82.9**	51 sec	**yes**
Vorometric-raw	79.7	44 sec	no
CE	80.9	14 hours	**yes**
MAMMOTH	80.8	1.6 hours	**yes**
3D-BLAST	76.2	14 sec	no
PSI-BLAST	61.8	**8 sec**	no

Classification

- ASTRAL-25: v1.65-v1.67 difference set

	Family	Superfam	Fold	TM	%N	rmsd
Vorometric-TM	**90.7**	**94.9**	97.6	0.74	**87.2**	2.43
Vorometric-raw	85.9	91.2	97.0	—	—	—
Vorolign	86.4	92.4	**97.7**	0.74	76.3	**1.9**
CE	84.6	91.9	94.1	**0.77**	78.2	1.95
SSEA	60.8	68.9	75.6	—	—	—
BLAST	48.9	52.5	52.8	—	—	—

Reference

Ahmet Sacan, I. Hakki Toroslu, and Hakan Ferhatosmanoglu. Distance-based Indexing of Residue Contacts for Protein Structure Retrieval and Alignment. *IEEE 8th International Symposium on Bioinformatics & Bioengineering. 2008.*