Geometric Suffix Tree: A New Index Structure for Protein 3-D Structures

Tetsuo Shibuya

Human Genome Center,

Institute of Medical Science,

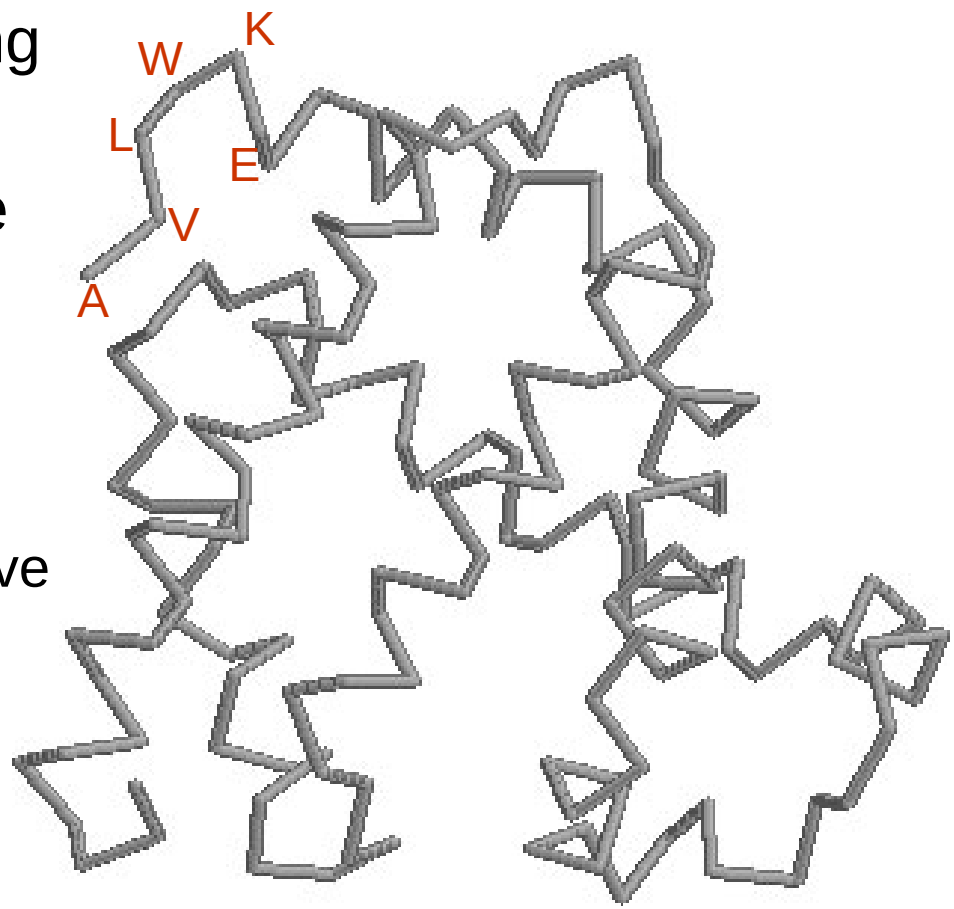
University of Tokyo

Today's Talk

- Backgrounds
  - Protein structures
  - Suffix Trees
- Geometric suffix tree
  - Generalization of suffix trees for indexing protein structures
  - Experiments
- Conclusions

Protein Structure

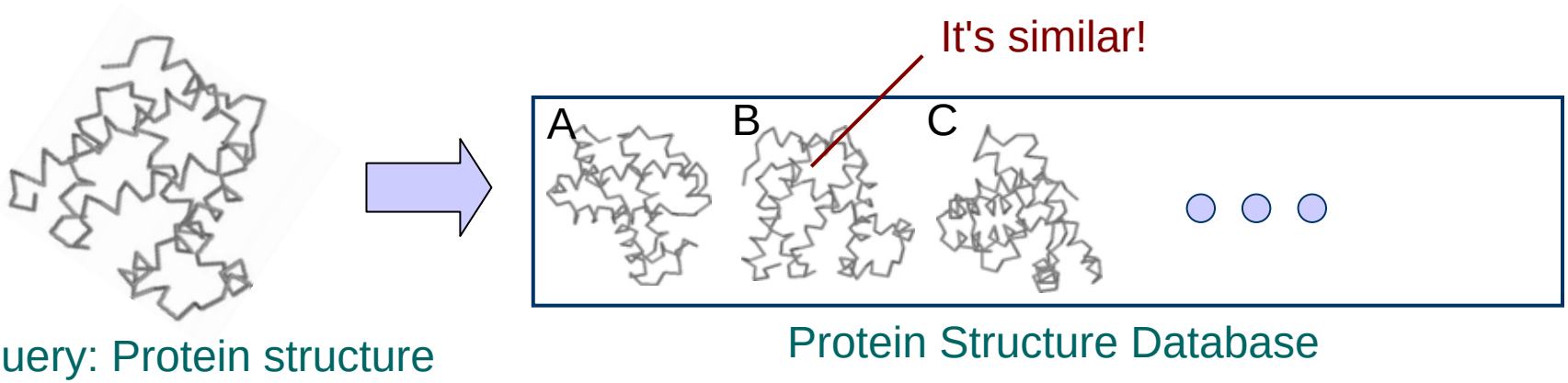
- Protein
  - A chain molecule consisting of 20 kinds of amino acids
  - Folded into some structure
- 3-D structure
  - Coordinates of C_α atoms (backbone)
    - C_α atom: The representative atom of an amino acid



Backgrounds

- Structurally similar proteins
  - tend to have similar functions
  - even if not similar in the residue level
- Structural search on a protein structure database
  - Functional analysis for proteins with newly solved structures
  - Increasing database size (PDB: 35,000~ entries)

→ Sophisticated index structure is desired!



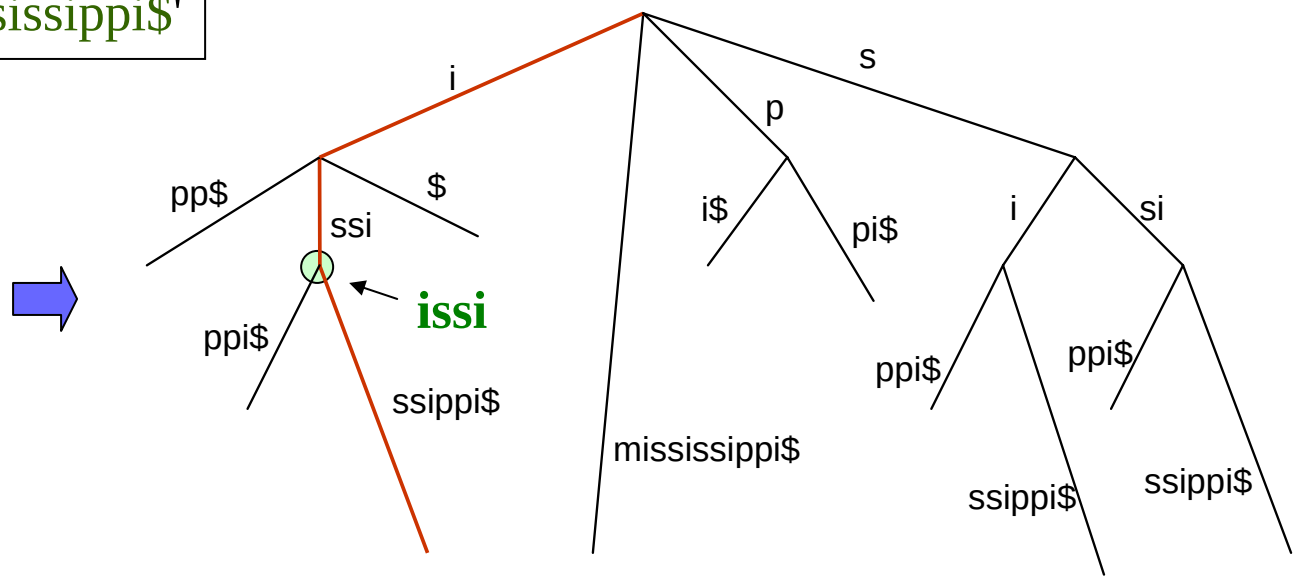
Suffix Tree [Weiner '73]

- A sophisticated index structure for strings
- Compacted trie of all the suffixes of a string *S*
  - Each leaf corresponds to a suffix of *S*
  - Enables efficient substring search

Suffix tree of 'mississippi$'

All the suffixes

mississippi$
ississippi$
ssissippi$
sissippi$
issippi$
ssippi$
sippi$
ippi$
ppi$
pi$
i$



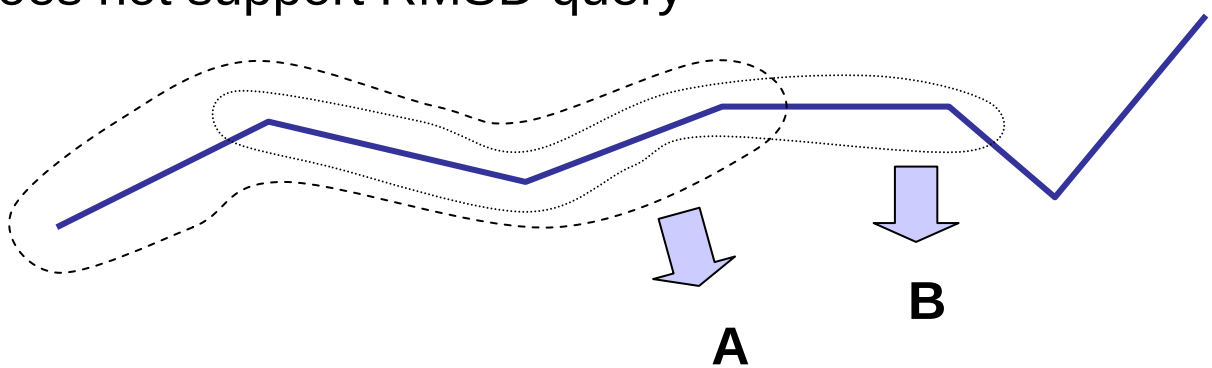
Suffix Tree Features

- Linear-time construction
- Various pattern matching applications
  - Motif finding
  - Repeat finding
  - Large-scale alignment
  - etc.

Good! ...But they are not for structures...

Today's Topic

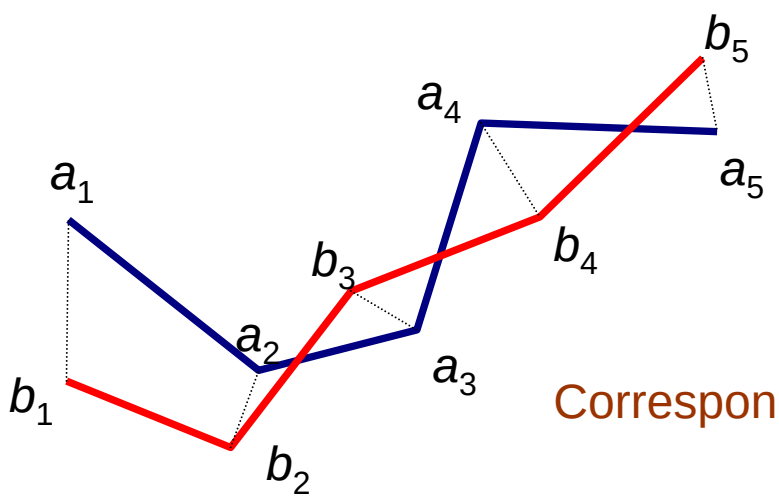
- Extend suffix trees for protein structures
  - "Geometric Suffix Tree"
    - based on the RMSD measures
- Related Work
  - Suffix trees for protein 3-D structure
    - PSIST [Gao et al. '05]
      - Covert structures into alphabetical strings
      - Does not support RMSD query



How to compare two proteins?

- RMSD: Root Mean Square Deviation
  - The most famous measure for protein structure comparison

$$RMSD(A,B) = \min_{R,v} \sqrt{\sum_{i=1}^{n} | a_i - R \cdot (b_i - v) |^2 / n}$$



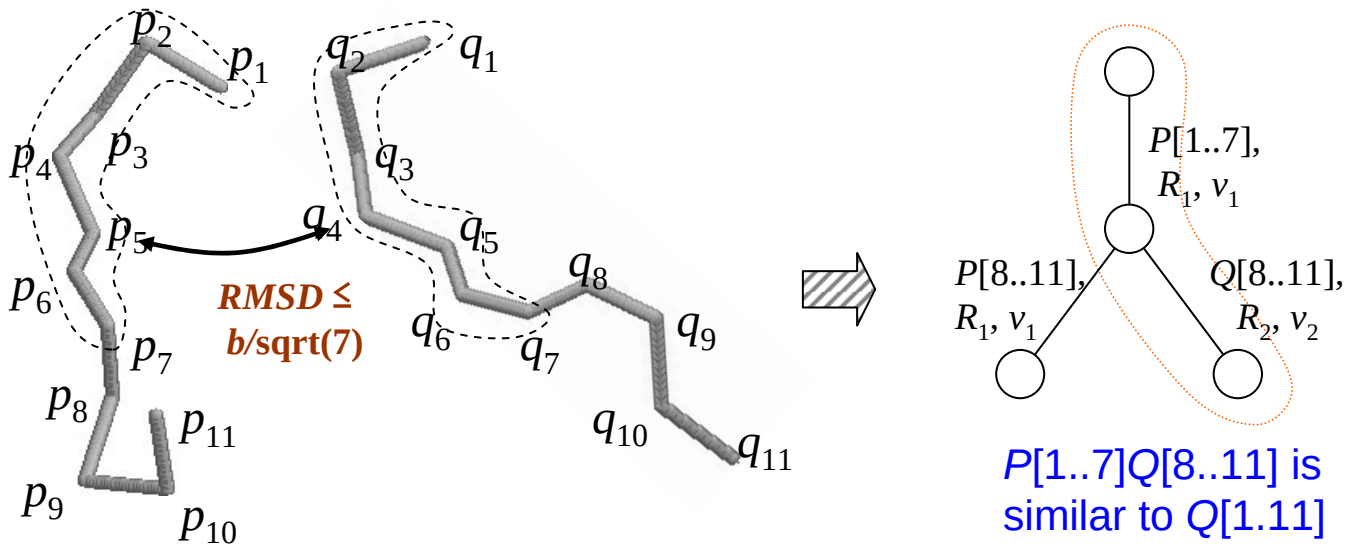
Correspondence of atoms is given

Problem

- Given
  - a substructure of a protein as a query
  - a structure DB
- Find *all* the similar substructures
  - *i.e.* RMSD ≤ some given bound *d*
  - '*All*' means no false negatives/positives

Geometric Trie

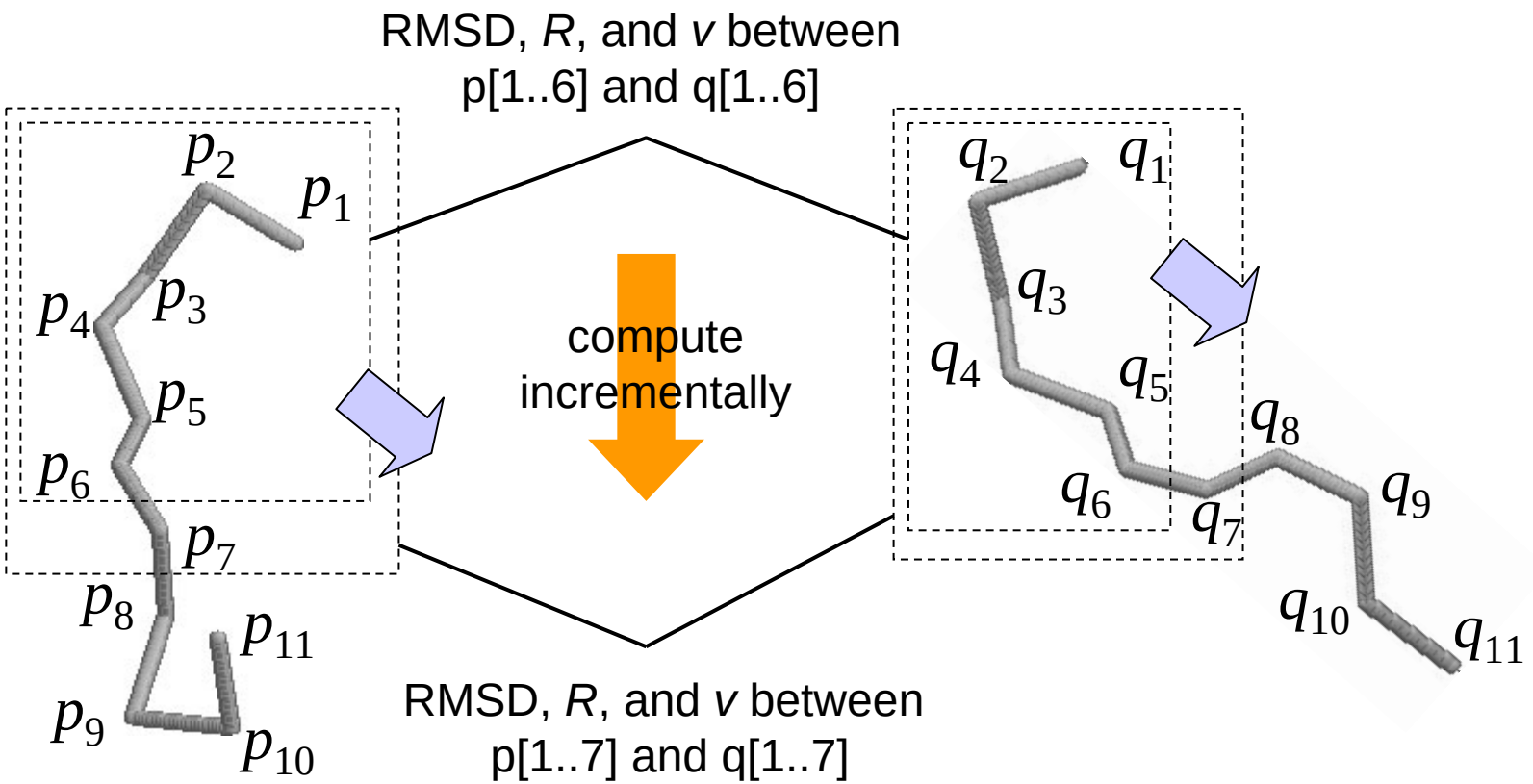
- Tree that represents multiple structures
- Similar prefix substructures are compacted into an edge
  - *i.e.* if sqrt(*l*)·RMSD ≤ *b* (*l*: length of the prefix, *b*: given bound)
    - RMSD(P_i,Q_i) is not always ≤ RMSD(P_{i+1},Q_{i+1})
    - But sqrt(*l*)·RMSD(P_i,Q_i) is always ≤ sqrt(*l*+1)·RMSD(P_{i+1},Q_{i+1})
- Edge information - O(1) size!
  - *i*, *j*: start / end indices (+sequence#)
  - *R*, *v*: the rotation matrix and the translation vector



Geometric Tree

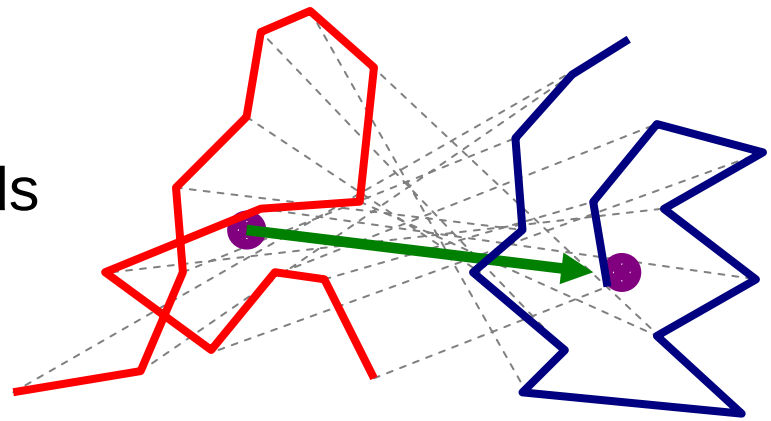
- Geometric trie over all the suffix substructures
- $O(n)$ space (though there are $O(n^2)$ substructures)

But how to compute the RMSD incrementally?



RMSD Computation

- Translation optimization
  - Independent from Rotation Optimization
  - Translate so that two centroids comes to the same position
- Rotation optimization
  - Rotate after translation



$$\underset{R}{minimize} \sum_{i=1}^{n} | \, p_i - R \cdot q_i \, |^2$$

Translated vectors

Optimal Rotation for Minimizing RMSD

[Arun et al. '87, Schwartz et al. '87]

- Problem

$$\underset{R}{minimize} \sum_{i=1}^{n} | \, p_i - R \cdot q_i \, |^2$$

- Solution by SVD (Singular Value Decomposition)
  - Computation time: $O(n)$
  - Post-processing is required in some rare degenerate cases

$R = VU^T$ where

$U\Sigma V$ is the SVD of $H = \sum_{i=1}^{n} q_i p_i^T$

3x3 matrix

Incremental RMSD Computation

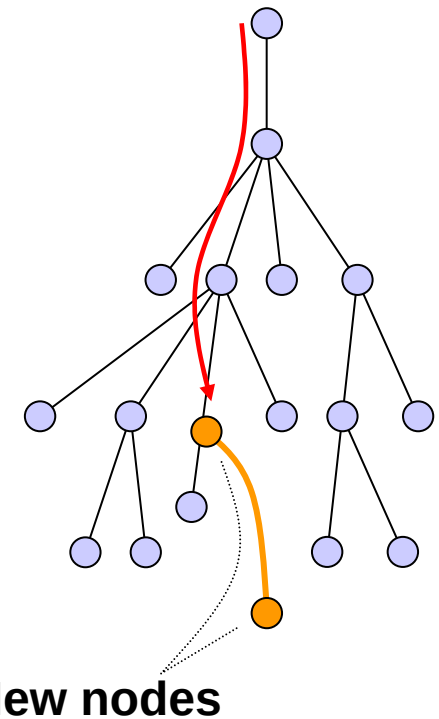
- **Theorem:** *T*he value of the RMSD, *R* and *v* can be computed in constant time if we are given the following values
  - which can be computed incrementally!

$$\sum_{i=1}^{n} q_i p_i^T \qquad \sum_{i=1}^{n} p_i p_i^T \qquad \sum_{i=1}^{n} q_i q_i^T$$

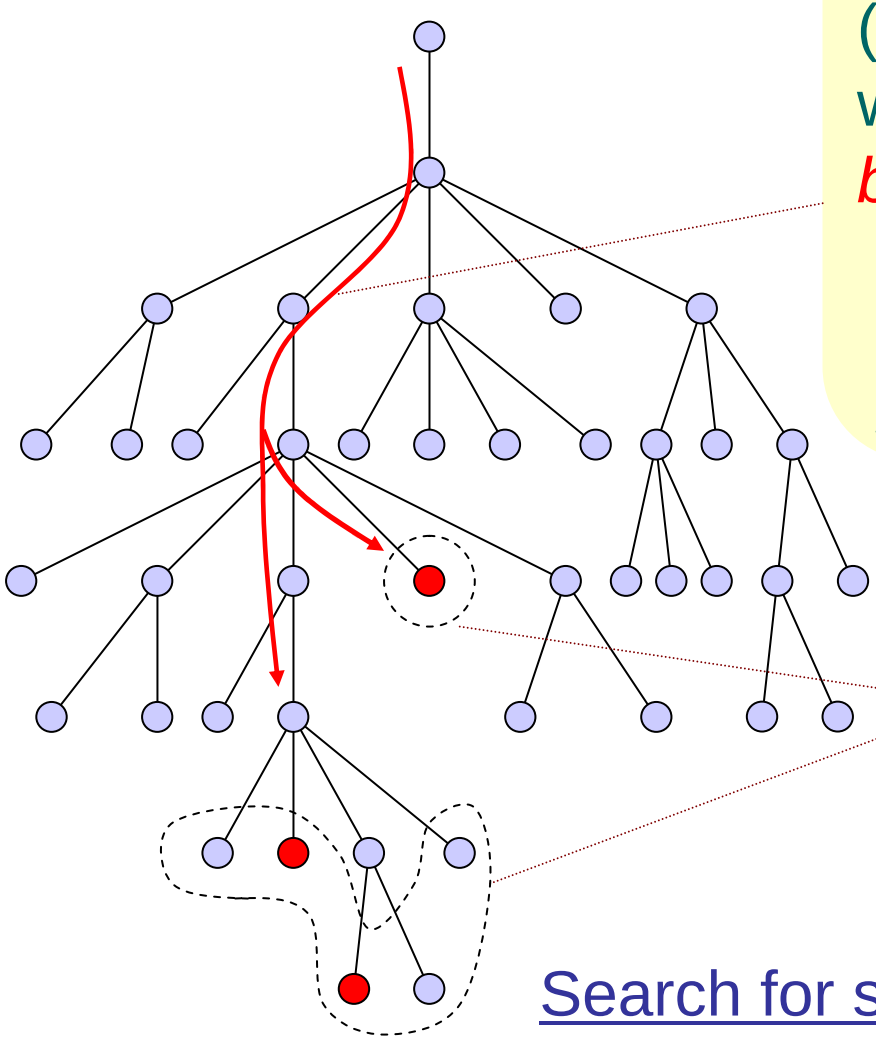
$$\sum_{i=1}^{n} p_i \qquad \sum_{i=1}^{n} q_i$$

Construction Algorithm

- Just add naively each suffix substructures
  - $O(n^2)$ time for a string of size n
    - *cf.* $O(n^3)$ time if we do not use incremental RMSD computation
  - $O(k \cdot n^2)$ time for k structures of sizes at most n



Search Algorithm



Search for all the nodes with (prefix) structures of length *l* whose RMSD to the query is ≤ *b*/sqrt(*l*) + *d*

where

l: query length

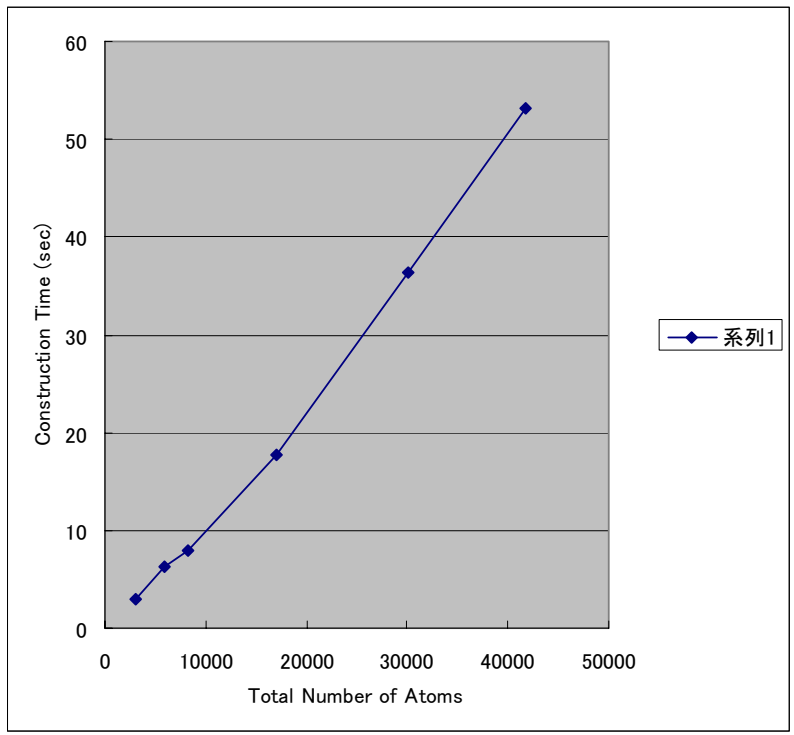
b: bound used in construction

Check whether RMSD ≤ *d*

Search for substructures with RMSD ≤ *d*

Computation Time

- Construction Time
  - Linear to DB size
    - Due to the protein length bound
- Search Time
  - Input
    - Query: 50
    - DB
      - 317 related structures
      - 41,719 atoms
    - RMSD Bound: 1.0Å
      - the bound most often used in protein analysis
  - Results
    - Search time 0.39 sec
      - About 3 times faster than the naive search
    - Reasonable bound
    - 19 hits found



CPU: 1.2GHz UltraSPARC III Cu

Conclusions

- Geometric suffix trees
  - Suffix trees extended for Protein 3-D structures
- Future work
  - More flexible similarity search
  - Faster algorithms (construction / query)
  - Bioinformatics applications
    - Motif finding / functional analysis / protein structure clustering

Thank you very much.

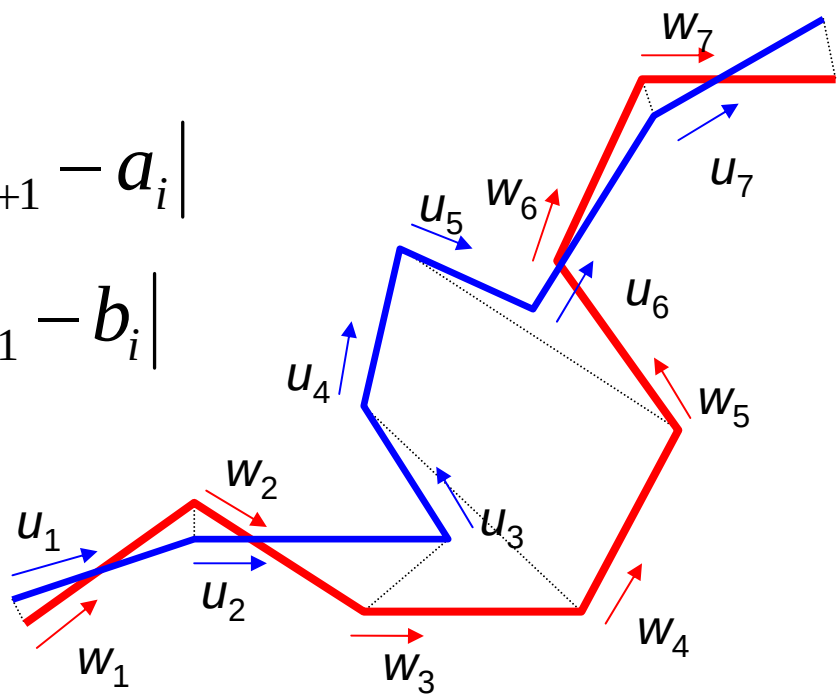
URMSD: Unit-Vector Root Mean Square Deviation

- Variation of the RMSD

$$URMSD(A,B) = \min_{R,v} \sqrt{\sum_{i=1}^{n-1} | u_i - R \cdot (w_i - v) |^2 / n}$$

where

$$\begin{cases} u_i = (a_{i+1} - a_i) / |a_{i+1} - a_i| \\ w_i = (b_{i+1} - b_i) / |b_{i+1} - b_i| \end{cases}$$



Optimizing Rotation

- Problem without translation

$$\underset{\text{rotation } R}{minimize} \sum_{i=1}^{n} | p_i - R \cdot q_i |^2$$

Translated vectors

- i.e.

$$\sum_{i=1}^{n} | p_i - R \cdot q_i |^2 = \sum_{i=1}^{n} \{ p_i^T p_i + q_i^T q_i - (p_i^T R q_i + q_i^T R^T p_i) \}$$

$$= \sum_{i=1}^{n} (p_i^T p_i + q_i^T q_i) - 2 \cdot trace (R \cdot \sum_{i=1}^{n} q_i p_i^T)$$

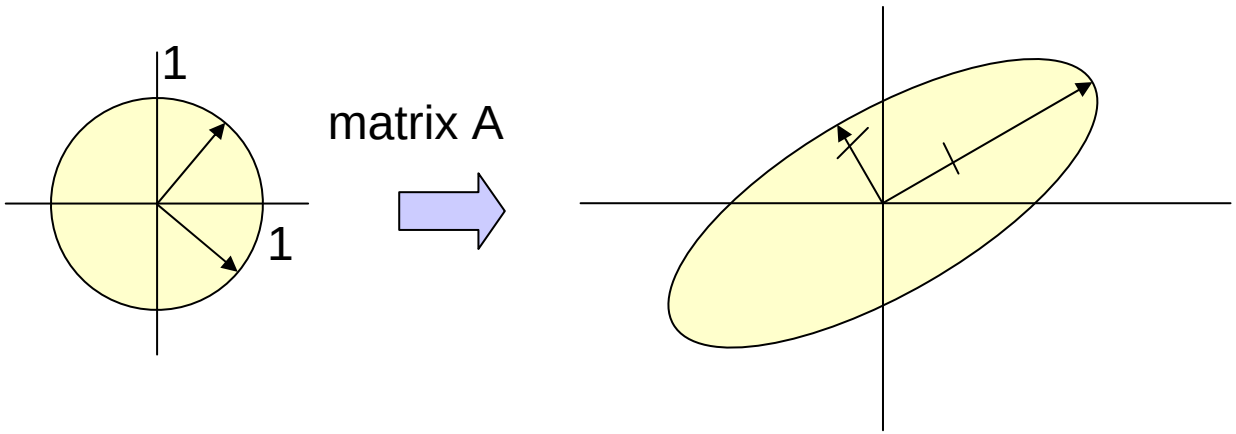
Let it be H

Tips for Optimizing Rotation

- Theorem
  - Given: Positive definite matrix M and orthogonal matrix Q
  - Property: trace(M) >= trace(QM)
- Then...
  - If RH is positive definite, It is the R and H to compute!
    - Note: There's a degenerate case that R is a symmetric matrix
  - Let $R=VU^T$ where $H=U \Sigma V^T$ (Singular value decomposition)
    - $RH=V \Sigma V^T$ is positive definite!
  - SVD of H can be computed in constant time (as H is a 3x3 matrix)
- How to guarantee that R is not a symmetric matrix?
  - det(R) should be 1 (It is -1 in case of a symmetric matrix)
  - If the object is on a 2-D plane, it is easy to compute the actual R by flipping R.
  - In other cases, it is difficult to compute the actual matrix... but it's a rare case. (Heuristically, the above flipped R could be used.)

Singular Value Decomposition

- Computation time: $O(n^3)$ (for an nxn matrix)



Orthonormal vectors v1, v2, ...

Orthonormalized translated vectors u1, u2, ...

$$A(v_1, v_2, ...) = (u_1, u_2, ...) \begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{pmatrix} \Leftrightarrow A = U \Sigma V^T$$

V U

Positive, diagonal matrix

URMSD: Unit-Vector Root Mean Square Deviation

- Variation of the RMSD

$$URMSD(A,B) = \min_{R,v} \sqrt{\sum_{i=1}^{n-1} | u_i - R \cdot (w_i - v) |^2 / n}$$

where

$$\begin{cases} u_i = (a_{i+1} - a_i) / |a_{i+1} - a_i| \\ w_i = (b_{i+1} - b_i) / |b_{i+1} - b_i| \end{cases}$$