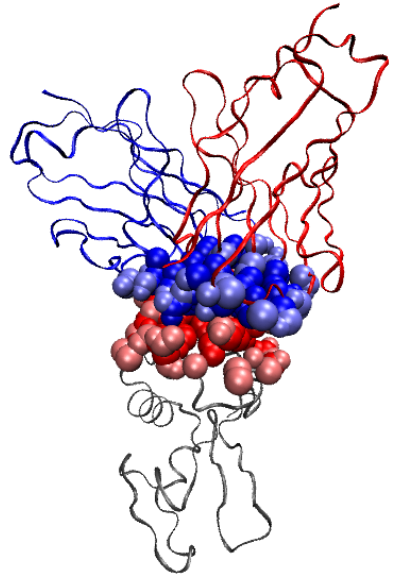
Algorithms - Biology - Structure

Frederic.Cazals@inria.fr

http://team.inria.fr/abs



informatics mathematics

ABS

Algorithms - Biology - Structure: Team and Vision

Modeling high-resolution structures

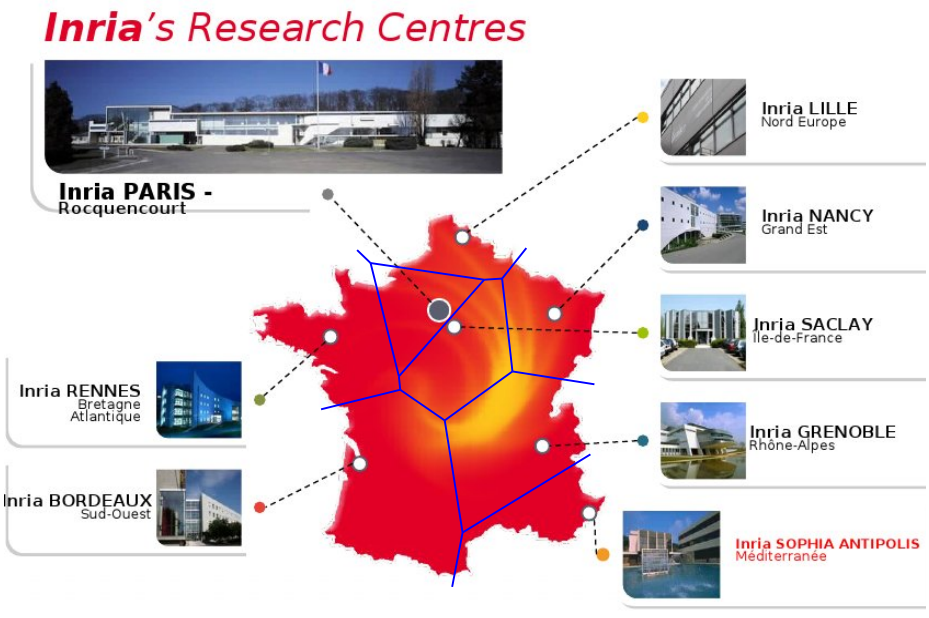
Modeling the flexibility of macro-molecules

Modeling large assemblies

Software

Research directions

Algorithms - Biology - Structure



▷ History
- Team created : July 2007

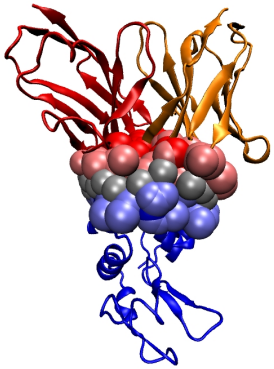
▷ Composition
- Permanent: D. Mazauric, F. Cazals
- (part time) Engineer: T. Dreyfus

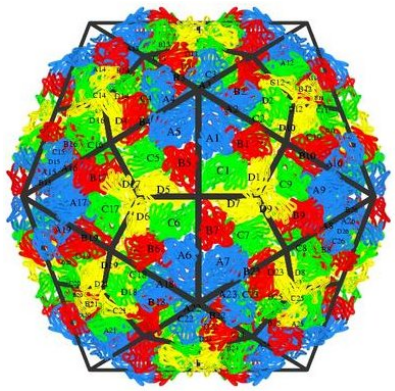
- PhD students
  A. Chevallier (Energy landscapes)
  R. Tetley (Structural alignments)
  D. Bulavka (Collective coordinates)
  M. Simsir (Modeling drug efflux in cancer)

▷ Graduated over the past 4 years
D. Agarwal: Native mass spectrometry; Harvard med school
A. Lhéritier: Machine learning/Two-sample tests; Amadeus SA
S. Marillet: Modeling antibody-antigen complexes; CHU Poitiers

The structure-to-function relationship

▷ Protein complexes and biological functions





– Understanding the stability and the specificity of macro-molecular interactions

– Exploiting structural information crystallography, NMR, EM, SAXS,...

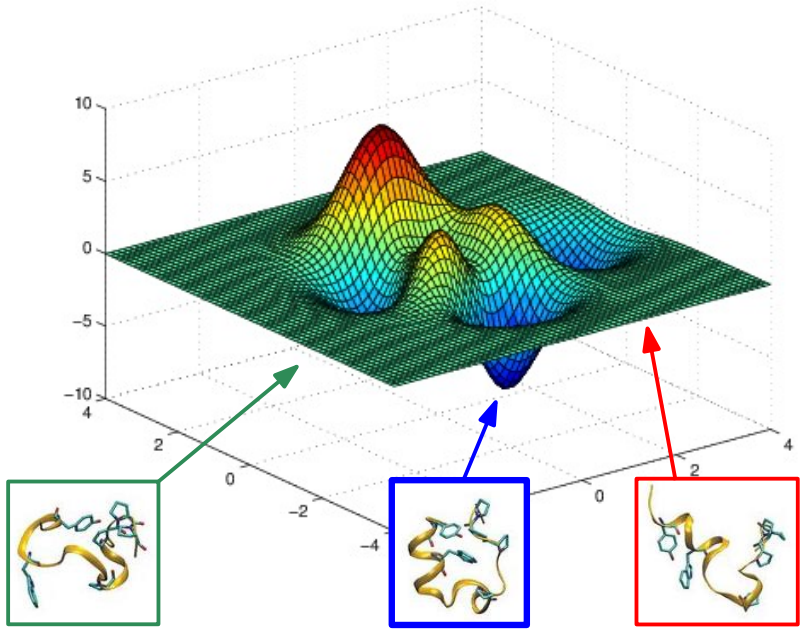
– Performing predictions with little/no structural information using remote homology information

▷ Structural information is scarce

▷Ref: Janin, Bahadur, Chakrabarti; Quart. reviews of biophysics; 2008

▷Ref: Levitt; PNAS 106; 2009

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

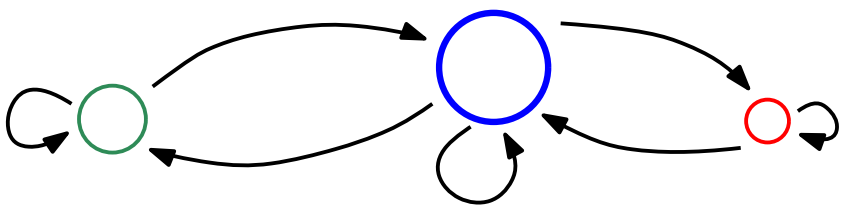


Potential Energy Landscape

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

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

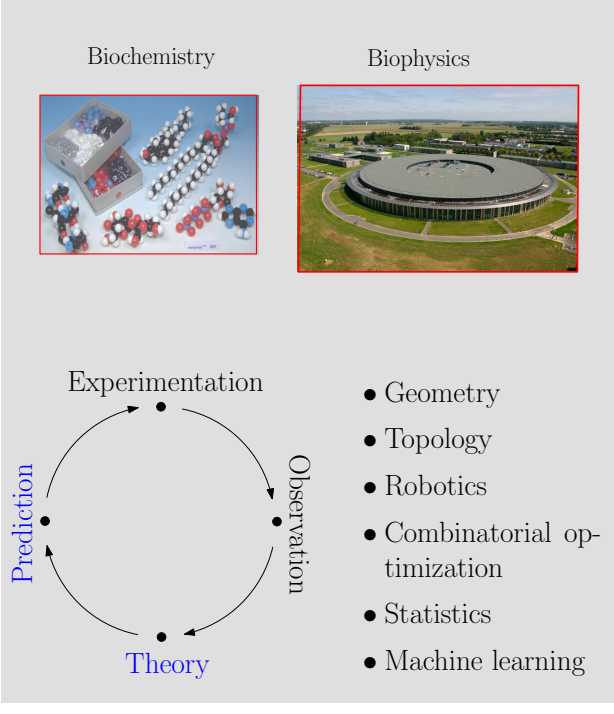
Thermodynamics: meta-stable conformations i.e. ensemble of conformations easily inter-convertible into one - another.



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

Vision: synergy computer science - structural biology

▷ Modeling: leveraging experimental data



▷ Complementary approaches

– Machine learning approaches:
  classification / regression

– Ab initio approaches:
  structure / thermodynamics / dynamics

▷ Work-packages at a glance

– Modeling high-resolution structures
– Modeling large assemblies
– Modeling the flexibility of proteins
– Algorithmic foundations

ABS

Algorithms - Biology - Structure: Team and Vision

Modeling high-resolution structures

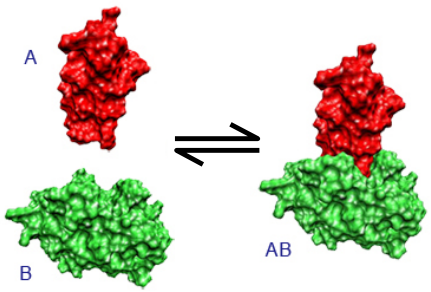
Modeling the flexibility of macro-molecules

Modeling large assemblies

Software

Research directions

Estimating binding affinities



▷ Dissociation constant and dissociation free energy:

$$K_d = [A][B]/[AB]$$
$$\Delta G_d = -RT \ln K_d/c^\circ = \Delta H - T\Delta S.$$

▷ Problem statement: estimate the binding affinity of two partners from

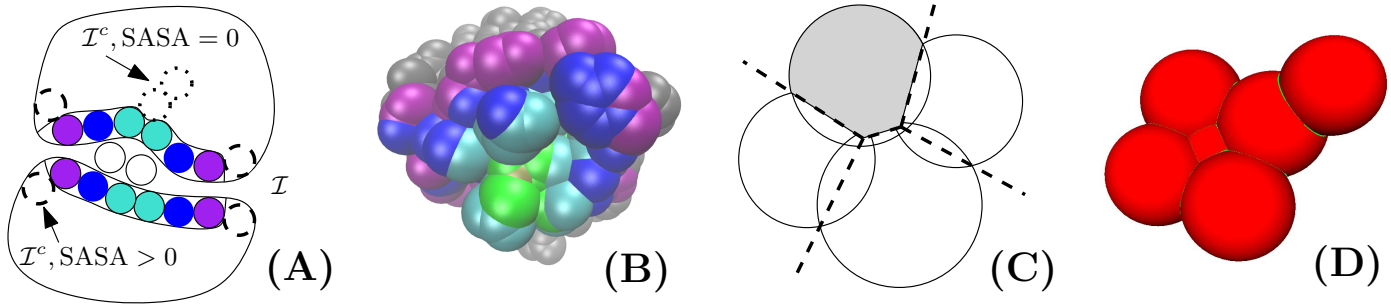
– High resolution crystal structures of partners and complex
– Specific conditions (pH, ionic strength, . . .)
– Key difficulty:
enthalpy - entropy compensation (K_d is of thermodynamic nature)
(!) predictions with $\Delta G_d < 1.4$ kcal/mol are hard

▷ State-of-the-art: numerous approaches

– Knowledge based approaches:
complex models face overfitting; sparse models may be overly restrictive
– Molecular mechanics based approaches:
require specific hypothesis. . . or massive calculations

▷Ref: `Kastritis et al, Protein science, 2011` (the SAB; 144 cases)
▷Ref: `Janin, Protein Science, 2014`

Estimating binding affinities



▷ Contributions: models combining novel parameters and supervised regression

– Novel variables coding enthalpic and entropic variations upon binding
– Model selection procedure based on cross validation
– State-of-the-art binding affinity estimates on the SAB:
  whole SAB: K_d within one and two *OOM* in 48% and 79% of cases
  high resolution (2.5Å): K_d within one and two *OOM* in 62% and 89%

▷ Assessment:

– Sensitivity to the resolution of crystal structures (cf Cruickshank's formula)
– Sensitivity to coverage of model space by learning set (supervised regression)
– Predicting is not explaining

▷Ref: Marillet, Boudinot, Cazals; Proteins 2015
▷Ref: Marillet, Lefranc, Boudinot, Cazals; Frontiers in Immuno., 2017
▷Ref: Vangone and Bonvin, eLIFE,2015

ABS

Algorithms - Biology - Structure: Team and Vision

Modeling high-resolution structures

Modeling the flexibility of macro-molecules

Modeling large assemblies

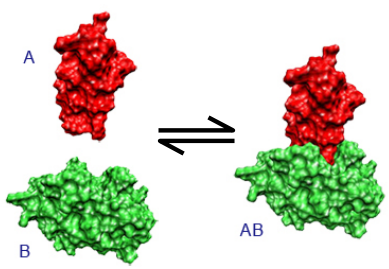
Software

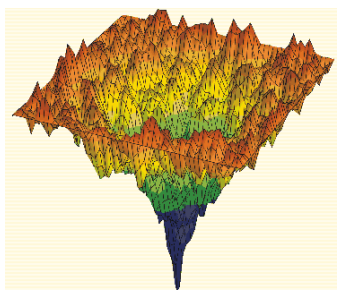
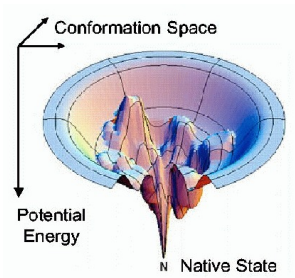
Research directions

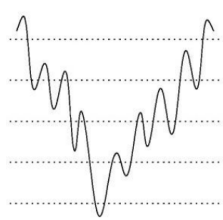
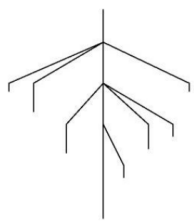
Energy landscapes:
structure – thermodynamics – dynamics

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

- ▷ Three (overlapping) classes of ab initio approaches:
  - Molecular dynamics (including REMD, metadynamics)
    Model reduction: dimensionality reduction (PCA, Isomap, diffusion maps)
  - Monte Carlo methods (MCMC, importance sampling, Wang-Landau)
    Model reduction: Markov state model design via lumping
  - Energy landscapes methods (the basin hopping lineage)
    Model reduction: superposition approach via coarse-graining
- ▷ Bottleneck: massive calculations required



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

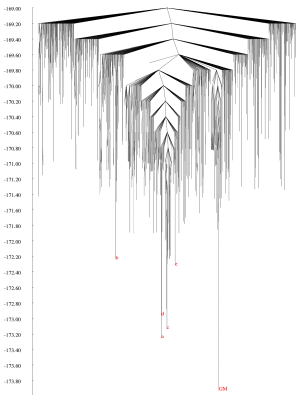
▷Ref: Wales; Energy Landscapes; 2003

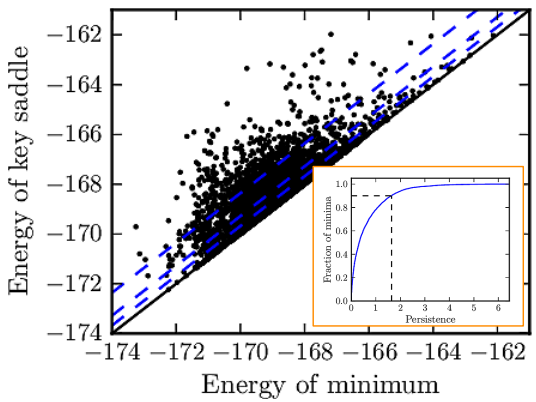
▷Ref: Chipot; Frontiers in free-energy calculations; 2014

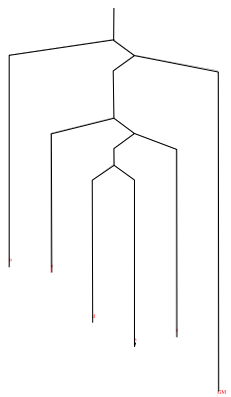
Analysis of sampled energy landscapes

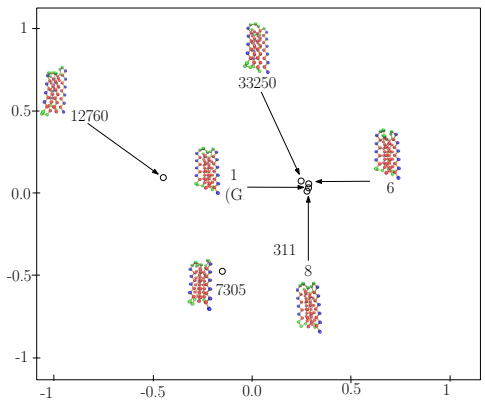
▷ Contributions: novel concepts and algorithms to

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









▷ Assessment:

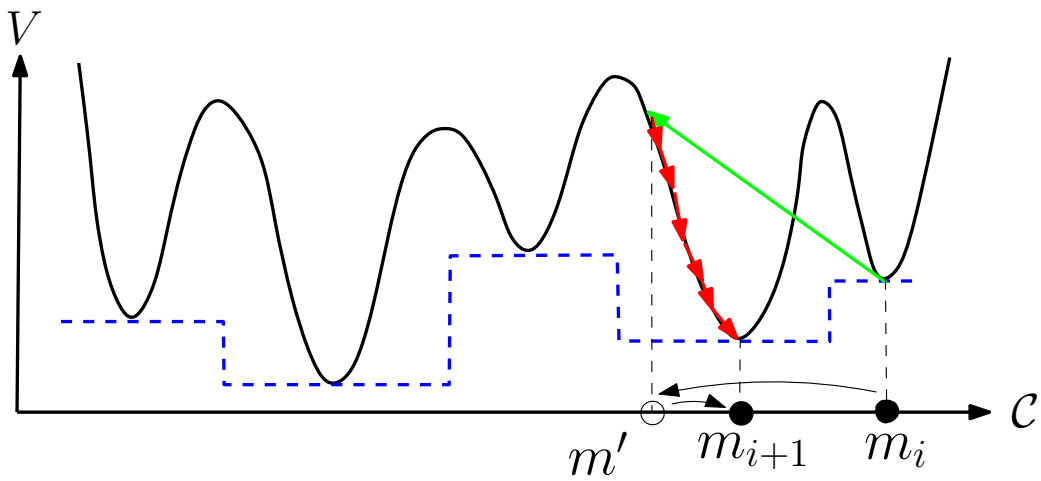
– State-of-the-art algorithms analysis/coarse-graining methods
– Most of the analysis geared towards potential energy landscapes
work ahead on free energy landscapes

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

Exploring Potential Energy Landscapes:

basin hopping

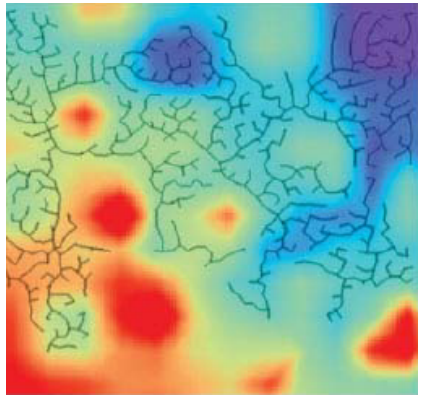
- ▷ Goal: enumerating low energy local minima
- ▷ Basin-hopping and the basin hopping transform
  - – Random walk in the space of local minima
  - – Requires a *move set* and an *acceptance test* (cf Metropolis)
    and the ability to descend the gradient (*quenching*)
    aka energy minizations
- ▷ Limitation: no built-in mechanism to escape traps



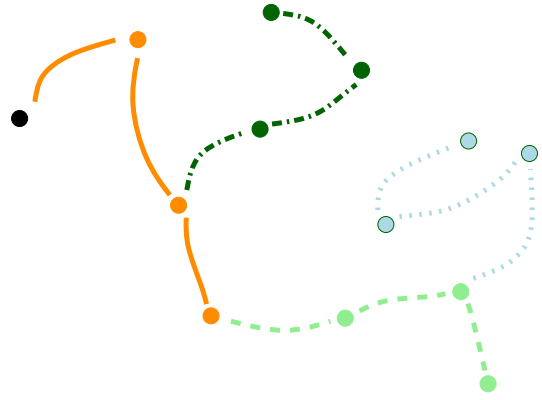
▷Ref: Li and Scheraga, PNAS, 1987

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

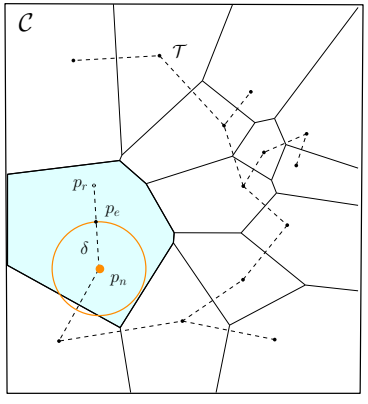
▷ Goal: crawl down the potential energy landscape



▷ Hybrid algorithm: alternate BH and T-RRT extensions



▷ Strategy: force the exploration of *empty space*



▷ Key ingredients:

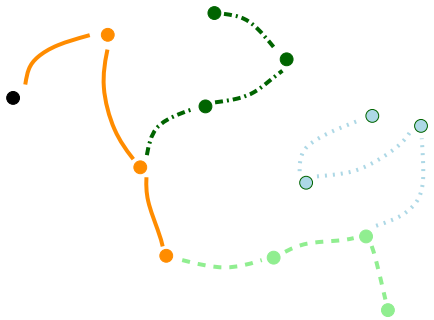
- Boosting the identification of low lying minima with the Voronoi bias
- Favoring spatial adaptation—local exploration parameters
- Handling distances efficiently

▷Ref: Roth, Dreyfus, Robert, Cazals; J. Comp. Chem.; 2016

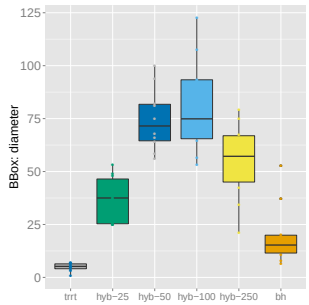
Exploring energy landscapes: performances of `Hybrid`

- ▷ Contributions: enhanced exploration of low lying regions of a complex landscape
- ▷ Protocol: on BLN69, a model protein with 207 d.o.f:
  - Contenders: `BH`, `T-RRT`, `Hybrid` for various parameter values b

• Algorithm

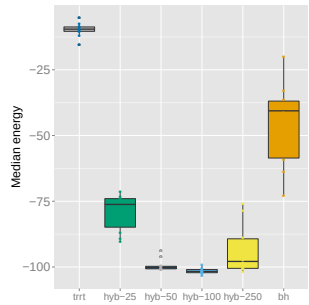


• BBox ∅: low lying mins



$BLN69 - min - E_{-100}$

• Median energies: all mins



$BLN69 - min - all$

- ▷ Assessment:

– PEL exploration:
  – doubled the num. of local mins. (458,082 minima to 1,044,118)
  – explored lower regions of the PEL
– Combines critical building blocks:
  minimization, spatial exploration *boosting*, nearest neighbor searches
– Ongoing: bridging the gap to thermodynamics via DoS calculations

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

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

ABS

Algorithms - Biology - Structure: Team and Vision

Modeling high-resolution structures

Modeling the flexibility of macro-molecules

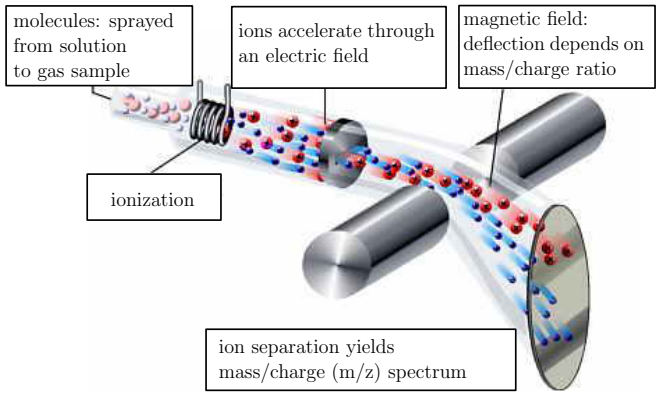
Modeling large assemblies

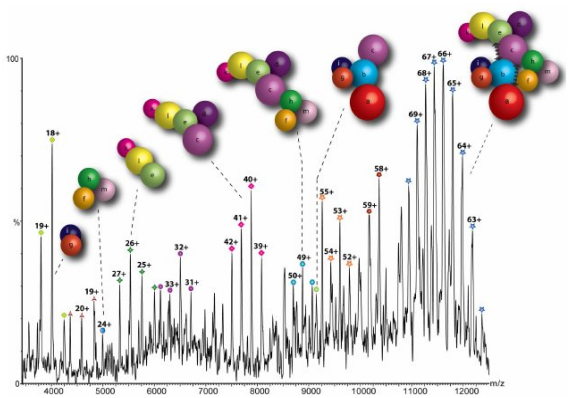
Software

Research directions

Large Assemblies: Native Mass Spectroscopy

▷ Input: mass spectrum of oligomers of a (large) assembly





(1) Disrupting an assembly into oligomers (from sub-units to bigger complexes)
(2) Mass spectrometry yields a m/z spectrum then a mass spectrum
(3) Decomposing an individual mass yields the list of proteins in a sub-complex

▷ Problem: reconstructing pairwise contacts from the composition of oligomers

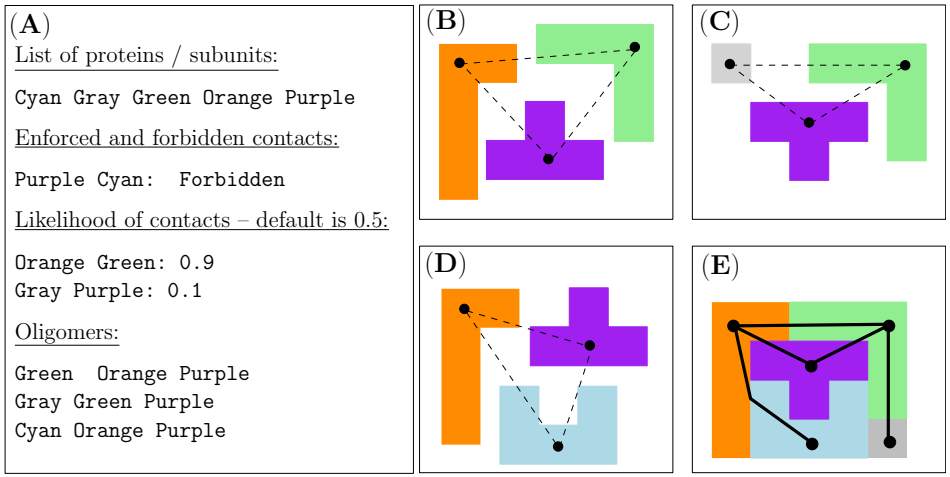
NB: coarse structural information (contacts) from combinatorial information

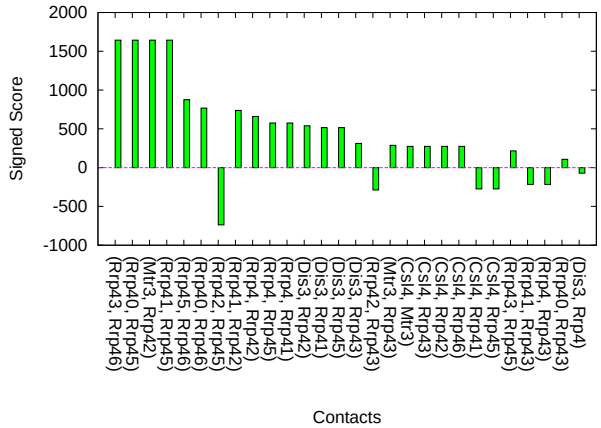
▷ State-of-the-art

– Experiments: recent techniques mastered by few groups (Robinson, Hecht)
– Data analysis: heuristics

▷Ref: Taverner, Robinson et al; Accounts of chemical research; 2008

Native Mass Spectrometry: Connectivity Inference from oligomers





▷ Contributions
- Hardness: problem is NP-complete and APX-hard (**P** $\neq$ **NP**: no PTAS)
- Exact algorithm based on Mixed Integer Linear Programming (MILP)
  $\rightarrow$ generates all solutions for $OPT + k$ ($k \geq 0$)
- Greedy polynomial algorithm with controlled approximation factor:
  $\rightarrow$ $2(\log n + \kappa)$, with κ max. # oligomers of a vertex
- Experiments on four of the biggest systems know to date:
  - more parsimonious solutions (than those of contenders)
  - edges reported in (almost) perfect agreement with known contacts

▷ Assessment: doubled the quality of predictions by contenders

▷Ref: Inria ABS + Inria COATI, European Symp. on Algorithms, 2013
▷Ref: Agarwal, Caillouet, Coudert, Cazals, Molecular and Cellular Proteomics, 2015

Connectivity inference with biophysical constraints

▷ Graph constraints reflecting biophysical and structural biology properties:

- subunit with limited number of neighbors $\rightarrow$ bounded maximum degree
- subunit with known contacts $\rightarrow$ family of admissible subgraphs
- presence of symmetries $\rightarrow$ symmetries of admissible graphs

▷ Generalized inference as minimum $\mathcal{F}$-Overlay: given a graph family $\mathcal{F}$:
Input: a hypergraph $H = (V, \mathcal{E})$ – with $\mathcal{E}$ the oligomers
Output: a graph $G = (V, E)$ with minimum $\mid E(G) \mid$ such that:

- $\forall S \in \mathcal{E}$: induced graph $G[S]$ has a spanning subgraph in $\mathcal{F}$

NB: $\mathcal{F} \equiv$ all trees $\Leftrightarrow$ $G[S]$ is connected $\Leftrightarrow$ previous inference problem

▷ Our results:

- Complexity dichotomy: for every $\mathcal{F}$, we can tell whether Minimum $\mathcal{F}$-Overlay is Polynomial or NP-complete.
- Parameterized algorithms: for almost every $\mathcal{F}$ for which the problem is NP-complete, we can tell whether the problem is FPT or W[1]-hard.

▷Ref: D. Mazauric et al, IWOCA 2017

ABS

Algorithms - Biology - Structure: Team and Vision

Modeling high-resolution structures

Modeling the flexibility of macro-molecules

Modeling large assemblies

Software

Research directions

The Structural Bioinformatics Library

http://sbl.inria.fr

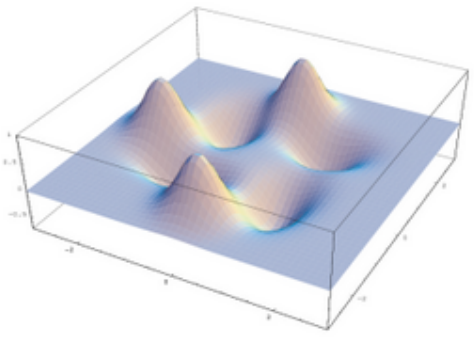
HOME WHAT IS THE SBL ? APPLICATIONS GETTING THE SBL DOCUMENTATION SBL COMMUNITY F.A.Q

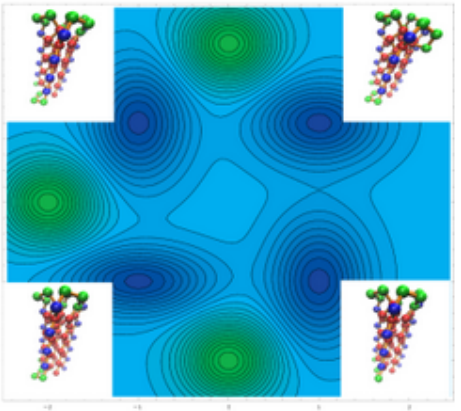
Structural Bioinformatics Library

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

Conformational analysis:

modeling energy landscapes





Why adopt the SBL ?

For Biologists:

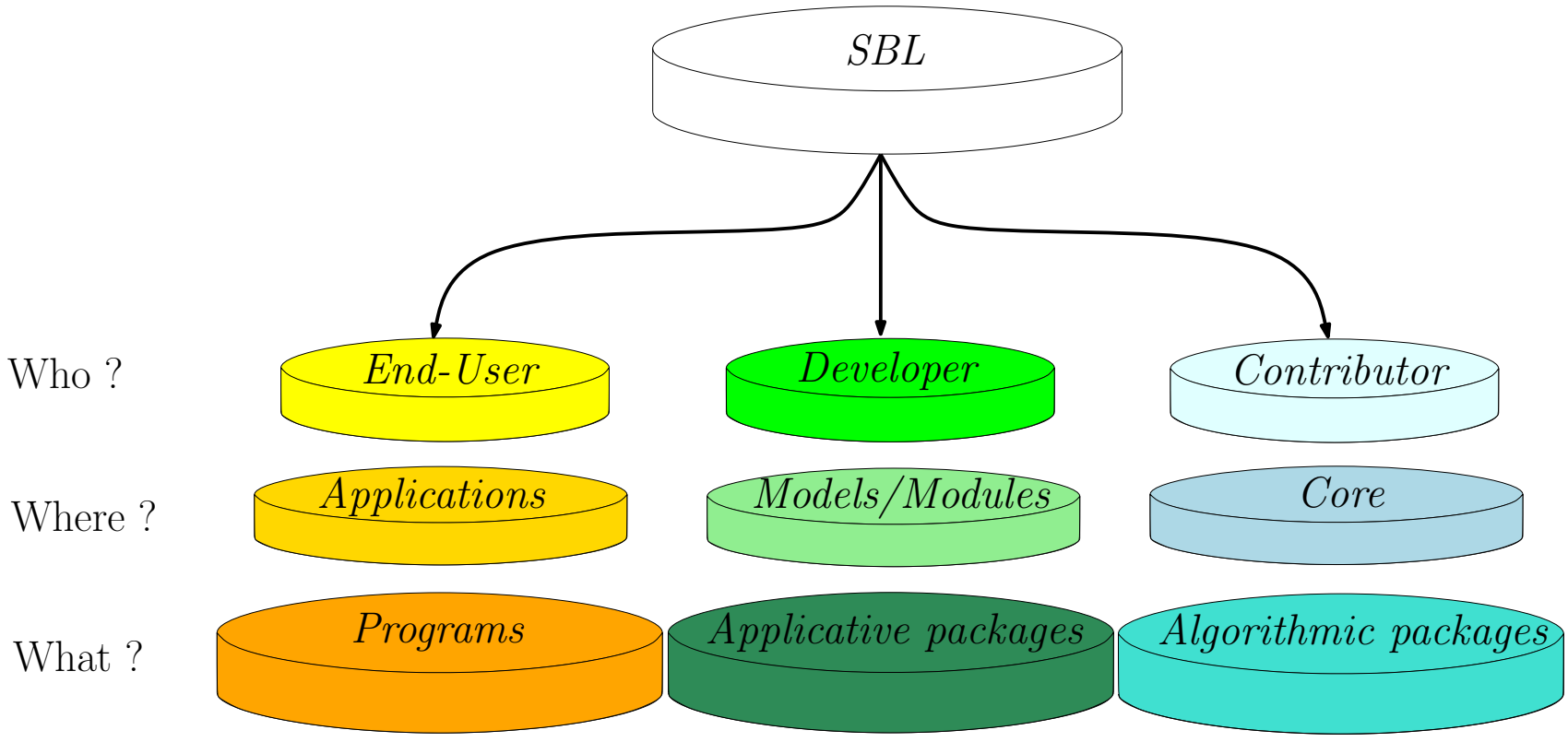
- comprehensive in silico environment providing applications,
- answering complex bio-physical problems,
- in a robust, fast and reproducible way.

For Developers:

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

▷Ref: Cazals and Dreyfus; Bioinformatics, 2017

The Structural Bioinformatics Library: Architecture



ABS

Algorithms - Biology - Structure: Team and Vision

Modeling high-resolution structures

Modeling the flexibility of macro-molecules

Modeling large assemblies

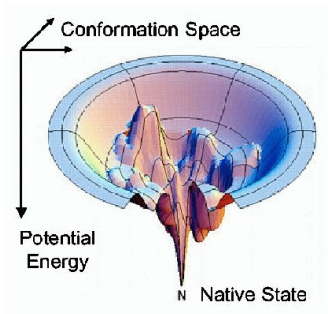
Software

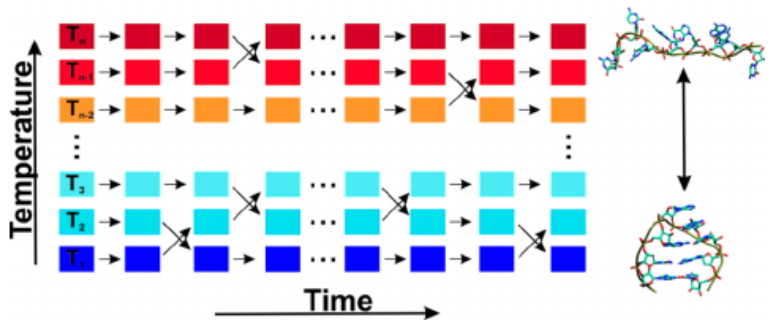
Research directions

Modeling the flexibility of macro-molecules

- Task: Enhanced sampling algorithms
  - Application(s): energy landscape exploration
  - Punchline: atomic move sets (correlated moves, collective coordinates)

- Task: Enhanced thermodynamics sampling algorithms
  - Application(s): thermodynamic sampling
  - Punchline: multi-canonical sampling DoS calculations (adaptive Wang-Landau)



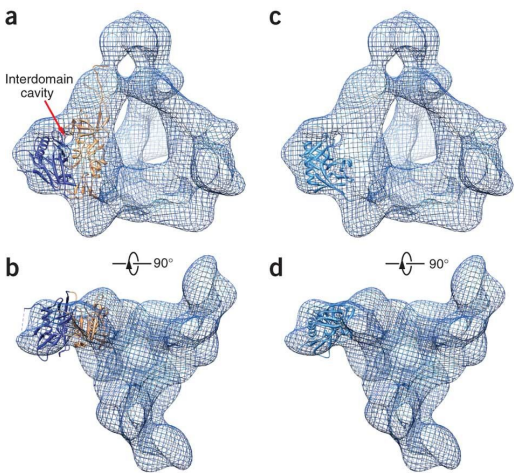


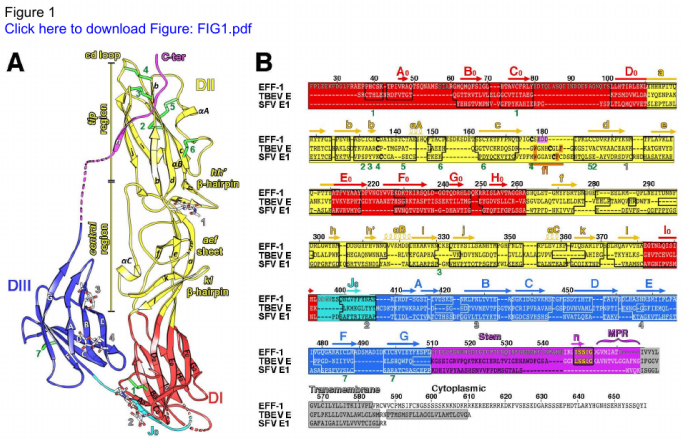
Modeling large assemblies

Modeling high-resolution structures

- ▷ Task: Reconstruction in integrative modeling (Xtallography, cryo-EM)
  - Punchline: continuous dynamic programming, enumerative algorithms
- ▷ Task: Enhanced functional annotations of proteins in sequence - structure studies
  - Punchline: probabilistic sequence HMM, biased with structural information
- ▷ Task: Towards understanding dynamics mechanisms
  - Punchline: identifying meta-stable states of dynamic molecular machines

    Example: class II fusion proteins, functions of the influenza polymerase





Algorithmic foundations

Software

- Task: Density of states calculations – high dimensional integration
  - Punchline: improve (polytope) volume calculations / Wang-Landau sampling
- Task: Graph algorithms techniques for structural biology
  - Punchline: graph decompositions, algorithms, guarantees
- Task: Structural Bioinformatics Library
  - Punchline: continue development + leverage the impact