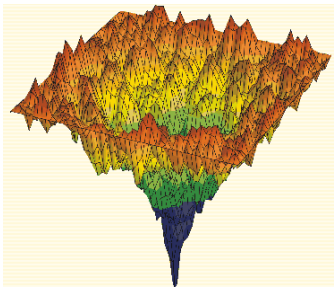
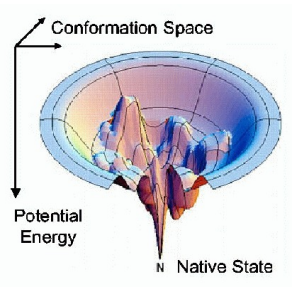
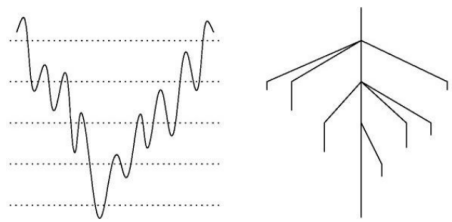
Energy landscapes:

sampling, analysis, and comparison







Frederic.Cazals@inria.fr

Energy landscapes: sampling and analysis

Introduction

Sampling landscapes

Software

References

Proteins: from structure to function across dynamics

- Demo vmd
- Given is the Potential Energy Landscape: a potential energy function i.e.

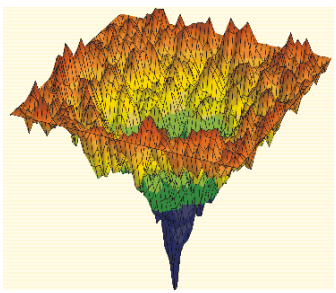
$$U : \mathcal{C} \to \mathbb{R} \tag{1}$$

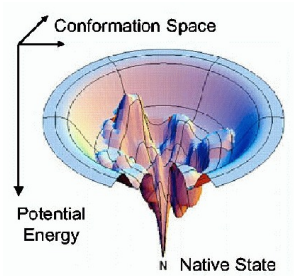
- Core questions pertain to the realms of:
  - Structure: stable states (conformations) / ensembles of coherent conformations
    $\to$ sampling the PEL: enumerating low lying local minima
  - Thermodynamics: probability for the stable states
    $\to$ integrating Boltzmann's factor on the basins of the PEL
  - Kinetics: dynamics between the stable states
    $\to$ building Markov state model on the PEL

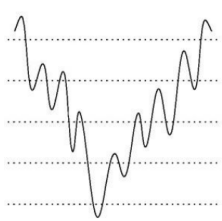
Energy landscapes and the trinity
Structure – Thermodynamics – Dynamics

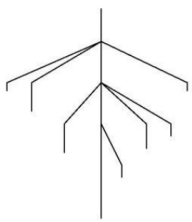
- ▷ Problem statement: emergence of function from structure and dynamics
  For proteins: understanding *minimal frustration*

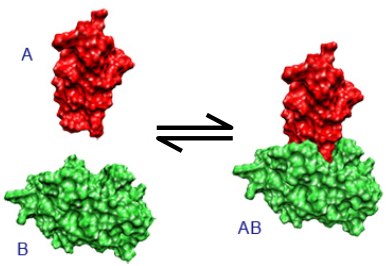
- ▷ State-of-the-art: contributions from various perspectives
  - Molecular dynamics (including REMD, metadynamics),
  - Energy landscapes methods (the basin hopping lineage),
  - Monte Carlo methods (MCMC, Wang-Landau, importance sampling)
  - Markov state models
  - Dimensionality reduction (PCA, Isomap, diffusion maps)











▷Ref: Becker and Karplus, The Journal of Chemical Physics, 1997

▷Ref: Wales; Energy Landscapes; 2003

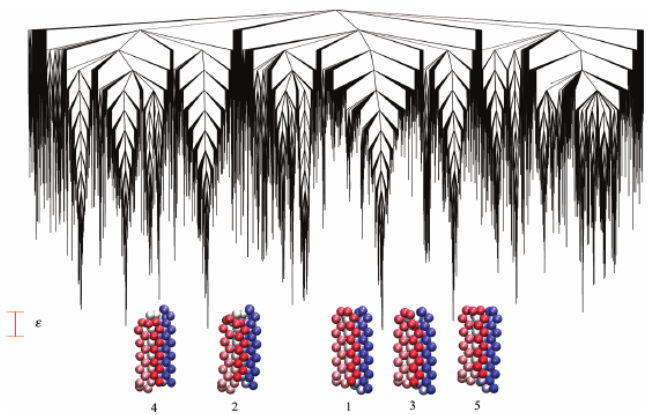
▷Ref: Chipot; Frontiers in free-energy calculations; 2014

BLN69: a Simplified Protein Model

- Description:
  - Three types of Beads: : hydrophobic(B), hydrophylic(L) and neutral(N)
  - Configuration space of intermediate dimension: 207
  - Challenging: frustrated system
  - Exhaustively studied: DB of $\sim 450k$ critical points (Industry)

$$V_{BLN} = \frac{1}{2} \cdot K_r \sum_{i=1}^{N-1} (R_{i,i+1} - R_e)^2 + \frac{1}{2} K_0 \sum_{i=1}^{N-2} (\theta_i - \theta_e)^2 + \epsilon \cdot \sum_{i=1}^{N-3} [A_i(1 + \cos\phi_i) + B_i(1 + 3\cos\phi_i)]$$
$$+4\epsilon \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} \cdot C_{ij}[(\frac{\sigma}{R_{i,j}})^{12} - D_{ij}(\frac{\sigma}{R_{i,j}})^6]$$

- Disconnectivity graph: describes merge events between basins



- Ref: Honeycutt, Thirumalai, PNAS, 1990
- Ref: Oakley, Wales, Johnston, J. Phys. Chem., 2011

Energy landscapes: sampling and analysis

Introduction

Sampling landscapes

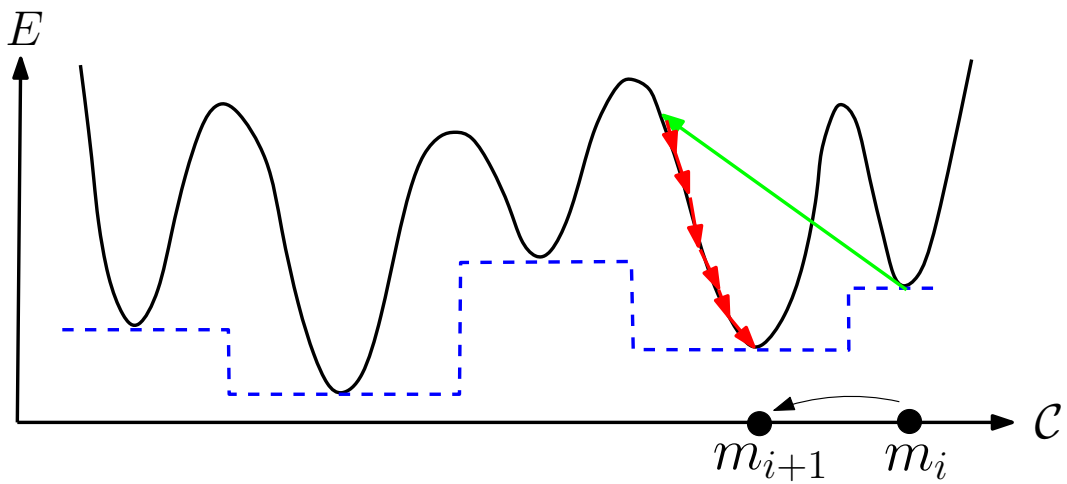
Software

References

Exploring Potential Energy Landscapes:

basin hopping

- Goal: enumerating low energy local minima
- Basin-hopping and the basin hopping transform
  - Random walk in the space of local minima
  - Requires a *move set* and an *acceptance test* (cf Metropolis)
    and the ability to descend the gradient (*quenching*)
    aka energy minizations

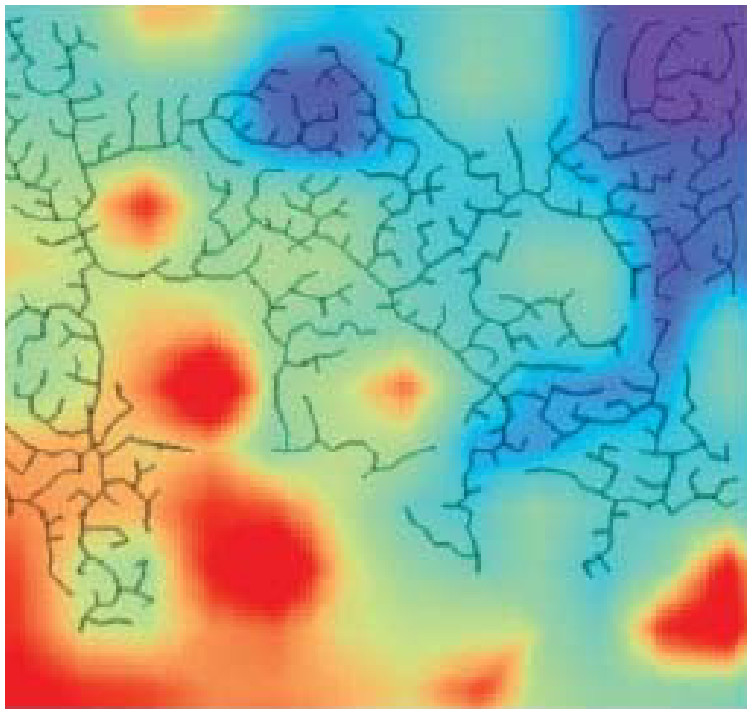


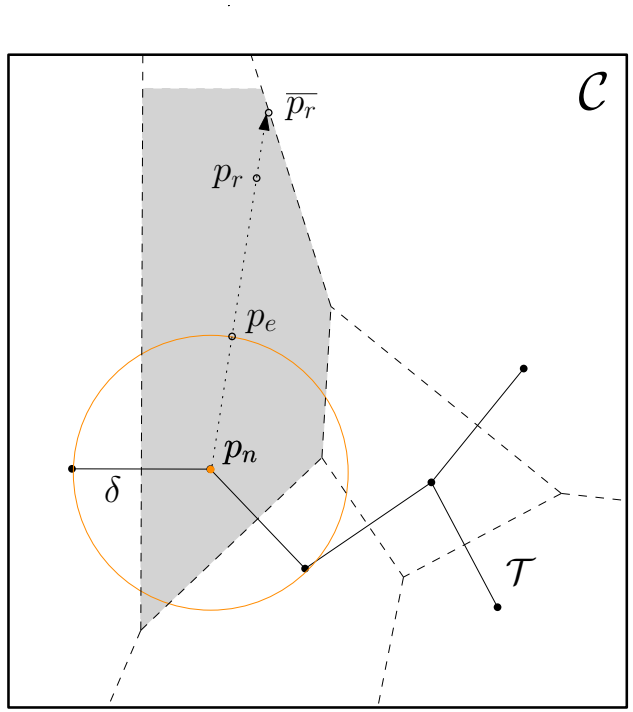
Ref: Li and Scheraga, PNAS, 1987

Exploring Potential Energy Landscapes:

transition based rapidly exploring random trees (T-RRT)

- ▷ Goal: sample basins and transitions
- ▷ Algorithm growing a random tree favoring yet unexplored regions
  - node to be extended selection: *Voronoi* bias
  - node extension: interpolation + Metropolis criterion (+temperature tuning)





▷Ref: LaValle, Kuffner, IEEE ICRA 2000

▷Ref: Jaillet, Corcho, Pérez, Cortés, J. Comp. Chem, 2011

Exploring energy landscapes:
a generic approach yielding BH, T-RRT,...

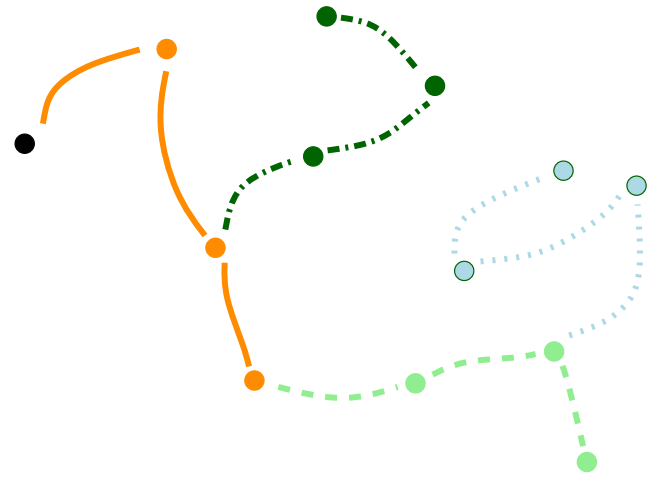
▷ Template:

Require: E($\cdot$): potential energy
Require: Parameters: T:
temperature; δ: step size

Initialize the set P with one conformation
while StopCondition= False **do**
$p_n \leftarrow$ SelectConformationToExtend(P)
$p_e \leftarrow$ ExtendConformation(p_n)
UpdateMoveSetParams(δ)
if AcceptConformation(p_n, p_e) **then**

RecordNewConformation(p_e, P)
UpdateAcceptanceTestParams

▷ Hybrid algorithm: alternate BH and T-RRT extensions



▷ Key ingredients:

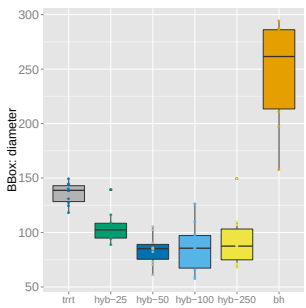
- Boosting the identification of low lying minima
- Favoring spatial adaptation—local exploration parameters
- Handling distances efficiently

▷Ref: Roth, Dreyfus, Robert, Cazals; J. Comp. Chem.; 2015

Exploring energy landscapes: performances of `Hybrid`

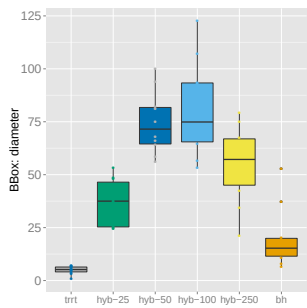
- ▷ Contributions: enhanced exploration of low lying regions of a complex landscape
- ▷ Protocol:
  - Contenders: `BH`, `T-RRT`, `Hybrid` for various parameter values b
  - Count and assess the local minima reported from two reference databases:
    $BLN69 - min - all$: 458,082 minima
    BLN69-min-E_{-100}: 5932 minima.

- Bounding box $\emptyset$: all mins



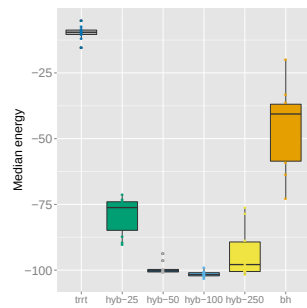
$BLN69 - min - all$

vs low lying



$BLN69 - min - E_{-100}$

- Median energies



$BLN69 - min - all$

- ▷ Assessment:

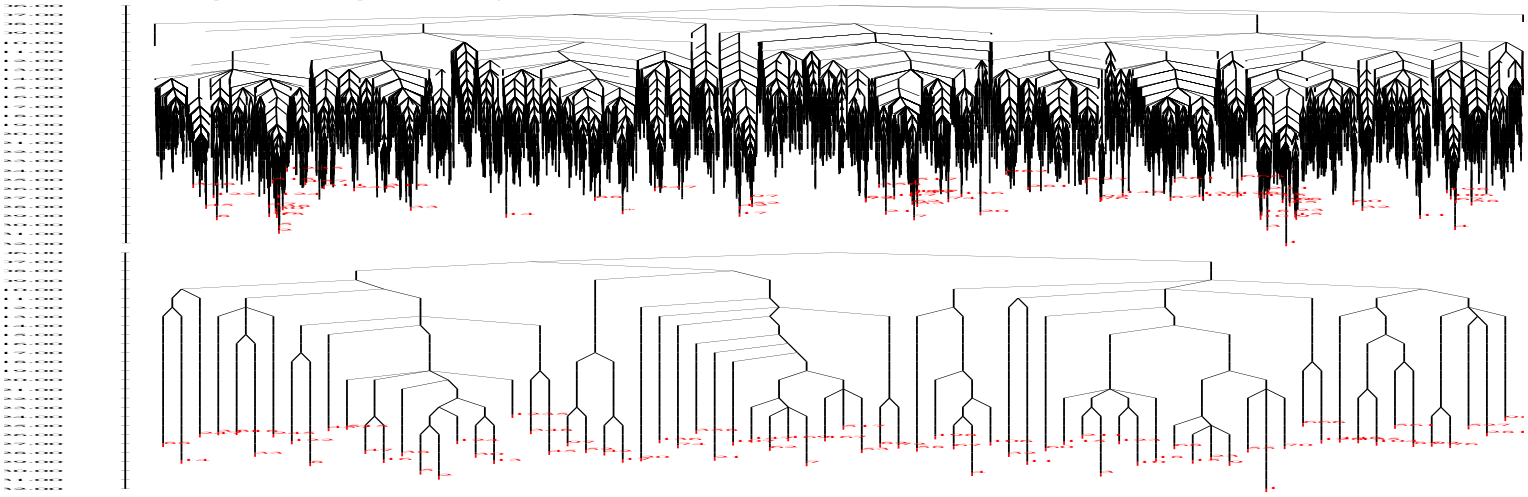
– Combines critical building blocks:
minimization, spatial exploration boosting, nearest neighbor searches
– Bridging the gap to thermodynamics

▷Ref: Oakley et al; J. of Physical Chemistry B; 2011

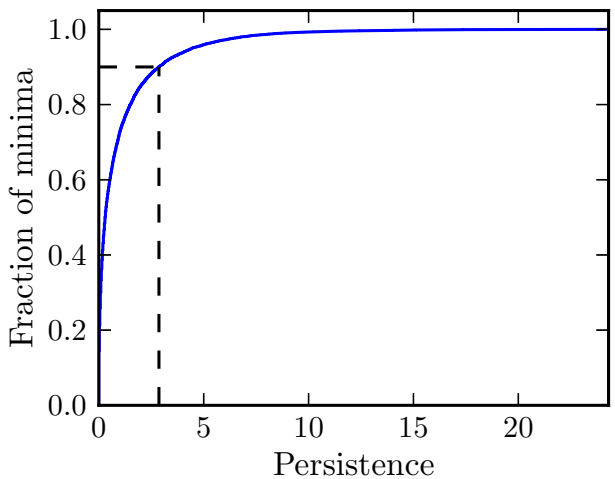
▷Ref: Roth, Dreyfus, Robert, Cazals; J. Comp. Chem.; 2015

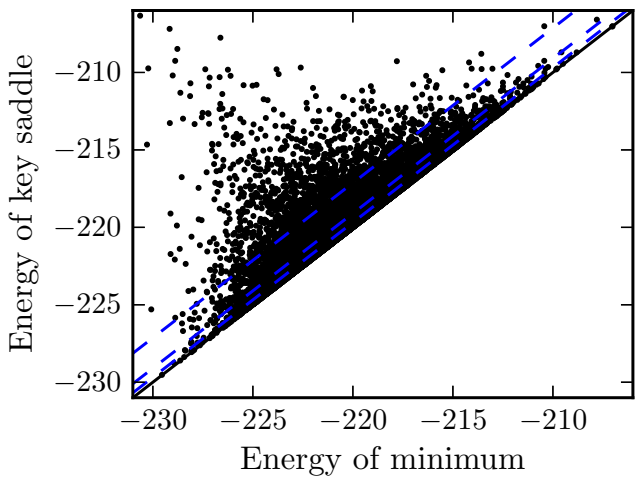
Lennard-Jonnes cluster LJ_{60}

- Coarse graining the system:



- Using the distribution of barriers' heights:





- Ref: Carr, Mazauric, Cazals, Wales; J. Chem. Phys.; 2016

Sampling: discussion

- Critical features
  - + - distance used – impacts the Voronoi bias
  - + - data structures used for nearest neighbor queries
  - + - move set
  - + - temperature and step size adaptation

- Open questions

  (parameterized) mathematical models for PEL

  output sensitive analysis for exploration algorithms

Energy landscapes: sampling and analysis

Introduction

Sampling landscapes

Software

References

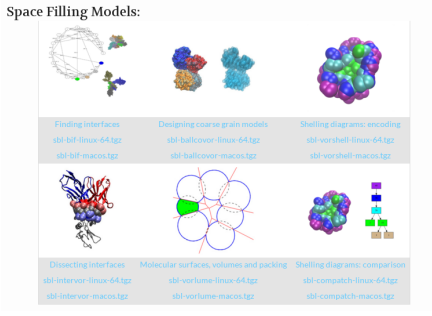
The Structural Bioinformatics Library (SBL): 101

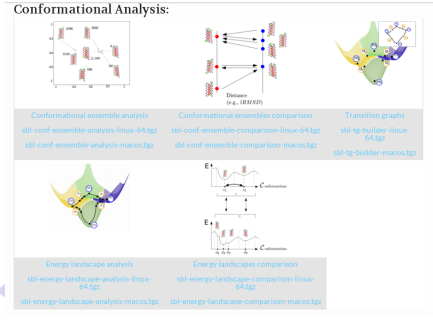
http://sbl.inria.fr

- What: generic **C++ / Python library** for Structural Bioinformatics
  - Combining **high level applications** and **low level algorithms** (combinatorial, topological and geometric)
- Who for:
  - **End-Users** : compiled binaries solving specific problems
    Space filling models / Conformational analysis / large assemblies
  - **Developers** : C++ framework to create novel applications
  - **Contributors** : contribute generic C++ packages "a la" CGAL

- Platforms: **Unix** Linux and MacOS (released) and Windows (pending)
- License: academia: **open source** like; industries: specific licence

- Getting the SBL: `http://sbl.inria.fr/downloads`
- Getting the pre-compiled applications: `http://sbl.inria.fr` > Applications





Energy landscapes: sampling and analysis

Introduction

Sampling landscapes

Software

References

Bibliography

F. Cazals, T. Dreyfus, D. Mazauric, A. Roth, and C.H. Robert.

Conformational ensembles and sampled energy landscapes: Analysis and comparison.

Journal of Computational Chemistry, 36(16):1213–1231, 2015.

J. Carr, D. Mazauric, F. Cazals, and D.J. Wales.

Energy landscapes and persistent minima.

The Journal of Chemical Physics, 144(5), 2016.

A. Roth, T. Dreyfus, C.H. Robert, and F. Cazals.

Hybridizing rapidly growing random trees and basin hopping yields an improved exploration of energy landscapes.

Journal of Computational Chemistry, 37(8):739–752, 2016.

F. Cazals and D. Mazauric.

Mass transportation problems with connectivity constraints, with applications to energy landscape comparison.

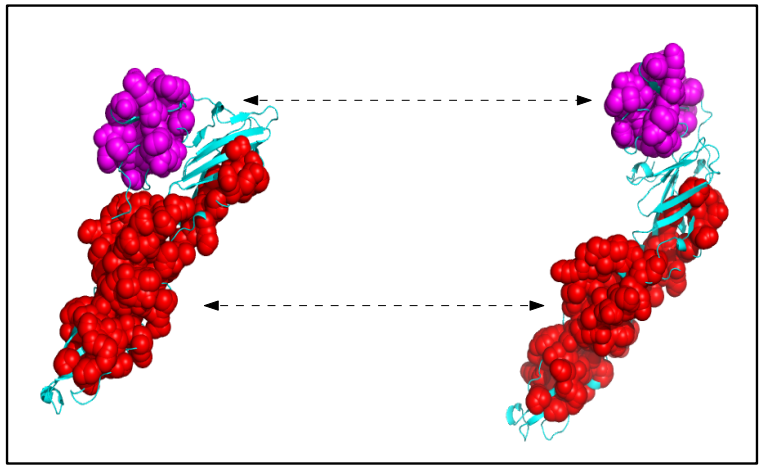
Submitted, 2016. Preprint: Inria tech report 8611.

F. Cazals and T. Dreyfus.

The Structural Bioinformatics Library.

Under revision, 2016.

A bootstrap method for detecting structurally conserved motifs



F. Cazals - R. Tetley

informatiques mathématiques
Inria

A bootstrap method for finding structurally conserved motifs

Motivation

Method

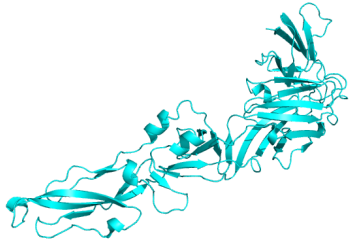
Application to class II fusion proteins

Structural similarity measures

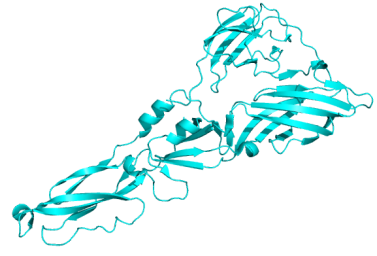
- Comparing conformations of:

  (PB1) the same molecule: mapping between atoms known (identical atoms!)
  $\rightarrow$ a geometric problem
  (PB2) two related molecules (e.g. two polypeptide chains of different length)
  $\rightarrow$ a dual combinatorial (common contacts) + geometric problem (how similar?)

- (PB1) Geometric comparison of the same molecule:
  - least RMSD, Cauchy-Binet score
  - issue #1: for large structures, small numbers $\sim 1\mathring{A}$ are fine; larger number are often meaningless.
  - issue #2 (related): a score does not give a mapping
- (PB2) Comparison of two related molecules:
  - contact map overlap
  - main issue: the longer the alignment the worse the geometric measure



TBEV pre-fusion



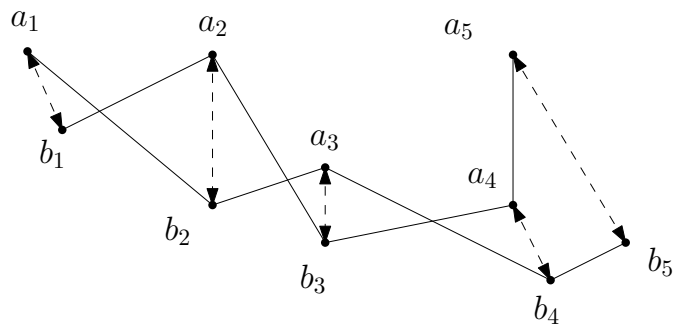
TBEV post-fusion

A geometric distance for two ordered point clouds:

the least Root Mean Square Deviation: IRMSD

- ▷ Data: two point sets $A = \{a_i\}_{i=1,\dots,n}$, $B = \{b_i\}_{i=1,\dots,n}$, with a 1-1 correspondence $a_i \leftrightarrow b_i$
- ▷ Root Mean Square Deviation:

$$\mathrm{RMSD}(A, B) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|a_i - b_i\|_2^2} \tag{1}$$



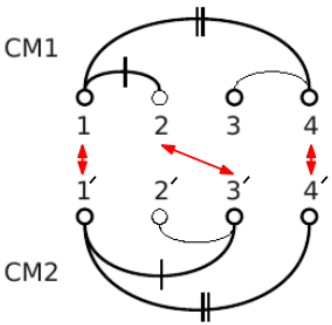
- ▷ least Root Mean Square Deviation:

$$\mathrm{IRMSD}(A, B) = \min_{g \in SE(3)} \mathrm{RMSD}(A, g \cdot B). \tag{2}$$

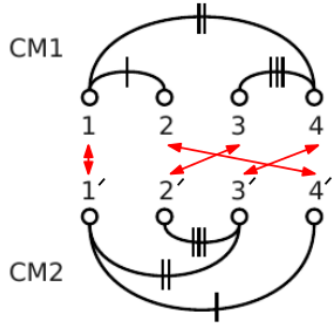
- ▷ Pros and cons:
  - pro: easy to compute (quadratic problem, SVD)
  - cons: medium range values for large structures tell nothing

Contact map overlap with Apurva

▷ Contact map of a polypeptide chain



A: Order MCES



B: MCES

A graph stating when two amino-acids (a.a.) are in close proximity (e.g. distance between their C_α carbons).

▷ Contact map overlap (CMO):

- Find subsets of vertices I and J yielding the largest set of common edges in their induced graphs
- Constraint: since amino-acids are linearly ordered, crossings are not allowed (Fig.)

▷ Hardness: decision problem is NP-hard.

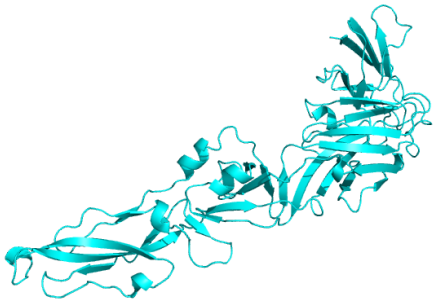
▷ Algorithm: integer programming model + branch-and-bound algorithm + Lagrangian relaxation.

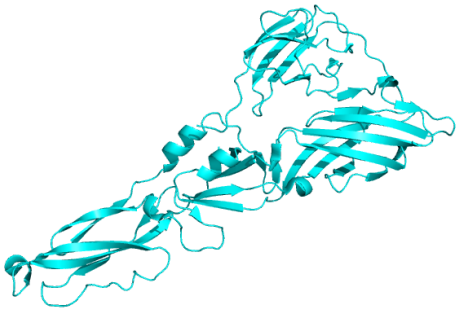
▷Ref: Papadimitriou et al, FOCS 1999

▷Ref: R. Andonov, N. Malod-Dognin, and N. Yanev, J. of Computational Biology, 2011

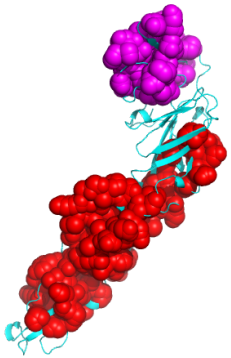
Ex: TBEV glycoprotein in two different conformations pre and post fusion

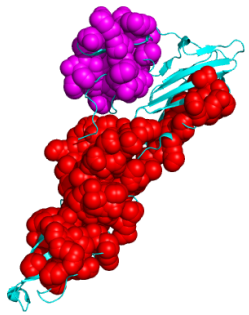
▷ Classical analysis:





▷ Our motifs:





Statistics from Apurva:

- 370 a.a. aligned
- IRMSD: 11.1Å

	pre-fusion	post-fusion
Motif	Alignment size	IRMSD
Red	88	1.69
Purple	40	0.38

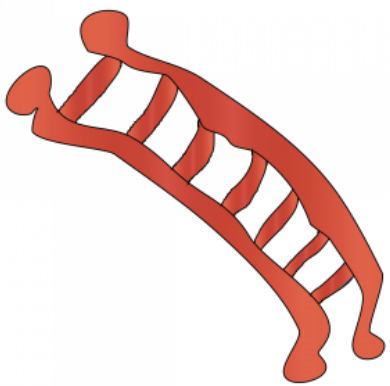
Structural Motif

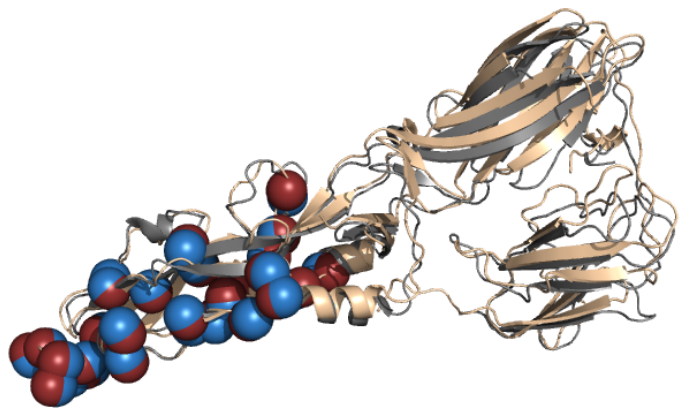
- Input: We are given two polypeptide chains S_A and S_B

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:

$$r_{\mathrm{lRMSD}}(M_A, M_B) = \mathrm{lRMSD}(M_A, M_B)/\mathrm{lRMSD}(S_A, S_B). \tag{3}$$

The sets M_A and M_B are called *structural motifs* provided that $|M_A| = |M_B| \geq s_0$ and $r_{\mathrm{lRMSD}}(M_A, M_B) \leq r_0$, for appropriate thresholds s_0 and r_0.





A bootstrap method for finding structurally conserved motifs

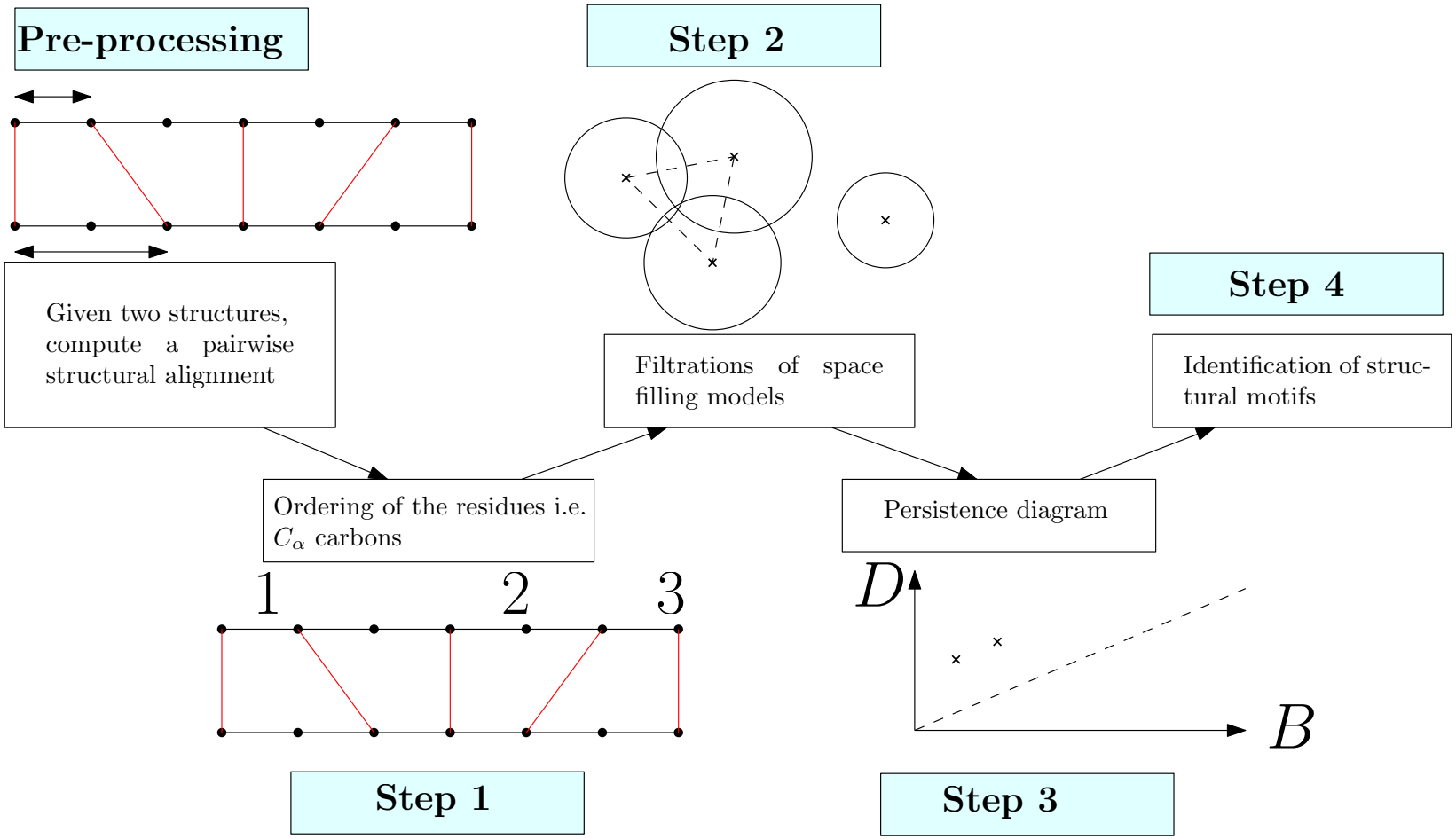
Motivation

Method

Application to class II fusion proteins

Detecting Motifs: overview

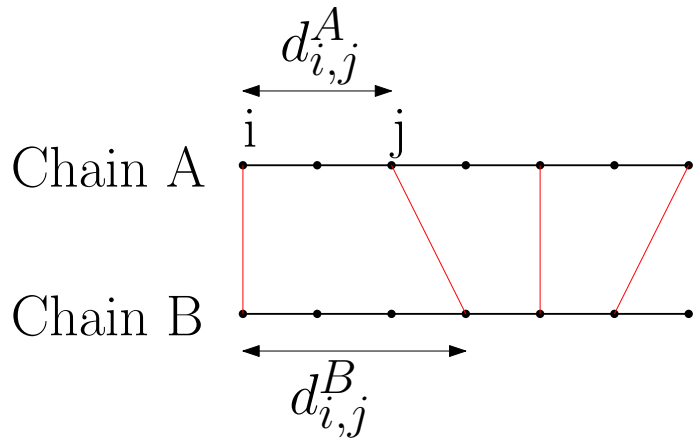
▷ Rationale: Using a criterion of structural conservation to order residues, the persistent connected components that arise upon inserting them in that order in a space filling model should correspond to structural motifs.



Step 1: computing C_α ranks for the polypeptide 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



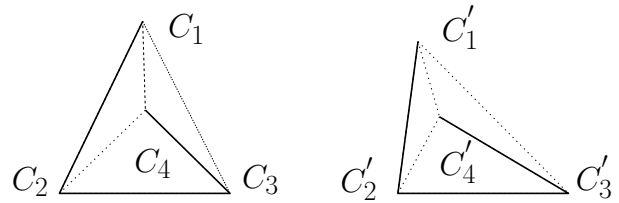
▷ Distance difference matrix between A and B:

$$s_{i,j} = \mid d^A_{i,j} - d^B_{i,j} \mid, i = 1, \ldots, N, j = 1, \ldots, N. \quad (4)$$

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

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.



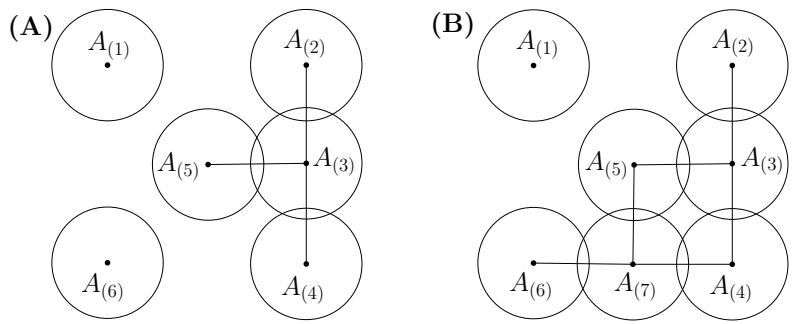
Sorted scores: $s_{12} < s_{34} < s_{23} < s_{13} < s_{14} < s_{24}$

Step 2: building filtrations of space filling models (NB: filtration = nested set)

- ▷ Model a collection of amino-acids with its Solvent Accessible Surface

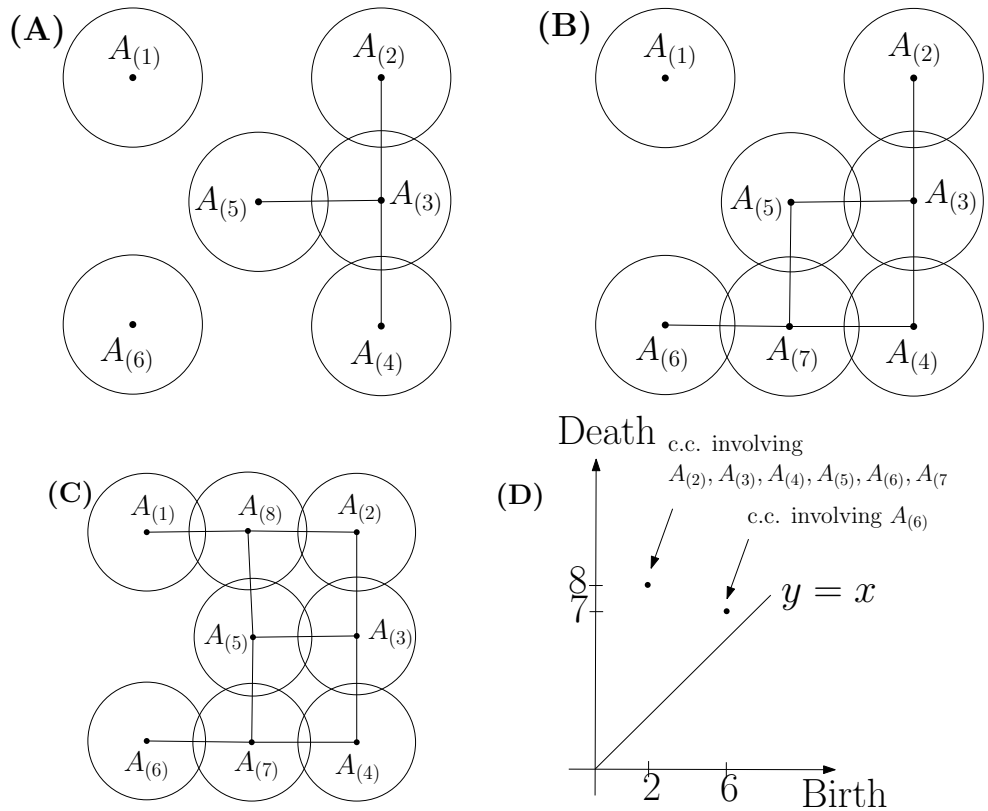
todo: add pict a.a. as lines / as vdw / as SAS

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



Step 3: compute the persistence diagram of the connected components of the space filling models

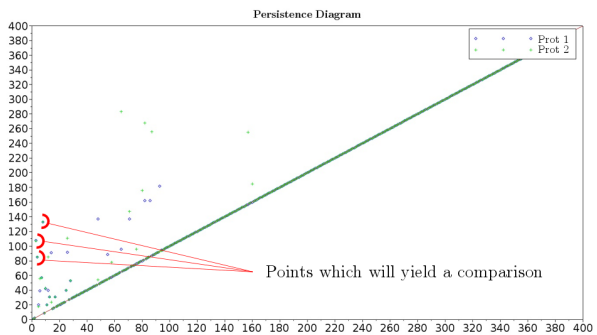
- ▷ Assessing the stability of conserved regions:
  - compute its connected components
  - maintain the associated persistence diagram



Step 4: identifying motifs – rationale

▷ Structure comparison yield motifs (def 1): connected components associated to the PD points:

- New structural alignment yields two motifs M_A and M_B
- if $r_{IRMSD} \leq r_0$ and $|M_A| = |M_B| \geq s_0$ record the structural motif



Comparing connected components associated with neighboring points in the PD

▷ 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 4: 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.

A bootstrap method for finding structurally conserved motifs

Motivation

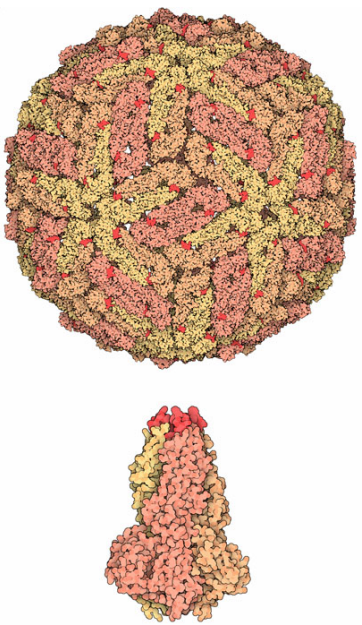
Method

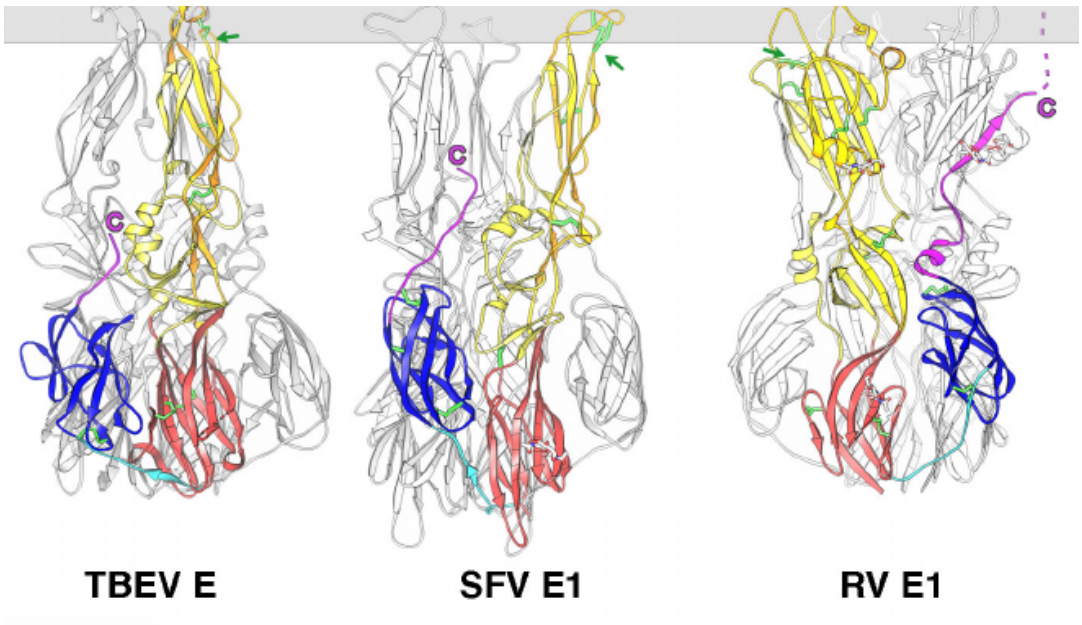
Application to class II fusion proteins

Class II fusion proteins

▷ Function: involved in membrane fusion of viruses–including dengue and zika.

▷ Hierarchical structure: secondary, tertiary, quaternary structures conserved
Organized in three domains.





▷ Main statistics: structural conservation $\sim 15\mathring{A}$; sequence identity $< 10\%$

▷Ref: Rey et al, Cell 157, 2014

Study

▷ Data: Consider N structures with mild atomic structure conservation and poor pairwise sequence identity.

▷ Questions:

- 1. can we identify structural motifs that would characterize the N structures?
- 2. are these motifs characterized by conserved sequence patterns, that would allow retrieving fusion proteins from databases of protein sequences?

Name	Family	Genus	PDB file
Semliki Forest virus	Togaviridae	Alphavirus	SFV-1RER.pdb
Dengue fever virus	Flaviviridae	Flavivirus	DFV.pdb
Tick-borne encephalitis virus	Flaviviridae	Flavivirus	TBEV.pdb
Hantaan river virus	Bunyaviridae	Hantavirus	HRV.pdb
Rift valley fever virus	Bunyaviridae	Phlebovirus	RVFV.pdb
Rubella virus	Togaviridae	Rubivirus	RBV-4ADI.pdb
C.Elegans	NA	NA	EFF1.pdb

Table: **Structures used in this study**

Structure of class II fusion proteins: details

Figure 1
Click here to download Figure: FIG1.pdf

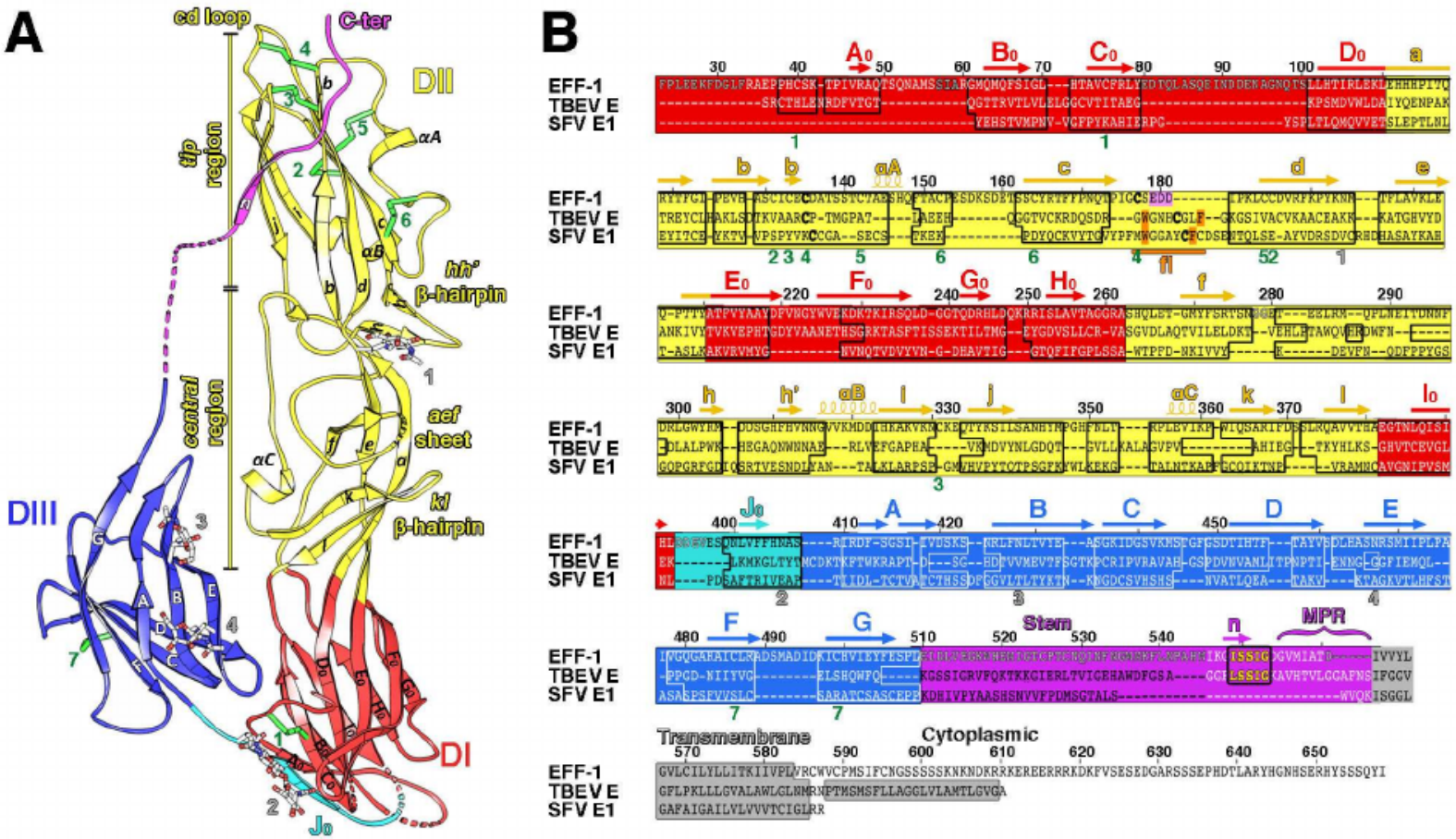
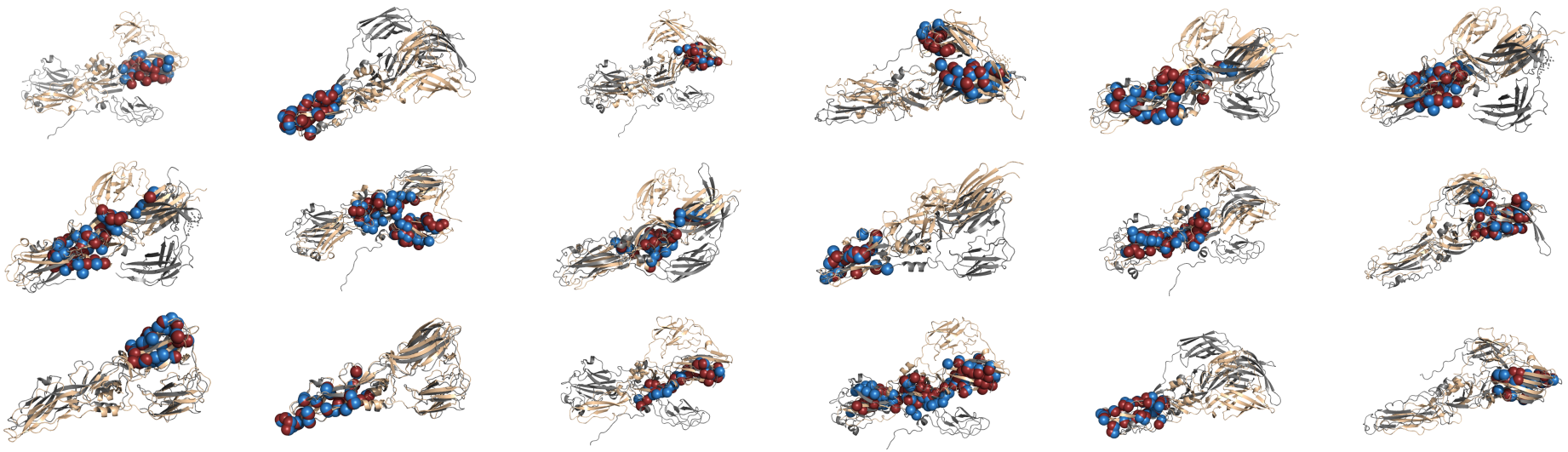


Figure: **SSE elements on fusion domains from Perez et al, 2014**

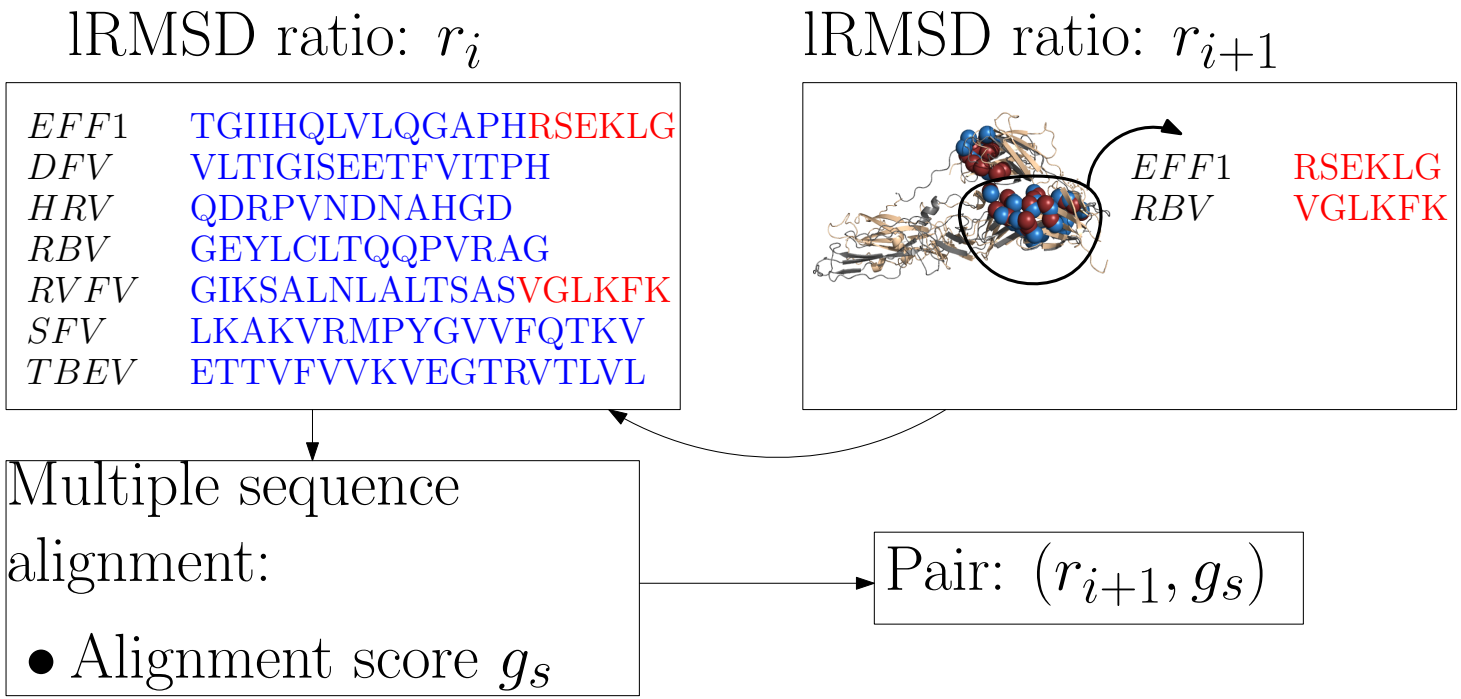
Structural motifs: results

▷ Summary: We uncover 124 structural motifs with sizes ranging from 20 to 153, 18 of which display and exceptionally good lRMSD ratio (≤ 0.5).

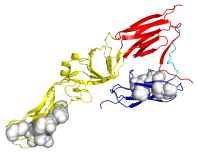


From structural motifs to sequence patterns

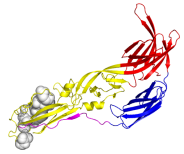
▷ Ordered structural motifs: Upon ordering the structural motifs with increasing lRMSD ratio ($r_1 < \cdots < r_i < r_{i+1} < \cdots < r_k$), we perform the following steps (on a per domain basis).



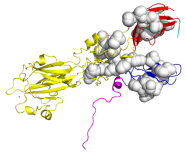
Results



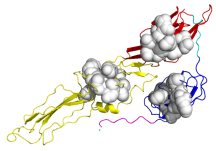
HRV-Hanta.



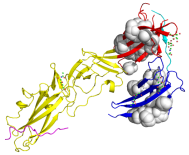
RVFV-Phlebo.



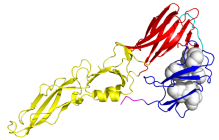
RBV-Rubi.



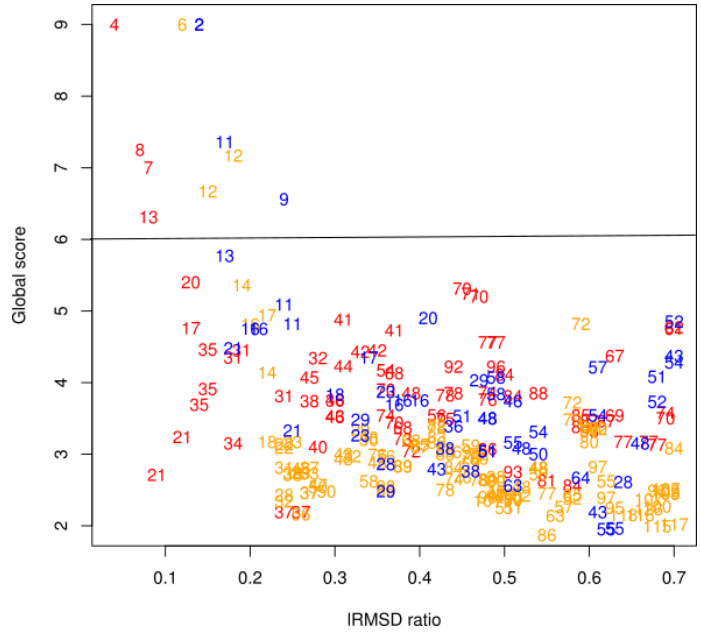
SFV-Alpha.



EFF-1



TBEV



Conclusions and further work

- Two main contributions:
  - A method to detect sub-regions of increased sequence and structural conservation in a set of structures.
  - Application of this method to the class II fusion proteins: yields structural motifs **significantly more conserved than the whole** + **correlation between this structural conservation** and the associated **sequence conservation**.
- Further work, applied:
  - Comparing proteins in different conformations – sampling energy landscapes
- Further work, theory:
  - When/why does our method work?
    - subtle interplay between the quality of the initial alignment, and the matching encoding in persistence diagrams
  - k-partition matching: NP-complete problem with polynomial time algorithms for specific (intersection) graphs