New perspectives on protein flexibility

High dimensional volumes and DoS

Move sets for polypeptide chains

Comparing energy landscapes

F. Cazals, Inria – Algorithms-Biology-Structure
http://team.inria.fr/abs

3IA Côte d'Azur
Axis 3

Overall perspective

- When is a well-posed computer science/modeling problem solved?
  - Intrinsic difficulty understood
  - (Almost) Optimal algorithms available
- Strategy:
  - Identify computationally tractable problems
    - Approximability is the issue, not NP-hardness
  - Develop efficient algorithms
    - Bias on the geometric/combinatorial side
  - Develop the corresponding software
    - Software: large research instrument
- Structure, thermodynamics, kinetics: will these problems get solved ?

New perspectives on protein flexibility

Volumes of polytopes

Tripeptide Loop Closure (TLC)

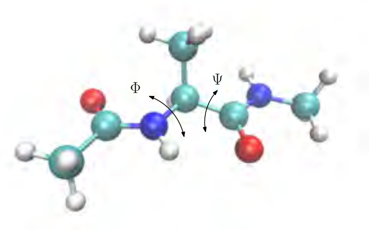
- TLC: background
- TLC steric constraints
- Loop sampling

Comparing energy lanscapes

Landscapes and thermodynamics

Density of states and partition functions

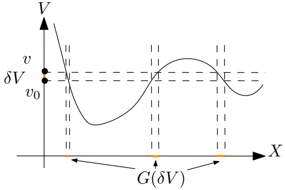
Dialanine



- Potential energy:

$$V_{\text{total}} = V_{\text{bonded}} + (V_{\text{vdw}} + V_{\text{electro}})$$

▷ Potential energy landscape:



▷ Density of states (DoS) for $A \subset X$:

- For any $v_0 < v$:

$$G([v_0, v]) = \int_A 1_{[v_0, v]}(V(x))dx$$

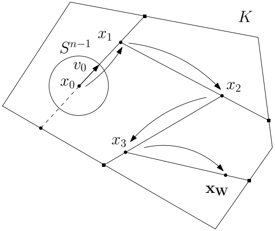
▷ Partition function for $A \subset X$ from DoS:

$$Z_A(T) = \int_A e^{-\beta v} dG(v)$$

▷ Nb: DoS calculation: volume calculation in phase space

Polytope volume calculations

▷ Problem statement: design effective algorithms to estimate the volume of high dimensional polytopes (dim. $\in [100 \ldots 1000]$)



▷ Unless P=NP: no polynomial time algorithm with approx factor $(cd / \log d)^d$

▷ State-of-the-art: multi-phase Monte Carlo methods embarking

- Rounding procedures to put the polytope in isotropic position
- Random walks: ball-walk, hit-and-run, billiard walk
    - Mixing times analysis – and heuristics for early stops

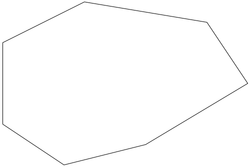
▷Ref: Cousins and Vempala, Math. Prog. Comp., 2016

▷Ref: Chalkis, Emiris, Fisikopoulos, arXiv:1905.05494, 2019

▷Ref: Chevallier et al, J. Computational Geometry, 2022

▷Ref: Chevallier et al, AISTATS, 2022

Volume of polytopes: hardness, randomized algorithms



▷ Hardness: no polynomial time algorithm with approx factor $(cd/\log d)^d$ – unless P=NP

▷ ε-approximation of the volume: for any parameter $\varepsilon > 0$, a number V

$$(1-\varepsilon)\mathrm{Vol}(K) \leq V \leq (1+\varepsilon)\mathrm{Vol}(K).$$

▷ (ε, δ)-approximation algorithm: algorithm returning an ε-approximation with a probability at least $1-\delta$.

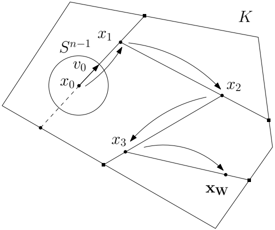
▷ Complexity, the $O^\star(n)$otation:

- $O(d^4)$: upper bound as a function of the dimension d
- $O^\star(d^4)$: term in $\log d, \varepsilon, \delta$ removed; focus on the dimension solely

▷Ref: Cousins, Vempala, SIAM J. Comp., 2018

Random walk: hit-and-run

- ▷ Goal: sample point in K according to a prescribed density f
- ▷ (Random-direction) hit-and-run: random point x_W after W steps



- ▷ Iteratively:
  - pick a random vector
  - move to random point on the chord $l \cap K$, chosen from the distribution induced by f on l
- ▷ Comments:
  - risk of being trapped near a vertex
  - large W helps *forgetting* the origin x_0

▷ Thm (Berbee et al) The limit distribution induced by HR is uniform in K.

▷ Thm (Vempala et al) HR can be modified to sample an isotropic Gaussian (restricted to K).

▷ Thm (Lovász) Let r and R denote the radii of the largest inscribed and circumscribed balls for K. One sample generation: $O^\star(d^3)$.

▷ NB: precise statement in terms of total variation distance omitted

▷Ref: Berbee et al, Math. Prog., 1987

▷Ref: Lovász, Math. Prog. Ser. A, 1999

Randomized algorithms: complexity

▷ Volume estimated using a sequence of isotropic Gaussians:

$$\mathrm{Vol}(K) = \int_K f_0(x)dx \frac{\int_K f_1(x)dx}{\int_K f_0(x)dx} \cdots \frac{\int_K dx}{\int_K f_{m-1}(x)dx} \equiv \int_K f_0(x)dx \prod_{i=1,\ldots,m} R_i \quad (1)$$

▷ Cooling schedule i.e. sequence of Gaussians $f_0, \ldots, f_m$:

- f_0: sharply peaked in K
- f_m: uniform distribution i.e. $a_m = 0$

▷ Thm. For a convex body K given by a membership oracle, and such that $B \subset K \subset RB$, an $(\varepsilon, \delta)-$ approximation can be obtained in time

$$O(\frac{d^4}{\varepsilon^2} \log^9 \frac{n}{\varepsilon\delta} + d^4 \log^8 \frac{n}{\delta} \log R) = O^\star(d^4) \quad (2)$$

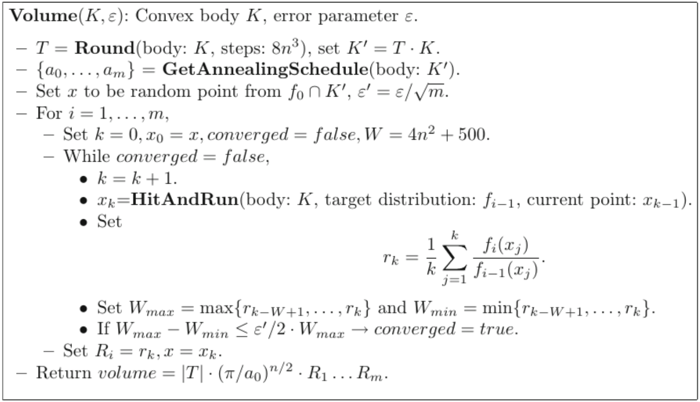
▷Ref: Lovász, Vempala, J Comp. Syst. Sciences, 2006

▷Ref: Cousins, Vempala, SIAM J. Comp., 2018

A practical algorithm: outline

▷ Method:

- multi-phase Monte-Carlo using $m = O(\sqrt{d})$ logconcave functions $\{f_0, \ldots, f_{m-1}\}$,
  - $f_i(x) \propto e^{-a_i^T x}$ or $f_i(x) \propto exp(-a_i \|x\|^2)$
- At each step: estimate $r_k \approx \int_K f_k(x)dx / \int_K f_{k-1}(x)dx$



```
Volume(K, ε): Convex body K, error parameter ε.
– $T$ = Round(body: $K$, steps: $8n^3$), set $K' = T \cdot K$.
– $\{a_0, \ldots, a_m\}$ = GetAnnealingSchedule(body: $K'$).
– Set $x$ to be random point from $f_0 \cap K'$, $\varepsilon' = \varepsilon/\sqrt{m}$.
– For $i = 1, \ldots, m$,
    – Set $k = 0, x_0 = x, converged = false, W = 4n^2 + 500$.
    – While $converged = false$,
        • $k = k + 1$.
        • $x_k$=HitAndRun(body: $K$, target distribution: $f_{i-1}$, current point: $x_{k-1}$).
        • Set
              $$r_k = \frac{1}{k} \sum_{j=1}^{k} \frac{f_i(x_j)}{f_{i-1}(x_j)}.$$
        • Set $W_{max} = \max\{r_{k-W+1}, \ldots, r_k\}$ and $W_{min} = \min\{r_{k-W+1}, \ldots, r_k\}$.
        • If $W_{max} - W_{min} \leq \varepsilon'/2 \cdot W_{max} \rightarrow converged = true$.
    – Set $R_i = r_k, x = x_k$.
– Return $volume = |T| \cdot (\pi/a_0)^{n/2} \cdot R_1 \ldots R_m$.
```

▷Ref: Cousins and Vempala, Math. Prog. Comp., 2016

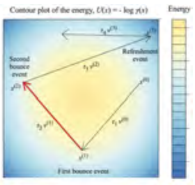
Piecewise deterministic Markov processes (PDMP)

the non-reversible Bouncy Particle Sampler (BPS)

▷ Notations: state space (position, velocity): $z = (x, v) \in E = \mathbb{R}^d \times \mathbb{R}^d$.

▷ PDMP z_t: a continuous time Markov process defined by:

1. a deterministic flow $\phi_t(z)$,
2. function determining the length of steps: jump kernel $\lambda(z)$
3. a jump kernel in phase (x, v) space: $q(\cdot|z)$



▷ BPS: PDMP to sample a distribution $\pi(x)$ in $\mathbb{R}^d$ using piecewise linear trajectories bouncing on high energy level set surfaces

1. Linear trajectories: $\phi_t(x, v) = (x + tv, v)$,
2. Arrival time of 1D inhomogeneous Poisson process of intensity $\lambda(x, v) = \max(0, -\langle \nabla_x (\log \pi)(x), v \rangle)$,
3. $q(\cdot|z)$: reflection w.r.t. the gradient of the potential:

$$(x, v') = \left(x, v - 2 \frac{\langle v, \nabla_x (\log \pi)(x) \rangle}{\| \nabla_x (\log \pi)(x) \|^2} \nabla_x (\log \pi)(x) \right) \tag{3}$$

4. +Refresh of velocity to ensure ergodicity

▷Ref: Doucet et al, Stats. and probability letters, 136, 2018

Extension: BPS on a bounded domain – a polytope

▷ Three types of events:

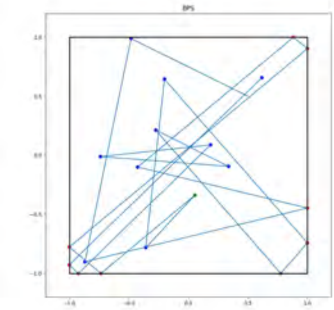
- PDMP events: as usual
- Reflexions on the boundary

$$v' = v - 2\frac{\langle n, v\rangle}{\|n\|^2}n, \qquad (4)$$

- Refresh events: velocity resampled from isotropic normal distribution

▷ Numerics: lazy update of linear algebra operations

▷ Example BPS trajectory in the 2d cube $[-1,1]^2$:



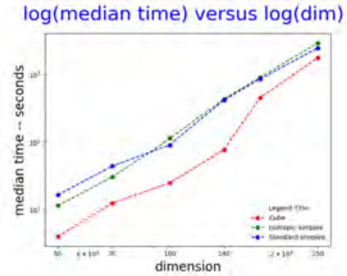
Blue: PDMP jump events, Red: reflections on the boundary, Green: refresh events

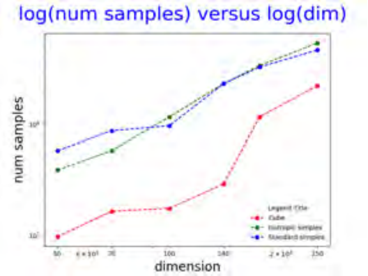
Nb: $\pi(x)$: Gaussian of variance $\sigma = 1$.

PDMP to compute volumes of polytopes: experiments

- Complexity: $C = O(d^c)$, dimension up to $d = 250$

- Protocol: find the smallest number of samples so that the estimated volume is within *err*% from the exact value


log(median time) versus log(dim)


log(num samples) versus log(dim)

- Linear regression in log log scale for the three polytopes:

	Time		Num. samples	
model	slope	R^2	slope	R^2
cube	3.77	0.96	1.94	0.88
Δ_{iso}	3.52	1.00	1.72	0.99
Δ_{std}	3.18	0.99	1.37	0.96

Computing volumes and DoS: outlook

- Polytopes: very efficient algorithms, provably correct
- Beyond polytopes: three classes of questions
  - Designing cooling schedules
  - Mixing times of RW – related to the conductance of the Markov chains i.e. narrow passages
  - Sample generation – beyond line-segments

Bibliography : volumes

A. Chevallier, F. Cazals, and P. Fearnhead.
Efficient computation of the the volume of a polytope in high-dimensions using piecewise deterministic markov processes.
In *AISTATS*, 2022.

A. Chevallier, S. Pion, and F. Cazals.
Improved polytope volume calculations based on Hamiltonian Monte Carlo with boundary reflections and sweet arithmetics.
J. of Computational Geometry, 13(1):55–88, 2022.

A. Chevallier and F. Cazals.
Wang-Landau algorithm: an adapted random walk to boost convergence.
J. of Computational Physics, 410(1):1–19, 2020.

New perspectives on protein flexibility

Volumes of polytopes

Tripeptide Loop Closure (TLC)

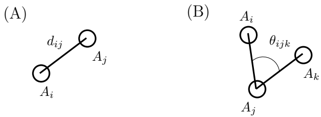
- TLC: background
- TLC steric constraints
- Loop sampling

Comparing energy lanscapes

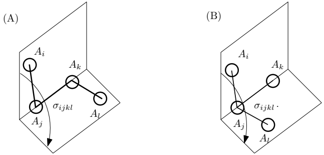
Geometric models: Cartesian and internal coordinates

- Cartesian versus internal coordinates: $\{x_i y_i z_i\}_i$ versus $\{d_{ij}, \theta_{ijk}, \sigma_{ijkl}\}$

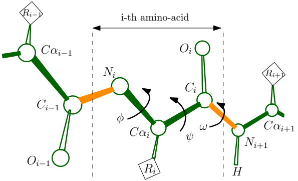
- Bond length and valence angle



- Dihedral angles



- Protein backbone

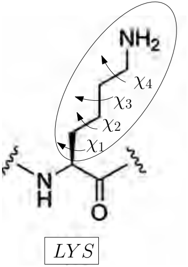


Ramachandran diagram, per a.a. type:

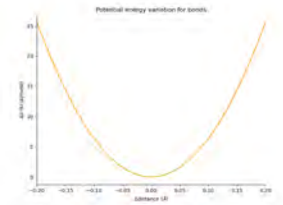
- bivariate distribution for (ϕ, ψ)

- Side chain: 20 natural amino acids

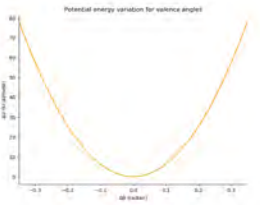
Exple: Lysine, 4 dihedral angles



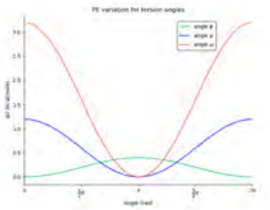
Softness of Internal coordinates –force constants from `CHARMM` 36



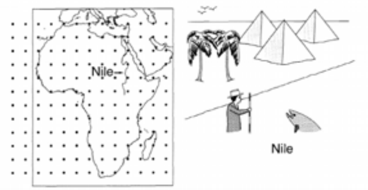
Bonds: $\delta d_{ij} \sim .2\text{Å} : \Delta V \sim 20 kcal/mol$



Valence angles: $\delta\theta_{ij} \sim 10^\circ : \Delta V \sim 20 kcal/mol$



Torsion angles:$\Delta V \sim 3 - 4 kcal/mol$

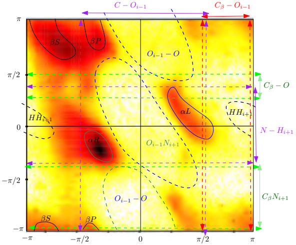


Quadrature vs importance sampling
(Frenkel and Smit, 2002)

⇒ Dihedral angles are indeed *soft* coordinates

The Ramachandran diagrams

▷ Ramachandran diagrams and populated regions

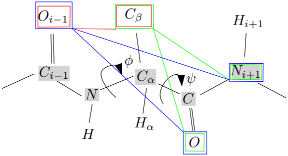


- Main regions: $\alpha L, \alpha R, \beta S, \beta P$
- Three prototypical diagrams
  - Glycine
  - Proline
  - Others – e.g. Aspartic acid

▷ Distance constraints and the Ramachandran tetrahedron

$C1 : C_\beta - O_{i-1}$ $C2 : C_\beta - O + C_\beta N_{i+1}$

$C3 : O_{i-1} - O + O_{i-1}N_{i+1}$

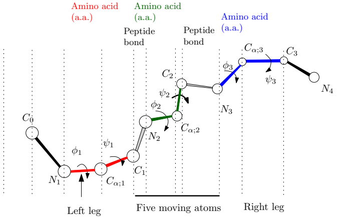


▷Ref: Stereochemistry of polypeptide chain configurations, JMB, 1963; Ramachandran et al

▷Ref: Revisiting the Ramachandran plot, Protein Science, 2003; Ho et al

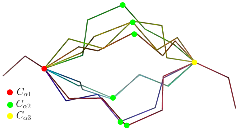
The Tripeptide loop closure – TLC

▷ TLC: for 3 amino acids, fix all internal coordinates BUT the $(\phi_i, \psi_i)_{i=1,2,3}$ angles

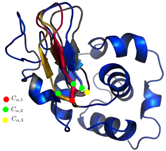


⇒ Find all possible values $(\phi_i, \psi_i)_{i=1,2,3}$ compatible with the fixed internal coordinates

▷ Theorem: at most 16 solutions



3 consecutive a.a.



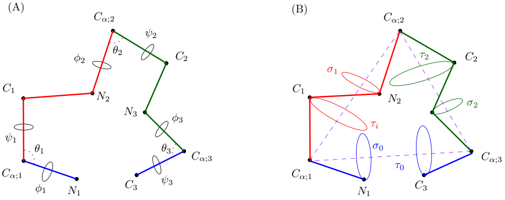
3 a.a. sandwiching SSE–CDRs

▷Ref: Gō and Scheraga, Macromolecules, 1970

▷Ref: Coutsias et al, J. Comp. Chem., 2004

TLC model: from six to three angles

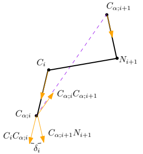
▷ Motions of the 3 rigid bodies: 6 angles



Nb: indices mod(3), e.g., $\sigma_0 = \sigma_3$

▷ . . . which are actually three

$$\sigma_i = \tau_i + \delta_i. \qquad (5)$$



$\delta_i = \angle\text{Plane}(C_{\alpha;i}C_{\alpha;i+1}C_i), \text{Plane}(C_{\alpha;i}C_{\alpha;i+1}N_{i+1})$

▷ Key ingredients of TLC:

- Initially: six dihedral angles $\{(\phi, \psi)\}_{\{i=1,2,3\}}$
- Then: three pairs $\{\delta_i, \tau_i\}$
- Finally: three angles τ_i

▷ The valence angle constraints: the θ_i angles at the $C_{\alpha;i}$s must remain constant.
$\Rightarrow$ It is the coupling introduced by the θ_i angles onto the rotation angles τ_i yields a degree 16 polynomial.

▷`Ref: Coutsias et al, 2004`

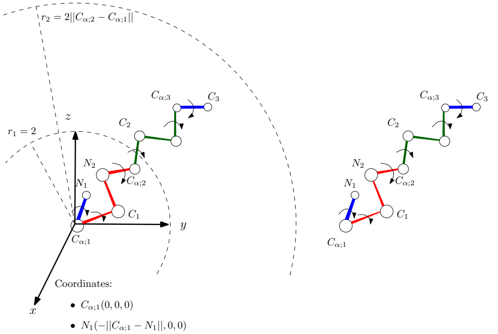
TLC with moving legs and embeddable tripeptides

▷ Geometric model:

- Tripeptide such that : left leg $N_i C_{\alpha;i}$ fixed, right leg $C_{\alpha;i+2} C_{i+2}$ free to move
- Six dihdedral angles $\{\phi_i, \psi_i\}$ free

▷ Question: provide necessary conditions on the position of the first and last segment–the **legs**, for the Tripeptide Loop Closure (TLC) algorithm to hold solutions.

▷ Nb: the relative position of legs suffices; in that case, position + orientation of $C_{\alpha;i+2} C_{i+2}$ yields a 5-dim search space.

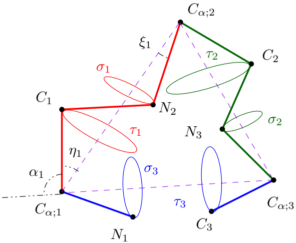


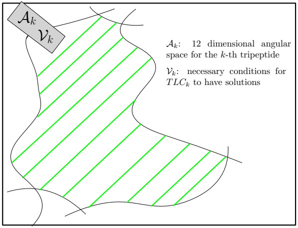
Coordinates:

- $C_{\alpha;1}(0,0,0)$
- $N_1(-||C_{\alpha;1} - N_1||, 0, 0)$

TLC: necessary conditions on the existence of solutions

▷ TLC problem for a tripeptide – say T_k: degree 16 polynomial parameterized by 12 angles defining the space $\mathcal{A}_k = \{\alpha_{k,i}, \eta_{k,i}, \xi_{k,i-1}, \delta_{k,i-1}\}, i \in \{1,2,3\}$.





$\mathcal{A}_k$: 12 dimensional angular space for the k-th tripeptide

$\mathcal{V}_k$: necessary conditions for TLC_k to have solutions

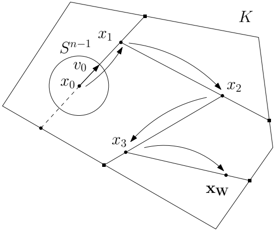
▷ Contribution: necessary conditions for TLC to admit solutions

- Based on the 12 angles in $\mathcal{A}_k$
- Defined by 24 hyper-surfaces in $\mathcal{A}_k$
- These hyper-surfaces: curved walls for Hit-and-Run

▷Ref: O'Donnell, Cazals; J. Comp. Chem., 2023

Polytope volume calculations

▷ Problem statement: design effective algorithms to estimate the volume of high dimensional polytopes (dim. $\in [100 \dots 1000]$)



▷ Unless P=NP: no polynomial time algorithm with approx factor $(cd/\log d)^d$

▷ State-of-the-art: multi-phase Monte Carlo methods embarking

- Rounding procedures to put the polytope in isotropic position
- Random walks: ball-walk, hit-and-run, billiard walk
  - Mixing times analysis – and heuristics for early stops

▷Ref: Cousins and Vempala, Math. Prog. Comp., 2016

▷Ref: Chalkis, Emiris, Fisikopoulos, arXiv:1905.05494, 2019

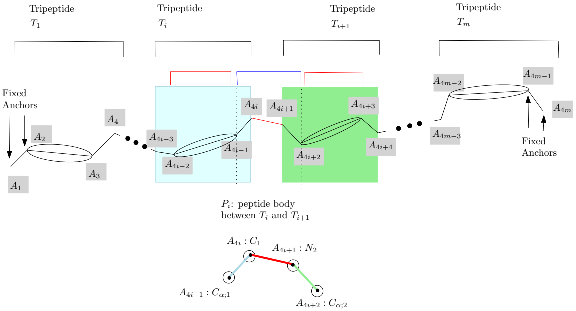
▷Ref: Chevallier et al, J. Computational Geometry, 2022

▷Ref: Chevallier et al, AISTATS, 2022

Global geometric model

- Loop studied L: $M = 3 \times m$ amino, m tripeptides: $L = T_1, \ldots, T_m$
- Loop decomposition: rigid peptide bodies and their complements

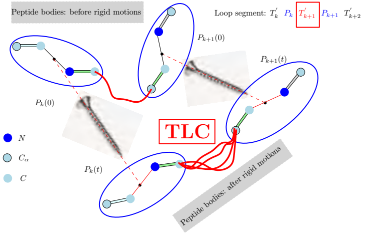
$$L = P_0\ T_1^{'}\ P_1\ \ldots\ P_{k-1}\ T_k^{'}\ P_k\ \ldots\ P_{m-1} T_m^{'} P_m. \tag{6}$$



- Parametric space:
  - For one peptide body: $SE(3) = SO(3) \times \mathbb{R}^3$
  - For one tripeptide: solution space of TLC. . . except that
    - The angular parameterization of TLC $\{\alpha, \xi, \eta, \delta\}$: depends on $SE(3) \times SE(3)$ since the left and right legs come from P_{i-1} and P_{i-1}

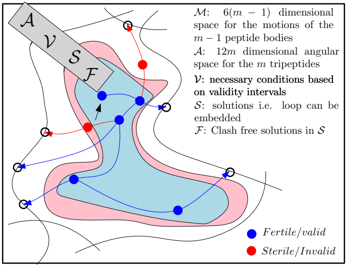
Loop sampling: spaces involved and solution sketch

- ▷ Loop decomposition into: rigid peptide bodies and tripeptides cores


Peptide bodies: before rigid motions
Loop segment: T'_k P_k T'_{k+1} P_{k+1} T'_{k+2}
$P_{k+1}(0)$
$P_{k+1}(t)$
$P_k(0)$
TLC
N
C_α
C
$P_k(t)$
Peptide bodies: after rigid motions

$$L = P_0\ T'_1\ P_1\ \ldots$$
$$P_k\ T'_{k+1}\ P_{k+1}\ \ldots$$
$$P_{m-1} T'_m P_m.$$

- ▷ Random sampling of loop conformations using Hit-and-Run:



$\mathcal{M}$: $6(m-1)$ dimensional space for the motions of the $m-1$ peptide bodies

$\mathcal{A}$: $12m$ dimensional angular space for the m tripeptides

$\mathcal{V}$: **necessary conditions based on validity intervals**

$\mathcal{S}$: solutions i.e. loop can be embedded

$\mathcal{F}$: Clash free solutions in $\mathcal{S}$

● *Fertile/valid*

● *Sterile/Invalid*

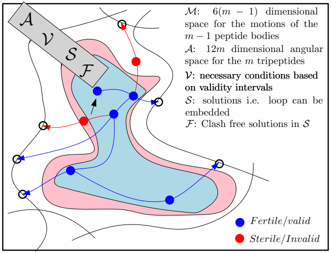
- Aim: perform rejection sampling in a region $\mathcal{V}$ containing all valid loop geometries.
- How: with Hit-and-Run in a domain characterizing necessary conditions – cf validity intervals

Loop sampling: spaces involved and solution sketch

▷ Global parameterization of the conformational space of the loop: based on rigid bodies associated with peptide bonds

- $\mathcal{M}$: motion space for the $m-1$ peptide bodies, essentially $(SE(3))^{m-1}$
- $\mathcal{A}$: $12m$-dimensional angular space coding the geometry of tripeptides
- $\mathcal{V}$: domain bounded by hyper-surfaces corresponding to Validity Constraints Necessary Constraints for TLC to admit solutions
- $\mathcal{S}$: the fertile space, where TLC admits one solution for each tripeptide
- $\mathcal{F}$: clash free solutions in $\mathcal{S}$ for $\{N, C_\alpha, C, O, C_\beta\}$ pairs

▷ Number of solutions: $\prod_i$(num solutions tripeptide i)



$\mathcal{M}$: $6(m-1)$ dimensional space for the motions of the $m-1$ peptide bodies

$\mathcal{A}$: $12m$ dimensional angular space for the m tripeptides

$\mathcal{V}$: necessary conditions based on validity intervals

$\mathcal{S}$: solutions i.e. loop can be embedded

$\mathcal{F}$: Clash free solutions in $\mathcal{S}$

● *Fertile/valid*

● *Sterile/Invalid*

Validity domain for the whole chain L with m tripeptides

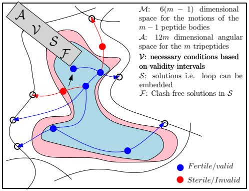
▷ Angles τ: $3m$ angles τ (3 for each tripeptide)

▷ Recap per angle τ:

- For one angle: at most 4 Depth One Validity Intervals (DOVI)
- For each DOVI: 2 sub-manifolds of $\mathcal{A}_k$ defined by the previous equations; yields (at most) 8 sub-manifolds in $\mathcal{A}_k$.

▷ For one tripeptide: 3 τ angles $\Rightarrow$ 24 constraint surfaces in the 12 dimensional angular space $\mathcal{A}_k$.

▷ For the whole loop: total of $24m$ constraint surfaces.



$\mathcal{M}$: $6(m-1)$ dimensional space for the motions of the $m-1$ peptide bodies

$\mathcal{A}$: $12m$ dimensional angular space for the m tripeptides

$\mathcal{V}$: necessary conditions based on validity intervals

$\mathcal{S}$: solutions i.e. loop can be embedded

$\mathcal{F}$: Clash free solutions in $\mathcal{S}$

Algorithms and parameters

▷ Unmixed loop sampler $\mathrm{ULS}^{N_V;N_{OR}}_{One|All;N_{ES}}[p_0]$:

- $One|All$ a flag indicating how many solutions are retained at each embedding step,
- N_{ES} the number of embedding steps,
- N_V the number of random trajectories followed in motion space,
- N_{OR} the output rate (the number of steps in-between the ones where conformations get harvested),
- p_0: the starting configuration.

▷ Mixed loop sampler $\mathbb{MLS}^{N_V;N_{OR}}_{One|All;N_{ES}}[p_0]$: every other step, the loop is shifted by 1 or 2 units to also sample the peptide bodies.

Loops sampling: ϕ, ψ and ω

- Typical values of the torsion angle ω:
  - SSE?
  - loops?

Loops sampling: ϕ, ψ and ω

▷ Typical values of the torsion angle ω:

- SSE? $\pi \pm 2 - 3^\circ$
- loops? $\pi \pm 15^\circ$

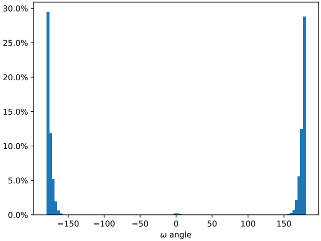
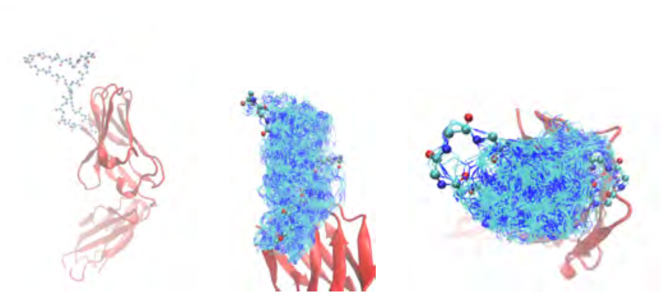


Illustration: CDR-H3-HIV, 30 amino acids

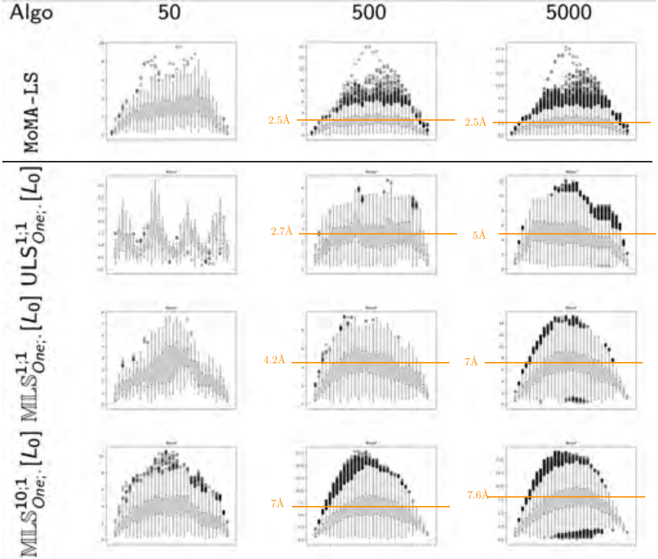
- System:
  - The loop is a complementarity-determining region (CDR-H3) from PG16, an antibody with neutralization effect on HIV-1.
  - pdbid: 3mme, chain A; residues: 93-100, 100A-100T, 101, 102.



Conformations generated by algorithm $\mathbb{MLS}^{\mathbf{1;1}}_{One;\mathbf{250}}$. **(A)** Variable domain (red) and the 30 a.a. long CDR3. **(B,C)** Side/top view of 250 conformations.

- Generation speed: $\sim$ 10 conformations per second

Results: sampling and study of fluctuations



Backbone RMSF (36 atoms) for the 12 amino acid long loop PTPN9-MEG2.

Bibliography : backbone move sets

T. O'Donnell, C.H. Robert, and F. Cazals.
Tripeptide loop closure: a detailed study of reconstructions based on Ramachandran distributions.
Proteins: structure, function, and bioinformatics, 90(3):858–868, 2022.

T. O'Donnell, V. Agashe, and F. Cazals.
Geometric constraints within tripeptides and the existence of tripeptide reconstructions.
J. Comp. Chem., 2023.

T. O'Donnell and F. Cazals.
Enhanced conformational exploration of protein loops using a global parameterization of the backbone geometry.
J. Comp. Chem., 2023.

Outlook

▷ Key features:

- First global parametric model of protein loops amenable to effective sampling strategies a-la Hit-and-Run
- Results: on par or better with state-of-the-art methods
  - Atomic fluctuations along the loop
  - Mutual reachability for existing conformations
- Insights on the intrinsic difficulty of the problem–via random walks and curved polytopes

▷ Open problems:

- Uniformity of sampling (Theorem)
- Connexion to micro-canonical ensembles and densities of states
- Sampling with side chains

New perspectives on protein flexibility

Volumes of polytopes

Tripeptide Loop Closure (TLC)

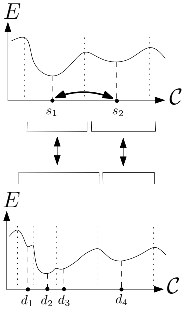
- TLC: background
- TLC steric constraints
- Loop sampling

Comparing energy lanscapes

Comparing (Sampled) Energy Landscapes: Motivation

▷ Comparing (sampled) landscapes:

- Assessing the coherence of two force2 fields for a given system (atomic, CG)
- Comparing two related systems: e.g. wild type/mutated proteins
- Comparing two simulations: different initial conditions and/or algorithms



▷ Idea: find a mapping between basins considering

- the similarity between the *native states* (one per basin)
- the coherence between the *volumes* of the basins (their probabilities)
- the connectivity between basins

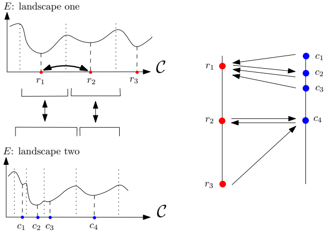
▷ Terminology: sampled (potential) energy landscape:

- portion revealed by a simulation
- given: minima, transitions between them, volumes of basins

Comparing Sets of Local Minima using a Minimum Oriented Spanning Forest (MSF): method

▷ Given two sets of local minima and a distance metric to compare them:

each local minimum chooses its nearest neighbor
cf One-sided Hausdorff distance



NB: local minima

- all those discovered during exploration
- persistent ones only (remove ruggedness)

▷ Statistics:

– ave. weight of edges from the first landscape to the second one: $\overline{w}^{MSF}_{1\to 2}$
– ave. weight of edges from the second landscape to the first one: $\overline{w}^{MSF}_{2\to 1}$

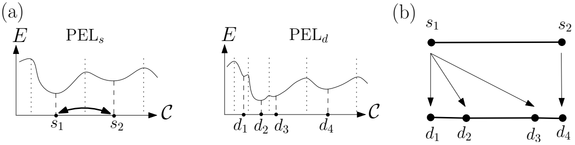
▷ Remarks:

– can be combined with topological persistence
– algorithm, cf MST: Borůvka/ distributed Kruskal

Comparisons without Connectivity Constraints:

the Earth Mover Distance yields a Linear Program

▷ Consider two landscapes : PEL$_s$ with n_s basins, PEL$_d$ with n_d basins



▷ Problem Earth-Mover-Distance (EMD):

find the transport plan of minimum cost, i.e. solution of the following linear program

$$LP \begin{cases} \text{Cost: Min} \sum_{i=1,\ldots,n_s, j=1,\ldots,n_d} f_{ij} \times d_{\mathcal{C}}(s_i, d_j) & \\ \sum_{i=1,\ldots,n_s} f_{ij} = w_j^{(d)} & \forall j \in 1,\ldots,n_d, \\ \sum_{j=1,\ldots,n_d} f_{ij} \leq w_i^{(s)} & \forall i \in 1,\ldots,n_s, \\ f_{ij} \geq 0 & \forall i \in 1,\ldots,n_s, \forall j \in 1,\ldots,n_d \end{cases}$$

▷ Property: in OPT, the number of edges carrying flow is $O(n_s + n_d - 1)$

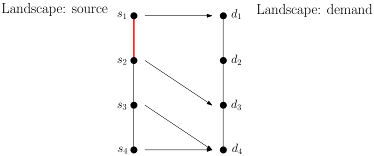
▷ Pros and cons:

- Information used: location of minima, weight of basins
- Linear program: solved in polynomial time
- Connectivity information not used

▷Ref: Chvátal, Linear programming, 1983; Rubner, Tomasi, Guibas, IJCV, 2000

Comparisons with Connectivity Constraints

▷ Earth Mover Distance: may violate the connectivity constraints



▷ Def: Transport plan with connectivity constraints: every connected subgraph of PEL$_s$ exports towards a connected subgraph of PEL$_d$

🦇There may exist an exponential number of connected subgraphs

▷ Problem EMD-CCC: maximum flow under constraints of {maximum cost, connectivity constraints (and transport plan size M)}

▷ Complexity results

- Decision versions of EMD-CC and EMD-CCC: NP-complete
- Optimization version of EMD-CC is not in APX
  If P $\neq$ NP: no polynomial algorithm with constant approx factor

▷ Algorithm `Alg-EMD-CCC-G`

- Greedy polynomial algorithm producing solutions i.e.
  respecting the connectivity constraints and the max cost.
  Complexity: $O(n^3m^2)$, with n and m the num. vertices of the graphs

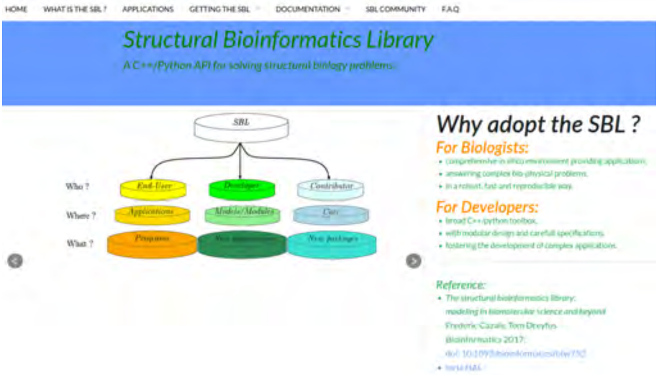
Bibliography : comparing landscapes

- F. Cazals, T. Dreyfus, D. Mazauric, A. Roth, and C.H. Robert.
  Conformational ensembles and sampled energy landscapes: Analysis and comparison.
  J. Comp. Chem., 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):4, 2016.

- F. Cazals and D. Mazauric.
  Mass transportation problems with connectivity constraints, with applications to energy landscapes comparison.
  Technical Report 8611, Inria, 2016.

The Structural Bioinformatics Library



Structural Bioinformatics Library

A C++/Python API for solving structural biology problems.

Why adopt the SBL ?

For Biologists:

- comprehensive in silico environment providing applications,
- answering complex bio-physical problems,
- in a robust, fast and reproducible way.

For Developers:

- broad C++/python toolbox,
- with modular design and carefull specifications,
- fostering the development of complex applications.

Reference:

- The structural bioinformatics library: modeling in biomolecular science and beyond
  Frederic Cazals, Tom Dreyfus
  Bioinformatics 2017;
  doi: 10.1093/bioinformatics/btw752
- [illegible] HAL

▷ Pointers:

- ▶ Frontpage
- ▶ Applications
- ▶ Online doc

▷ Upates

- ▶ Conda channels for linux and macos
- ▶ Online demos for applications
- ▶ Next: plugins for VMD and pymol

▷Ref: Cazals and Dreyfus; Bioinformatics, 2016

▷Ref: Le Breton, Sarti, Cazals; In preparation

Acknowledgments

- Collaborators
    - Dorian Mazauric, Timothee O'Donnell (Inria)
    - Augustin Chevallier (Univ. Lancaster)
    - Edoardo Sarti and Côme Le Breton (Inria)
    - Charles Robert (CNRS IPBC)
    - Juan Cortés (CNRS LAAS)
    - David Wales (Cambridge University)

    - Inria
    - 3IA Côte d'Azur
    - Université Côte d'Azur