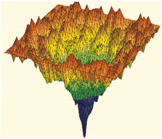
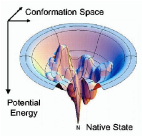
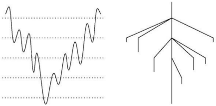
Exploration and characterization of energy landscapes: selected tools from the Structural Bioinformatics Library

`http://sbl.inria.fr`







Frederic.Cazals@inria.fr

Introduction

Software: the SBL

Energy Landscapes analysis

Energy landscapes comparison

Comparing two clusterings

Exploration and characterization of energy landscapes: tools from the SBL

Introduction

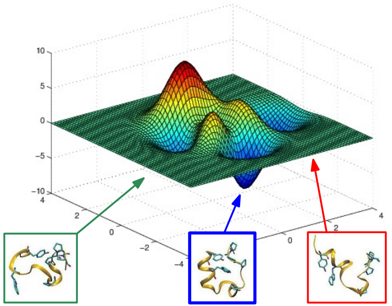
Software: the SBL

Energy Landscapes analysis

Energy landscapes comparison

Comparing two clusterings

Emergence of macromolecular function(s) from Structure – Thermodynamics – Dynamics

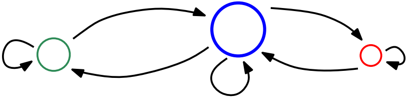


Potential Energy Landscape

- large number of local minima
- enthalpic barriers
- entropic barriers

Structure: stable conformations i.e. local minima of the PEL

Thermodynamics: meta-stable conformations i.e. ensemble of conformations easily inter-convertible into one - another.



Dynamics: transitions between meta-stable conformations e.g. Markov state model

Exploration and characterization of energy landscapes: tools from the SBL

Introduction

Software: the SBL

Energy Landscapes analysis

Energy landscapes comparison

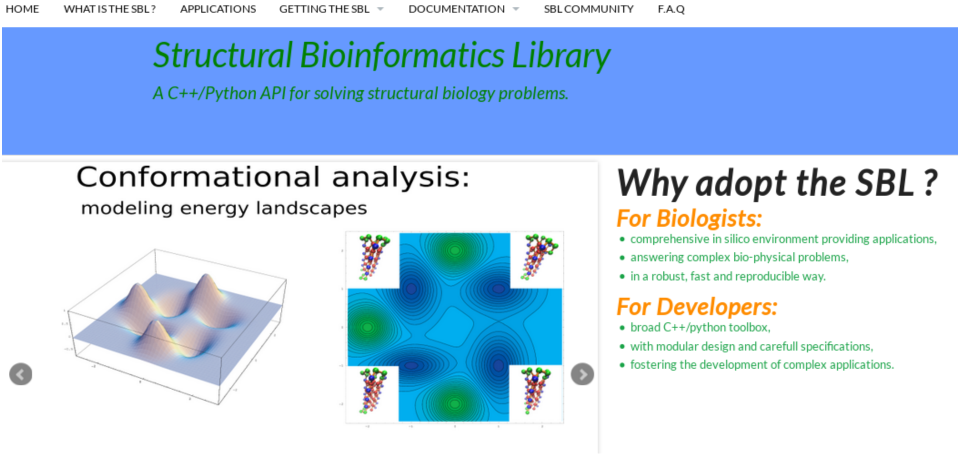
Comparing two clusterings

The Structural Bioinformatics Library

- Rationale for starting yet another library:
  - Many excellent environment / libraries, but low level algorithms and end-user applications entangled
  - Often hard / impossible to play legos with software components
  - Often hard / impossible to hybridize biophysical models (atomic, CG) and low level algorithms/applications

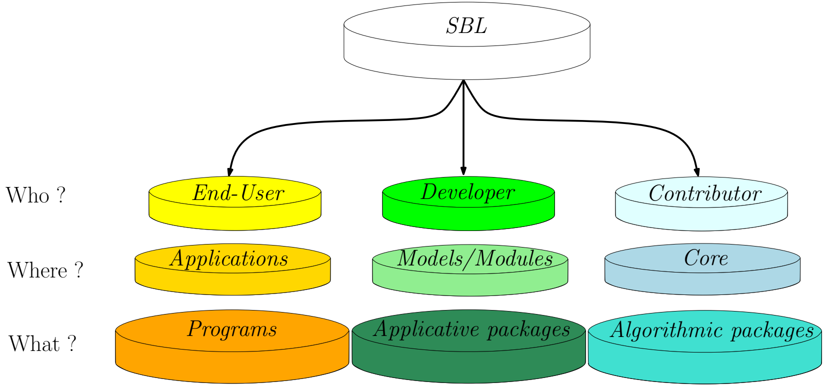
The Structural Bioinformatics Library

http://sbl.inria.fr



▷Ref: Cazals and Dreyfus; Bioinformatics, 2017

The Structural Bioinformatics Library: Architecture



Tour of the web site

- Web site: `http://sbl.inria.fr`
  - Applications
  - Online documentation

Using the SBL: the easy way

- Install with conda from the conda https://anaconda.org/sbl/
  - Delivers: executables (sbl-*.exe), documentation, demo / data sets
  - Currently: only linux package provided; MacOS . . . soon.
- Run demos provided in the Jupyter notebooks
  - Voronoi interfaces: https://sbl.inria.fr/doc/Space_filling_model_interface-user-manual.html
  - Lanscape explorers: https://sbl.inria.fr/doc/Landscape_explorer-user-manual.html

Advanced software design: C++ templates

- Template: parameter class used to instantiate a generic class/algorithm
- Traits class: template parameter with specific requirements

```
template <class Traits> class Explorer{
 public:

  typedef typename Traits::Conformation         Conformation;
  typedef typename Traits::Move_set_class       Extender;
  typedef typename Traits::Metropolis_criterion Accept_criterion;
};
```

- Template classes yields an un-challenged flexibility:
  - SBL-CORE: templated algorithmic classes coming with requirements
  - SBL-MODELS: models (geometric, biophysical) used to assemble traits classes
  - SBL-APPLICATIONS: *glued* models and core algorithms

Advanced software design C++: illustrations

▷ Example 1: force fields: three packages yield all force fields (CHARMM, AMBER, MARTINI,. . .)

– Rationale: a force field generically requires iterating and typing pairs / triples / quadruples of (pseudo-)atoms

– Base classes:
http://sbl.inria.fr/doc/group__Molecular__covalent__structure-package.html
http://sbl.inria.fr/doc/group__Molecular__coordinates-package.html
http://sbl.inria.fr/doc/group__Molecular__potential__energy-package.html

– Configuration files:
individual force fields in ffXML format

– Applications:
http://sbl.inria.fr/doc/group__Landscape__explorer-package.html
http://sbl.inria.fr/doc/group__Energy__landscape__analysis-package.html
http://sbl.inria.fr/doc/group__Energy__landscape__comparison-package.html

▷ Example 2: interfaces of macro-molecular complexes

– Rationale: an interface in a space filling model (atomic or CG) generically requires *facing* atoms

– Base classes:
http://sbl.inria.fr/doc/group__Molecular__structure__classifier-package.html
http://sbl.inria.fr/doc/group__Pointwise__interactions-package.html
http://sbl.inria.fr/doc/group__Molecular__interfaces-package.html

– Applications:
http://sbl.inria.fr/doc/group__Space__filling__model__interface-package.html
http://sbl.inria.fr/doc/group__Space__filling__model__interface__finder-package.html

Exploration and characterization of energy landscapes: tools from the SBL

Introduction

Software: the SBL

Energy Landscapes analysis

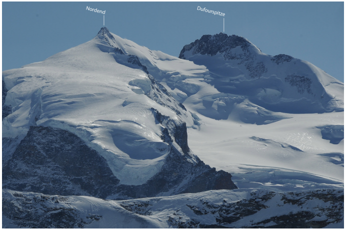
Energy landscapes comparison

Comparing two clusterings

Landscapes: Significant Features versus Noise

When is a Local Maximum/minimum a Significant Peak/Lake?

- Key features in a landscape: lakes , peaks, passes
  - local minima, maxima, and *saddles* of the elevation function
- Defining a peak . . . a matter of scales
  - prominence: closest distance to the nearest local maximum with higher elevation
  - **culminance**: elevation drop to the saddle leading to a higher local maximum
- Some well known peaks have tame statistics: the Norden peak
  - fourth highest peak of the Mont Rose massif, 4609 meters
  - prominence: 575 meters; culminance: 94 meters



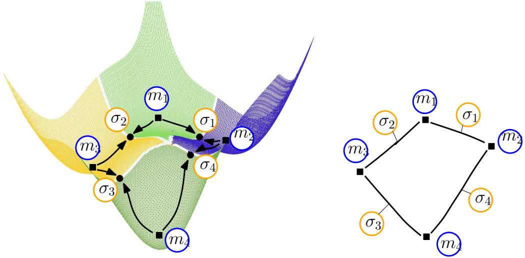
▷Ref:

http://www.zermatt.ch/en/page.cfm/zermatt_matterhorn/4000er/nordend

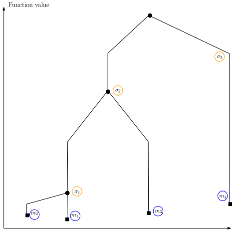
Height Functions and their Critical points: Key Concepts from Morse Theory

▷ Example: the Himmelblau function:

$$f(x, y) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2.$$



Transition graphs



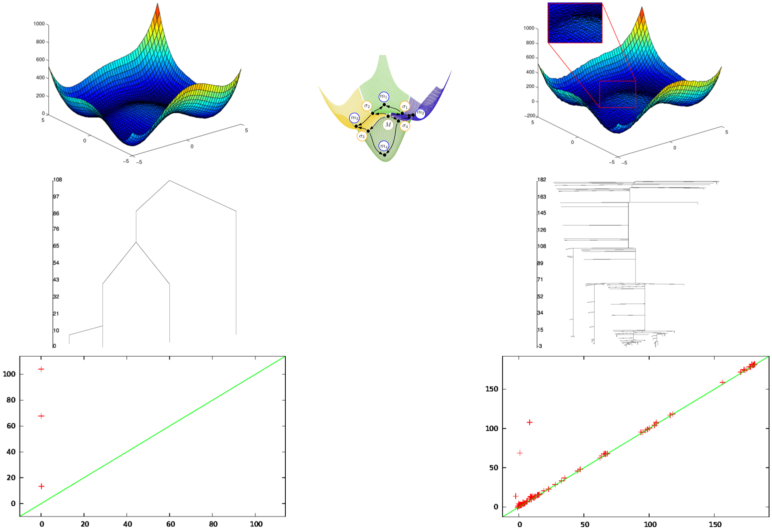
Disconnectivity graph

▷ NB: critical points and their connexions define the Morse-Smale-Witten (MSW) complex

```
▷Ref:  Milnor, Morse Theory, 1963
▷Ref:  Banyaga and Hurtubise, Morse homology, 2004
```

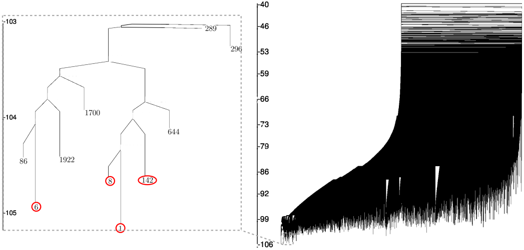
Landscape Analysis at a Glimpse: topological persistence

The Himmelblau function: $f(x, y) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2$

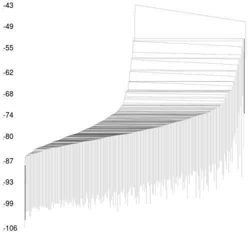


BLN69: Persistence reveals Frustration

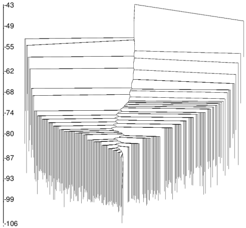
▷ Whole disconnectivity graph for the 458,082 local minima



▷ Persistence reveals frustration: persistence threshold 15 ϵ



▷ Persistence + clubbing saddles in energy slices of 0.5ϵ

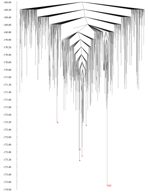


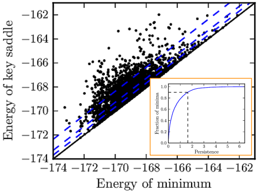
▷Ref: Cazals et al, J. Comp. Chem., 2015

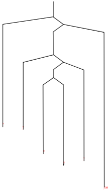
Analysis of sampled energy landscapes

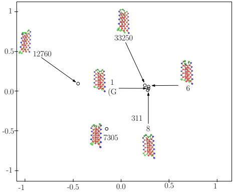
- ▷ Contributions: novel concepts and algorithms to

– Analyze conformational ensembles
– Analyze sampled energy landscapes: coarse graining with topological persistence









- ▷ Assessment:

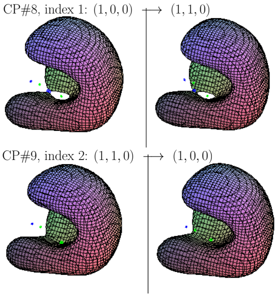
– State-of-the-art algorithms analysis/coarse-graining methods
– Most of the analysis geared towards potential energy landscapes
work ahead on free energy landscapes

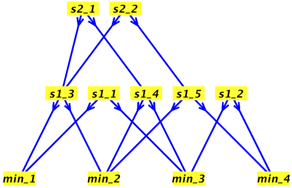
▷Ref: Cazals, Dreyfus, Mazauric, Roth, Robert; J. Comp. Chem., 2015
▷Ref: Carr, Mazauric, Cazals, Wales; J. Chem. Phys.; 2016

Morse Homology: Illustration

▷ Example: evolving homology of a 3D landscape defined by a polynomial

$$P = \left(x^2 + y^2 + z - 1\right)^2 + \left(z^2 + y^2 + x - 3\right)^2 + \left(x^2 + z^2 + y - 2\right)^2$$





NB: max at infinity not represented.

▷ Key construction: the **Morse-Smale-Witten chain complex** i.e. the connections between critical points whose indices differ by one is sufficient to compute the Betti numbers

▷Ref: R. Tom, Sur une partition en cellules...; CRAS; 1449
▷Ref: S. Smale; Differentiable dynamical systems; Bull. AMS; 1967
▷Ref: R. Boot, Morse theory indomitable, Pub. IHES, 1988

Intermezzo: higher order homology

- Stability of basins: order 0 homology
- Stability of loops: order 1 homology
  - Loops in RNA?
- ...

SBL packages

- Transition_graph_of_energy_landscape_builders
  - http://sbl.inria.fr/doc/Transition_graph_of_energy_landscape_builders-user-manual.html
- Energy_landscape_analysis
  - http://sbl.inria.fr/doc/Energy_landscape_analysis-user-manual.html

Exploration and characterization of energy landscapes: tools from the SBL

Introduction

Software: the SBL

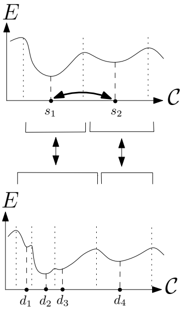
Energy Landscapes analysis

Energy landscapes comparison

Comparing two clusterings

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

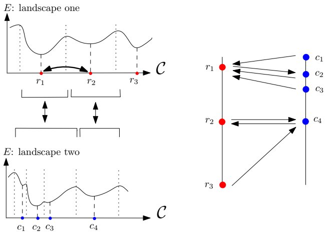
- Terminology: sampled (potential) energy landscape:
  - portion revealed by a simulation
  - given: minima, transitions between them, volumes of basins

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 (e.g. in the lRMSD sense)

- ▷ Example: comparing local minima of two landscapes



NB: local minima

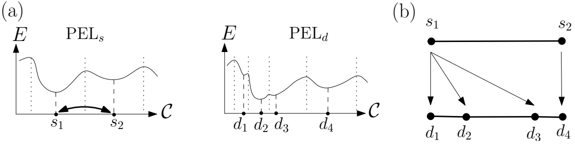
- all those discovered during exploration
- persistent ones only (remove ruggedness)

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

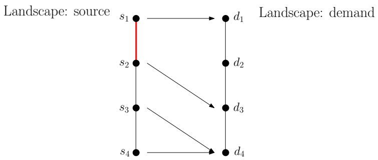
$$LP \begin{cases} \text{Cost: Min} \sum_{i=1,\ldots,n_s, j=1,\ldots,n_d} f_{ij} \times d_{\mathcal{C}}(s_i, d_j) & \\ \sum_{i=1,\ldots,n_s} f_{ij} = w_j^{(d)} & \forall j \in 1,\ldots,n_d, \\ \sum_{j=1,\ldots,n_d} f_{ij} \leq w_i^{(s)} & \forall i \in 1,\ldots,n_s, \\ f_{ij} \geq 0 & \forall i \in 1,\ldots,n_s,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

- ▷ 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,

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 $\mathbf{P} \neq \mathbf{NP}$: 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: $O(n^3 m^2)$, with n and m the num. vertices of the graphs

SBL packages

- Energy_landscape_comparison
  - http://sbl.inria.fr/doc/Energy_landscape_comparison-user-manual.html

Exploration and characterization of energy landscapes: tools from the SBL

Introduction

Software: the SBL

Energy Landscapes analysis

Energy landscapes comparison

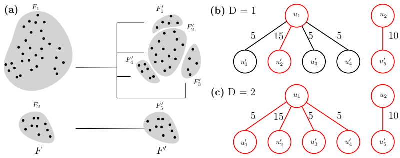
Comparing two clusterings

Comparing two clusterings using matchings between clusters of clusters

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

Preprint 2017

https://sbl.inria.fr/doc/D_family_matching-user-manual.html



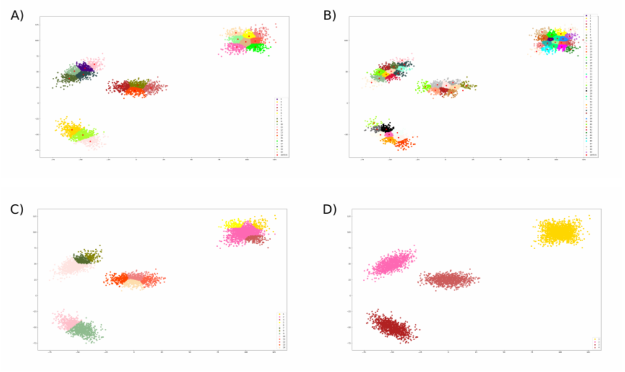
Comparing clusterings: at which *scale* do clusters merge?

What is the *right* number of clusters?

▷ Example:

- Using `k-means++` to cluster 5000 samples from five Gaussian blobs
- Using D-family matching to infer the *right/natural* # of clusters

(A) k-means++, $k = 20$ (B) k-means++, $k = 50$

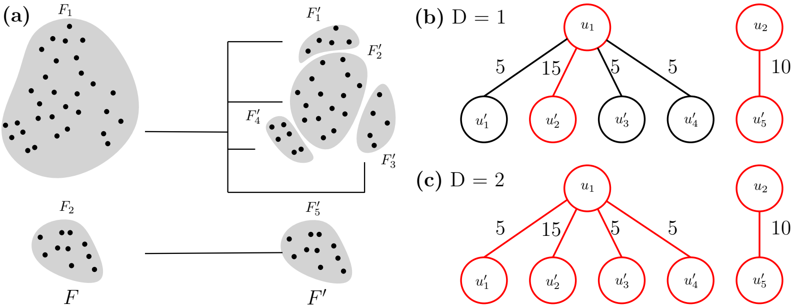


(C) $D = 3$, 17 meta clusters, $\Phi_D(G) = 4068$ (D) $D = 4$, 4 meta clusters, $\Phi_D(G) = 5000$

On the stability of clusterings

- Clustering methods:
  Come in many guises
  Have (a plethora of) parameters

- Key questions: are clusterings stables / what is the right clustering?



- Comparing clusterings via clusters of clusters
  - Find a matching between clusters of clusters
  - Meta-cluster: induced connected component of the *intersection graph*
  - Matching complexity governed by the *diameter* of the metaclusters

Comparing clusterings using matchings between clusters of clusters

▷ Contributions:

– Formalization of the D-family matching problem

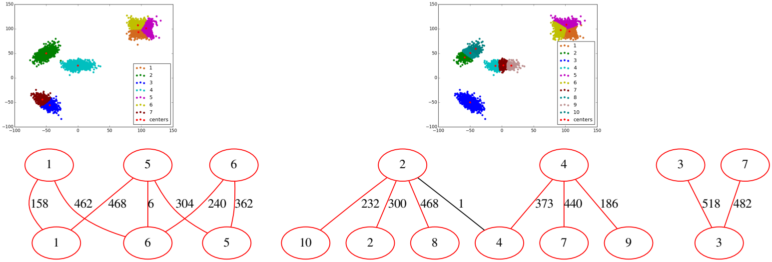
– NP-completeness results and unbounded approximation ratio for simple strategies
  Open: is the problem APX hard?

– Exact polynomial time algos. for selected intersection graphs (trees)

– Heuristics for general graphs

– Extensive experiments (vs. the variation of information)

▷ Stability of kmeans++:



▷Ref: Cazals, Mazauric, Tetley, Watrigant; submitted

SBL packages

- D_family_matching:
  - http://sbl.inria.fr/doc/D_family_matching-user-manual.html

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 T. Dreyfus.

The Structural Bioinformatics Library.

Bioinformatics,33(7): 1–8, 2017.

F. Cazals and D. Mazauric.

Optimal transportation problems with connectivity constraints.

Submitted, 2016. Preprint: Inria tech report 8991.

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

Comparing two clusterings using matchings between clusters of clusters.

Submitted, 2017.

A. Chevallier and F. Cazals.

Boosting the convergence of the Wang-Landau Algorithm for density of states calculations. In preparation.

Enjoy. . .

- ▷ The science

Enjoy. . .

- ▷ The scenery

Enjoy. . .

▷ The SBL

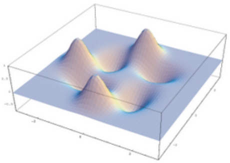
HOME WHAT IS THE SBL ? APPLICATIONS GETTING THE SBL DOCUMENTATION SBL COMMUNITY F.A.Q

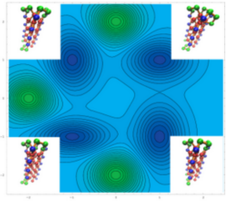
Structural Bioinformatics Library

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

Conformational analysis:

modeling energy landscapes





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:

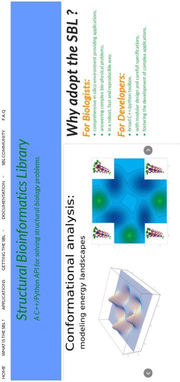
- broad C++/python toolbox,
- with modular design and carefull specifications,
- fostering the development of complex applications.

Enjoy. . .

- Find the right mixture. . .

The Science

The Scenery



The SBL