

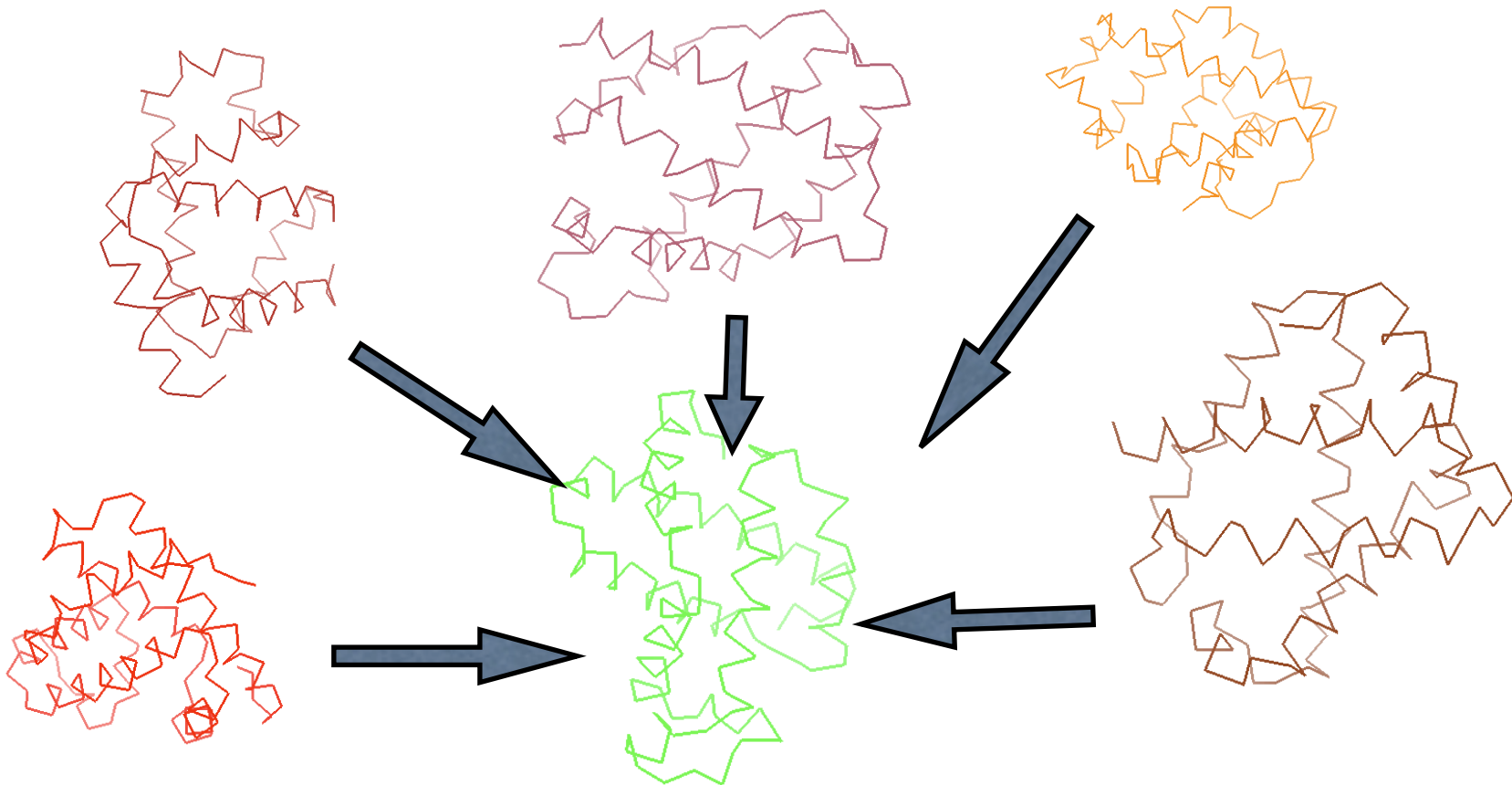
Mean Shape and Protein Backbone Alignment

Jeff M Phillips

(w/ Pankaj Agarwal + Yusu Wang)

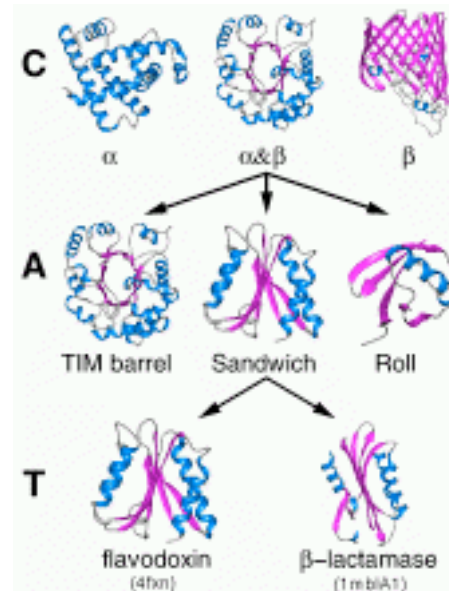
Goal: Create Mean Protein Shape

- Align protein structures
- Create “Mean” protein shape
- Define similarity measure of protein shapes



Why Compare Protein Structures?

Structural classification (SCOP, CATH)



Can provide mean shape / template protein.

- common structure and common function
- template for protein folding (threading)
- identify key elements for protein engineering

Prior Work

Invariant descriptor(dist. matrix) / Geometric Hashing:

- [DALI], [Protein3Dfit], [SSAP], [MAMMOTH], [PRIDE], [Multiprot], [MUSTA], [MALECON], [MASS], [TETRADA], [Russel, Guibas]
- **transformation invariant**

Align SSE (α -helices and β -sheets):

- [PrISM], [VAST], [SARF], [K2], [DEJAVU], [SSM], [Matras], [Singh/Brutlag]
- **can construct mean shape**

Align backbones (usually $C\alpha$) by RMSD:

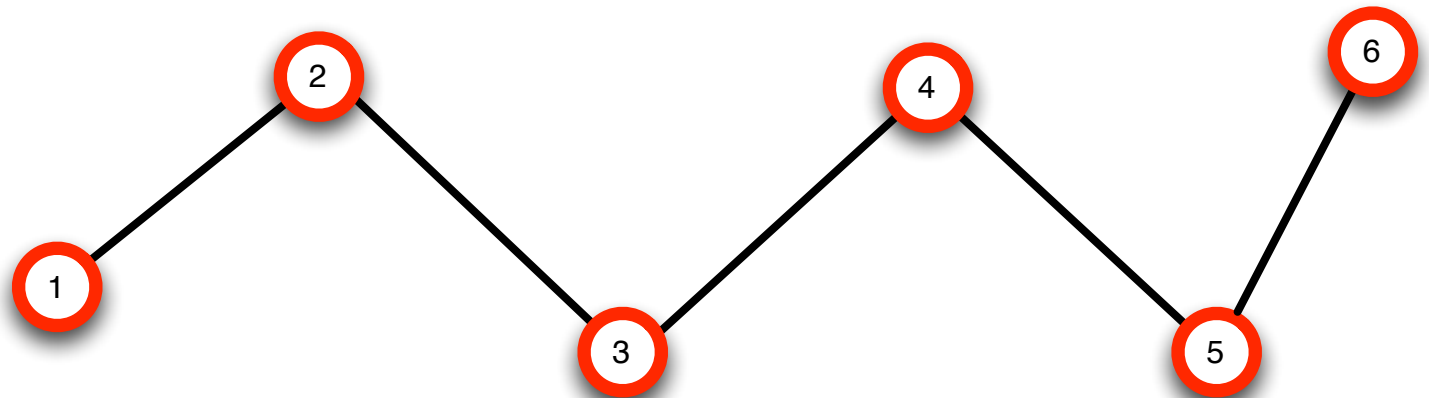
- [CE], [POSA], [STAMP], [SCALI], [SHEBA], [LGA], [PRIDE], [COMPARER], [Falicov, Cohen], [Wu, Schmidler, Hastie, Brutlag]
- **can construct mean shape**

Curve Matching

What ideas from curve matching can be applied to protein backbone matching?

Point-wise matching.

- faster?
- easier



Continuous matching.

- often similar algorithms as point-wise
- more robust of segment length varies
- harder...

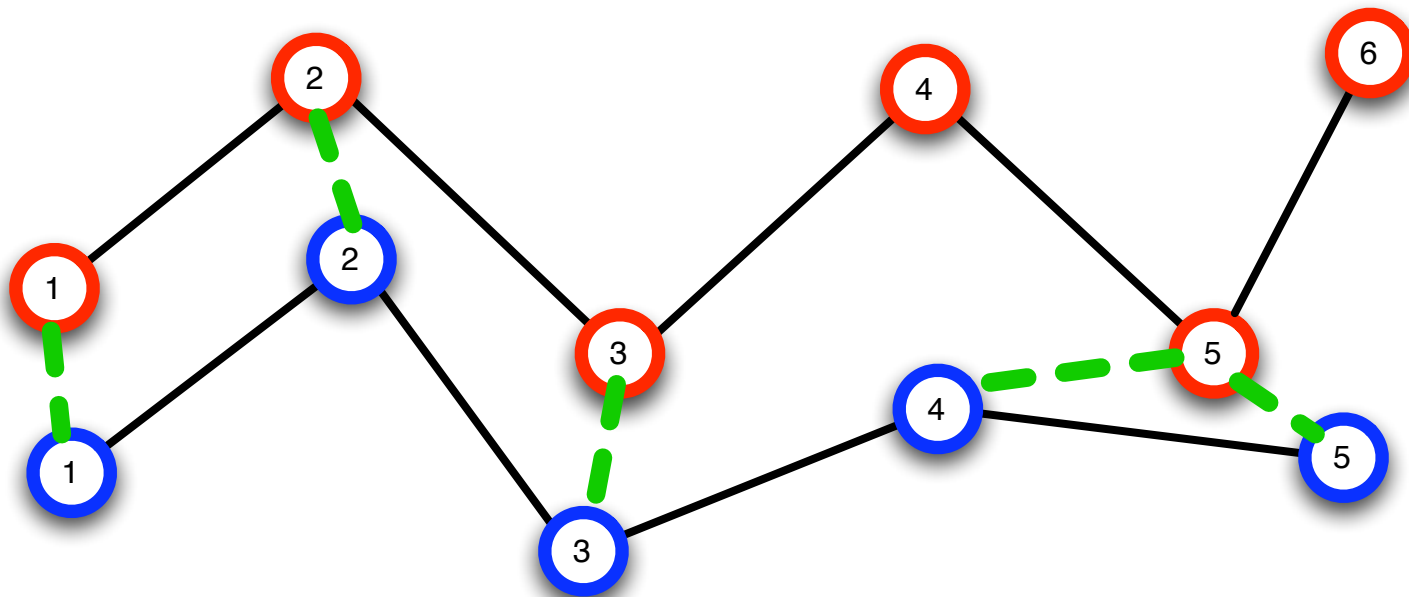
Backbone Matching

rms-Frechet (minimize RMSD): $O(n^2)$

- minimize average squared distance
- gives matching

Frechet (**new - Yusu**): $O(n \log n)$

- minimize maximum distance



Optimal Transformation

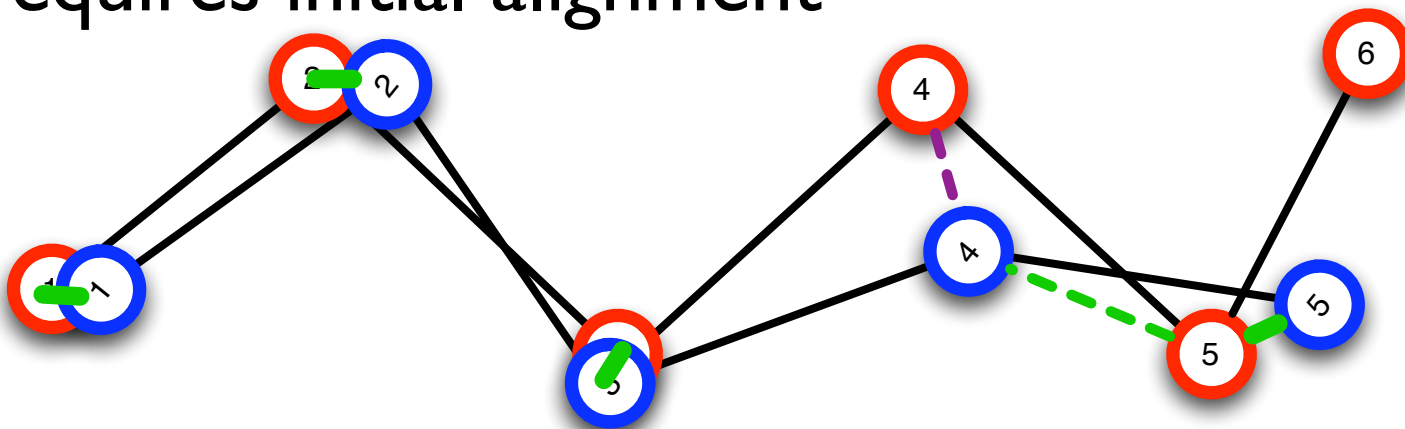
Rotation and Translation to minimize rmsd.

Horn '87 (many others) closed form solution.

- can weight by importance/confidence of points

Iterate between **optimal transform** and **rms-Frechet** is ICP (iterative closest point) algorithm.

- converges quickly
- requires initial alignment



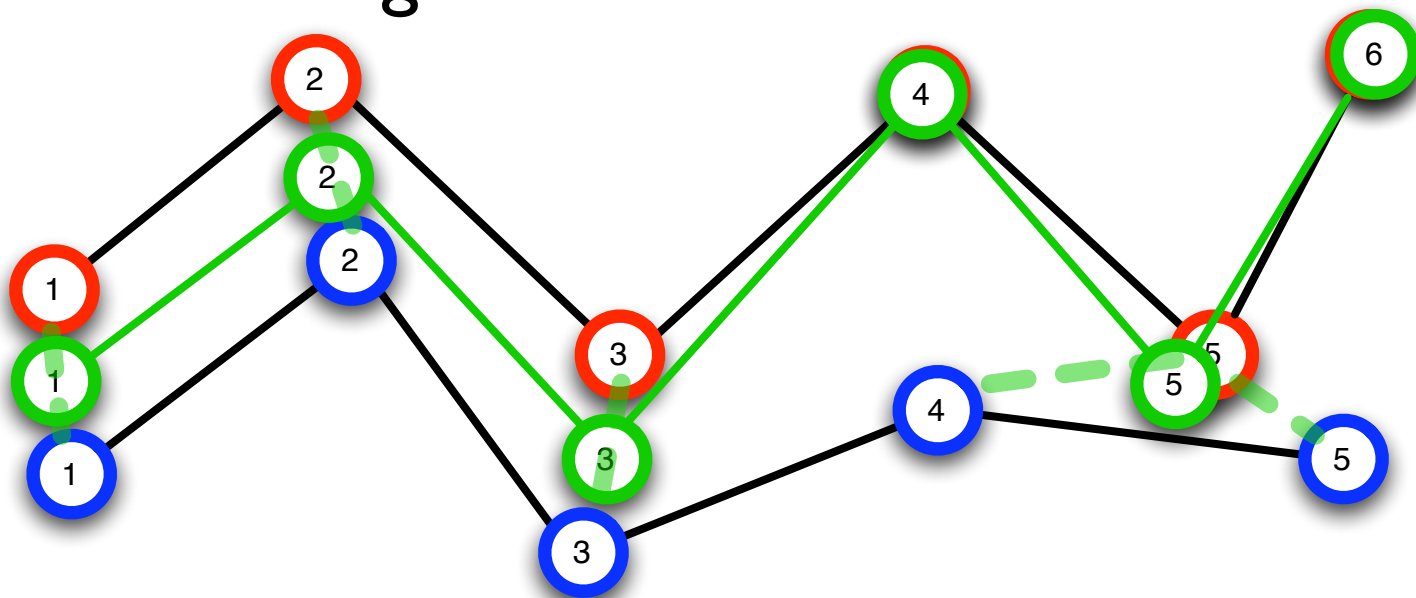
Mean Shape

Mean shape = average of rms-Frechet matching.

Extend to matching multiple curves to red curve.

- or match all curves to mean curve
- can assign confidence by standard deviation

Iterate between **mean shape** and **rms-Frechet** is k-means clustering.



Mean Shape Algorithm

Iterate between:

- **rms-Frechet matching**
- **optimal transform**
- **mean shape**

Initialize by aligning to one backbone by order.

Works well when shapes similar.

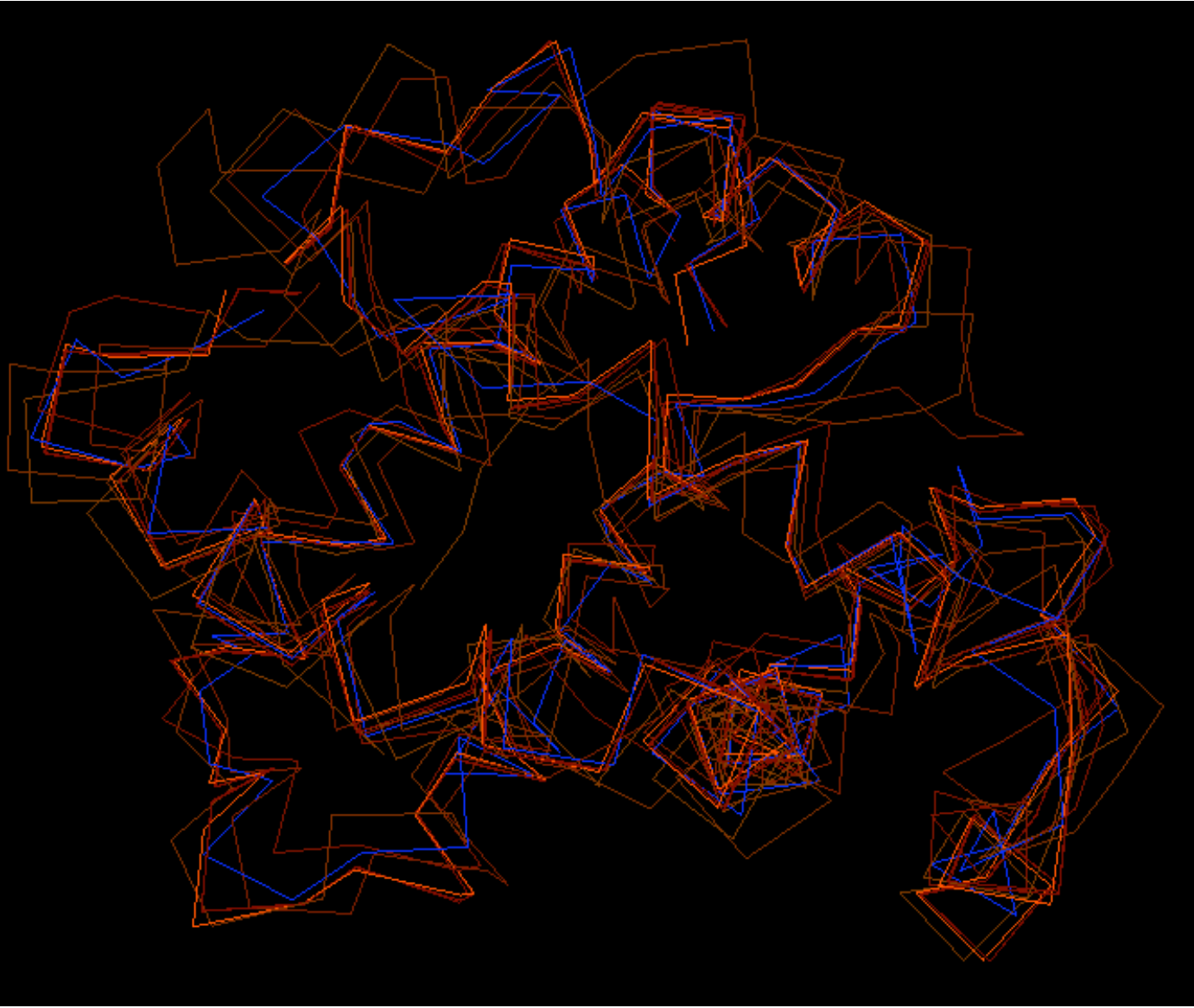
Weighting by standard deviation improves robustness.

Not order dependent.

Case Study: Globins

mean shape

7 globins



1.37 rmsd
<0.1 seconds

How is this new?

Many algorithms:

1. compute all pairwise alignments
2. pick best pair - merge to create mean shape
3. align all remaining to new mean shape
4. goto 2

Problem:

- When two are aligned, they are fixed relative to each other and mean shape.
- When mean shape changes, their alignment to it does not.

How fast is this algorithm?

- Both k-means clustering [Har-Peled, Sadri '04] and ICP [Phillips] have linear lower bounds in size of point sets.
- Thus this algorithm takes at least $\Omega(kn)$ time for k backbones with n atoms.
- ICP has linear convergence [Pottmann03]. k-means has linear convergence in random instances [Vassilvitskii] (...I heard).
- Would imply linear convergence for this algorithm. But variations have quadratic convergence.

Secondary Structure Alignment

future work

We propose to extend our rmsd mean shape algorithm.

- each SSE is a colored point
- use optimal transform extension to segments
- match with continuous rms-Frechet (Agarwal, Phillips) or DP on all SSE
- create mean shape - approx

