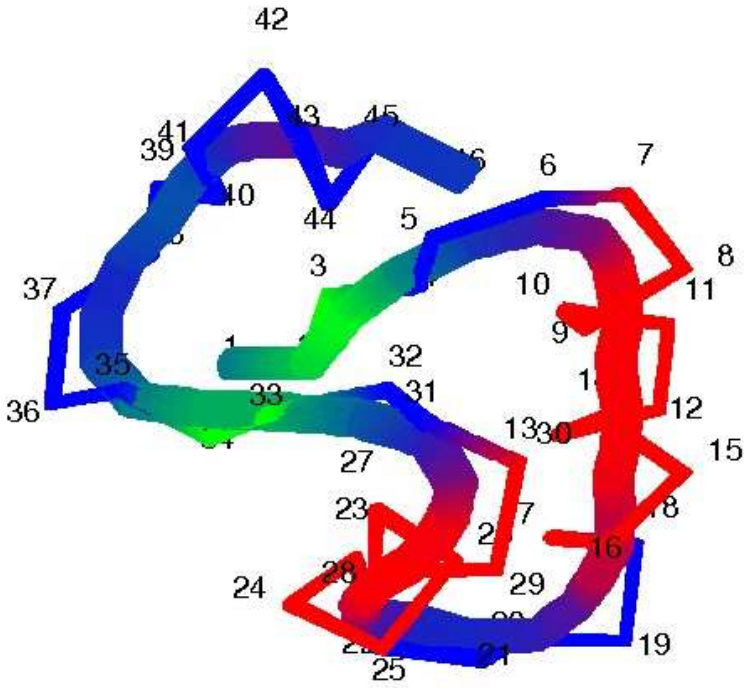
Description and Comparison of Protein 3d-Structures with Emphasis on (Bio)-Topology

Peter Røgen

Technical University of Denmark
Department of Applied Mathematics and Computer Science

13. juni 2017

Representations of protein chains in this talk



The carbon alpha curve (thin)

A smoothened representation (thick)

Lindorff-Larsen, Røgen, Paci, Vendrusscolo & Dobson, Trends in Biochemical Sciences, 30(1), 13-19(2005)

Røgen, J. Phys. Condens. Matter, 17:1523-1538(2005)

A C^2-coloring of the smooth curve

Røgen; Karlsson, Geometriae Dedicata, 134, 91-107, 2008

Methods for comparing protein structures are needed

Methods for comparing protein structures are needed developed

Structure is more conserved than sequence

Even sequence identical structures are highly flexible

Kosloff, Kolodny, Proteins - Struc. Func. and Bioinf. 71(2):891-902(2008)

Current alignments are topologically blind

and structural comparison guide and steer our efforts to predict protein structures

my claim

Mutations cause plastic deformation

New structural alignment methods double every five years for three decades

Hasegawa, Holm, Advances and pitfalls in protein structure alignment, Curr. Opin. in Struc. Biol. 19(3):341-348(2009)

Typical numbers

Known native protein 3d structures: $> 10^5$ chains and fast growing

Number of known protein folds: $> 10^3$ but constant

Protein structure models (generated for structural prediction): 10^5 to 10^6

Typical numbers

Known native protein 3d structures: $> 10^5$ chains and fast growing

Number of known protein folds: $> 10^3$ but constant

Protein structure models (generated for structural prediction): 10^5 to 10^6

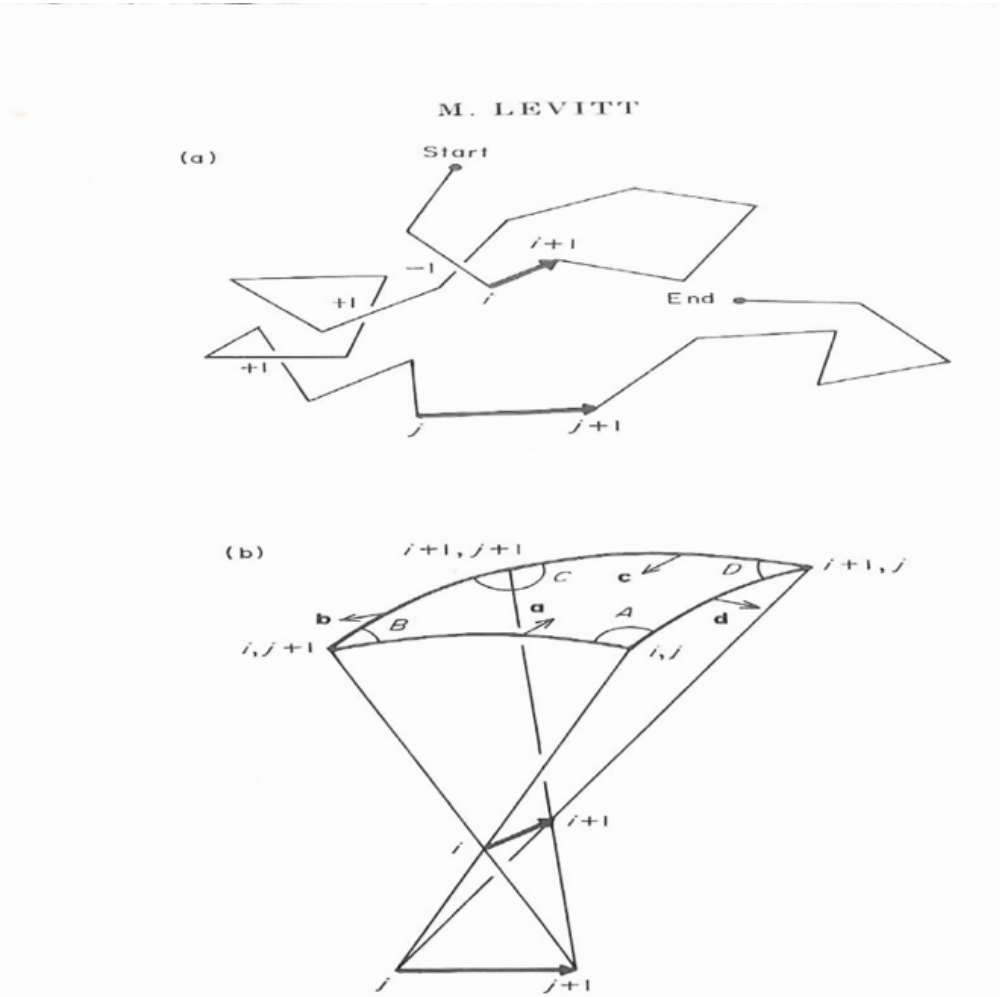
The number of known sequences: huge

Alignment free and alignment methods

Part 1: Separation of protein folds using descriptors
- *Mathematical terms: separation of points*.

Part 2: Introducing Reidemeister moves to structural pair-alignment
- *Mathematical terms: from immersions to embeddings*.

Part 1: The first method to separate protein topologies



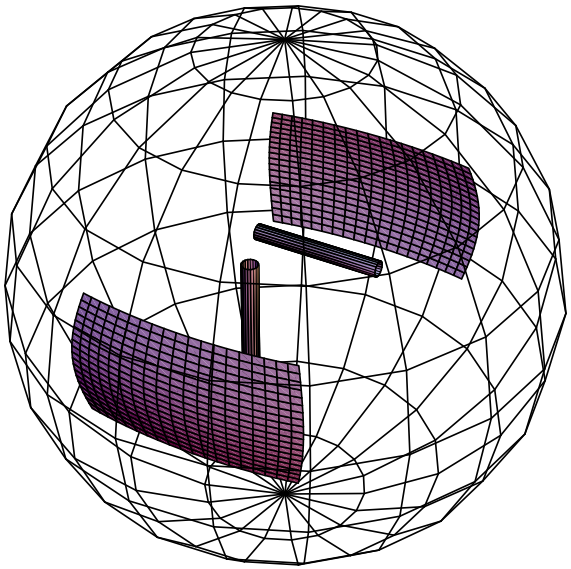
The total writhing number W can be calculated as a sum of the
$, i+1)$ and $(j, j+1)$ overlap when the structure is viewed from a rand
from the positions of the segments. The 2 segments are seen to o
n a direction that passes through the spherical quadilateral shown.
hat give the area of the quadrilateral in terms of angles, A, B, C

Michael Levitt separated protein models using the writhe.

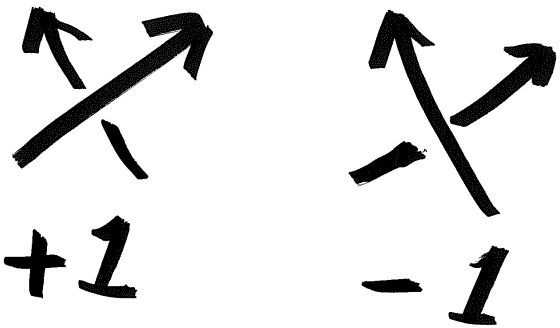
Distinct threaded models where separated by 1.6 in writhe.

Levitt, J. Mol. Biol., 170, 723-764(1983)

Writhe



$|W_{i,j}|$ denotes the probability to see the i'th and the j'th line segment cross when looking from an arbitrary direction.



$W_{i,j}$ equals $|W_{i,j}|$ times the sign of the crossing.

Writhe and generalized Gauss integrals

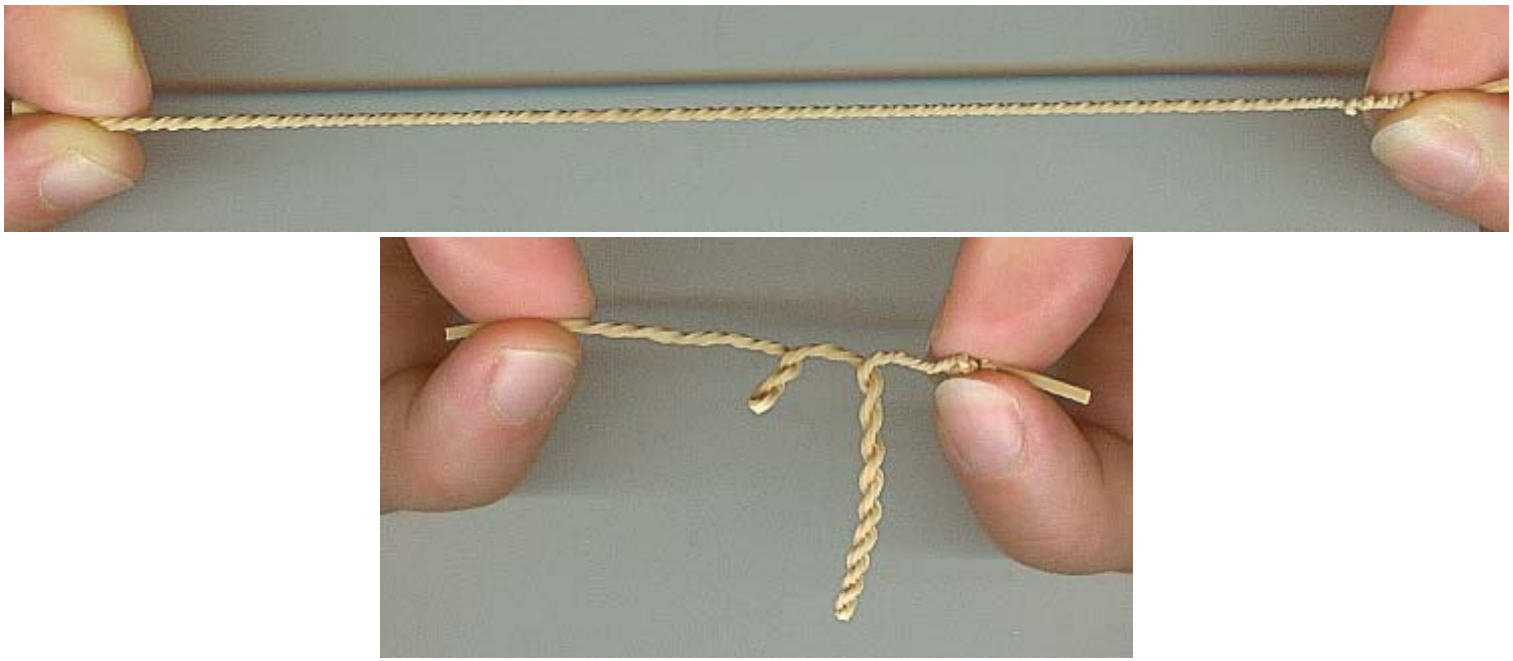
$$I_{(1,2)} = \sum_{i_1<i_2} W_{i_1,i_2} \qquad \text{Writhe}$$

$$I_{|1,2|} = \sum_{i_1<i_2} |W_{i_1,i_2}| \qquad \text{ACN}$$

$$I_{(1,2)|3,4|} = \sum_{i_1<i_2<i_3<i_4} W_{i_1,i_2}|W_{i_3,i_4}|$$

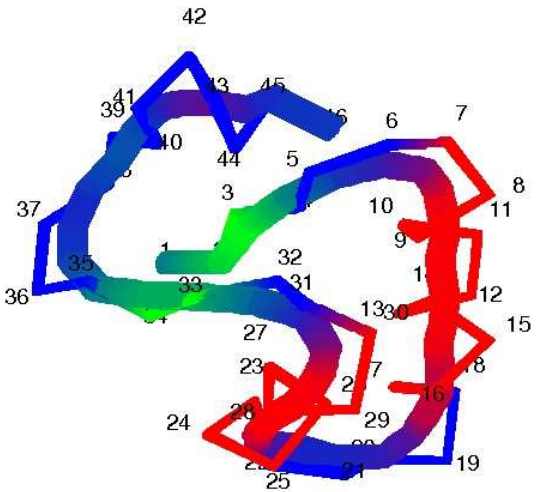
$$I_{(1,4)(2,5)(3,6)} = \sum_{i_1<i_2<i_3<i_4<i_5<i_6} W_{i_1,i_4} W_{i_2,i_5} W_{i_3,i_6}$$

What does writhe look like?



Does writhe grow as $\left(\frac{\text{Length}}{\text{Radius}}\right)^1$ for a fixed knot type?

The parabolic section



The writhe of a space curve $\mathcal{C}$ may be written as

$$\mathrm{Wr}(\mathcal{C}) = \frac{1}{4\pi} \int \int_{\boldsymbol{x},\boldsymbol{x}' \in \mathcal{C}} \frac{\left(d\boldsymbol{x} \times \boldsymbol{r}/r\right) \cdot d\boldsymbol{x}'}{r^2},$$

where $\boldsymbol{r} = \boldsymbol{x} - \boldsymbol{x}'$ and $r = |\boldsymbol{r}|$.

The parabolic section

$$2\,\mathrm{PS}\left(\mathcal{C}, \mathcal{C}'\right) = \int_{\mathcal{C}} \int_{\mathcal{C}'} \frac{\left(d\boldsymbol{x} \times \boldsymbol{r}/r\right) \cdot \left(d\boldsymbol{x}' \times \boldsymbol{r}/r\right)}{r}.$$

The distance excess

$$\mathrm{DE}\left(\mathcal{C}\right) = \mathrm{PS}\left(\mathcal{C}, \mathcal{C}\right).$$

Parabolic section

Let $\mathcal{C}$ and $\mathcal{C}'$ be continuous piecewise C^1-curves starting in $\boldsymbol{a}$ and $\boldsymbol{a}'$ and ending in $\boldsymbol{b}$ and $\boldsymbol{b}'$ respectively. If $\mathcal{C} \cap \mathcal{C}'$ is finite, then

$$\begin{aligned} 2\,\mathrm{PS}\left(\mathcal{C}, \mathcal{C}'\right) &= \int_{\mathcal{C}} \int_{\mathcal{C}'} \frac{(d\boldsymbol{x} \times \boldsymbol{r}) \cdot (d\boldsymbol{x}' \times \boldsymbol{r})}{r^3} \\ &= \left|\boldsymbol{a} - \boldsymbol{b}'\right| + \left|\boldsymbol{b} - \boldsymbol{a}'\right| - \left|\boldsymbol{a} - \boldsymbol{a}'\right| - \left|\boldsymbol{b} - \boldsymbol{b}'\right|. \end{aligned}$$

Røgen; Karlsson, Geometriae Dedicata, 134, 91-107, 2008

Distance excess

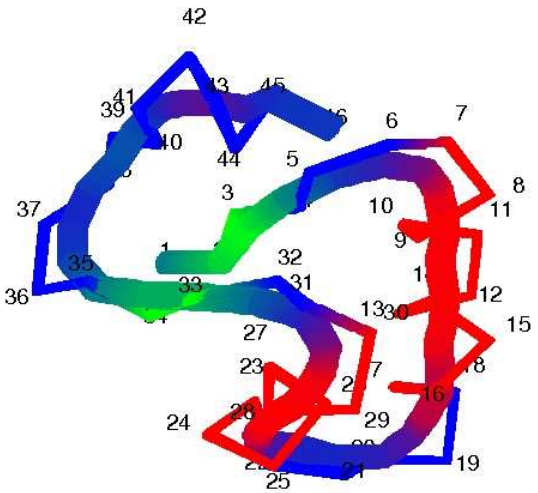
For a continuous piecewise $C^{1,\epsilon}$-curve $\mathcal{C}$ that starts at $\boldsymbol{a}$, ends at $\boldsymbol{b}$, has length $\left|\mathcal{C}\right|$, and has only finitely many points of self-intersection,

$$\mathrm{DE}\left(\mathcal{C}\right) = \left|\boldsymbol{b} - \boldsymbol{a}\right| - \left|\mathcal{C}\right|.$$

"The total action of curvature."

Røgen; Karlsson, Geometriae Dedicata, 134, 91-107, 2008

Colored descriptors



Using the coloring

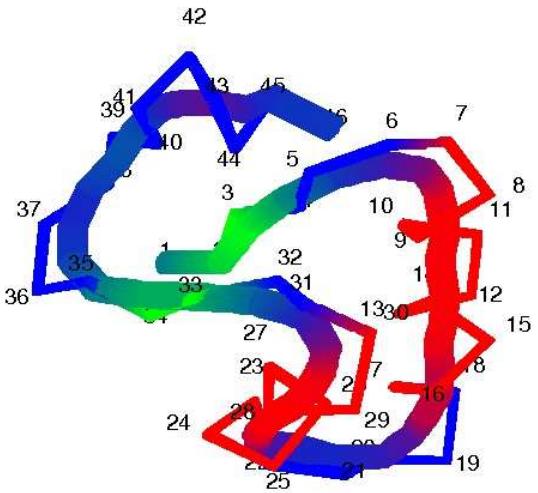
$$\sum_{i<j} \begin{pmatrix} \alpha_i\alpha_j & \alpha_i\beta_j & \alpha_i\gamma_j \\ \beta_i\alpha_j & \beta_i\beta_j & \beta_i\gamma_j \\ \gamma_i\alpha_j & \gamma_i\beta_j & \gamma_i\gamma_j \end{pmatrix} \mathrm{PS}(i,j)$$

we get 9 measures of how "parallel" the secondary structure elements of a protein are.

Likewise a coloring of writhe gives 9 descriptors.

The colored parabolic section is invariant under plastic deformation preserving secondary structure type and terminal points.

Colored descriptors



Using the coloring

$$\sum_{i<j} \begin{pmatrix} \alpha_i\alpha_j & \alpha_i\beta_j & \alpha_i\gamma_j \\ \beta_i\alpha_j & \beta_i\beta_j & \beta_i\gamma_j \\ \gamma_i\alpha_j & \gamma_i\beta_j & \gamma_i\gamma_j \end{pmatrix} \mathrm{PS}(i,j)$$

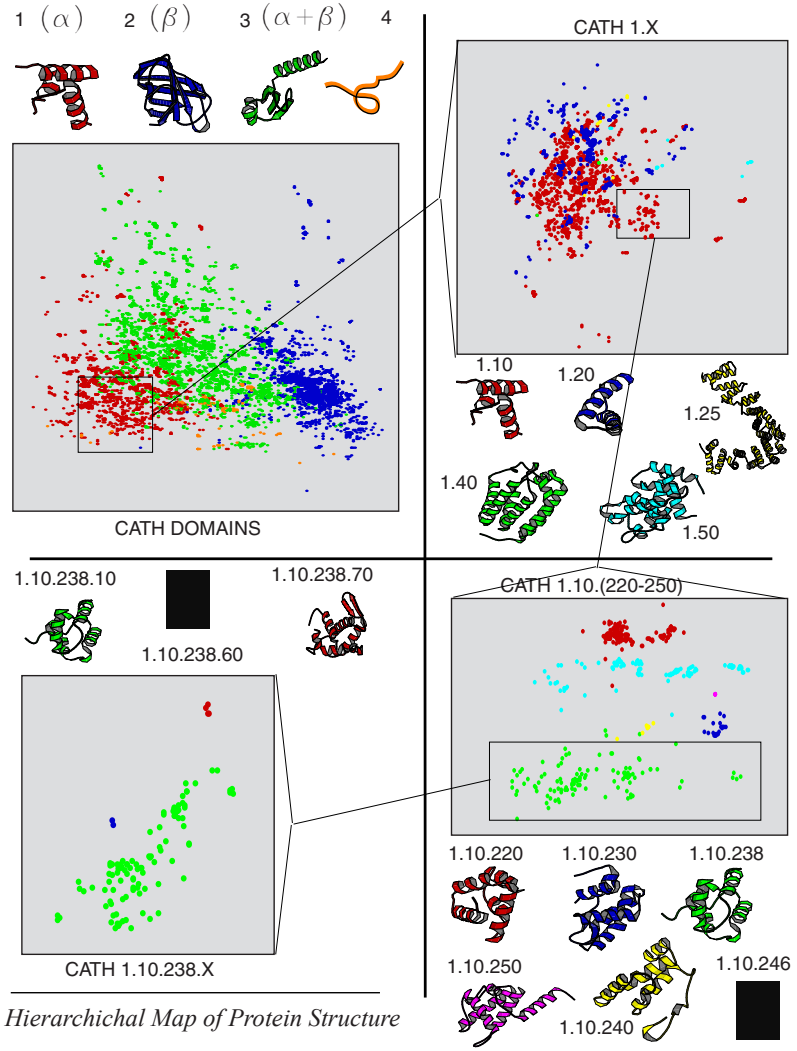
we get 9 measures of how "parallel" the secondary structure elements of a protein are.

Likewise a coloring of writhe gives 9 descriptors.

The colored parabolic section is invariant under plastic deformation preserving secondary structure type and terminal points.

Similarly, the writhe is almost invariant.

Separation of folds



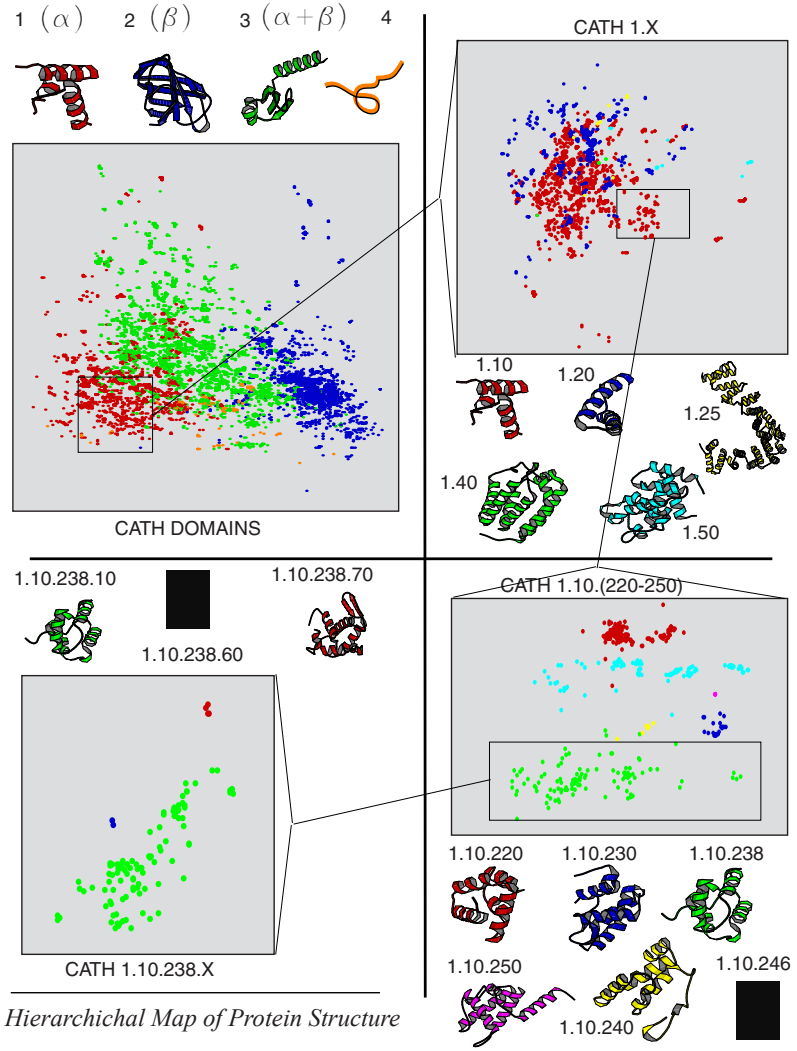
Hierarchichal Map of Protein Structure

using 29 generalized Gauss integrals and length.

96% of 20937 connected CATH2.4 protein domains are automatically classified.

Røgen; Fain, PNAS 100(1), 119-124,(2003).

Separation of folds



Hierarchichal Map of Protein Structure

using 29 generalized Gauss integrals and length.

96% of 20937 connected CATH2.4 protein domains are automatically classified.

Røgen; Fain, PNAS 100(1), 119-124,(2003).

This number seems unchanged for other descriptor families and when optimizing metric properties.

Problems: a) continuous nature of parts of fold space and

b) sub-fold separation.

Descriptors allow fast search (for fixed protein length)

Protein domain level:

- Fat graph: e.g. Penner, Knudsen, Wiuf, Andersen, Comm. on Pure and App. Math., LXIII, 1249-1297(2010)
- Automatic classification `GI.c GIT.c`: Røgen; Fain, PNAS 100(1), 119-124,(2003) with writhe decomposition Zhi, Shatsky & Brenner, Bioinformatics 26(9),1176-1184(2010)
- Clustering (models): `Pleiades`, Harder et al. , Bioinformatics, 28(4), 510-515(2011)
- Prediction of fold class from amino acid sequence: Nielsen, Røgen & Bohr, Math. and Comp. Model., 43(3-4), 401-412(2006)

Local geometry:

- 7-mer sequence to PS$(i, i+1)$: Røgen & Koehl, Proteins Struc. Func. and Bioinfor. 81(5), 841-851(2013)

Descriptors allow fast search (for fixed protein length)

Protein domain level:

- Fat graph: e.g. Penner, Knudsen, Wiuf, Andersen, Comm. on Pure and App. Math., LXIII, 1249-1297(2010)
- Automatic classification `GI.c GIT.c`: Røgen; Fain, PNAS 100(1), 119-124,(2003) with writhe decomposition Zhi, Shatsky & Brenner, Bioinformatics 26(9),1176-1184(2010)
- Clustering (models): `Pleiades`, Harder et al. , Bioinformatics, 28(4), 510-515(2011)
- Prediction of fold class from amino acid sequence: Nielsen, Røgen & Bohr, Math. and Comp. Model., 43(3-4), 401-412(2006)

Local geometry:

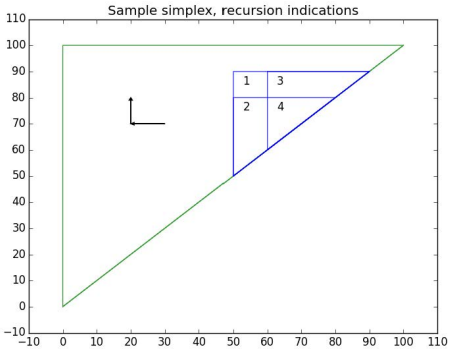
- 7-mer sequence to PS$(i, i+1)$: Røgen & Koehl, Proteins Struc. Func. and Bioinfor. 81(5), 841-851(2013)

There are many other descriptor metods e.g.

`FragBag` Budowski-Tal, Nov & Kolodny, PNAS 107: 3481-3486(2010)

but can a priory not detect topological changes.

A note on writhe

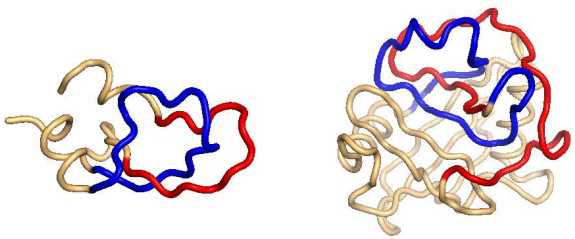


The writhe of all connected sub-chains takes $\mathcal{O}(n^2)$.

Now the mutual writhe of two sub-chains is $\triangle 1 - \triangle 2 - \triangle 3 + \triangle 4$.

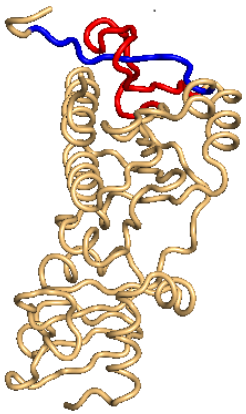
Mutual writhe finds rare configurations including "links" and pokes. Grønbæk, Hamelryck & Røgen, In prep.

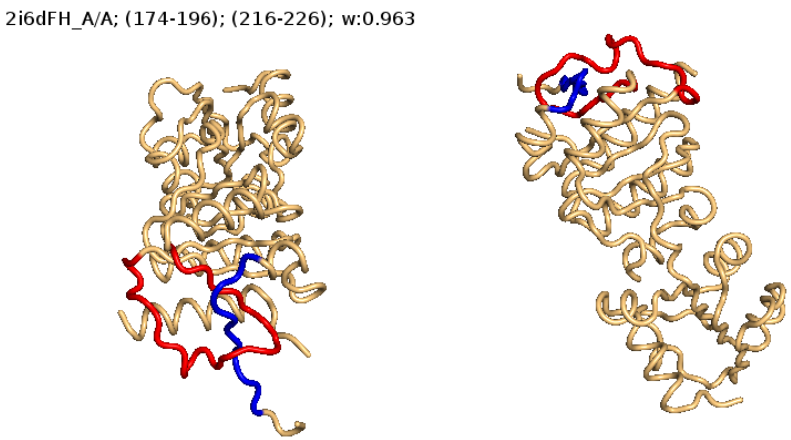
Dabrowski-Tumanski & Sulkowaska, PNAS,114(13)3415-3420(2017)



`GIsquared` calculates second order Gauss integrals of all connected sub-chains - $\mathcal{O}(n^3)$.

Grønbæk, Hamelryck & Røgen





Descriptors allow fast search (for any length-scale)

Intermediate or any length-scale

- Alignment using writhe of a fixed window length:

  Chang et al. BMC Bioinformatics, 7, 346-356(2006)

- Calculating descriptors on all connected sub-chains:
  `GIsquared` Grønbæk, Hamelryck & Røgen, not released
- Detection of rare threading motifs (links, pokes),
  `GIsquared`: Grønbæk, Hamelryck & Røgen, In preparation
- Current research: Descriptor-based structural alignment using both descriptors of fixed windows and **of pairs of windows**.

  Se also: Erdmann, J. Comput. Biol. 12(6) 609-37(2005)

Part 2: Reidemeister moves for structural alignment

`ProteinAlignmentObstruction` - a software for detecting and quantifying

steric and
topological

obstructions to structural alignments of proteins.

Alignment and superposition defines a morph

Coordinate based structural alignment (RMSD, TM-align, GDT-TS, CE, TopMatch, . . .) of two protein structures returns:

- a structure based sequence alignment and
- a superposition of the aligned subsets,

obtained by optimizing a carefully chosen score function

$$F\big(\text{distances between aligned pairs after superposition}\big).$$

Alignment and superposition defines a morph

Coordinate based structural alignment (RMSD, TM-align, GDT-TS, CE, TopMatch, . . .) of two protein structures returns:

- a structure based sequence alignment and
- a superposition of the aligned subsets,

obtained by optimizing a carefully chosen score function

$$F\big(\text{distances between aligned pairs after superposition}\big).$$

This is to score the direct Euclidean distances in the morph

$$(1-t) * \text{Structure}_1 + t * \text{Structure}_2, \text{ for } 0 \leq t \leq 1.$$

Implied assumption:

Alignment and superposition defines a morph

Coordinate based structural alignment (RMSD, TM-align, GDT-TS, CE, TopMatch, . . .) of two protein structures returns:

- a structure based sequence alignment and
- a superposition of the aligned subsets,

obtained by optimizing a carefully chosen score function

$$F\big(\text{distances between aligned pairs after superposition}\big).$$

This is to score the direct Euclidean distances in the morph

$$(1-t) * \text{Structure}_1 + t * \text{Structure}_2, \text{ for } 0 \leq t \leq 1.$$

Implied assumption: there is room for this morph.

Morph analysis

Question 1: Does the morph cause steric problems?
Question 2: Does the morph change the topology?

`ProteinAlignmentObstruction` quantifies answers for alignments and superposition provided by RMSD, TM-align, ...

Morph analysis

Question 1: Does the morph cause steric problems?
Question 2: Does the morph change the topology?

`ProteinAlignmentObstruction` quantifies answers for alignments and superposition provided by RMSD, TM-align, ...

Re question 1: Overlap(i, j) a contact map of the morph based on

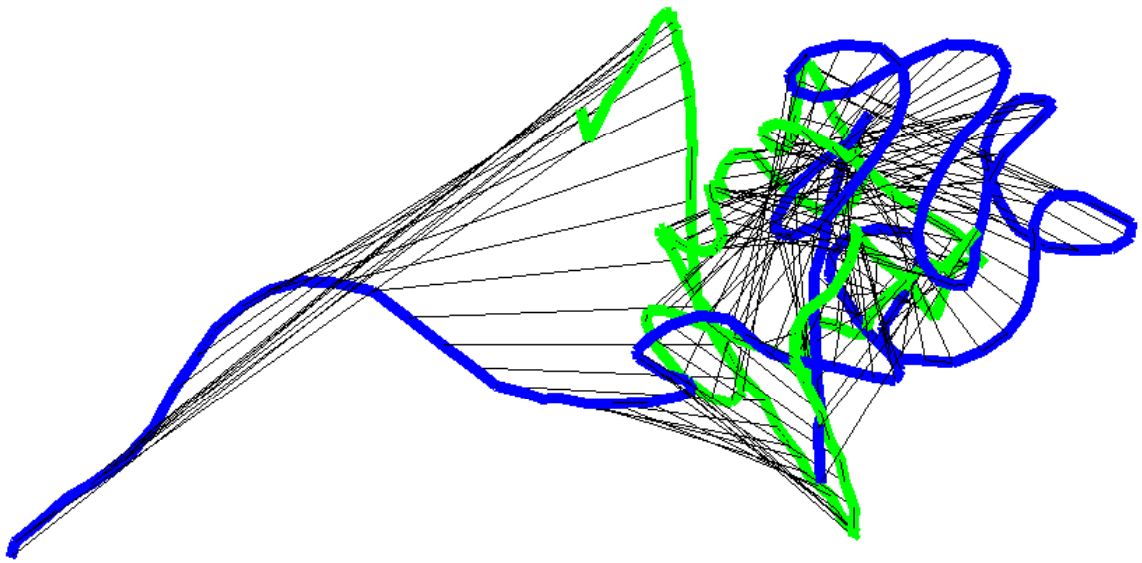
- the shortest distance between aligned pair (i, j) during the morph
- $d_{min}(|i - j|)$ for pairs $|i - j|$ apart along the backbone.

Overlap examples

The A and B chains of the open and closed forms of adenylate kinase, 4AKE and 1ANK both have RMSD $\approx$ 7Å.

These morphs cause **no overlap**. Unusual for an 7Å morph.

The morph between the un- and phosphorylated forms of Odhl 2KB3 to 2KB4 has RMSD=23Å and mean overlap of 1Å (0.6Å smooth curve).



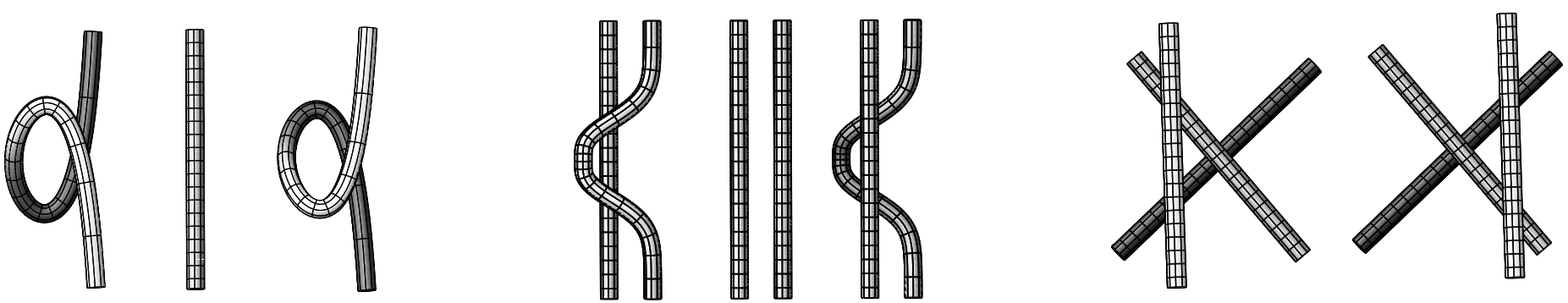
Question 2: quantifying change in (bio)-topology

Mathematically all protein chains share topology when considered as open curves.

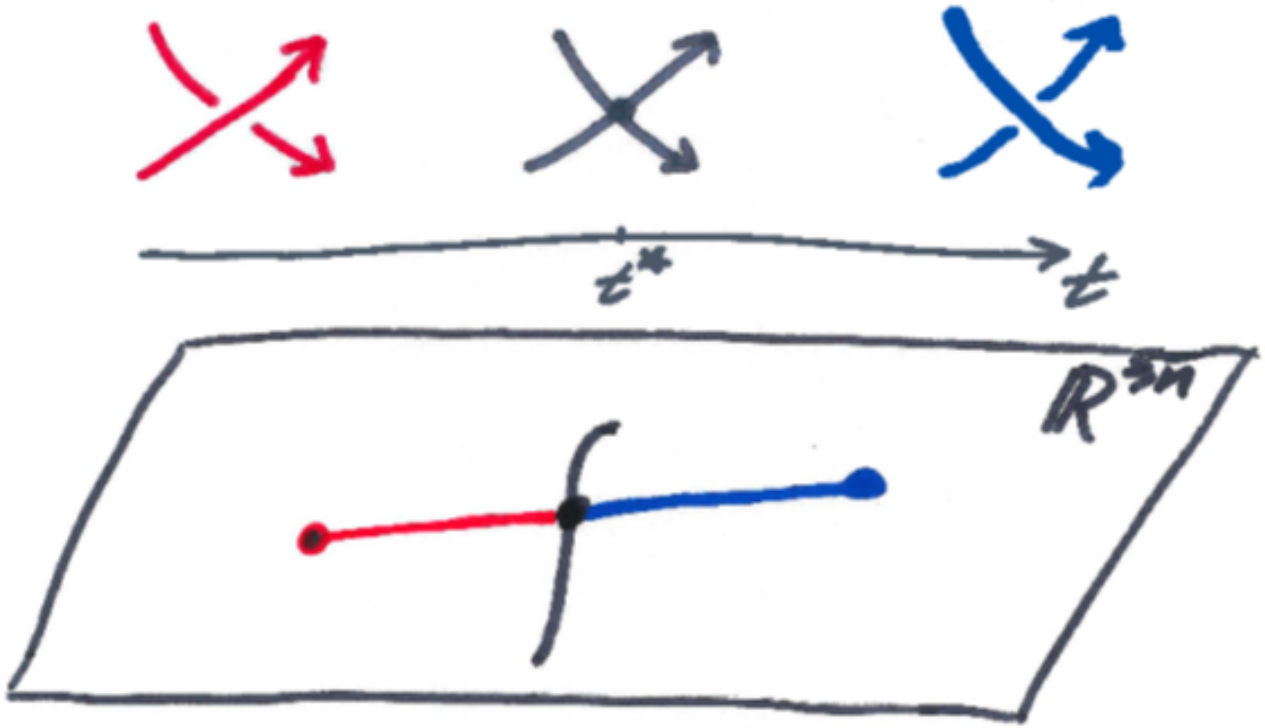
Which deformations result in a change in topology?

An answer, I expect, depends on the contest.

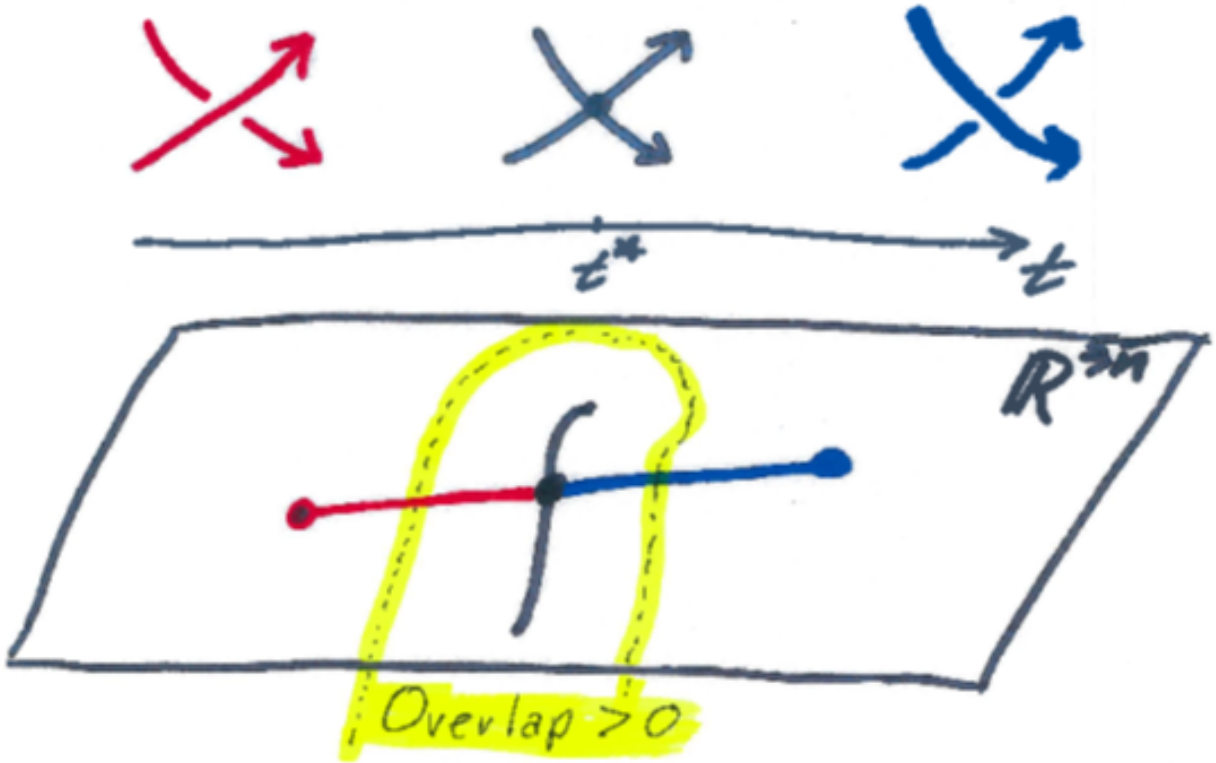
I offer a modelling tool based on a 3d interpretation of Reidemeister Moves



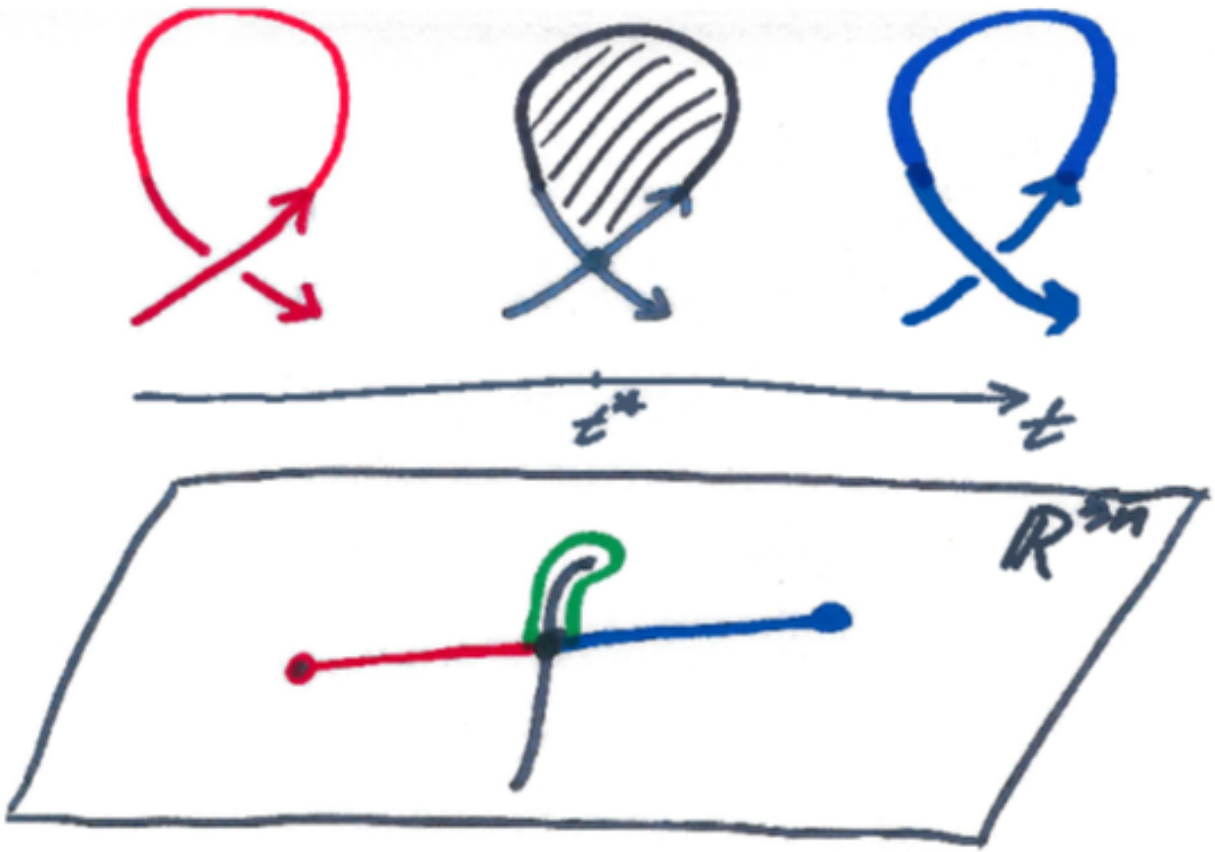
Defining a morph self-intersection



Filtering potential morph self-intersections



Avoiding one morph self-intersection
- Reidemeister move Ω_1



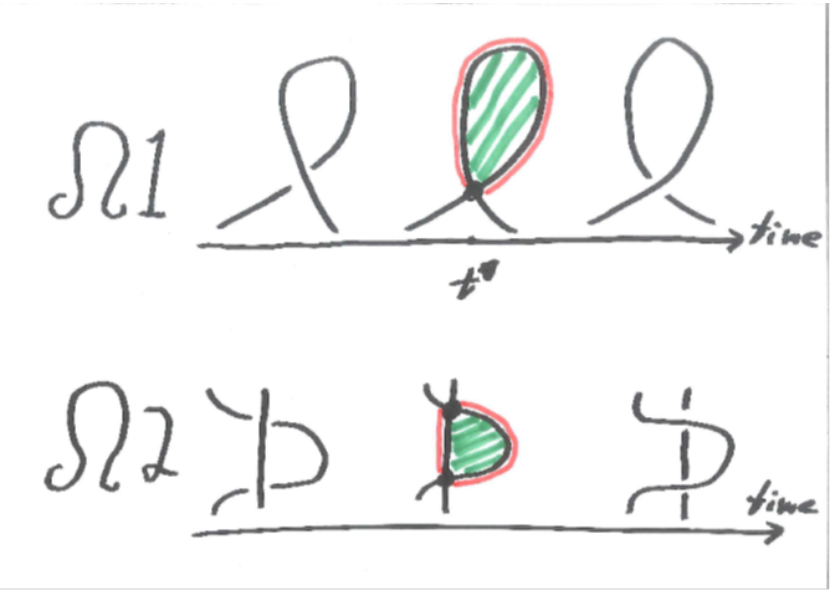
ProteinAlignmentObstruction

A morph $(1-t)*\mathrm{Structure}_1 + t*\mathrm{Structure}_2$ is given by a structural alignment program.

A) Overlap detects steric morph problems.
B) All morph self-intersections of the backbone curve are found.

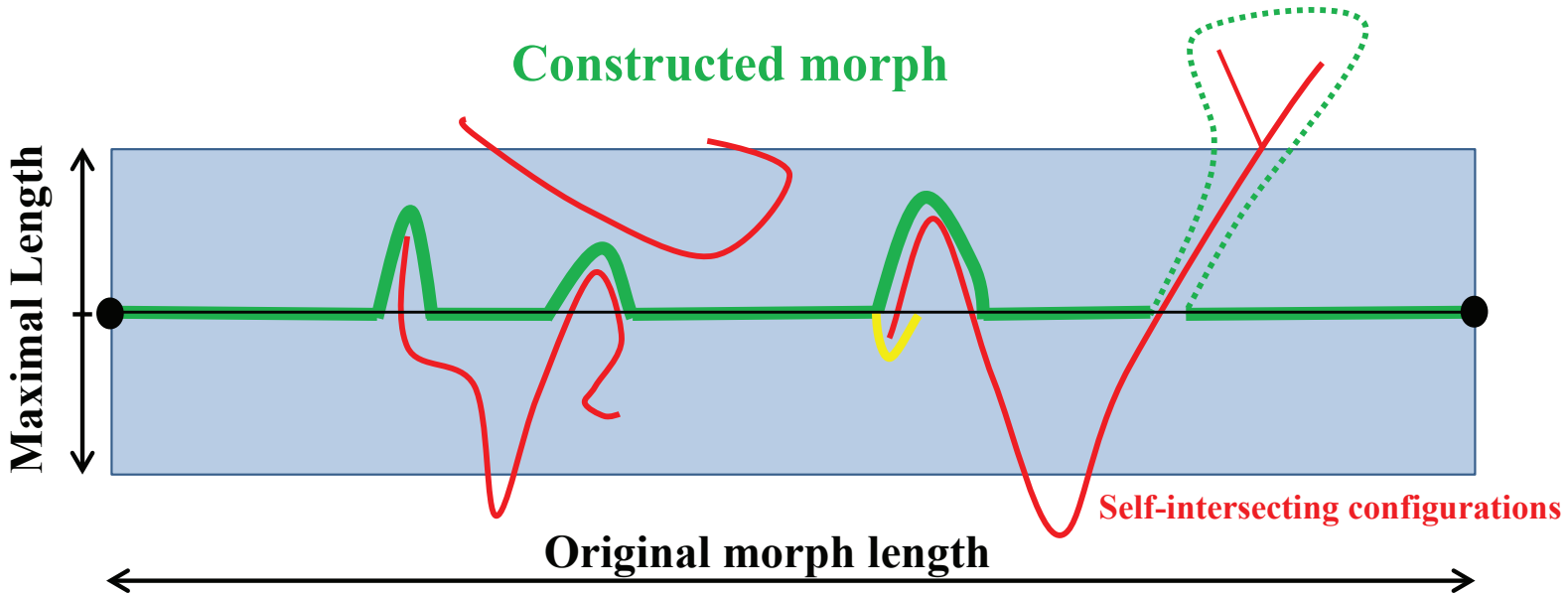
Some self-intersections can be avoided semi-locally and may if the backbone length $<$ MaxLength (user-defined).

End-contractions may be allowed.



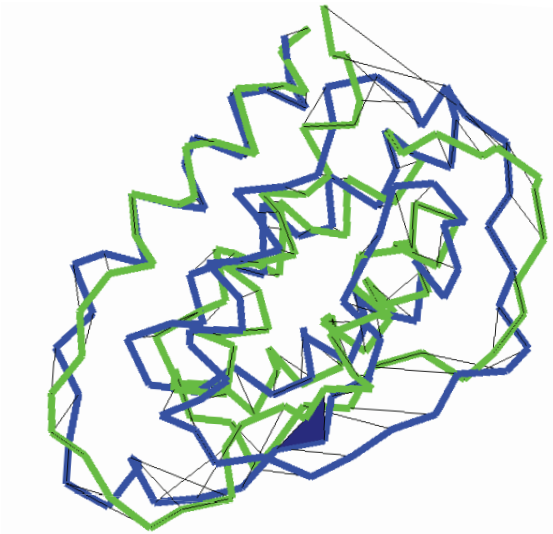
C) Avoid the maximal number self-intersections at the lowest possible price.
D) The remaining self-intersections are called essential self-intersections.
E) A self-avoiding morph is constructed using end-contractions.

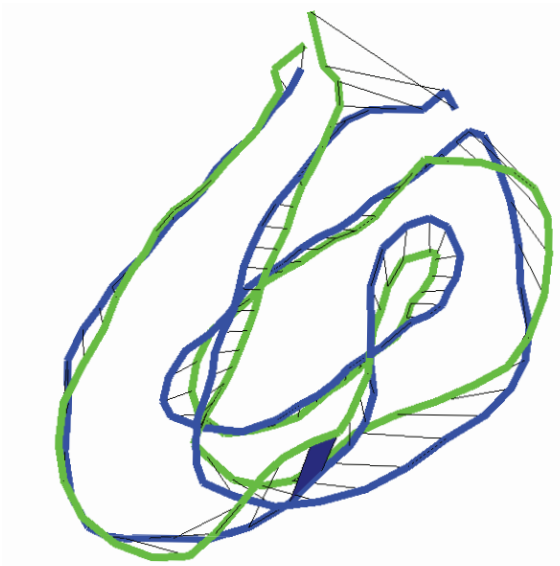
A self-avoiding path



An example

CATH-domains 1emd01 and 1jli00 share homology-class 1.20.120.200 but their 5.6Å RMSD morph have one (essential?) self-intersection.





Output:

- Mean overlap
- ♯ self-intersections
- ♯ essential self-intersections
- self-avoiding morph length - including length of Reidemeister moves and end-contraction
- and a torsional effect

Can RMSD, TM, GDT detect morph self-intersections?

Yes - due to proteins thickness. And no ...
Assume a morph self-intersection requires:

4 carbon alpha pairs moving 5Å and
4 carbon alpha pairs moving 2.5Å.

On an n-residue chain this gives

$$\begin{aligned} \mathrm{RMSD}(n) &< \frac{12\text{Å}}{\sqrt{n}} \\ \mathrm{GDT\_TS}(n) &\geq 1 - \frac{5}{n} \\ \mathrm{TM}(n) &> 1 - \frac{4}{n} \end{aligned}$$

Can RMSD, TM, GDT detect morph self-intersections?

Yes - due to proteins thickness. And no ...
Assume a morph self-intersection requires:

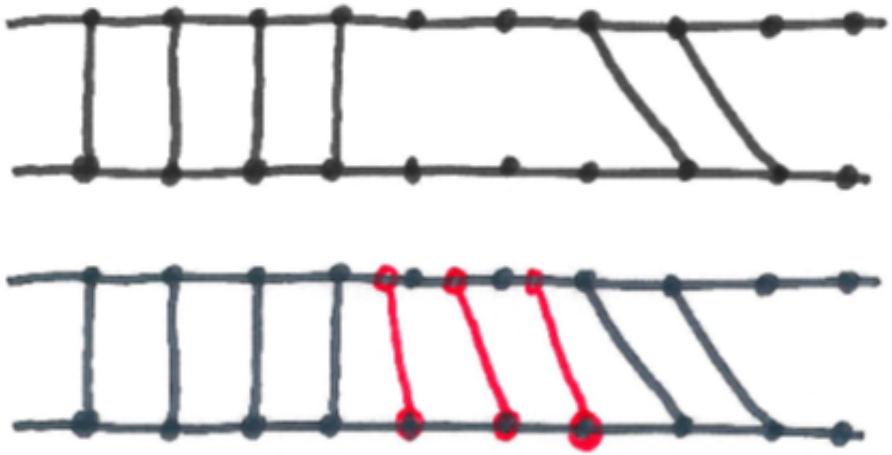
4 carbon alpha pairs moving 5Å and
4 carbon alpha pairs moving 2.5Å.

On an n-residue chain this gives

$$\begin{aligned} \mathrm{RMSD}(n) &< \frac{12\text{Å}}{\sqrt{n}} &&= 1.2\text{Å} \\ \mathrm{GDT\_TS}(n) &\geq 1 - \frac{5}{n} &&= 0.95, \text{ for } n = 100. \\ \mathrm{TM}(n) &> 1 - \frac{4}{n} &&= 0.96 \end{aligned}$$

Treating alignment gaps

Original alignment with a gap.



Filling the gap by linear interpolation.

Needed for Ridemeister type 2 pairs of self-intersections.

Self-intersections are divided into: aligned-aligned, gap-aligned and gap-gap.

Detecting a knot on a protein

2FG6 residues D171 to D238 contains a right handed trefoil knot.

Lai YL, Chen CC, Hwang JK. pKNOT v.2: the protein KNOT web server. Nucl. Acids Res.(2012)

We align the 83 residues D162-D244 of 2FG6 to 408 sequence class representatives of CATH 2.4 with 75 to 100 residues.

Global RMSD alignment. The entire smaller domain is aligned allowing one gap.

- All morphs have self-intersections; 8.4 (3.6 smooth curve) on average.
- All except 3[11] morphs have essential self-intersections for MaxLenght=10[20]

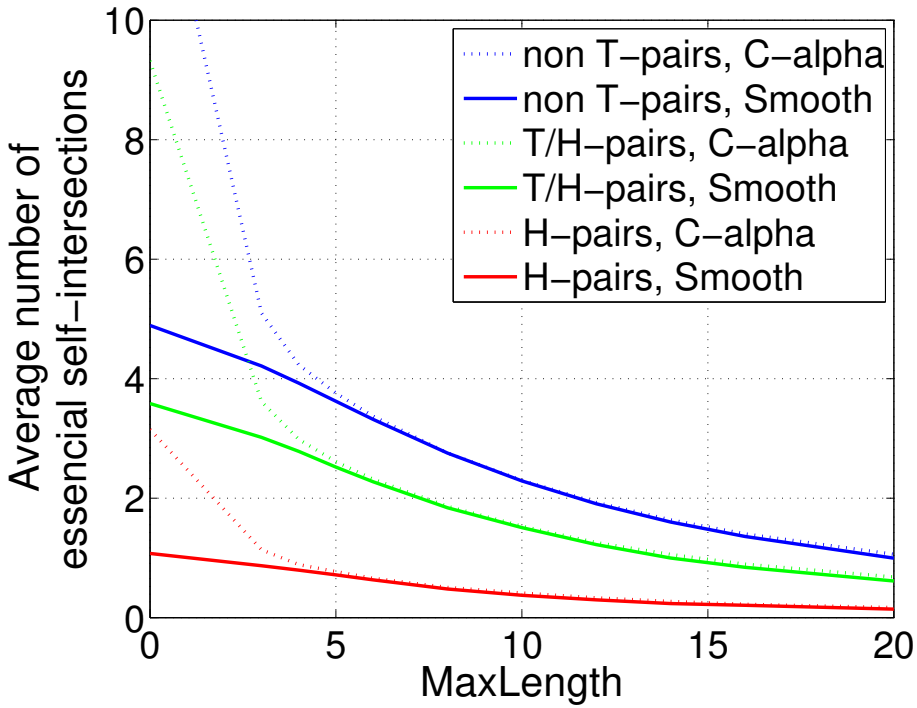
Allowing end-contractions of length MaxLength/2,

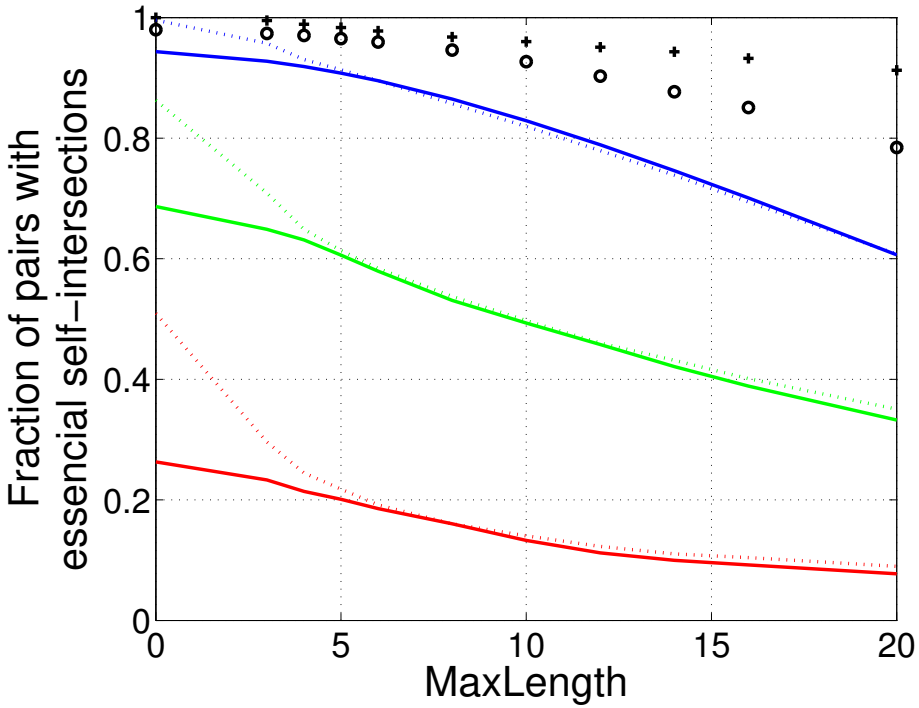
- no essential self-intersections are found in 6, 34, 109, 177, 325 of the 408 cases for MaxLength = 5, 10, 14, 16, 20.

Most TM-alingments are to short to capture the knot.

Fold classification and "Robustness" of essential self-intersections

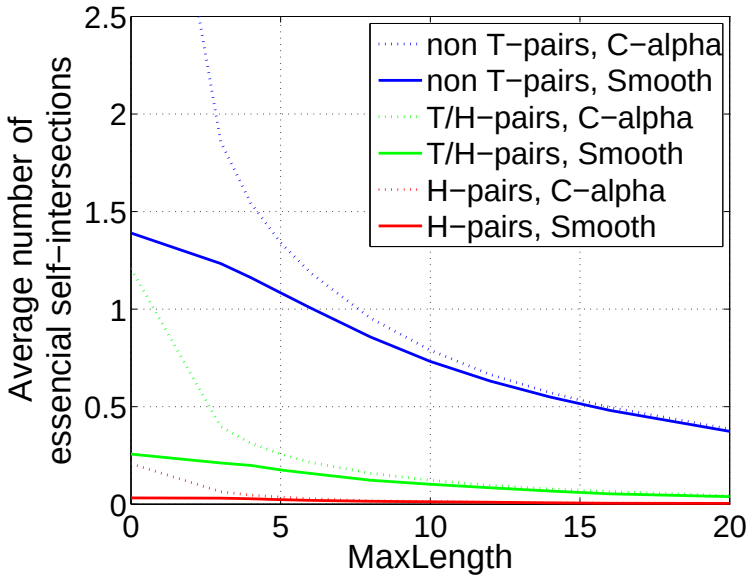
- 1034 sequence family representative CATH2.4 domains with 75-150 residues (clustered at 60% sequence identity) represent
- 521 homology families and 281 topologies
- Aligned with global RMSD alignment if length difference $< 10\%$

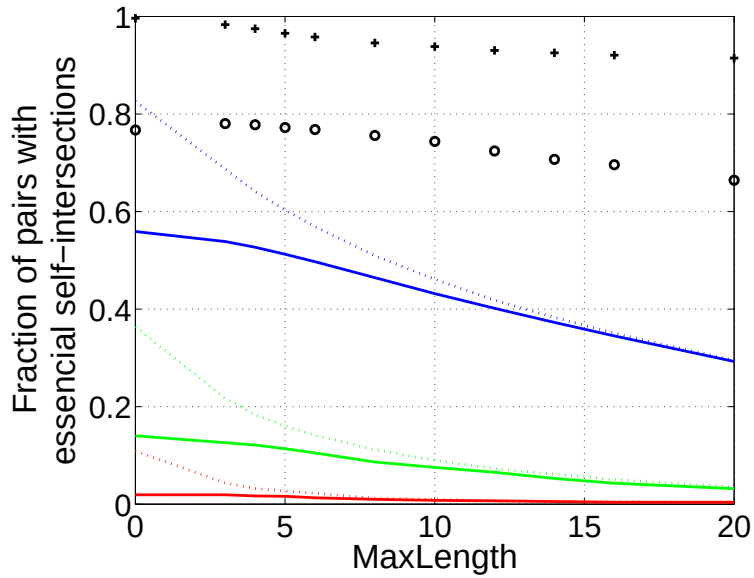




Fold classification and "Robustness" of essential self-intersections

- Using TM-align fewer essential self-intersections are found.
- Caused by shorter aligned windows and better alignment.

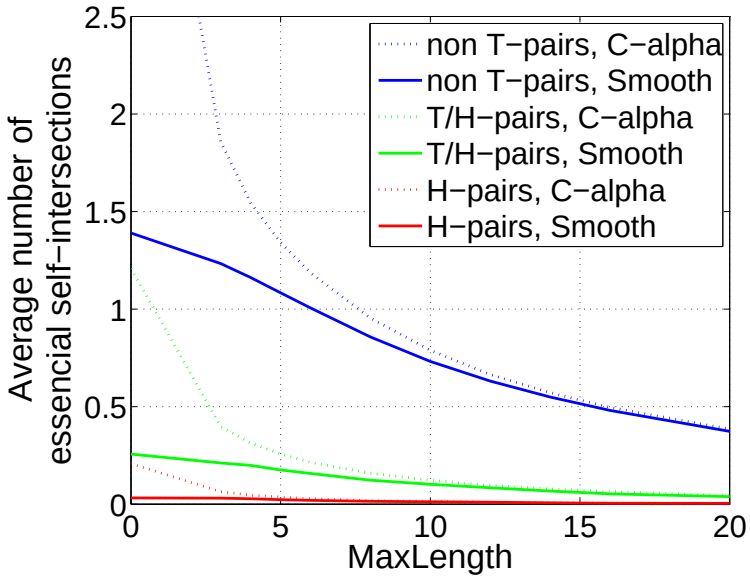


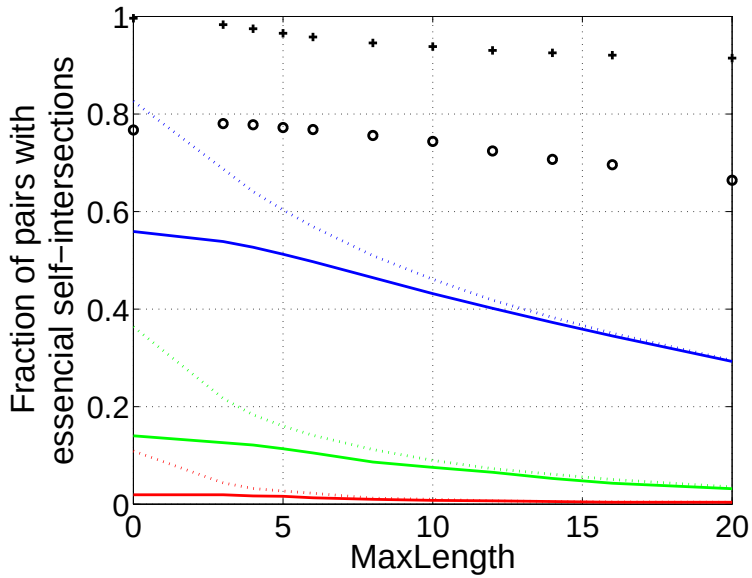


TM-align aligns residues within 5Å in superposition.

Fold classification and "Robustness" of essential self-intersections

- Using TM-align fewer essential self-intersections are found.
- Caused by shorter aligned windows and better alignment.





TM-align aligns residues within 5Å in superposition.

TM-aligned line-segments aligned cause 42165 (12103 smooth curve) self-intersections in 142068 morphs.