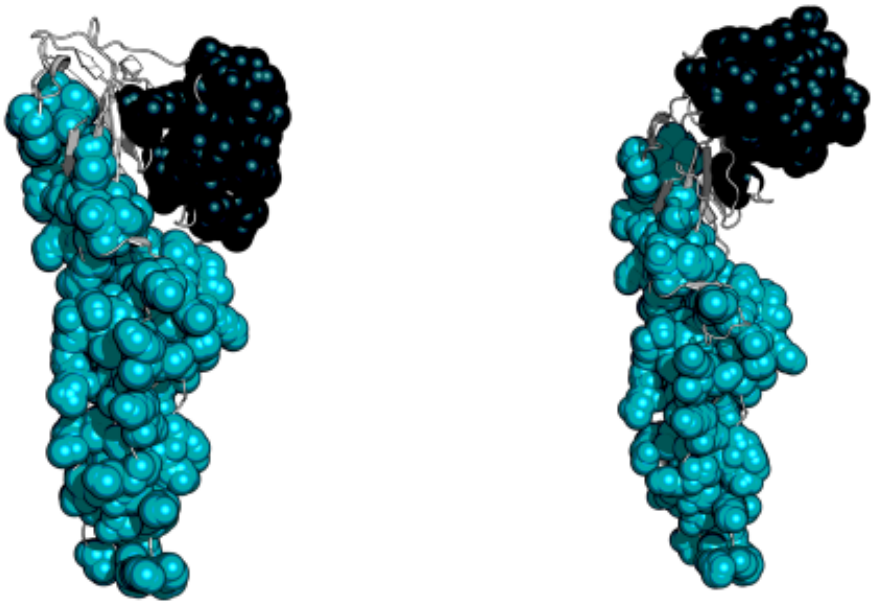
Mining molecular flexibility: novel tools, novel insights



F. Cazals, Inria – Algorithm-Biology-Structure
Joint work with
(Methods) R. Tetley, Inria – Algorithm-Biology-Structure
(Class II fusion) F. Rey, Institut Pasteur Paris

Mining molecular flexibility: novel tools, novel insights

Challenge *Dynamics of proteins*: specification

- Input: structure(s) of biomolecules + potential energy model
- Output
  - Thermodynamics: meta-stable states and observables
  - Dynamics: Markov state model – requires rare transition events
- Time-scales
  - Biological time-scale $>$ millisecond
  - Integration time step in molecular dynamics: $\Delta t \sim 10^{-15} s$

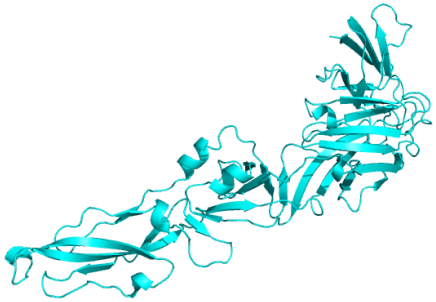
- 5.058ms of simulation time;
- $\sim$ 230 GPU years on NVIDIA GeForce GTX 980 proc.

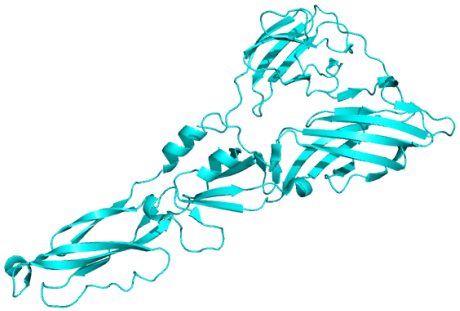
Ref: Chodera et al, eLife, 2019

Mining molecular flexibility: novel tools, novel insights

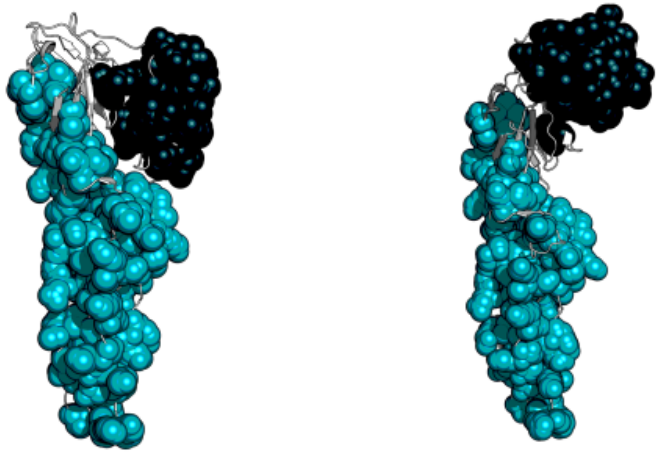
Combined RMSD : TBEV glycoprotein in two different conformations pre and post fusion

▷ Classical analysis:





▷ Our motifs:



Motif	Alignment size	lRMSD
Large	88	1.69
Small	40	0.38

Statistics from Apurva:

- 370 a.a. aligned
- lRMSD: 11.1Å

Structural Motif

▷ Input: We are given two polypeptide chains S_A and S_B

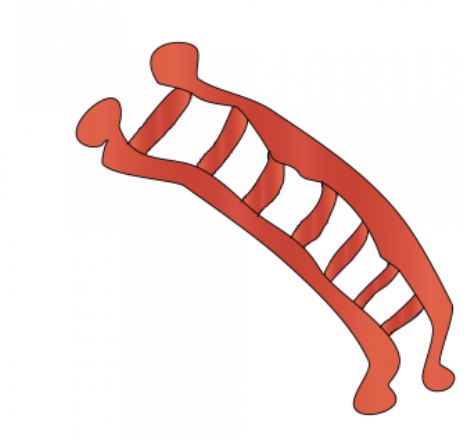
Definition 1. Given two sets of a.a. $M_A = \{a_{i_1}, \ldots, a_{i_s}\} \subset S_A$ and $M_B = \{b_{i_1}, \ldots, b_{i_s}\} \subset S_B$, and a one-to-one alignment $\{(a_{i_j} \leftrightarrow b_{i_j})\}$ between them, we define the ***least RMSD ratio*** as follows:

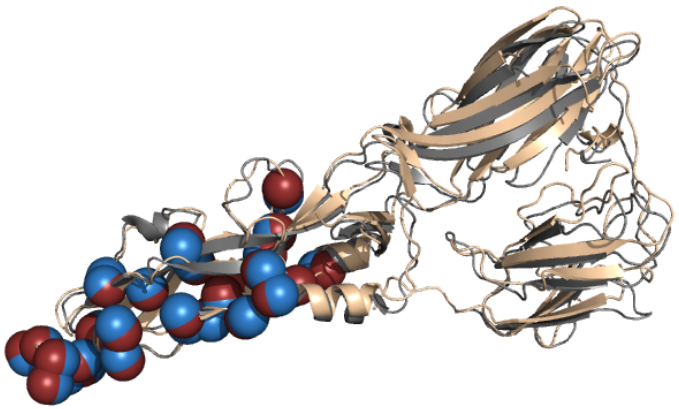
$$r_{\mathrm{lRMSD}}(M_A, M_B) = \mathrm{lRMSD}(M_A, M_B)/\mathrm{lRMSD}(S_A, S_B). \tag{1}$$

The sets M_A and M_B are called *structural motifs* provided that

$$|M_A| = |M_B| \geq s_0 \text{ and } r_{\mathrm{lRMSD}}(M_A, M_B) \leq r_0,$$

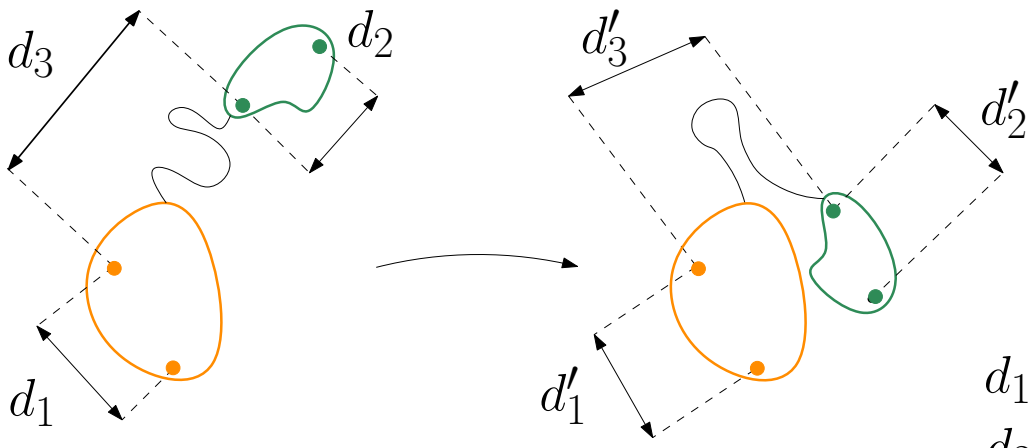
for appropriate thresholds s_0 and r_0.





Key idea: exploiting quasi-isometric deformations to identify **almost rigid** | **isometric** regions in structures

- Quasi-isometric deformation: (selected) distances (almost) preserved



$$d_1 \sim d'_1$$
$$d_2 \sim d'_2$$
$$d_3 \neq d'_3$$

- Tracking such deformation may be done at two scales:
  - Global preservation: maximal cliques – NP-hard problem.
  - Local preservation: spanning trees connecting atoms whose relative distances are conserved.

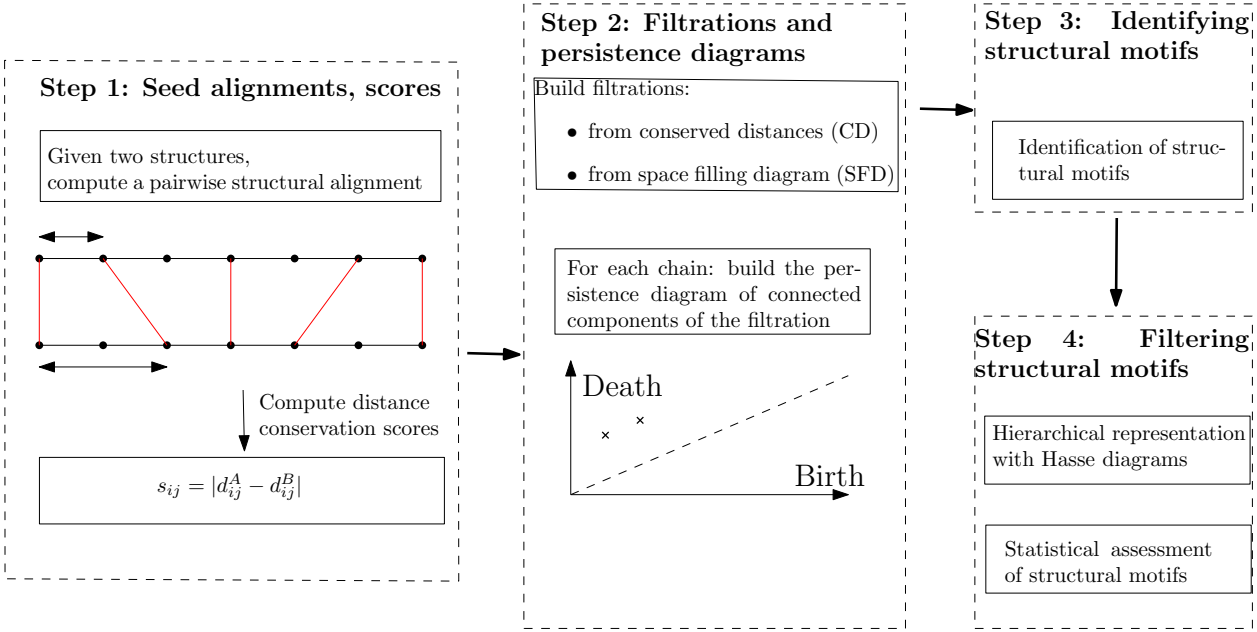
Multi-scale rigidity: embodied in the notion of filtration

▷ Key ideas

- Filtration: sequence of nested topological space – read: sequence of nested sets of amino-acids
- Ordering of a.a.: by decreasing *rigidity index* – those involved in rigid blocks come first

Motifs for two structures A and B: a generic approach

- Step 1: use an aligner for the seed alignment and scores
  - (A and B) Compute a seed alignment
  - (A, then B) Sort residues by decreasing structural conservation
- Step 2: use a filtration to perform a multiscale analysis
  - (A, then B) Identify structurally conserved regions
- Step 3: reuse the aligner to bootstrap the alignment
  - (A and B) Re-compute a structural alignment between pairs of regions

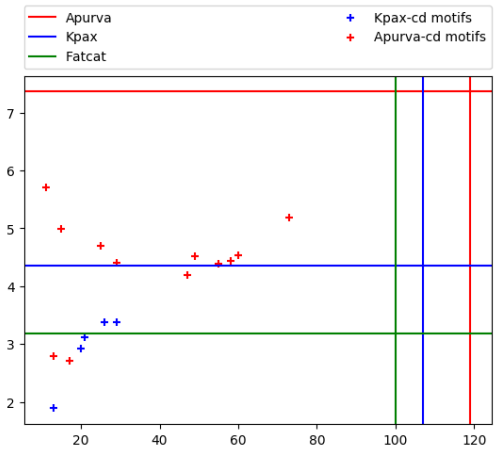


Generic method: instantiations

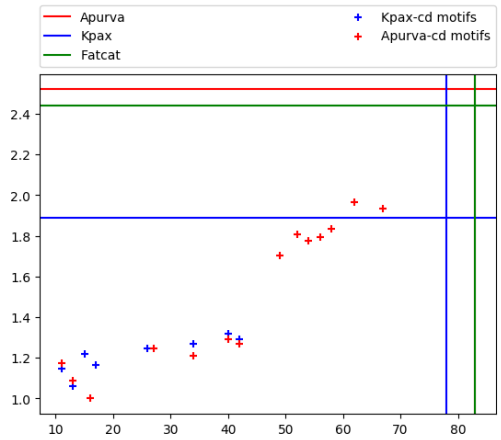
- Main steps:
  - step 1 $\equiv$ alignment to rigidity scores;
  - step 2 $\equiv$ rigidity scores to filtrations;
  - step 3 $\equiv$ filtrations to motifs via local alignments.
- Ingredient 1: an aligner for steps 1 and 3
  - Options: `Kpax`, `Apurva`, (`FATCAT`)
- Ingredient 2: filtration encoding based on rigidity scores
  - Option 1: based on conserved distances (cf Kruskal's MST algorithm)
  - Option 2: based on space filling diagrams (Voronoi / α-shapes)
- Resulting programs: `Align-Kpax-CD`, `Align-Kpax-SFD`, `Align-Apurva-CD`, `Align-Apurva-SFD`
- Nb: conformation vs homologous proteins: (trivial) alignment

Motifs reveal the multi-scale structural conservation within global alignments

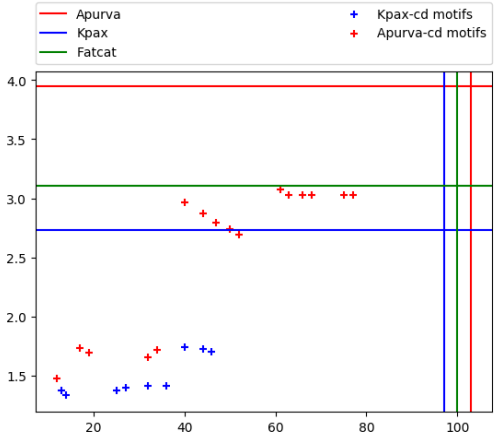
▷ Size of motifs vs lRMSD on challenging cases



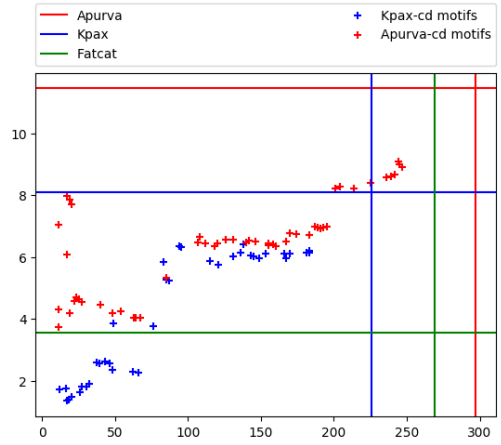
1BGE vs 2GMF



1CEW vs 1MOL



1CID vs 2RHE

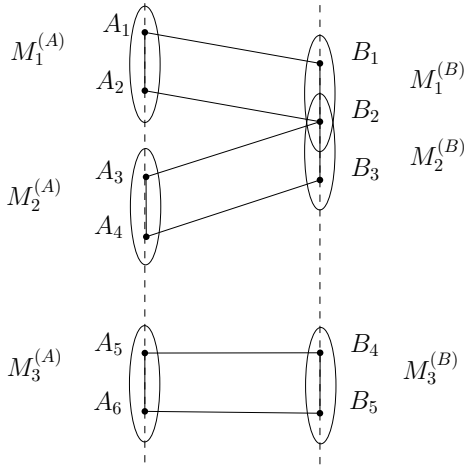


1CRL vs 1EDE

▷Ref: Pairs of structures: from Godzik et al, Bioinformatics, 2003

Comparing two molecules: the combined RMSD

- Rationale: use one rigid motion for each *rigid/structurally conserved* region
- Motifs for two molecules A and B, and their intersection graph



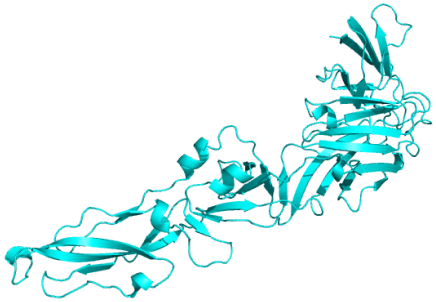
Definition 2. Consider two structures A and B for which non-overlapping domains $\{C_i^{(A)}, C_i^{(B)}\}_{i=1,\ldots,m}$ have been identified. Assume that a lRMSD has been computed for each pair $(C_i^{(A)}, C_i^{(B)})$. Let w_i be the weights associated with an individual lRMSD . The ***combined RMSD*** is defined by

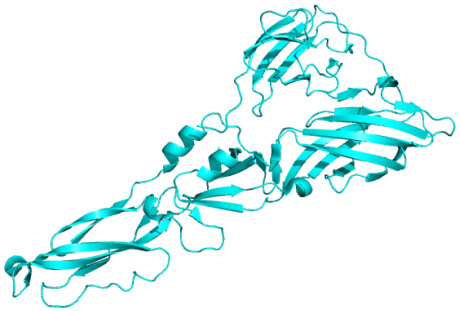
$$\mathrm{RMSD}_{\mathrm{Comb.}}(A,B) = \sqrt{\sum_{i=1}^{m} \frac{w_i}{\sum_i w_i} \mathrm{lRMSD}^2(C_i^{(A)}, C_i^{(B)})}. \tag{2}$$

- Rmk: comes into two guises, namely vertex weighted and edge weighted

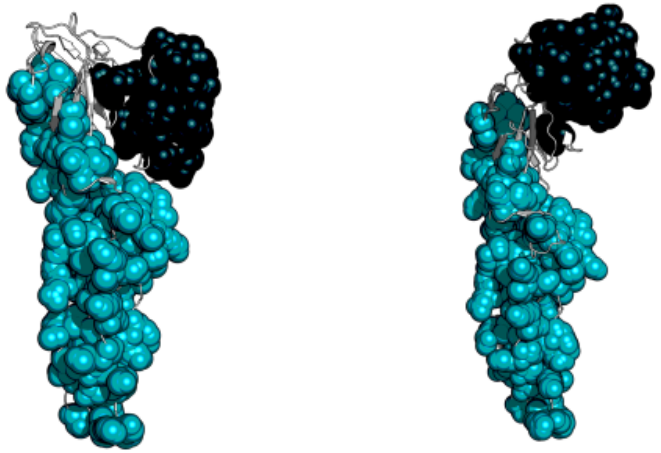
Combined RMSD : TBEV glycoprotein in two different conformations pre and post fusion

▷ Classical analysis:





▷ Our motifs:

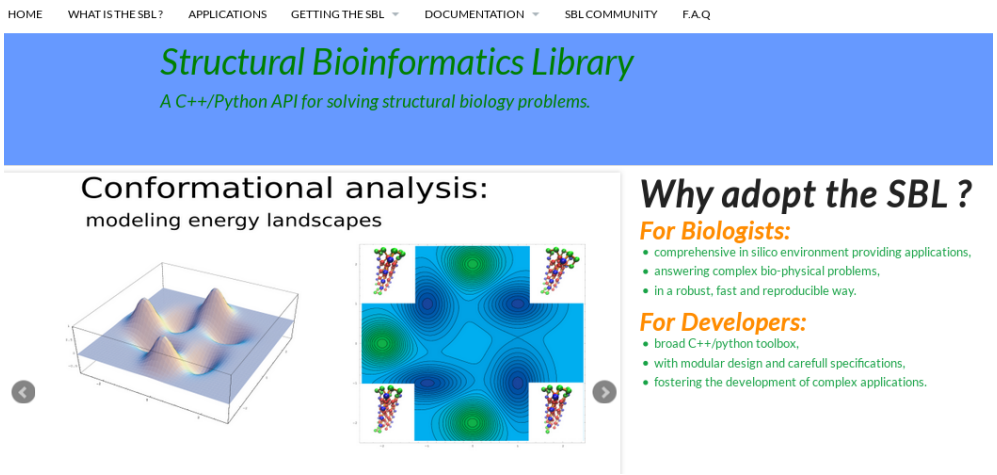


Motif	Alignment size	lRMSD
Large	88	1.69
Small	40	0.38

Statistics from Apurva:

- 370 a.a. aligned
- lRMSD: 11.1Å

The Structural Bioinformatics Library



http://sbl.inria.fr

▷Ref: Cazals and Dreyfus; Bioinformatics, 2016

SBL and Jupyter notebooks: guided tour

http://sbl.inria.fr/applications

Mining molecular flexibility: novel tools, novel insights

Summary and outlook

- Combined RMSD – $\text{RMSD}_{\text{Comb.}}$
  - Structural comparisons based on (relatively) independent sets
- Multiscale analysis of structural conservation
  - Segregating dof (internal coords.) into active and passive
  - Towards more efficient algorithms for thermodynamics - dynamics
- Software: all tools in the `SBL`
- Ongoing
  - Design of move sets
  - Applications to energy landscapes: exploration, thermodynamics

Bibliography

- Combined RMSD: [1]
- Structural motifs: [2]
- Software: [3]
- Partition functions [4]
- Cluster matching: [5]

F. Cazals and R. Tetley.

Characterizing molecular flexibility by combining lRMSD measures.
Proteins, 87(5):380–389, 2019.

F. Cazals and R. Tetley.

Multiscale analysis of structurally conserved motifs.
2019.
Submitted.

F. Cazals and T. Dreyfus.

The Structural Bioinformatics Library: modeling in biomolecular science and beyond.
Bioinformatics, 7(33):1–8, 2017.

A. Chevallier and F. Cazals.

Wang-landau algorithm: an adapted random walk to boost convergence.
J. of Computational Physics (Under revision), 2019.

F. Cazals, D. Mazauric, R. Tetley, and R. Watrigant.

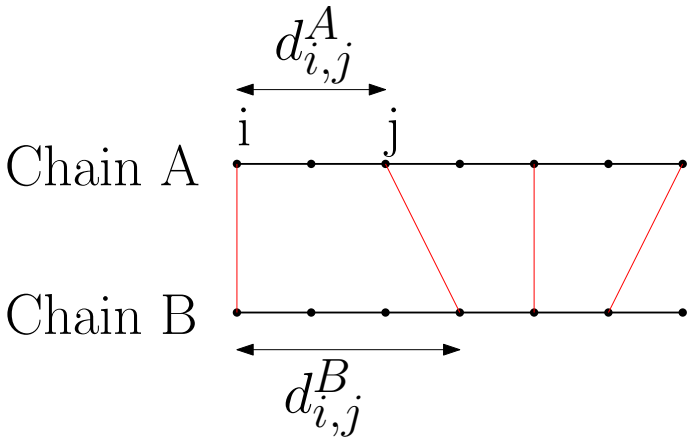
Comparing two clusterings using matchings between clusters of clusters.
ACM J. of Experimental Algorithms, 24(1):1–42, 2019.

Mining molecular flexibility: novel tools, novel insights

Mining molecular flexibility: novel tools, novel insights

Step 1: rigidity score as C_α ranks for chains A and B

- Input: a structural alignment yields
  - $d^A_{i,j}$: dist. between C_α i and j on chain A
  - $d^B_{i,j}$: dist. between C_α i and j on chain B



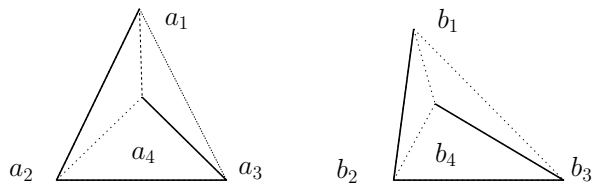
- Distance difference matrix between A and B:

$$s_{ij} = \mid d^A_{i,j} - d^B_{i,j} \mid, i = 1, \ldots, N, j = 1, \ldots, N. \tag{3}$$

- C_α rank of residue i: index of the smallest s_{ij} involving this residue in the sorted sequence Sorted$\{s_{ij}\}$.

Assuming the ordering of scores depicted, the ranks are as follows:

- one for C_1 and C_2
- two for C_3 and C_4
- likewise for the second chain.

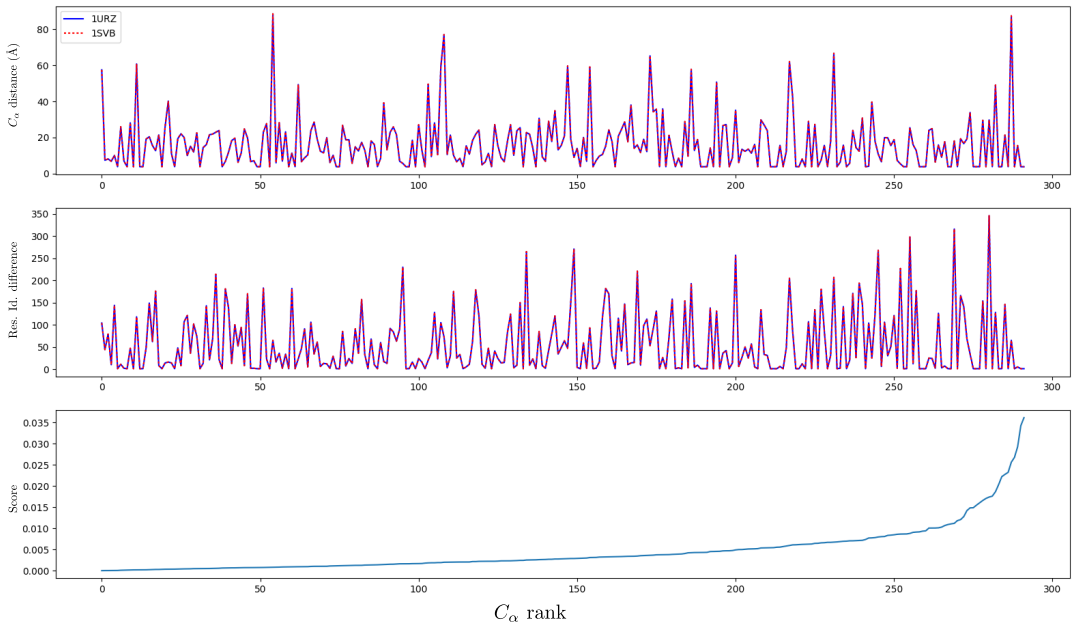


Sorted scores: $s_{12} < s_{34} < s_{23} < s_{13} < s_{14} < s_{24}$

Step 1: illustration for 1SVB - 1URZ

▷ Plots:

- C_α distance plot: for chain A, the function $d^A_{i,j}$ (or $d^B_{i,j}$) as a function of the C_α rank.
- Sequence shift plot: for chain A (or chain B), the function $j - i$ as a function of the C_α rank.
- Score plot: score s_{ij} as a function of the C_α rank.

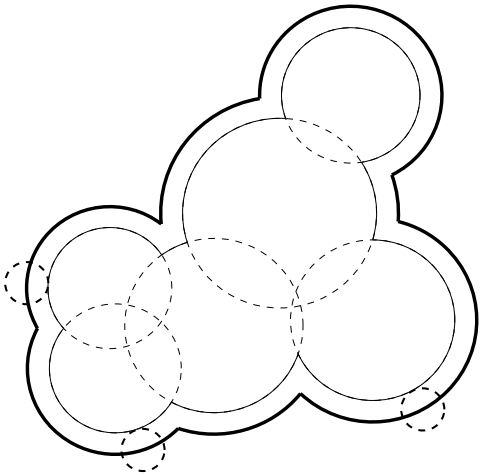


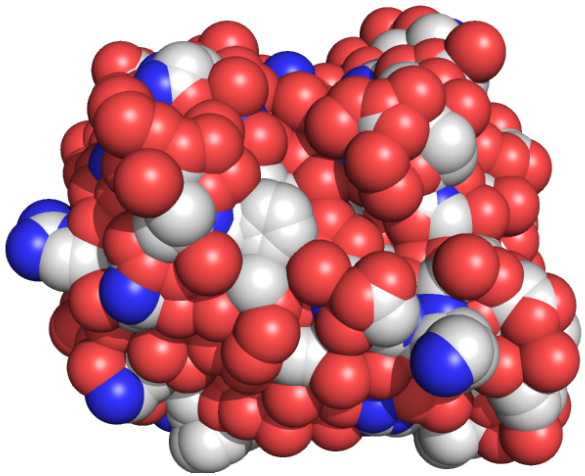
Step 2a – filtration using Space Filling Diagrams

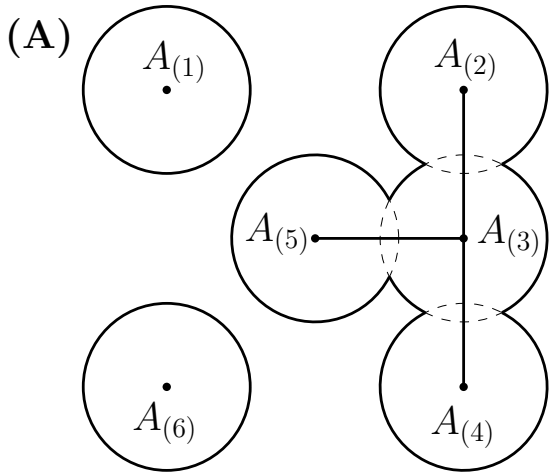
building the filtration

- Filtration = sequence of nested sets
- Model a collection of amino-acids with its Solvent Accessible Surface

- For both structures, independently:
  - insert a.a. by increasing C_α ranks,
  - maintain the corresponding space filling model of the Solvent Accessible Model



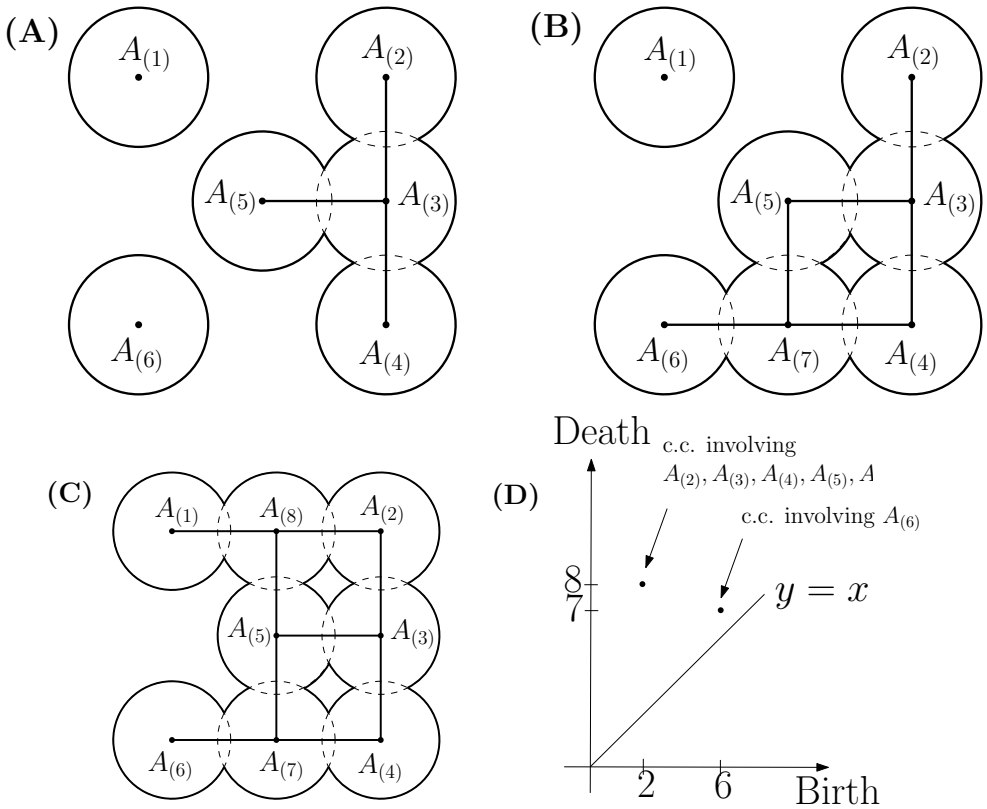




Step 2a – filtration using Space Filling Diagrams

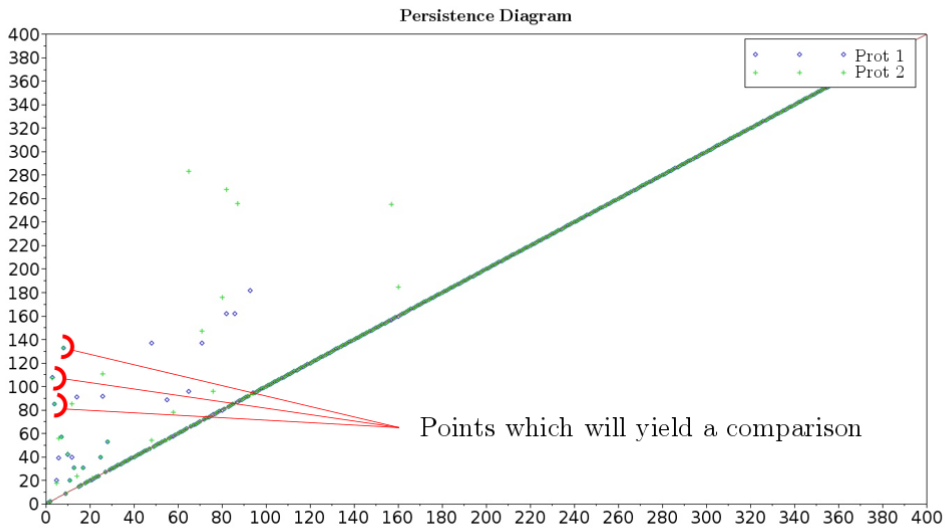
persistence diagram of the connected components

- Assessing the stability of conserved regions:
  - compute its connected components
  - maintain the associated persistence diagram



Step 3: identifying motifs – rationale

- Motifs from local structural alignments inferred from the PD:
  - points nearby in the pers. diag. have a *comparable* rigidity signature
  - each such point corresponds to a set of a.a. in one structure
  - therefore: run a local alignment between these regions
    - motif: $r_{IRMSD} \leq r_0$ and $|M_A| = |M_B| \geq s_0$



- Topological changes and accretion:
  - accretion: insertion of an a.a. connected to an already existing connected component.
  - concomitant birth and death i.e. 0-persistence i.e. point on the diagonal of the PD for c.c.
  - pitfall: accretion may be such that a PD has very few points!

Step 3: identifying motifs – details

▷ Identifying motifs:

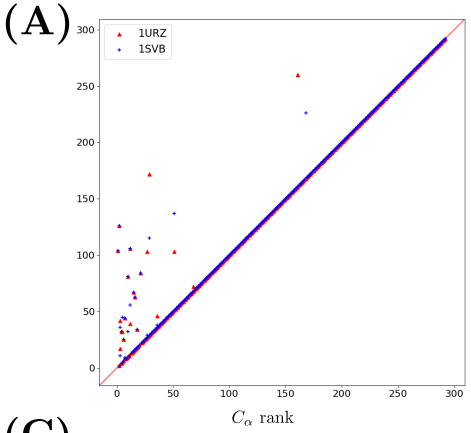
– For each critical value (death date) t of either persistence diagram:
  – compute the c.c. $F_A = \{c_1, \ldots, c_{n_A}\}$ of $\mathcal{F}_t^A$
  – compute the c.c. $F_B = \{c'_1, \ldots, c'_{n_B}\}$ of $\mathcal{F}_t^B$
  – (simple) compute a structural alignment for each pair $(c_i, c'_j) \in F_A \times F_B$
  – (involved) solve a k-partition matching for F_A and F_B,
    and run a structural alignment on the resulting meta-clusters

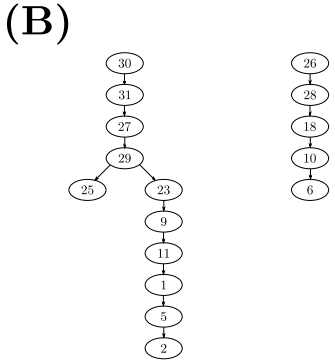
▷ Filtering motifs:

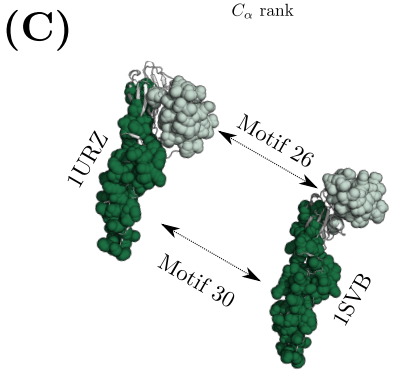
- compute the Hasse diagram (for the inclusion) of the motifs found
  NB: inclusion owes to the nested-ness of sublevel ets.
- retain the roots of the Hasse diagrams only.

Steps 2-3: illustration for 1SVB - 1URZ

- ▷ Step 2, Building the filtration and its persistence diagram (`Align-Identity-CD`)
- ▷ Step 3, Computing structural motifs with bootstrap: run a local alignment for regions associated with connected components defined by critical values in the persistence diagram







(D)

Motif index	lRMSD	length
30	1.82	109
31	1.81	107
27	1.66	85
29	1.62	83
25	0.59	11
23	0.59	58
9	0.59	39
11	0.61	36
1	0.28	18
5	0.29	17
2	0.29	16
26	3.77	43
28	3.92	39
18	4.23	32
10	4.12	24
6	4.51	18