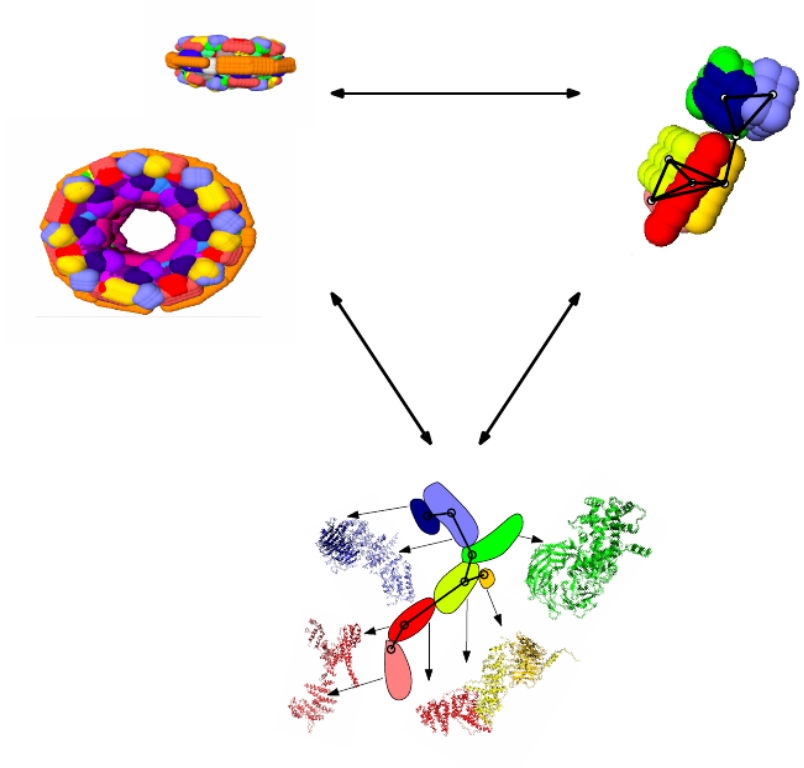
Modeling Contacts in Macro-molecular assemblies: from Inference to Assessment

Frederic.Cazals@inria.fr



Overview

PART 1:Connectivity Inference from Native Mass Spectrometry Data

PART 2:Building Coarse Grain Models

PART 3:Handling uncertainties in Macro-molecular Assembly Models

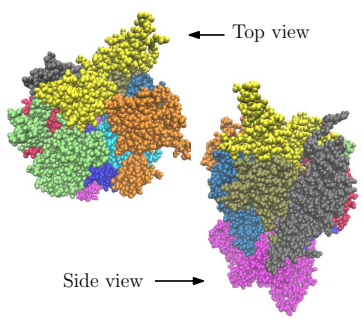
PART 4:Conformational Ensembles and Energy Landscapes: Analysis

PART 5:Conformational Ensembles and Energy Landscapes: Comparison

Connectivity Inference in Mass Spectrometry based Structure Determination

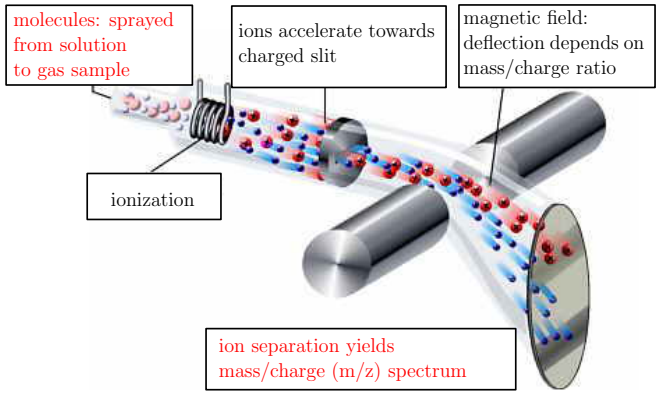
D. Agarwal and J. Araujo and C. Caillouet
and F. Cazals and D. Coudert and S. Pérennes

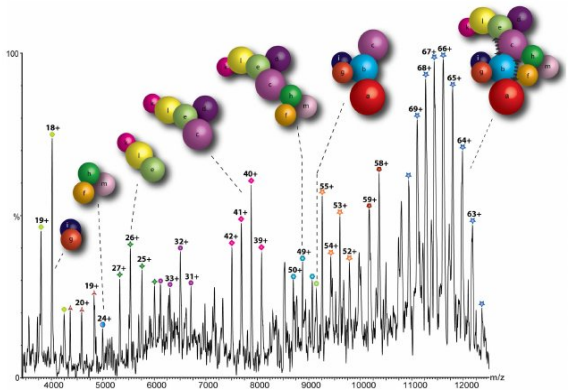
Algorithms-Biology-Structure, Inria Sophia
http://team.inria.fr/abs
COATI, Inria and Univ. Nice Sophia Antipolis and CNRS
http://team.inria.fr/coati



Modeling Contacts in Macro-molecular Assemblies

Mass Spectroscopy of Protein Complexes: 101





▷ Analyzing a mixture of sub-complexes: a three step process

(1) Mass spectrometry yields a m/z spectrum
(2) Processing the m/z spectrum yields a mass spectrum
(3) Decomposing an individual mass yields the list of proteins in a sub-complex

▷ Generating a mixture of sub-complexes by varying the chemical conditions

– Stringent conditions: full decomposition yields isolated proteins
– Milder conditions: overlapping complexes (oligomers)

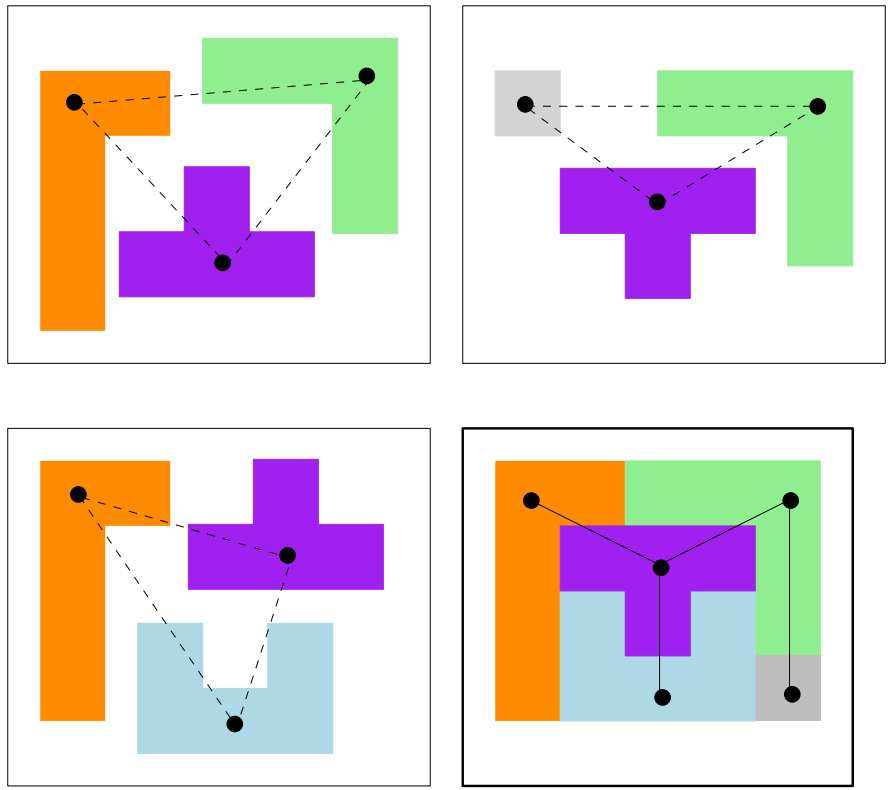
▷Ref: Taverner, Robinson et al; Accounts of chemical research; 2008

Checkpoint

- Consider an oligomer of size 4, involving four different proteins.
- In how many different ways can it be connected?

The Lego Example

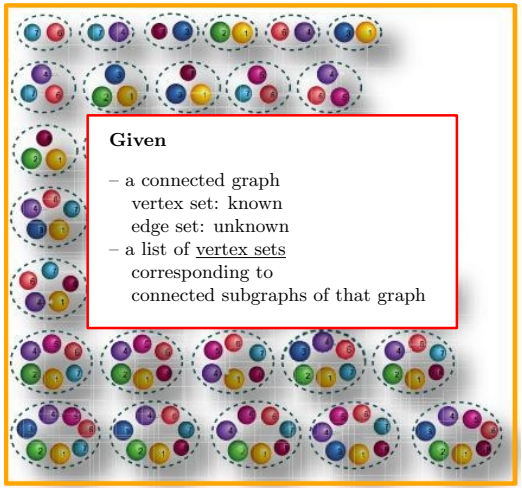
▷ Reconstruction contacts for an assembly of five proteins, given three complexes of size three

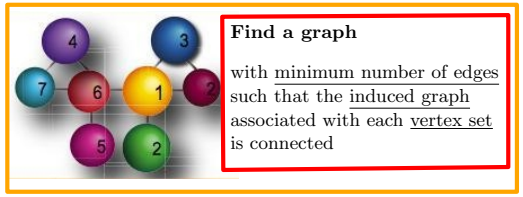


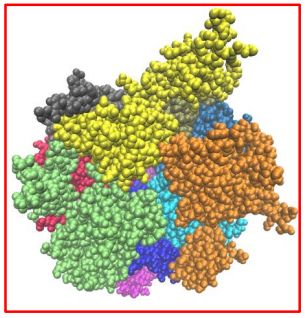
▷ Comments about Minimum connectivity:

- The pool of candidate edges is defined by the oligomers
- MCI yields a well posed problem
- MCI avoids speculating on the number of contacts
- Solutions in general not unique

Minimum Connectivity Inference: Problem Specification







▷ Formal specification:

– Input:

A set V of vertices (Vertex: protein)

A set C of vertex sets $\{V_i \subset V\}, i \in I$ (Vertex set: protein sub-complex)

– Goal: Find a graph $G = (V, E)$, (Edge: protein contact)

with E of minimal cardinality

– Constraints: the induced graph $V_i[E]$ is connected, $\forall i \in I$

▷ NB: edges of the complete graph on V: $\mathcal{E}$

▷ Previous work: Network Inference algorithm by Robinson et al.

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

Modeling Contacts in Macro-molecular Assemblies

Problem Statement

Hardness and Algorithms — Computer Science

Results — Structural Biology

Outlook

Hardness: Overview

▷ Decision version of the Connectivity Inference problem:

Inputs:	Set V of vertices	(proteins)
	Set of subsets $C = \{V_i \mid V_i \subset V \text{ and } i \in I\}$	(complexes)
	Integer $k > 0$	(budget)
Constraints:	Given $G = (V, E)$: the induced graph $G[V_i]$ is connected $\forall i \in I$	
Question:	Does there exist a feasible edge set E such that $\lvert E \rvert \leq k$?	

▷ Using a reduction of the **Set Cover** problem:

- The decision version of the Connectivity Inference problem is NP-complete
- Minimum Connectivity Inference is APX-hard
  $\exists\, \mu > 0$ such that approximating MCI within $1 + \mu$ is NP-hard

Mixed Integer Linear Programming (MILP) Formulation

- Objective function minimizing the number of edges:

$$\forall e \in \mathcal{E},\ \text{consider } y_e \in \mathbb{Z}_2 : \min \sum_{e \in \mathcal{E}} y_e$$

- Formulation uses flow variables on arcs (oriented edges):

$$\forall i \in I \text{ and } u, v \in V :\ f^i_{uv}, f^i_{vu} \in \mathbb{R}^+$$

- Constraints:
  - Connectivity of the ith complex: some $s_i \in V_i$ expels $|V_i| - 1$ units of flow, each other vertex collecting one unit

$$\sum_{a \in A_i^+(u)} f^i_a - \sum_{a \in A_i^-(u)} f^i_a = \begin{cases} |V_i| - 1 & \text{if } u = s_i \\ -1 & \text{if } u \neq s_i \end{cases}$$

  - Arc capacity

$$\left.\begin{array}{l} f^i_{uv} \leq |V_i| \cdot y_{uv} \\ f^i_{vu} \leq |V_i| \cdot y_{uv} \end{array}\right\} \qquad \forall i \in I,\ \forall e = uv \in \mathcal{E}$$

- An edge is selected if one of its two arcs carries some positive flow

MILP: Enumerating all Optimal Solutions

- MILP and decision problem: replace the objective function by $\sum_{e \in \mathcal{E}} y_e \leq k$
- Incremental constraint generation for solution enumeration:
  - E_ℓ is the ℓ-th solution (set of edges)
  - The solution E_ℓ gets excluded when adding the constraint

$$\sum_{e \in E_\ell} y_e \leq |E_\ell| - 1$$

- $\mathcal{S}_{\text{MILP}}$: ensemble of optimal solutions reported by MILP

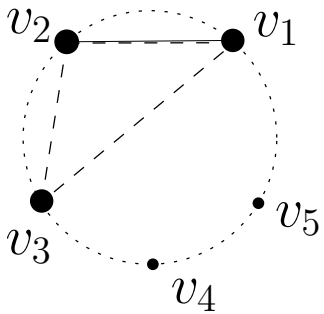
while MILP has a feasible solution E_ℓ s.t. $|E_\ell| \leq \text{OPT}$ **do**
  Add E_ℓ to $\mathcal{S}_{\text{MILP}}$
  Add constraint $\sum_{e \in E_\ell} y_e \leq |E_\ell| - 1$ to MILP
return $\mathcal{S}_{\text{MILP}}$

- NB: can also be used to report all solutions with at most k edges

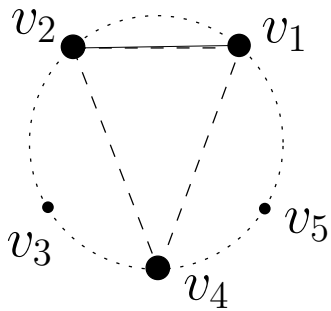
Approximation Strategy: Greedy Algorithm

- Greedy: iteratively pick the edge best at reducing
  the number of connected components, across all complexes
  → priority of edge e: # of c.c. merged upon picking e

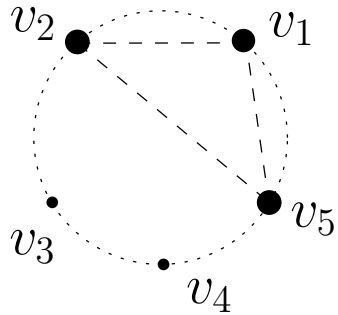
Complex #1



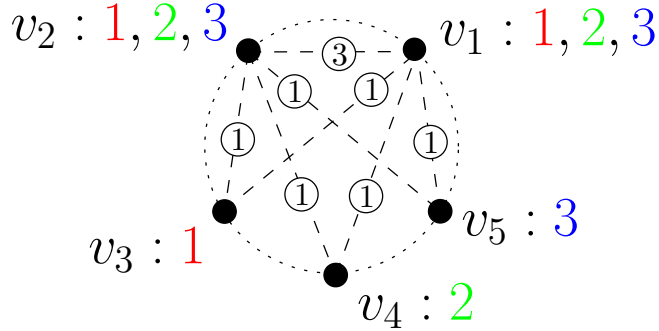
Complex #2



Complex #3



Complexes as colors



- Thm. `Greedy` yields a $2\log_2(\sum_{i\in I}|V_i|)$-approximation
- Implementation: priority queue + Union-Find data structures

  queue: to select the edge with best priority
  union-find data structures: maintaining the disjoint sets

Greedy Analysis (I)

- Notations:
  - Edge set incrementally built: $E^t \subset \mathcal{E}$, with $E^0 = \emptyset$
    yields the graph $G^t = (V, E^t)$
  - Induced graph associated to a complex: $V_i[E^t]$
    # connected components of $V_i[E^t]$: $|V_i[E^t]|$

Definition (Priority of edge e w.r.t. $F \subset \mathcal{E}$)

Number of c.c. that get merged upon selecting e:

$$\text{priority}(e, F) = \sum_{i \in I} |V_i[F]| - \sum_{i \in I} |V_i[F \cup \{e\}]|$$

- Trivial fact : The priority of an edge decreases along time.

$$OPT \geq \frac{\sum_{i \in I} |V_i[\emptyset]|}{Max_{e \in E} \text{priority}(e, \emptyset)}$$

Lemma

$$\forall F \subset \mathcal{E} : OPT \geq \frac{\sum_{i \in I} |V_i[F]|}{Max_{e \in E} priority(e, F)}$$

Greedy Analysis (II)

- Edge selected matches the best priority i.e.

$$e_{max}(t) = \max_{e \in \mathcal{E}} \text{priority}(e, E^t)$$

- Phase: sequence of steps $t, t+1, \ldots, t'$ with $e_{max}(t') \geq \frac{1}{2} e_{max}(t)$

- During a phase :
  - We merge at least $\frac{1}{2} e_{max}(t) \times (t' - t)$ components.

    This yields the following lower bound on the # of c.c. at time t:
    $\Longrightarrow \sum_{i \in I} |V_i(E^t)| \geq \frac{1}{2} e_{max}(t) \times (t' - t)$
  - And by the previous lemma: $OPT \geq \frac{1}{2}(t' - t)$

During a phase we pay at most twice the optimal

- Priority is halved at each phase: $\#\text{phases} \leq \mathbf{log}_2(\sum_{i \in I} |V_i|)$

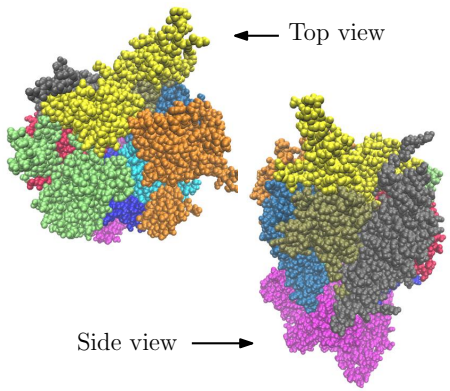
$\Longrightarrow \mathbf{2} log_2(\sum_{i \in I} |V_i|)$ approximation

Modeling Contacts in Macro-molecular Assemblies

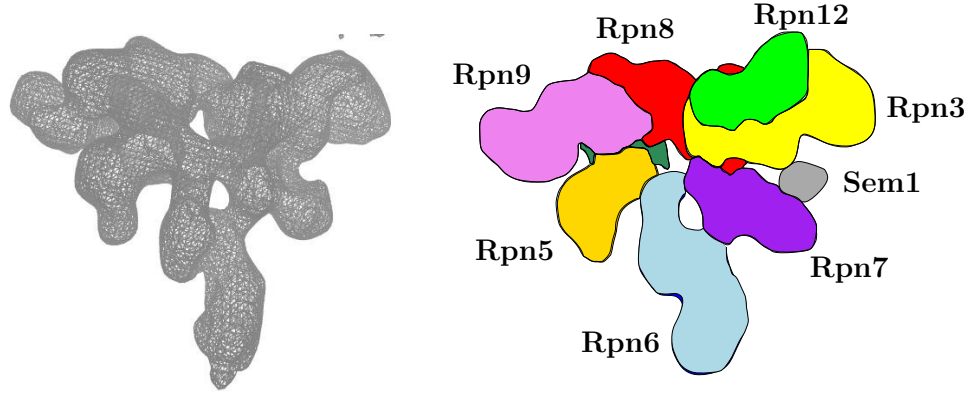
Example Complexes Under Scrutiny

- Yeast exosome
  - exonuclease complex involved in RNA processing and degradation
  - 10 distinct proteins: RNA processing and degradation
  - Input from mass spectrometry: 21 vertex sets
- Yeast 19S proteasome lid
  - Proteasomes: elimination of damaged / misfolded / short-lived proteins
  - 9 distinct proteins: degradation of damaged or misfolded proteins
  - Input from mass spectrometry: 14 vertex sets

- Yeast exosome: crystal structure



- Proteasome lid: cryo EM map



Assessing a Solution Set:

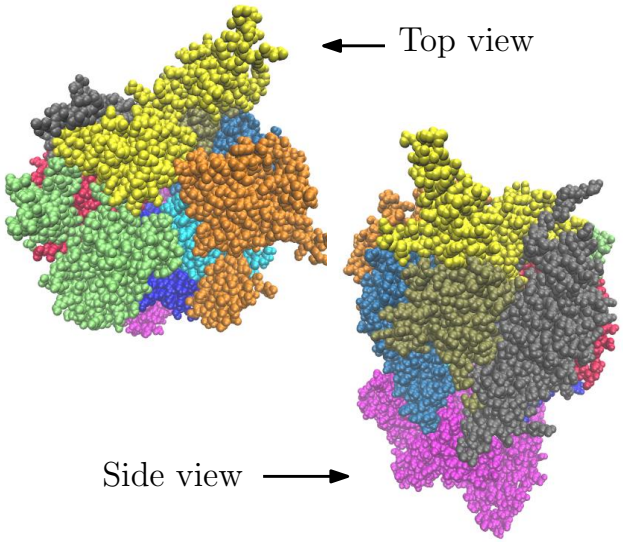
Comparing predicted edges versus experimentally observed protein contacts

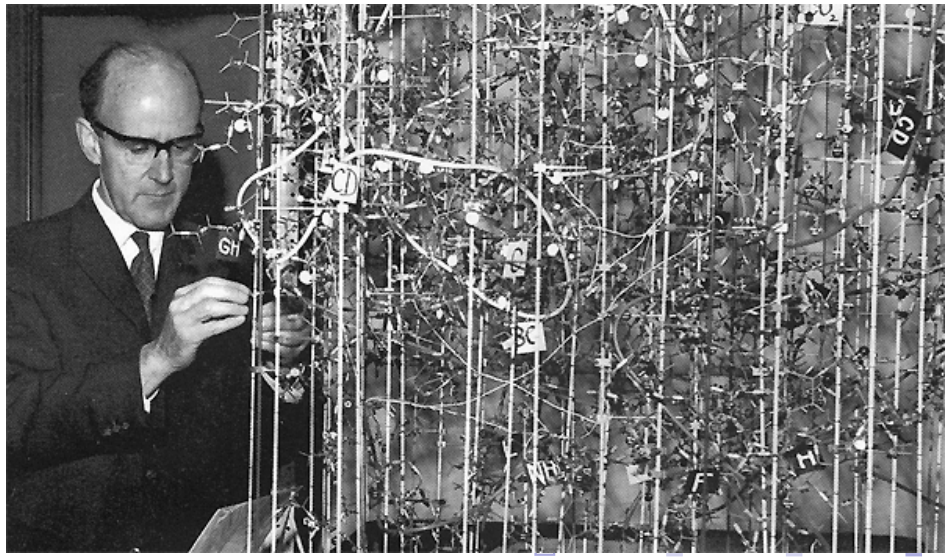
- Consider a contact (v_i, v_j) from solution $S \in \mathcal{S}_{\text{MILP}}$: true or false positive?

  $\rightarrow$ assessing a contact requires an exhaustive - reference set of contacts E_{Ref}

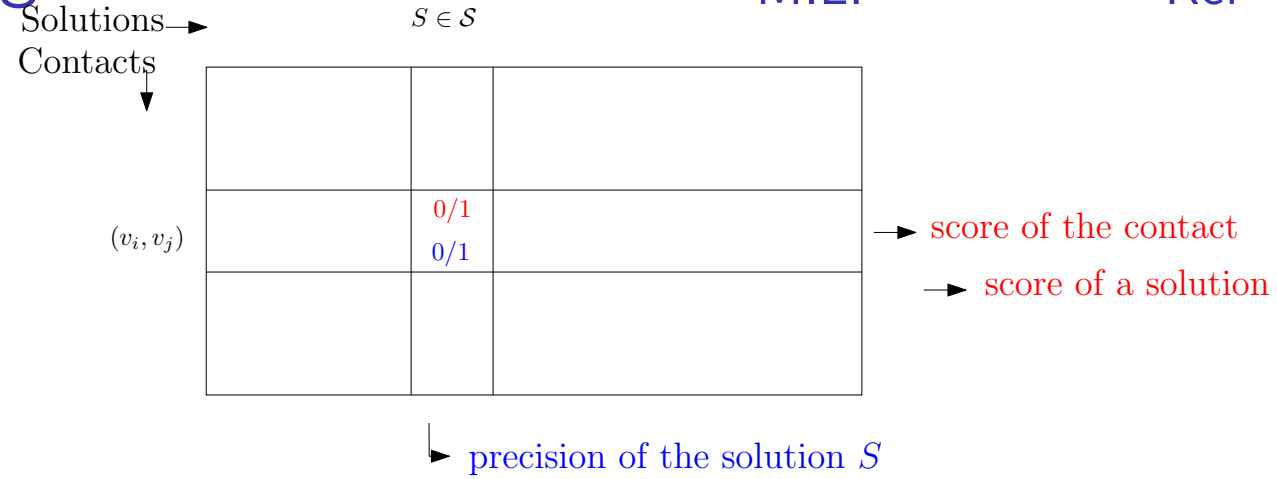
- Reference contact sets from various experiments

[Crystallography]	C_{Xtal}
[Bio-chemistry]	C_{Dim}: (TAP, etc)
[Cross-linking]	C_{XL}
[Combined]	$C_{\text{Xtal}} \cup C_{\text{Dim}} \cup C_{\text{XL}}$





Assessing a Solution Set $\mathcal{S} \subset \mathcal{S}_{\text{MILP}}$ w.r.t. E_{Ref}



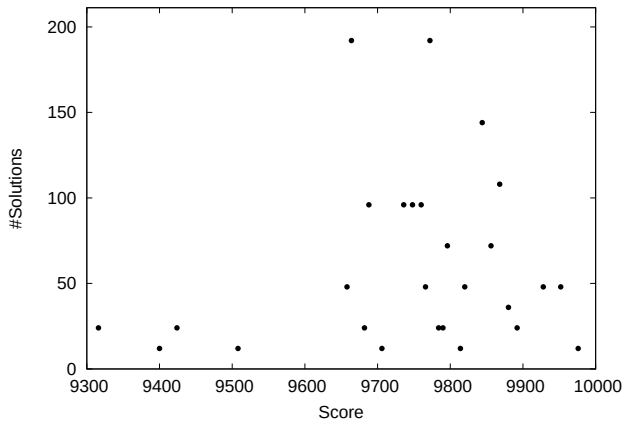
- Precision with respect to the reference set of contacts E_{Ref}
  - precision of solution $S \in \mathcal{S}$ wrt E_{Ref}: $P_{\text{MILP};\text{E}_{\text{Ref}}}(S) = |S \cap \text{E}_{\text{Ref}}|$
    $\rightarrow$ precision is maximum if $S \subset \text{E}_{\text{Ref}}$ i.e. no false positive
  - precision $P_{\text{MILP};\text{E}_{\text{Ref}}}(\mathcal{S})$ of an ensemble of solutions $\mathcal{S}$:
    (min, median,max) of the precisions of the solutions $S \in \mathcal{S}$

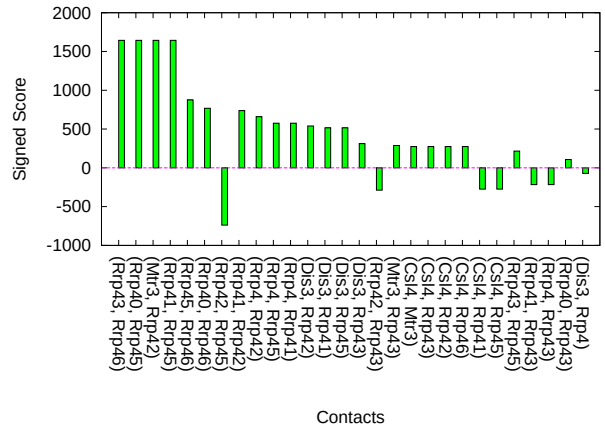
- Scores for contacts and solutions
  - score of a contact: # solutions from $\mathcal{S}$ it belongs to
  - signed score of contact: score $\times \pm 1$ depending on whether true/false positive

- Scores for contacts and consensus solutions:
  - score of a solution $S \in \mathcal{S}$: the sum of the scores of its contacts
  - consensus solutions $\mathcal{S}^{cons.}_{\text{MILP}}$: solutions achieving the maximum score

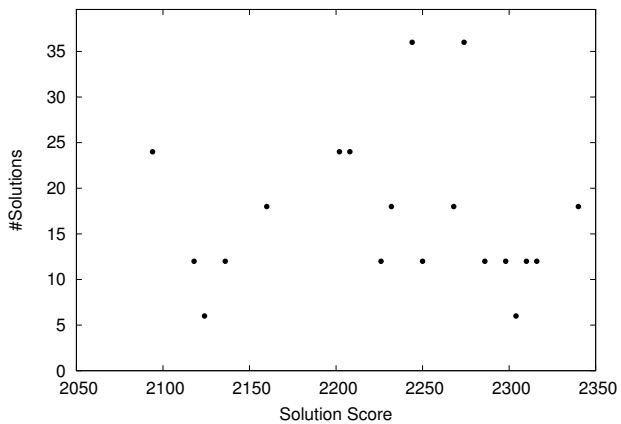
Signed Scores for Contacts and Solutions in $\mathcal{S}_{\mathrm{MILP}}$

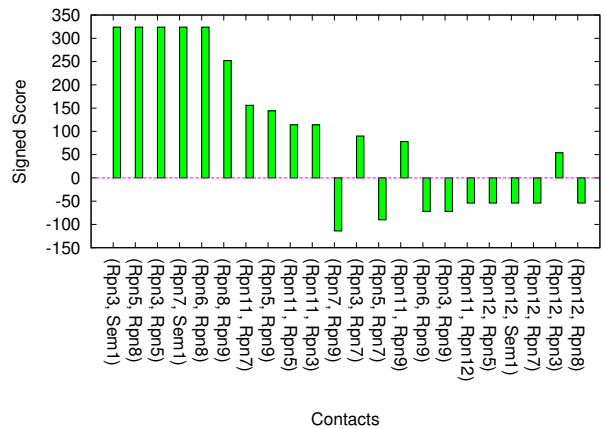
- Exosome ($E_{Ref} = C_{Xtal}$): scores for solutions and signed contact scores





- Proteasome (E_{Ref}): signed contact scores, and scores for solutions





- Take-home message: very few false positives ... and yet for good reasons.

Parsimony and Precision for Individual Solutions in $\mathcal{S}_{\text{MILP}}$:

Yeast Exosome

▷ Algorithm `NI` : *genetic algorithm* by Robinson et al.

Complex	#types	E_{Ref}	$\lvert E_{Ref}\rvert$	$\lvert S_{NI}\rvert$	$P_{NI;E_{Ref}}(S_{NI})$
Exosome	10	C_{Xtal}	26	12	12
19*S Lid*	9	$C_{Cryo} \cup C_{Dim} \cup C_{XL}$	19	9 (*NC**)	8
eIF3	12	$C_{Cryo} \cup C_{Dim} \cup C_{XL}$	17	17**	14

▷ `MILP`

Complex	#types	E_{Ref}	$\lvert E_{Ref}\rvert$	$\lvert S_{MILP}\rvert$	$\lvert\mathcal{S}_{MILP}\rvert$	$P_{MILP;E_{Ref}}(\mathcal{S}_{MILP})$	$\lvert\mathcal{S}^{cons.}_{MILP}\rvert$	$P_{MILP;E_{Ref}}(\mathcal{S}^{cons.}_{MILP})$
Exosome	10	C_{Xtal}	26	10	1644	(7, 9, 10)	12	(8, 9, 10)
19*S Lid*	9	$C_{Cryo} \cup C_{Dim} \cup C_{XL}$	19	10	324	(7, 8, 10)	18	(8, 9, 10)
eIF3	12	$C_{Cryo} \cup C_{Dim} \cup C_{XL}$	17	13	180	(8, 10, 12)	36	(9, 10, 11)

▷ `Greedy`

Complex	#types	E_{Ref}	$\lvert E_{Ref}\rvert$	$\lvert S_{G}\rvert$	$\lvert\mathcal{S}_{Greedy}\rvert$	$P_{Greedy;E_{Ref}}(\mathcal{S}_{Greedy})$	$\lvert\mathcal{S}^{cons.}_{Greedy}\rvert$	$P_{Greedy;E_{Ref}}(\mathcal{S}^{cons.}_{Greedy})$
Exosome	10	C_{Xtal}	26	10	756	(7, 9, 10)	756	(7, 9, 10)
19*S Lid*	9	$C_{Cryo} \cup C_{Dim} \cup C_{XL}$	19	10	324	(7, 8, 10)	18	(8, 9, 10)
eIF3	12	$C_{Cryo} \cup C_{Dim} \cup C_{XL}$	17	13	108	(9, 10, 12)	36	(9, 10, 11)

▷ Take-home message:

– MILP is more parsimonious than NI
– more than 80% of edges in consensus solutions: true positives

Precision for the Union of Solutions in $\mathcal{S}_{\mathrm{MILP}}$

- For each protein: union of neighborhood versus contacts in the assembly
- Symmetric difference between two sets S and R:

$$S\Delta_s R = (|S\backslash R|, |S \cap R|, |R\backslash S|). \tag{1}$$

- Applied to the union of neighborhoods vs reference contacts:

$$N(p, \mathcal{S}_{\mathrm{A}})\Delta_s N(p, R) \equiv (\bigcup_{S \in \mathcal{S}_{\mathrm{A}}} N(p, S))\Delta_s N(p, R) \tag{2}$$

- Results (false positives, true positives, missed contacts)

Protein	Ref. Degree	$N(p, \mathcal{S})\Delta_s N(p, R)$
Dis3	4	(1, 4, 0)
Rrp4	5	(2, 3, 2)
Rrp43	6	(3, 6, 0)
Rrp45	7	(2, 6, 1)
Rrp46	5	(0, 4, 1)
Rrp41	4	(2, 4, 0)
Rrp40	4	(0, 3, 1)
Csl4	6	(2, 4, 2)
Rrp42	5	(2, 5, 0)
Mtr3	6	(0, 3, 3)

Modeling Contacts in Macro-molecular Assemblies

Outlook

- Structural Biology
  - Mass spec. for protein complexes: about to revolutionize structural biology
    $\rightarrow$ reference algorithms for connectivity inference
  - Excellent agreement with experimental data
  - Solutions more parsimonious than previously computed ones
  - For current examples: `MILP` always succeeds
  - Software: about to be released (`MILP` , `Greedy`)

- Computer science: selected open questions
  - MILP has a hard time to outperform Greedy: is the approx. factor tight?
  - Structure of the solution set depending on
    structural properties of the unknown graph (min cuts)
    structure of the Hasse diagram of vertex sets (*hierarchical* vs *flat*)
  - Problem size: moving from ~ 10 to ≤ 500 vertices
    multiplicity issues appear : multiples copies per protein
  - Beyond topological information: 3D embedding of the solutions?
    minimum connectivity, degree of nodes

References

- Connectivity Inference in Mass Spectrometry based Structure Determination D. Agarwal, and J. Araujo, and C. Caillouet, and F. Cazals, and D. Coudert, and S. Perennes European Symposium on Algorithms (LNCS 8125), 2013
- Unveiling Contacts within Macro-molecular assemblies by solving Minimum Weight Connectivity Inference Problems D. Agarwal, and C. Caillouet, and F. Cazals, and D. Coudert submitted, 2014

Overview

PART 1:Connectivity Inference from Native Mass Spectrometry Data

PART 2:Building Coarse Grain Models

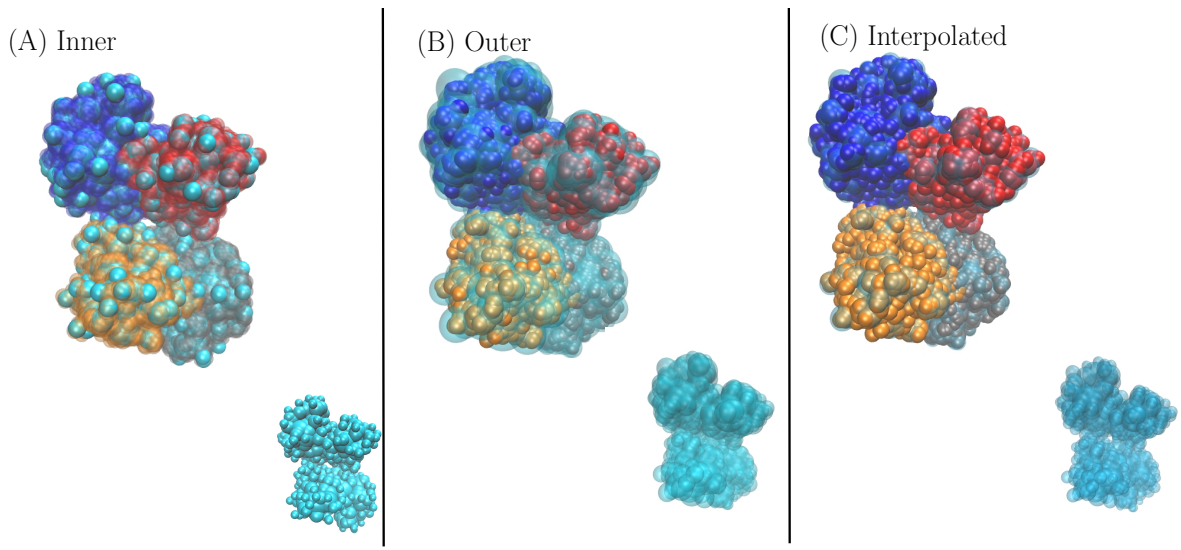
PART 3:Handling uncertainties in Macro-molecular Assembly Models

PART 4:Conformational Ensembles and Energy Landscapes: Analysis

PART 5:Conformational Ensembles and Energy Landscapes: Comparison

Greedy Geometric Algorithms for Collections of Balls, with Applications to Geometric Approximation and Molecular Coarse-Graining

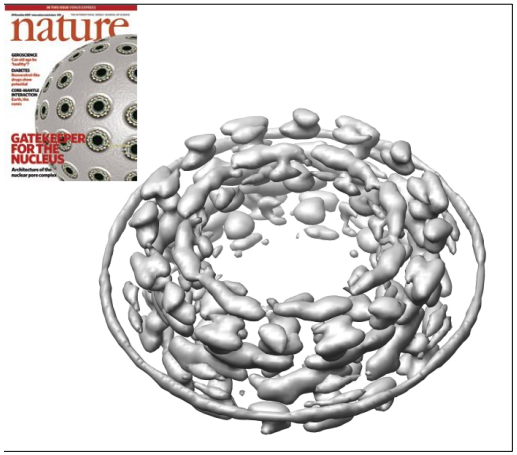
F. Cazals and T. Dreyfus and S. Sachdeva and N. Shah



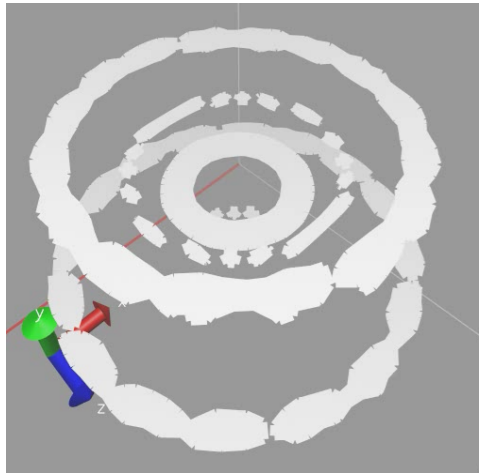
Modeling Contacts in Macro-molecular Assemblies

Separating the Molecules: Finding (Thick) Cracks Within a Map

- NPC: probability density maps



- Cryo-EM density maps



- Antelope canyon, AZ, USA

Checkpoint

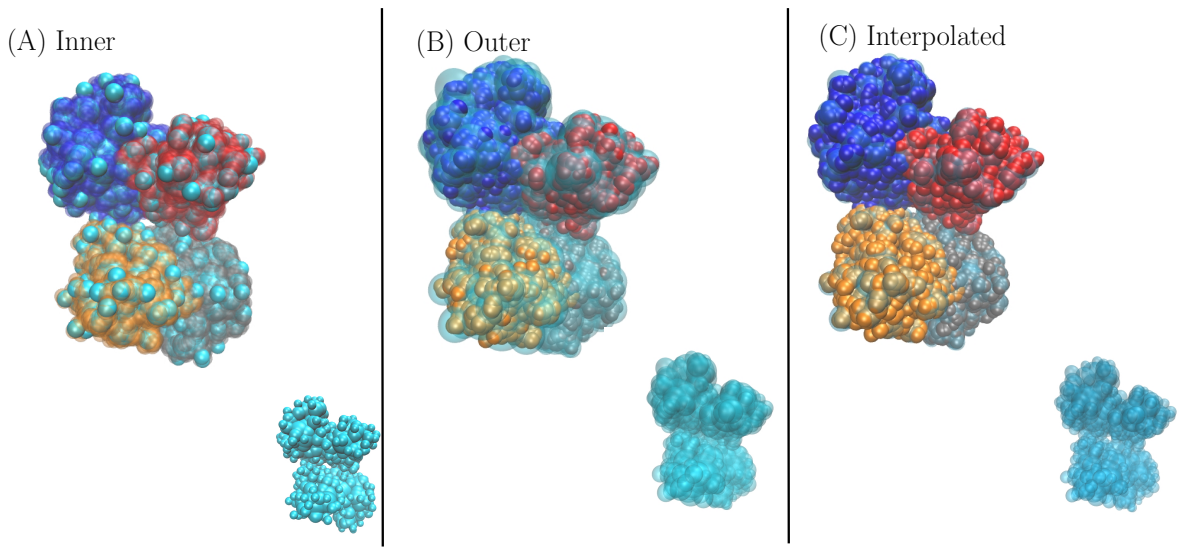
▷ Consider a planar domain D defined by a simple curve. To cover domain D with balls, where should these balls be centered?

Coarse Graining with a Fixed Budget of k balls: Overview

- Three approximation problems of a given input shape:
  - inner approximation with largest volume
  - outer approximation with least extra volume
  - volume preserving approximation

- From crystal structure: inner / outer / interpolated approximations

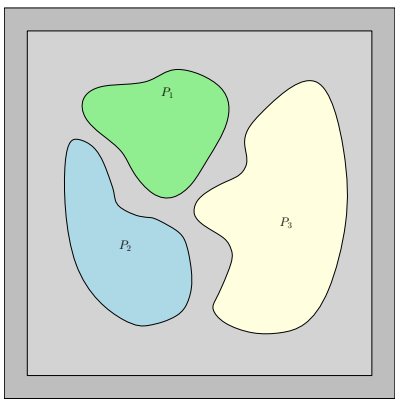
  3sgb (1690 atoms), approximated with 85 balls (5% of atoms)

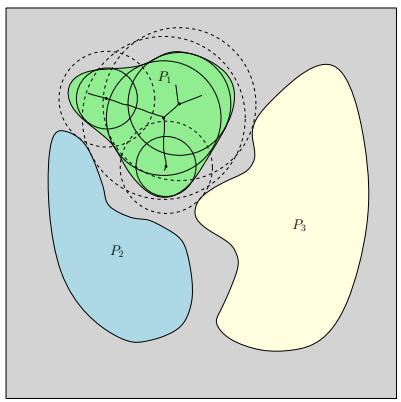


- NB: weighted versions accommodated too

Coarse Graining with a Fixed Budget of k balls: Problems

- Input: $\mathcal{F}_{\mathcal{O}}$ defined by a union of n balls
- Output: $k < n$ balls defining the approximation $\mathcal{F}_{\mathcal{S}}$
- Three problems:
  - *inner approximation*: $\mathcal{F}_{\mathcal{S}} \subset \mathcal{F}_{\mathcal{O}}$
  - *outer approximation*: $\mathcal{F}_{\mathcal{O}} \subset \mathcal{F}_{\mathcal{S}}$
  - *interpolated approximation*: an approximation sandwiched between the inner and outer approximations.
    - *Volume preserving approximation:* $\mathrm{Vol}(\mathcal{F}_{\mathcal{S}}) = \mathrm{Vol}(\mathcal{F}_{\mathcal{O}})$





Modeling Contacts in Macro-molecular Assemblies

Greedy Assessment: Volume Covered

Incidence of the Topology

- Input domain versus domain of the selection: volume comparisons

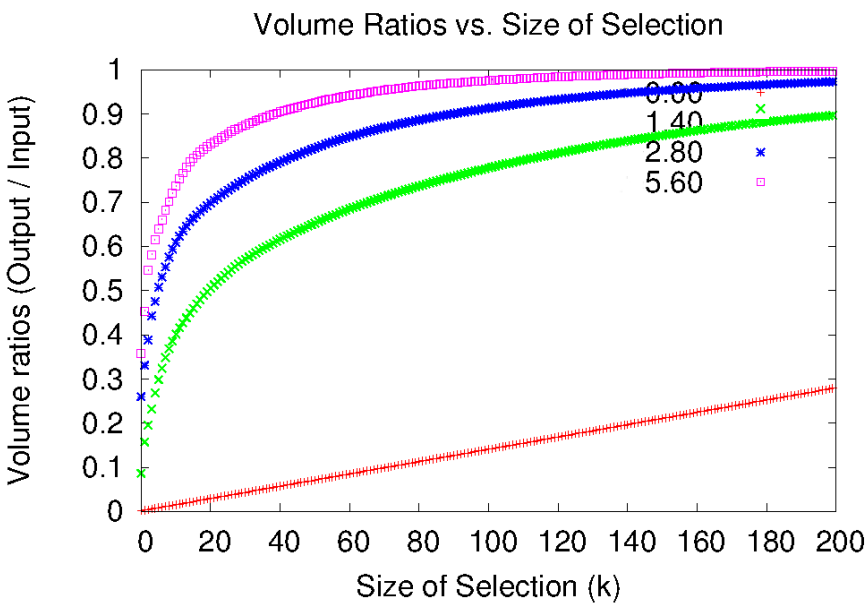
  $\mathcal{F}_{\mathcal{O}}^{r}$: input balls expanded by a quantity r

  $\rightarrow r = 0$: input model

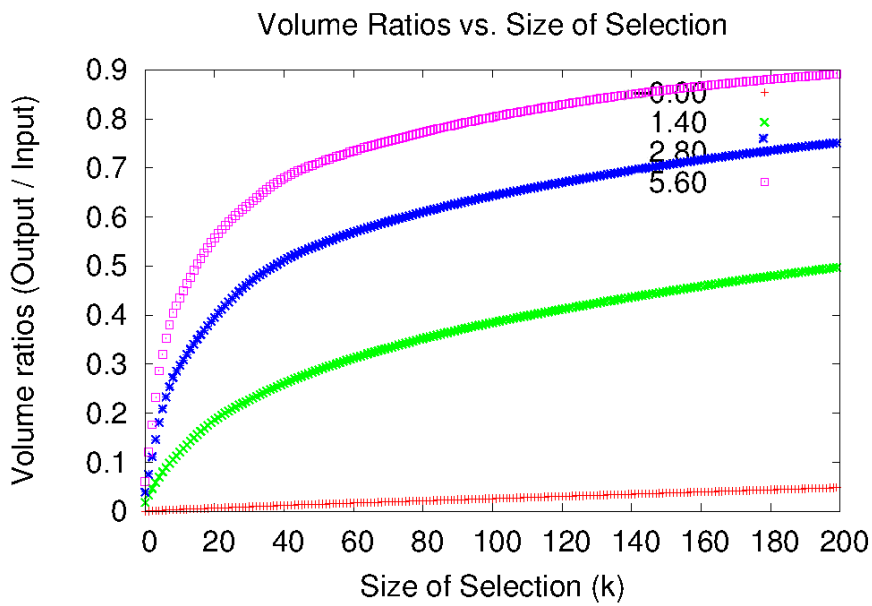
  $\mathcal{F}_{\mathcal{S}}^{r}$: domain of the selection for the expanded model

  Assessment: $\mathrm{Vol}(\mathcal{F}_{\mathcal{S}}^{r})/\mathrm{Vol}(\mathcal{F}_{\mathcal{O}}^{r})$ for increasing r

- PDB code 1igt: 1690 balls



- PDB 1igt: 10416 balls



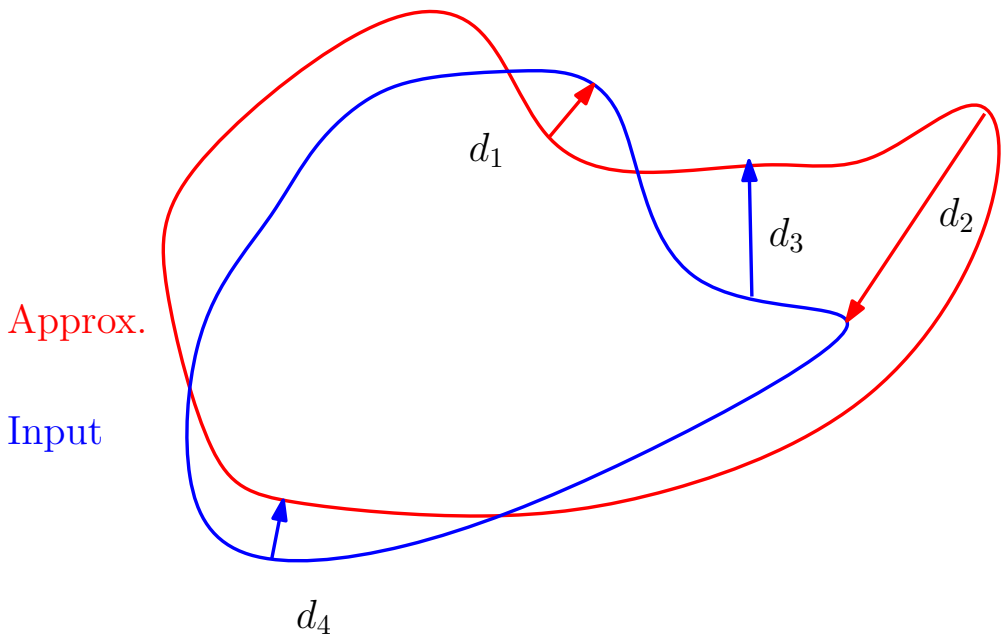
Greedy Assessment: (Signed) Hausdorff Distance

- Signed dist. of point p w.r.t. compact domain $\mathcal{F}$:

$$s(p, \partial\mathcal{F}) = \begin{cases} -\min_{q \in \partial\mathcal{F}} d(p,q) \text{ if } p \in \mathcal{F}, \\ +\min_{q \in \partial\mathcal{F}} d(p,q) \text{ otherwise,} \end{cases}$$

- Distance between boundaries: input domain $\partial\mathcal{F}_\mathcal{O}$ vs selection $\partial\mathcal{F}_\mathcal{S}$:

$$S_H(\partial\mathcal{F}_\mathcal{O}, \partial\mathcal{F}_\mathcal{S}) = [\min_{p \in \partial\mathcal{F}_\mathcal{S}} s(p, \partial\mathcal{F}_\mathcal{O}), \max_{p \in \partial\mathcal{F}_\mathcal{S}} s(p, \partial\mathcal{F}_\mathcal{O}); \min_{p \in \partial\mathcal{F}_\mathcal{O}} s(p, \partial\mathcal{F}_\mathcal{S}), \max_{p \in \partial\mathcal{F}_\mathcal{O}} s(p, \partial\mathcal{F}_\mathcal{S})]$$



- Assessment on a set of 96 protein complexes (1008 -13214 atoms)

Volume Preserving Approximations: Results

e	k/n	d_1	d_2	d_3	d_4
r_w	0.01	-8.39 ± 1.76	7.26 ± 1.74	-6.12 ± 1.77	5.54 ± 1.38
r_w	0.02	-7.64 ± 1.76	5.46 ± 1.11	-7.11 ± 2.41	4.89 ± 1.63
r_w	0.05	-5.61 ± 1.63	2.94 ± 0.85	-7.43 ± 2.38	4.76 ± 2.44
r_w	0.10	-4.05 ± 1.71	2.77 ± 1.52	-7.80 ± 1.80	5.25 ± 2.23
r_w	mean	-6.48 ± 2.42	4.66 ± 2.30	-7.10 ± 2.21	5.11 ± 1.98
5.6	0.01	-3.17 ± 0.88	3.49 ± 0.34	-4.36 ± 0.78	2.43 ± 0.24
5.6	0.02	-2.25 ± 1.54	2.58 ± 0.22	-3.55 ± 0.61	1.49 ± 0.15
5.6	0.05	-0.91 ± 0.35	1.68 ± 0.14	-2.77 ± 1.11	0.65 ± 0.91
5.6	0.10	-0.38 ± 0.12	1.08 ± 0.13	-1.68 ± 0.47	0.28 ± 0.07
5.6	mean	-1.92 ± 1.44	2.41 ± 0.89	-3.33 ± 1.20	1.38 ± 0.94

- Take home message: with a number of balls $\sim 5\%$ of atoms

  molecular volume exactly preserved
  distance between surfaces $\sim 2 - 3$ atoms (SAS model)

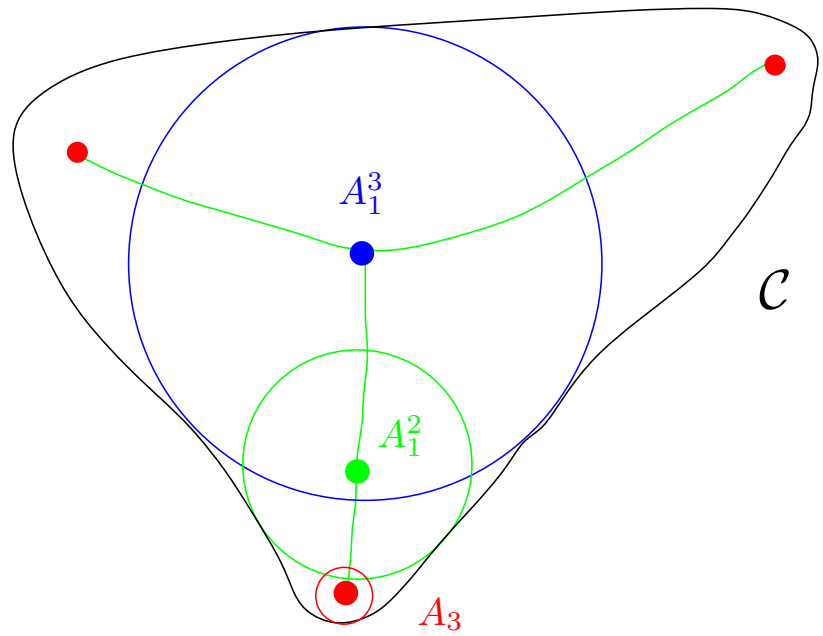
Modeling Contacts in Macro-molecular Assemblies

Medial Axis and Relatives

- For any open set $R \subset \mathbb{R}^n$:
  - Medial axis: points with at least two nearest neighbors in $\overline{R}$
  - Skeleton: centers of maximal balls
  - Singular set: points where the distance function is not differentiable
- For a smooth curve/surface:

$$\overline{MA} \subset \textit{Skeleton}$$

- Skeleton and local thickness:
  - Local: curvature properties
  - Global: related to bi/tri/tetra-tangent balls
- Medial axis transform: MAT



Max k-cover and the Greedy Strategy

▷ max k-cover:

$\mathcal{A}$: alphabet of m

$\mathcal{C}$: collection of subsets of $\mathcal{A}$

Select k subsets from $\mathcal{C}$
maximizing the number of points
from $\mathcal{A}$ which are covered

▷ Hardness:

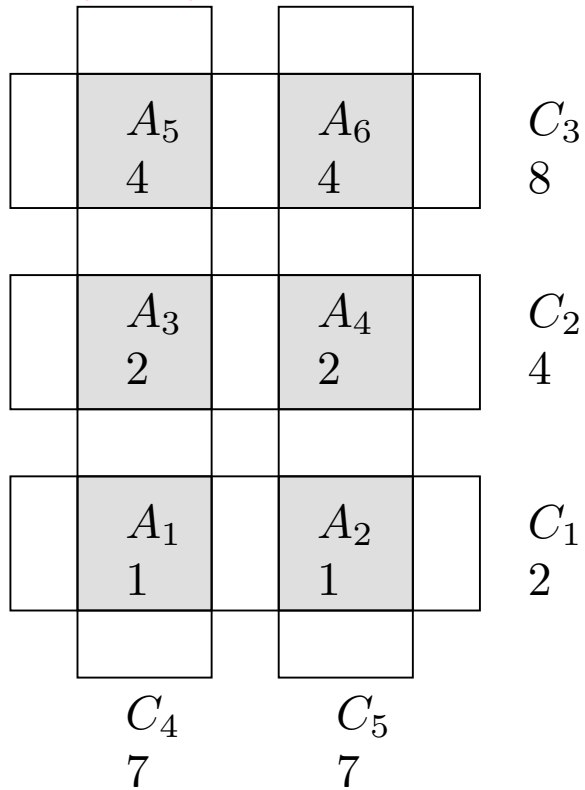
– problem is **NP**-complete

– OPT cannot be approximated within $1 - 1/e + \varepsilon$
unless $P = NP$

– Greedy algorithms achieve the $1 - 1/e$ bound

▷Ref: Feige; J. ACM; 1998

▷ Greedy may fail:

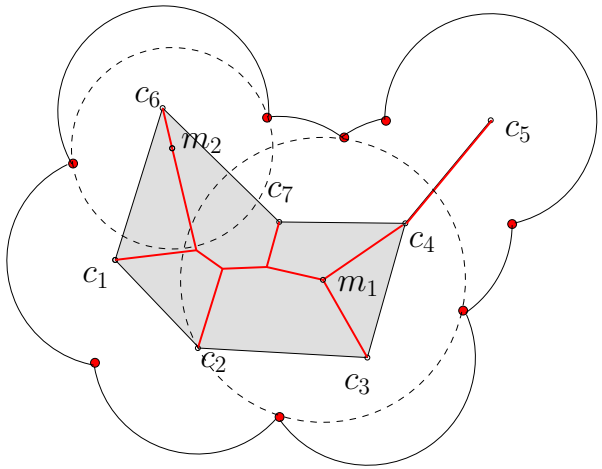


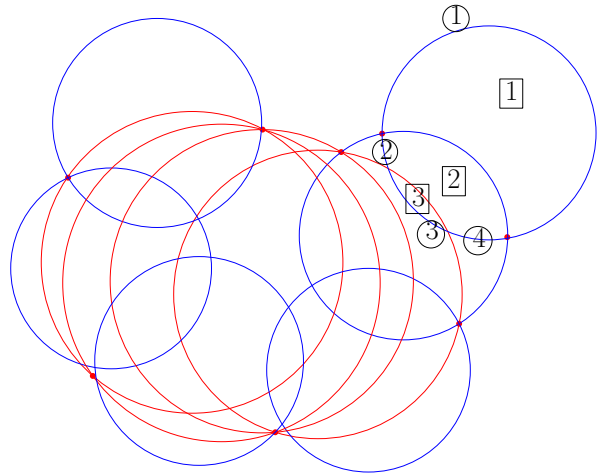
Greedy: $C_3 + C_2 = 12$

OPT: $C_4 + C_5 = 14$

Geometric Max k-cover for Balls

▷ Medial axis of the domain $\mathcal{F}_{\mathcal{O}}$, associated covering $\mathcal{F}_{\mathcal{C}}$, and induced arrangement of balls $\mathcal{A}$





▷ Given a function defined on the cells of $\mathcal{A}$:

- Maximize the weight of a selection of k cells
- Two cases: volume vs surface arrangements

  For the latter: cf role of the MA w.r.t. $\mathcal{F}_{\mathcal{C}} = \cup_i B_i$

▷ Complexity: geometric versions of max k-cover

▷Ref: Amenta, Kolluri; CGTA; 2001

▷Ref: Feige; J. ACM; 1998

Inner Approximation

▷ Punchline:
  – The first provably correct volume-based approximation algorithm of 3D shapes, which works in a finite setting ($\neq$ the ε-sample framework)

▷ Thm. The MAT of a union of balls is discrete in the following sense:

$$\mathcal{F}_{\mathcal{C}} = \bigcup_i B_i = \bigcup_{v \in \mathcal{V}} B_v^*. \tag{3}$$

with $\mathcal{V}$ the vertices of the medial axis.

▷ Corr. The 3D arrangement induced by balls in $\mathcal{V}$ can be used to run greedy algorithms.

▷ Thm. The Greedy strategy for positive volume weights has the following approximation ratios:

$$\begin{cases} 1-(1-1/k)^k > 1-1/e & \text{wrt to OPT weight (volume)} \\ 1-(1-1/n)^k & \text{wrt the total weight (volume)} \end{cases} \tag{4}$$

▷ Obs. The Greedy strategy for positive surface weights can be as bad ad $1/k^2$.

▷Ref: Cazals, Dreyfus, Sachdeva, Shah; Comp. Graphics Forum, 2014

Robust Implementation of Greedy for the Volume Case: *A High-profile Implementation*

- Delaunay triangulation (DT) *DTB* of the input balls
- Delaunay triangulation *DTV* of the boundary points of $\partial\mathcal{F}_{\mathcal{C}}$
  - Points have degree two algebraic coordinates
  - Degeneracies to be handled (e.g. $n > 3$ coplanar points)
- Medial axis of the input balls
  - Voronoi diagram DTV^* clipped by the α-shape of *DTB*
- MAT restricted to vertices of the MA
- Volume computations to run greedy

Ref: De Castro and F. Cazals and S. Loriot and M. Teillaud; CGTA; 2009

Ref: Cazals and H. Kanhere and S. Loriot; ACM TOMS; 2011

Modeling Contacts in Macro-molecular Assemblies

Outlook

- Pros
  - Flexible framework to design approximations
    - Inner / outer / volume preserving approximations
    - The molecule or complex can be processed as a whole
      - or can be decomposed into regions processed independently

- Geometric models produced can be complemented by
  - Connectivity information
  - Biophysical properties

References

- F. Cazals and T. Dreyfus and S. Sachdeva and N. Shah, Greedy Geometric Algorithms for Collections of Balls, with Applications to Geometric Approximation and Molecular Coarse-Graining, Computer Graphics Forum, 2014.

Overview

PART 1:Connectivity Inference from Native Mass Spectrometry Data

PART 2:Building Coarse Grain Models

PART 3:Handling uncertainties in Macro-molecular Assembly Models

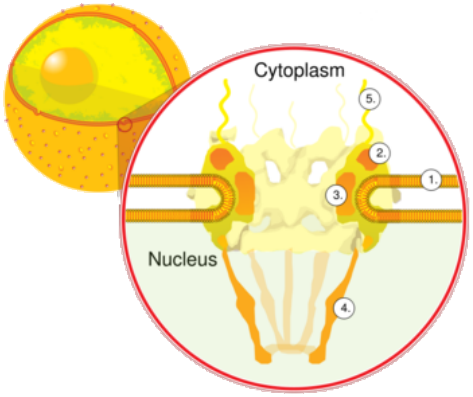
PART 4:Conformational Ensembles and Energy Landscapes: Analysis

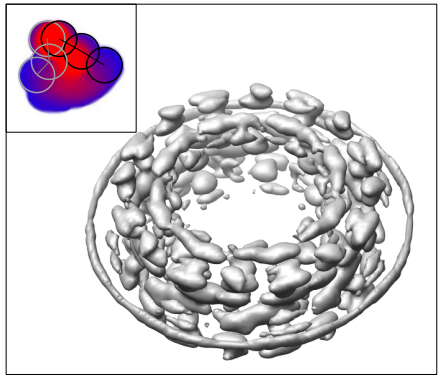
PART 5:Conformational Ensembles and Energy Landscapes: Comparison

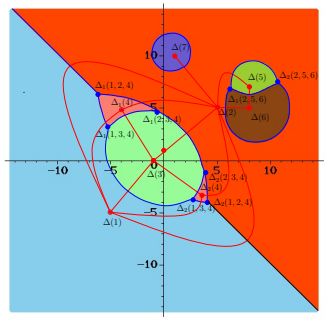
Assessing the Reconstruction of Macro-molecular Assemblies with Toleranced Models

Frederic Cazals, Tom Dreyfus, Inria ABS
Valerie Doye, Inst. J. Monod

Algorithms - Biology - Structure project-team
INRIA Sophia Antipolis France



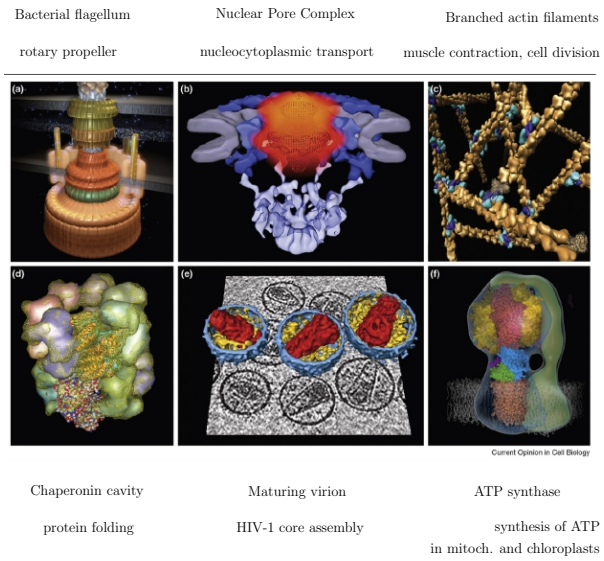




Modeling Contacts in Macro-molecular Assemblies

Structural Dynamics of Macromolecular Processes

Reconstructing Large Macro-molecular Assemblies



- Molecular motors
- NPC
- Actin filaments
- Chaperonins
- Virions
- ATP synthase

▷ **Difficulties**

Modularity
Flexibility

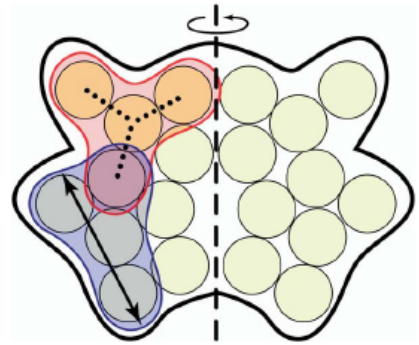
▷ **Core questions**

Reconstruction / animation
  Integration of (various) experimental data
Coherence model vs experimental data

▷`Ref: Russel et al, Current Opinion in Cell Biology, 2009`

Reconstructing Large Assemblies: a NMR-like Data Integration Process

▷ Four ingredients
– Experimental data
– Model: collection of balls
– Scoring function: sum of restraints
  restraint : function measuring the agreement
    ≪model vs exp. data≫
– Optimization method (simulated annealing,. . .)



▷ Restraints, experimental data and . . . ambiguities:

Assembly	: shape	cryo-EM	fuzzy envelopes
Assembly	: symmetry	cryo-EM	idem
Assembly	: sub-systems	mass spec.	stoichiometry
Complexes:	: interactions	TAP (Y2H, overlay assays)	stoichiometry
Instance:	: shape	Ultra-centrifugation	rough shape (ellipsoids)
Instances:	: locations	Immuno-EM	positional uncertainties

▷Ref: Alber et al, Ann. Rev. Biochem. 2008 + Structure 2005

Checkpoint

- Consider a real valued function:

$$f(x, y, z) : \mathbb{R}^3 \longrightarrow \mathbb{R} \tag{5}$$

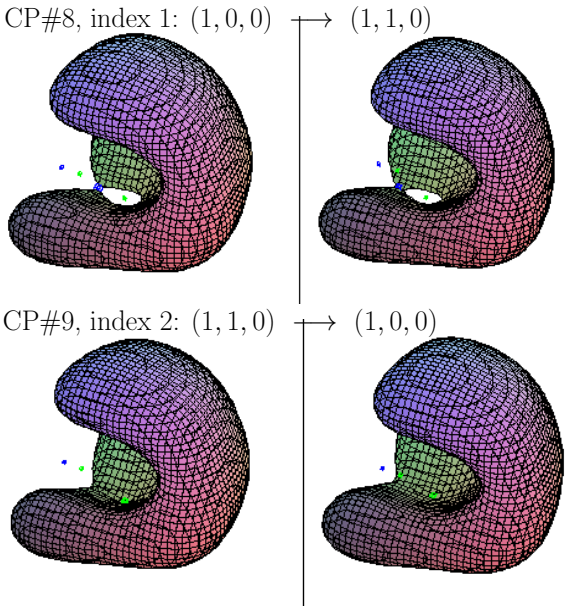
What is, in general, the locii of point defined as follows:

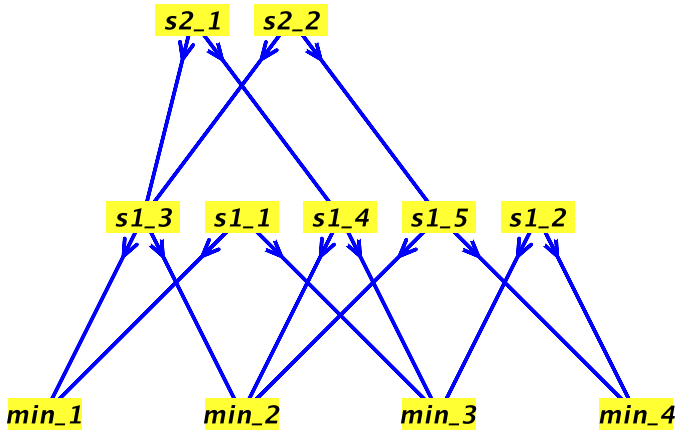
$$S = \{p = (x, y, z) \in \mathbb{R}^3 \mid f(p) = c\} \tag{6}$$

Morse Homology: Illustration

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

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

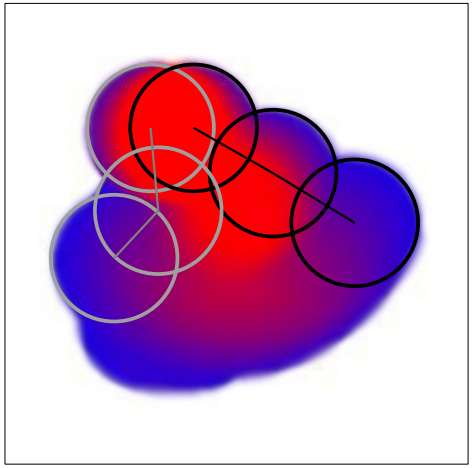


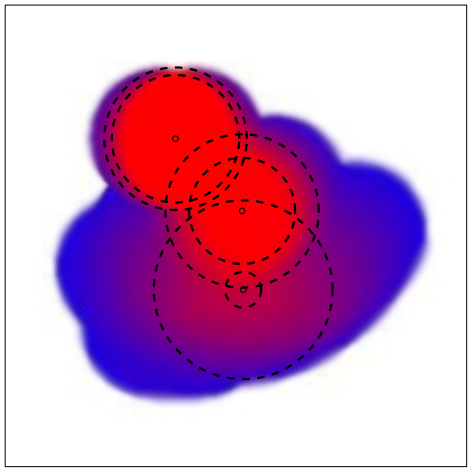


- Key construction: the **Morse-Smale(-Witten) chain complex** i.e. the connections between critical points whose indices differ by one is sufficient to compute the Betti numbers

- 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

Uncertain Data and Toleranced Models: the Example of Molecular Probability Density Maps

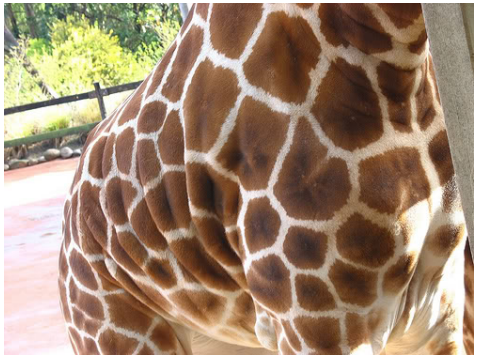


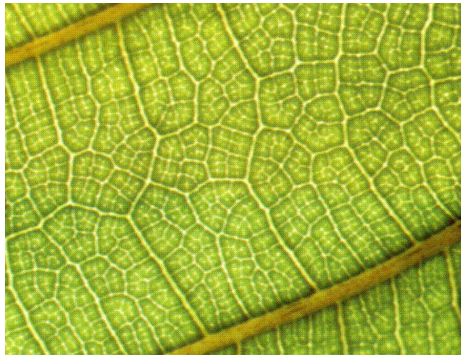


- **Probability Density Map of a Flexible Molecule**
  - Each point of the probability density map: probability of **being covered** by a conformation
- **Question:**
  How does one accommodate high/low density regions?
- **Toleranced ball** $\overline{S_i}$
  - Two **concentric** balls of radius $r_i^- < r_i^+$:
    inner ball $\overline{S_i}[r_i^-]$: high confidence region
    outer ball $\overline{S_i}[r_i^+]$: low confidence region
- **A continuum of models**
  - **Linear interpolation** of radii: $r_i(\lambda) = r_i^- + \lambda(r_i^+ - r_i^-)$
  - **Tracking intersections** of $\overline{S_i}[r_i(\lambda)]$:
    $\rightarrow$ **Voronoi diagram** of toleranced balls

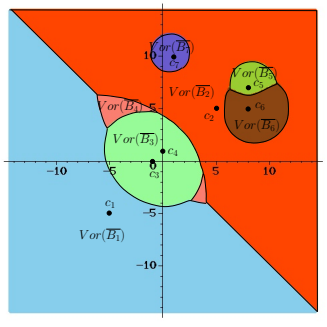
Modeling Contacts in Macro-molecular Assemblies

Voronoi diagrams in Biology, Geology, Engineering





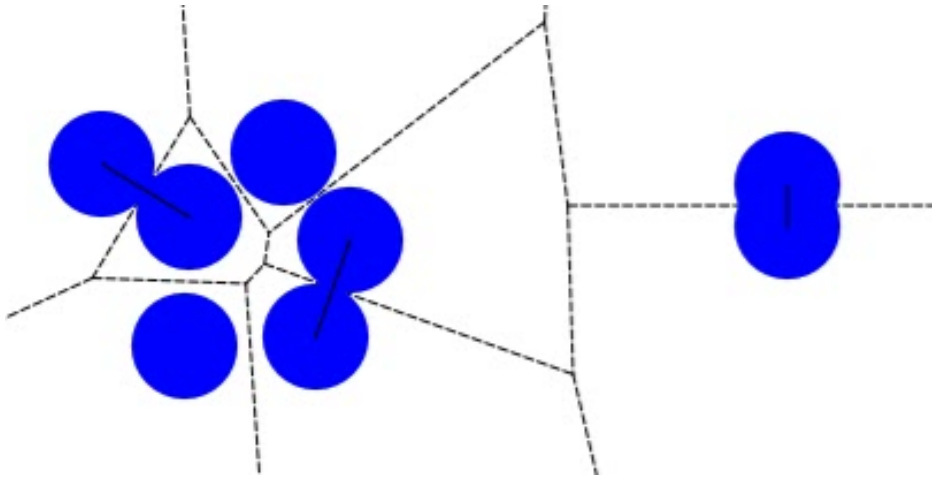




▷Ref: Cazals, Dreyfus; Symp. on Geometry Processing, 2010

The α-complex: Demo

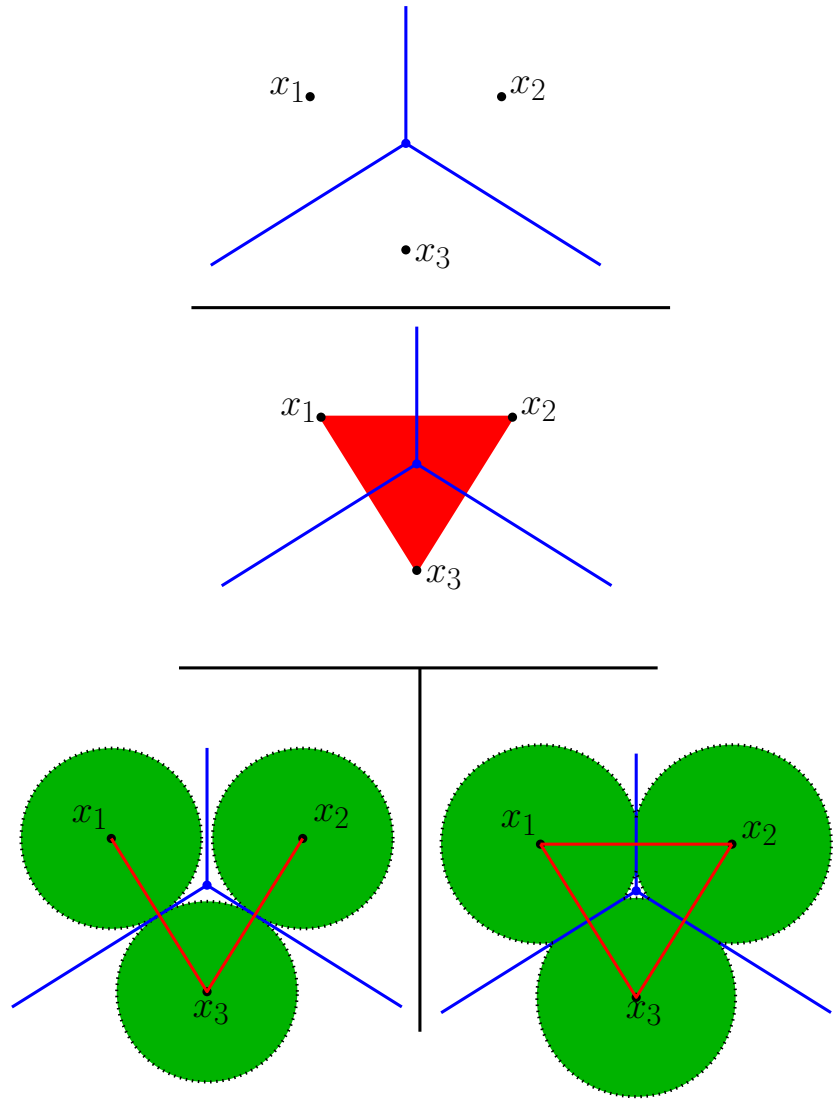
VIDEO/ashape-two-cc-cycle-video.mpeg



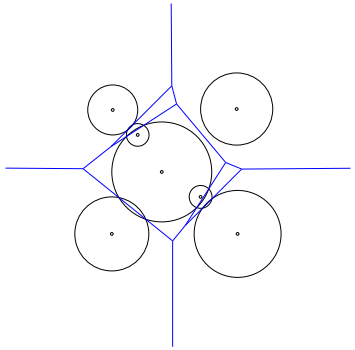
- α-complex
  - simplicial complex encoding the topology of growing balls
  - multi-scale analysis of a collection of balls
    - how many clusters / clusters' stability?
    - topology of the clusters?

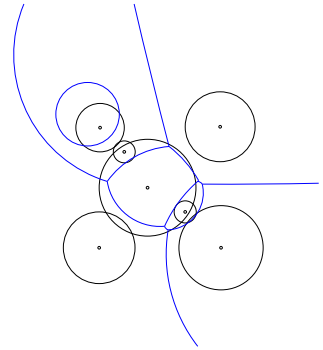
Euclidean Voronoi diagram and α-complex

- Voronoi diagram of $\mathcal{S} = \{x_i\}$
  - Voronoi region $Vor(x_i)$:
    $\{p \mid d(p, x_i) < d(p, x_j), i \neq j\}$
- Dual complex $K(\mathcal{S})$
  - Delaunay triangulation (Euclidean case)
  - Simplex Δ: dual of $\bigcap_{x_i \in \Delta} Vor(x_i) \neq \emptyset$
- α-complex $K_\alpha(\mathcal{S})$
  - Grown spheres:
    $S_{i,\alpha} = S_i(x_i, \alpha)$
  - Restricted Voronoi region:
    $R_{i,\alpha} = S_{i,\alpha} \cap Vor(x_i)$
  - $\Delta \in K_\alpha(\mathcal{S})$:
    $\bigcap_{x_i \in \Delta} R_{i,\alpha} \neq \emptyset$
- α-complex: topological changes
  induced by a **growth** process



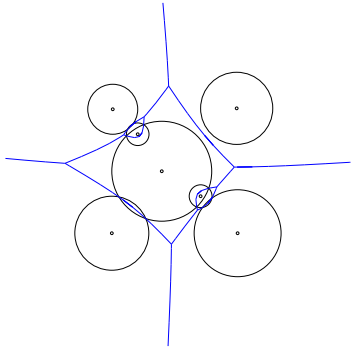
Growth Processes and Curved Voronoi diagrams

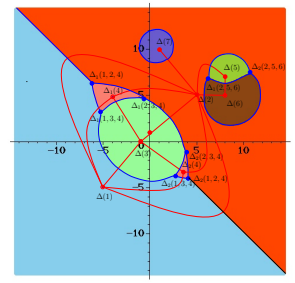




▷ Power diagram:
$d(S(c,r),p) = \|c-p\|^2 - r^2$

▷ Mobius diagram:
$d(S(c,\mu,\alpha),p) = \mu\|c-p\|^2 - \alpha^2$





▷ Apollonius diagram:
$d(S(c,r),p) = \|c-p\| - r$

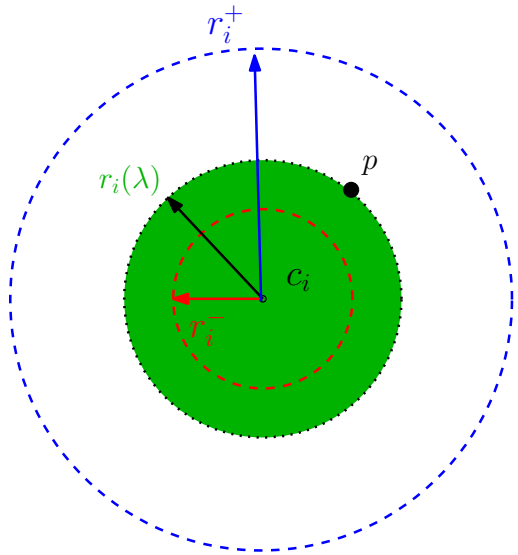
▷ **Compoundly Weighted Voronoi diagram**:
$d(S(c,\mu,\alpha),p) = \mu\|c-p\| - \alpha$

▷Ref: Boissonnat, Wormser, Yvinec; in *Effective Comp. Geom.*; 2006

Modeling Contacts in Macro-molecular Assemblies

From Toleranced Balls to Compoundly Weighted Points and Compoundly Weighted Voronoi Diagrams

- **Toleranced ball** $\overline{S_i}(c_i; r_i^-; r_i^+)$ and radius interpolation:
  - **Radius discrepancy**: $\delta_i = r_i^+ - r_i^-$
  - **Grown ball** $\overline{S_i}[\lambda](c_i, r_i(\lambda))$ with $r_i(\lambda) = r_i^- + \lambda\delta_i$



- **Growing ball** `swallowing` **a point** p**:**
  - p is at the surface of $\overline{S_i}[\lambda]$
    $\Leftrightarrow r_i(\lambda) = \| c_i p \|$
    $\Leftrightarrow \lambda = \frac{\|c_i p\| - r_i^-}{\delta_i}$

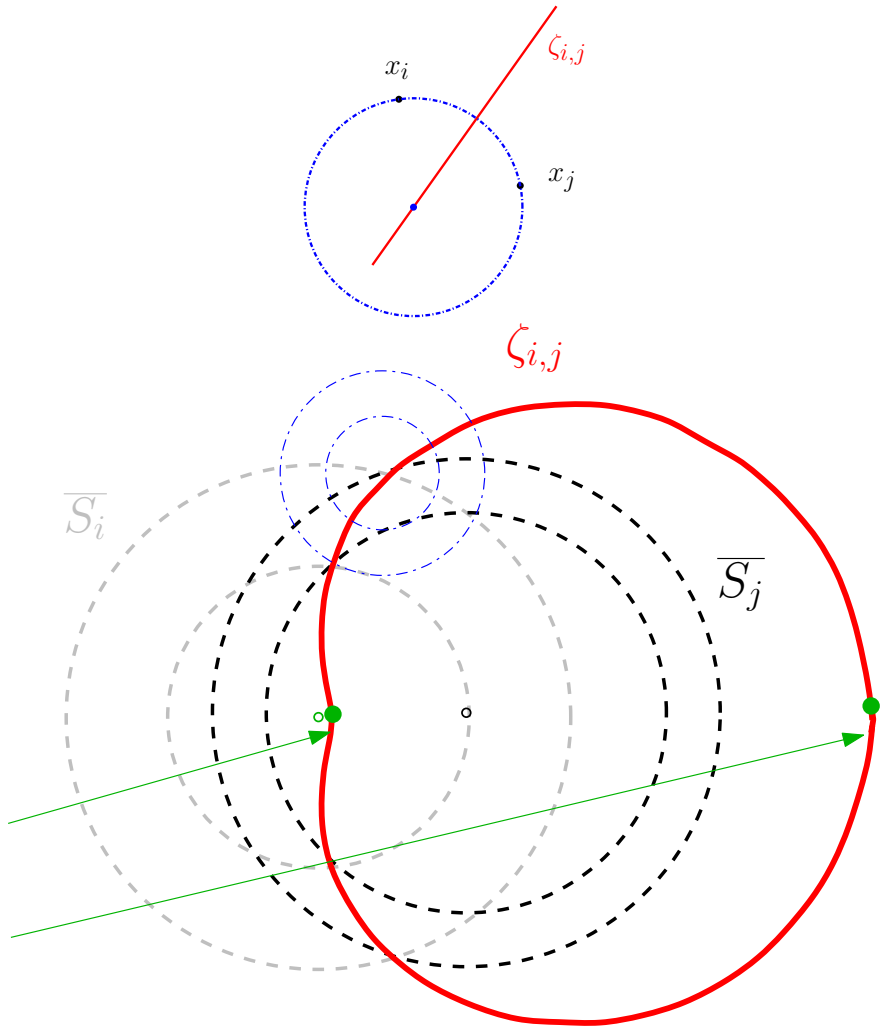
- **From Toleranced Ball to Compoundly Weighted Point:**
  - $S_i(c_i; \mu_i = \frac{1}{\delta_i}, \alpha_i = \frac{r_i^-}{\delta_i})$
  - $\lambda(S_i, p) = \frac{1}{\delta_i} \| c_i p \| - \frac{r_i^-}{\delta_i}$

The Voronoi Diagram induced by **Toleranced Balls** is the **Compoundly Weighted** one !

Bisectors

- Rationale from the Euclidean Voronoi diagram:
  - Bisector $\zeta_{i,j}$ of (x_i, x_j)
    centers of circumscribed balls to x_i and x_j

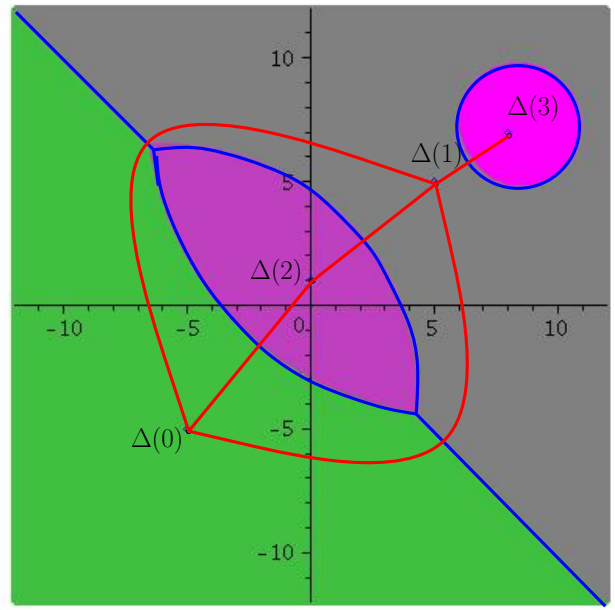
- Generalization to the CW case:
  - Bisector $\zeta_{i,j}$ of $(\overline{S_i}, \overline{S_j})$
    centers of toleranced tangent balls to $\overline{S_i}$ and $\overline{S_j}$
    $\Rightarrow$ **degree four** algebraic surface
  - Extremal toleranced tangent balls
    smallest one of radius $\underline{\rho}$
    $\Rightarrow$ first intersection of $\overline{S_{i_0}}[\underline{\rho}], \ldots, \overline{S_{i_k}}[\underline{\rho}]$
    largest one of radius $\overline{\rho}$
    $\Rightarrow$ last intersection of $\overline{S_{i_0}}[\overline{\rho}], \ldots, \overline{S_{i_k}}[\overline{\rho}]$



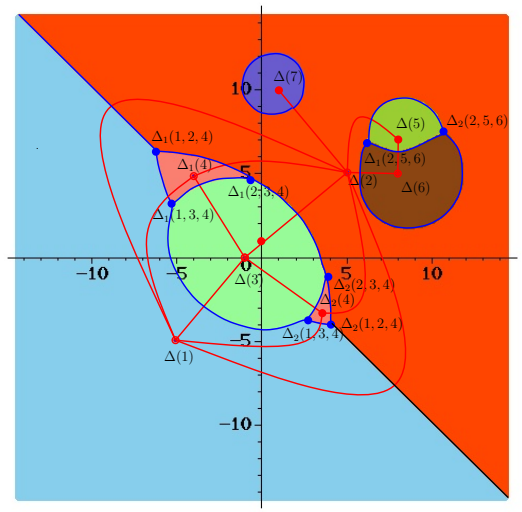
Voronoi Diagram and its Dual Complex: Topological Complications

- **Partition of the ambient space:**
  $Vor(\overline{S_i}) = \{p \in \mathbb{R}^3 \mid \lambda(\overline{S_i}, p) \leq \lambda(\overline{S_j}, p)\}$

- Voronoi region – in all generality:
  - Neither connected : collection of **faces**
  - Nor simply connected

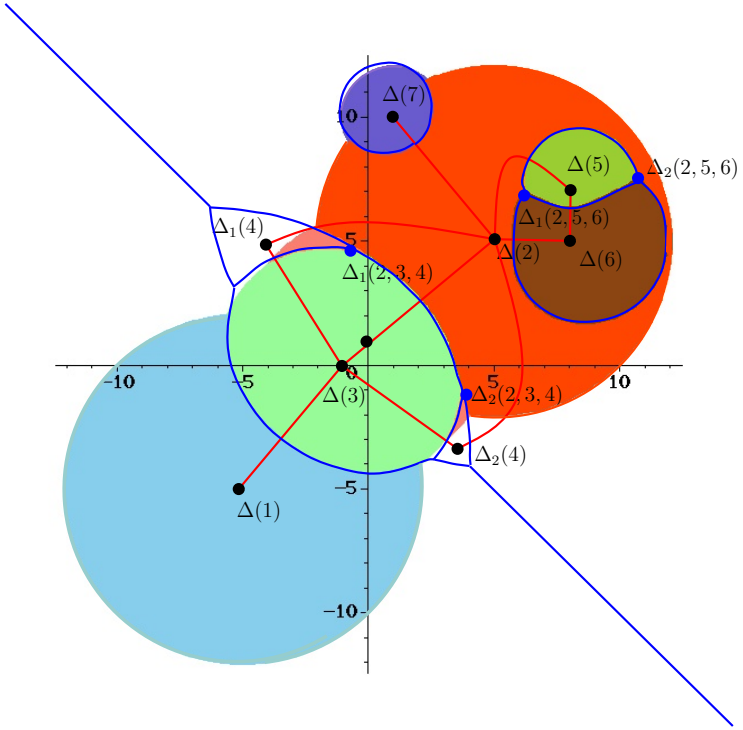


- Dual complex:
  - Not a **triangulation**
    $\rightarrow$ abstract representation with a **Hasse diagram**
  - abstract edges without triangle
    Hole in Voronoi region
    Ex. (**Top**): $\Delta(1,3)$
  - $\neq$ abstract triangles sharing two edges
    Lens sandwiched Voronoi region (Apollonius case)
    Ex. (**Top**): $\Delta_1(0,1,2)$ and $\Delta_2(0,1,2)$
  - $\neq$ abstract triangles sharing the same edges
    Composed hole in Voronoi region
    Ex. (**Bottom**): $\Delta_1(1,4,5)$ and $\Delta_2(1,4,5)$



Compoundly Weighted Filtration: the λ-complex

- Definition. λ-complex K_λ:
  - sub-complex of the dual complex
  - $\Delta \in K_\lambda$: $\bigcap_{\overline{S_i} \in \Delta} R_{i,\lambda} \neq \emptyset$
    $\rightarrow$ map λ to Δ



- **Status** of $\Delta \in K_\lambda$ and **boundary** $\partial\overline{S}[\lambda]$:
  - **singular**: $\bigcap_{\overline{S_i} \in \Delta} \overline{S_i}[\lambda] \in \partial\overline{S}[\lambda]$. Ex. $\Delta_{1,3}$
  - **regular** : $\bigcap_{\overline{S_i} \in \Delta} R_{i,\lambda} \in \partial\overline{S}[\lambda]$. Ex. $\Delta_{3,4}$
  - **interior** : $\bigcap_{\overline{S_i} \in \Delta} R_{i,\lambda} \notin \partial\overline{S}[\lambda]$. Ex. $\Delta_{2,3}$

- **Classification** of $\Delta(T_k)$:

	singular	regular	interior
(1) $\Delta(T) \in CH(\overline{S})$,Gabriel, non dominated/dominant	$(\underline{\rho}_{\Delta(T)}, \underline{\mu}_{\Delta(T)}]$	$(\underline{\mu}_{\Delta(T)}, +\infty]$	
(2) $\Delta(T) \in CH(\overline{S})$,non Gabriel, non dominated/dominant		$(\underline{\mu}_{\Delta(T)}, +\infty]$	
(3) $\Delta(T) \notin CH(\overline{S})$ Gabriel, non dominated/dominant	$(\underline{\rho}_{\Delta(T)}, \underline{\mu}_{\Delta(T)}]$	$(\underline{\mu}_{\Delta(T)}, \overline{\mu}_{\Delta(T)}]$	$(\overline{\mu}_{\Delta(T)}, +\infty]$
(4) $\Delta(T) \notin CH(\overline{S})$,non Gabriel, non dominated/dominant		$(\underline{\mu}_{\Delta(T)}, \overline{\mu}_{\Delta(T)}]$	$(\overline{\mu}_{\Delta(T)}, +\infty]$
(5) $\Delta(T) \notin CH(\overline{S})$ Gabriel, dominant	$(\underline{\rho}_{\Delta(T)}, \underline{\mu}_{\Delta(T)}]$	$(\underline{\mu}_{\Delta(T)}, \overline{\rho}_{\Delta(T)}]$	$(\overline{\rho}_{\Delta(T)}, +\infty]$
(6) $\Delta(T) \notin CH(\overline{S})$,non Gabriel, dominant		$(\underline{\mu}_{\Delta(T)}, \overline{\rho}_{\Delta(T)}]$	$(\overline{\rho}_{\Delta(T)}, +\infty]$
(7) $\Delta(T) \notin CH(\overline{S})$ Gabriel, dominated	$(\underline{\rho}_{\Delta(T)}, \underline{\mu}_{\Delta(T)}]$	$(\underline{\mu}_{\Delta(T)}, \gamma_{\Delta(T)}]$	$(\gamma_{\Delta(T)}, +\infty]$
(8) $\Delta(T) \notin CH(\overline{S})$,non Gabriel, dominated		$(\underline{\mu}_{\Delta(T)}, \gamma_{\Delta(T)}]$	$(\gamma_{\Delta(T)}, +\infty]$

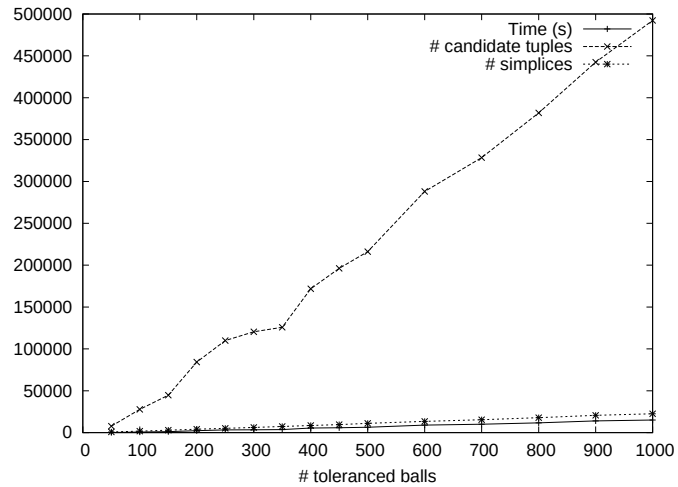
Algorithms

- **Naively enumerating candidate tuples:**
  - a **tuple** of toleranced balls:
    a pair, triple or quadruple
  - **candidate**: possibly **contributing simplices**

- **Computing the CW Dual Complex:**
  - Iterative construction of the skeleton,
    from tetrahedra to vertices

- **Time complexity:** $O(n(n^2 + \tau))$
  τ: number of candidate tuples

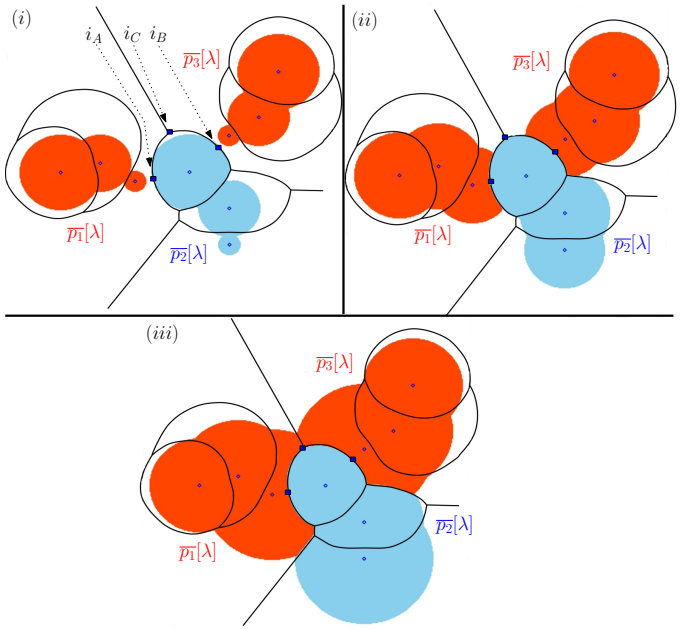
- **Difficulties:**
  - comparing roots of **degree four** polynomial
    checking that extremal TT balls are conflict-free
  - computing the dual of **non connected Voronoi region**:
    disambiguating the neighborhood of dual simplices

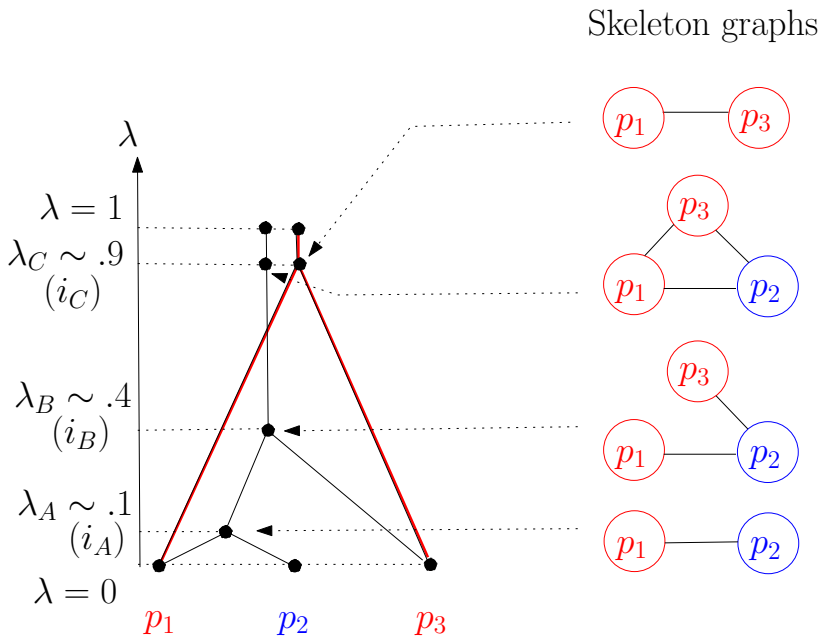


(**Random** Toleranced balls)

Modeling Contacts in Macro-molecular Assemblies

Multi-scale Analysis of Toleranced Models: Protein Contact History Encoded in the Hasse Diagram



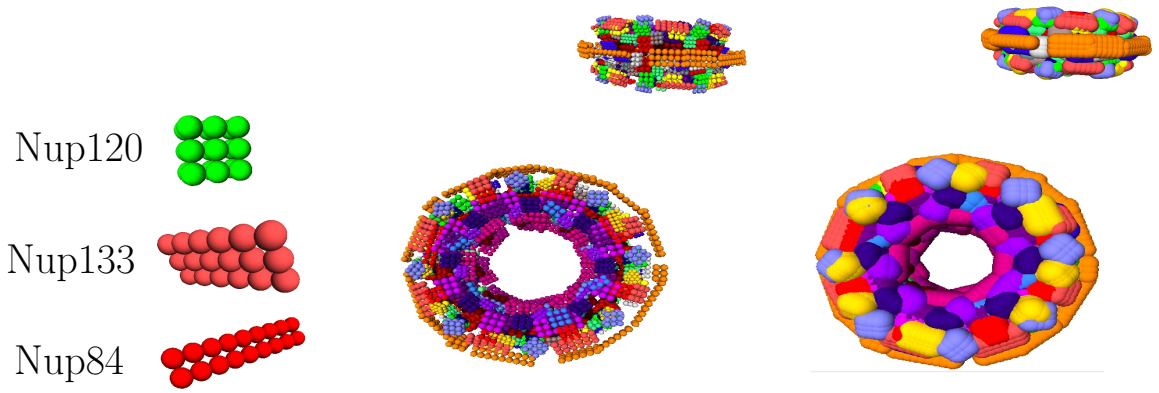


- **Red-blue bicolor setting:** red proteins are types singled out (e.g. TAP)
- Protein contact history: **Hasse diagram**
- **Finite set of topologies:** encoded into a Hasse diagram
  - **Birth and death** of a complex
  - **Topological stability** of a complex $s(c) = \lambda_d(C) - \lambda_b(C)$
- **Computation:** via intersection of Voronoi restrictions

Voratom: Assessing Contacts in the Toleranced Model of a Large Assembly

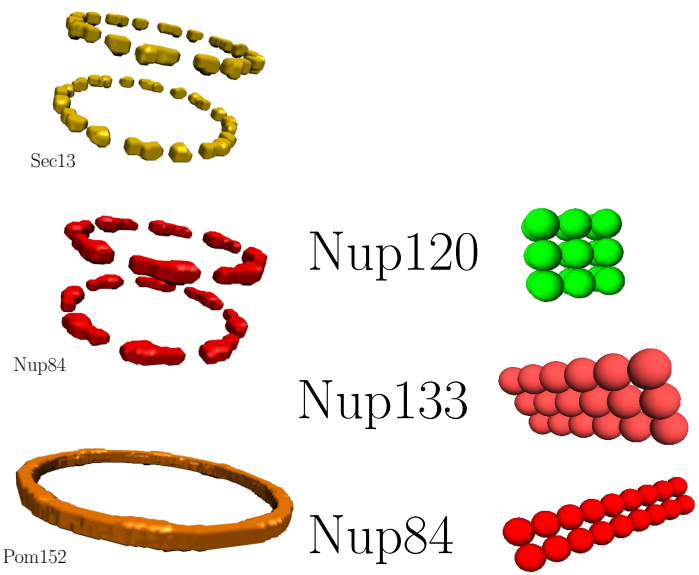
- 3 steps:
  - Building **occupancy volumes**
  - Building a **Toleranced Model**
  - Inferring the **Hasse diagram** encoding protein contacts

VIDEO/voratom-y-complex-long.mpeg

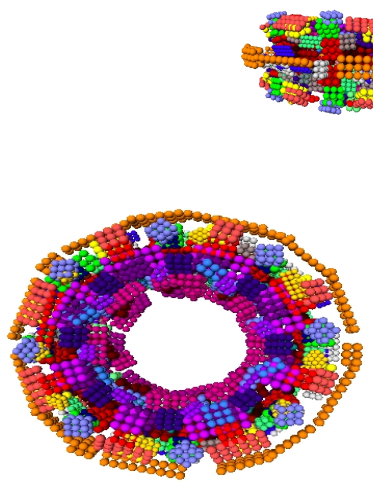


Toleranced Models for the NPC

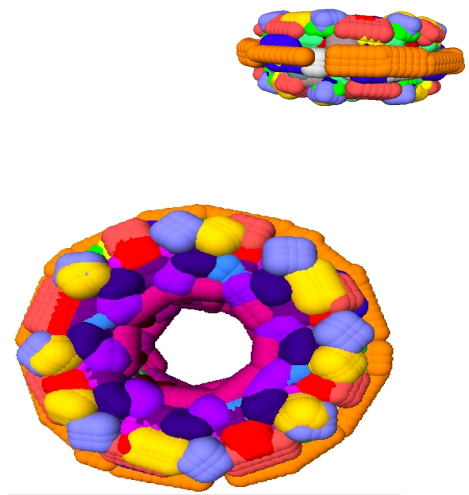
- Input: 30 probability density maps from Sali et al.
- Output: 456 toleranced proteins
- Rationale:
  $\rightarrow$ assign protein instances to *pronounced local maxima* of the maps
- **Geometry of instances:**
  – four canonical shapes
  – controlling $r_i^+ - r_i^-$: w.r.t volume estimated from the sequence



(i) Canonical shapes



(ii) NPC at $\lambda = 0$

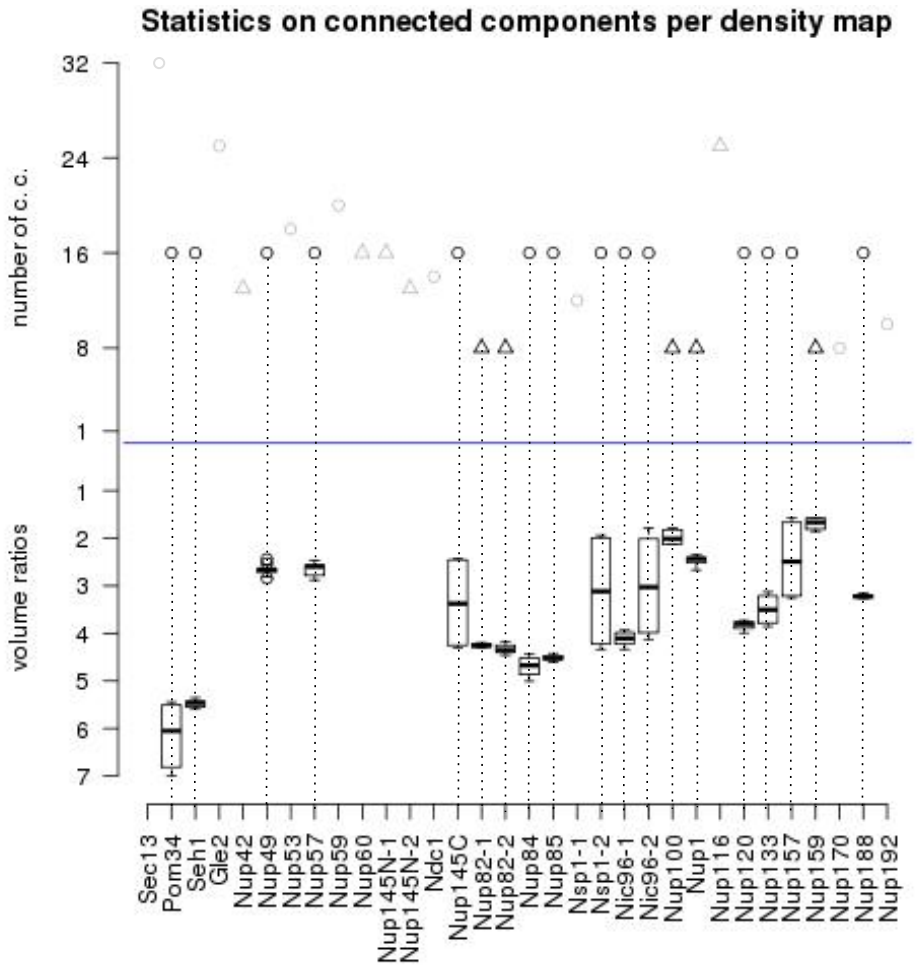


(iii) NPC at $\lambda = 1$

Stopping the Growth Process

Matching the Uncertainties on the Input Data

- **Uncertainty** of a density map: $\frac{\text{Volume of voxels with probability} > 0}{\text{Stoichiometry} \times \text{Reference volume}}$



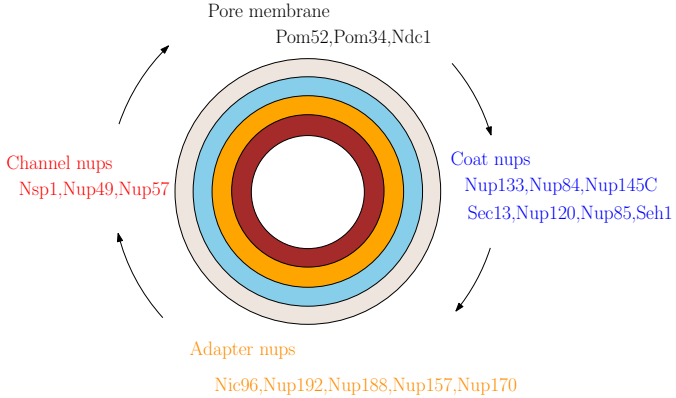
Probability density maps sorted by molecular weight

Three Analysis of the Toleranced Model of an Assembly

- **Local:**
  - Tracking copies of sub-complexes in the assembly
    $\rightarrow$ **Hasse diagram**
- **Global:**
  - Inspecting pairwise protein contacts
    $\rightarrow$ **Contact probabilities**
  - Controlling the volume of evolving complexes
    $\rightarrow$ **Volume ratio**

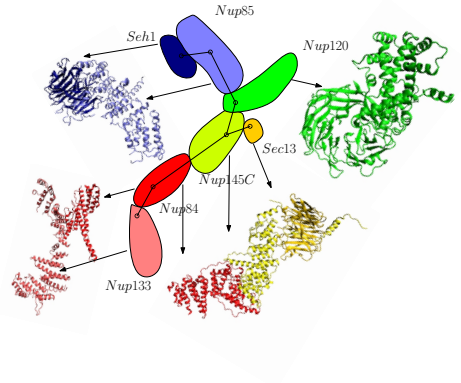
Putative Models of Sub-complexes: the Y-complex

▷ **Symmetric core of the NPC**



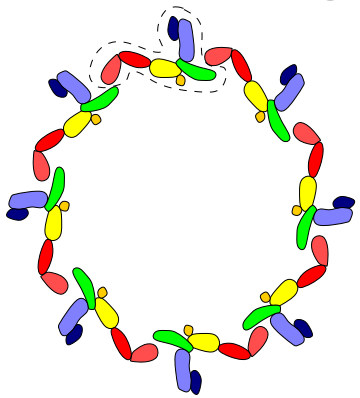
▷Ref: Blobel et al; Cell; 2007

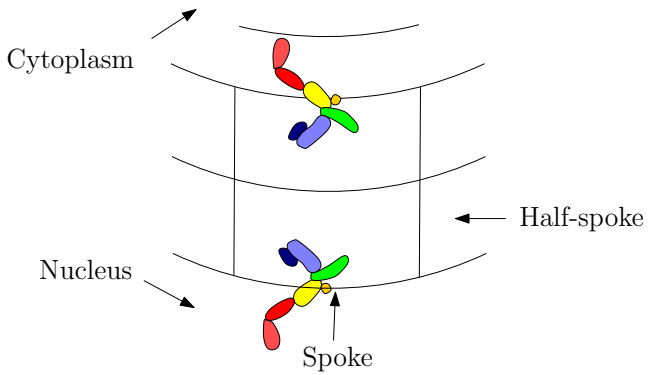
▷ **The Y-complex: pairwise contacts**



▷Ref: Blobel et al; Nature SMB; 2009

▷ **Y-based head-to-tail ring vs. upward-downward pointing**





▷Ref: Seo et al; PNAS; 2009

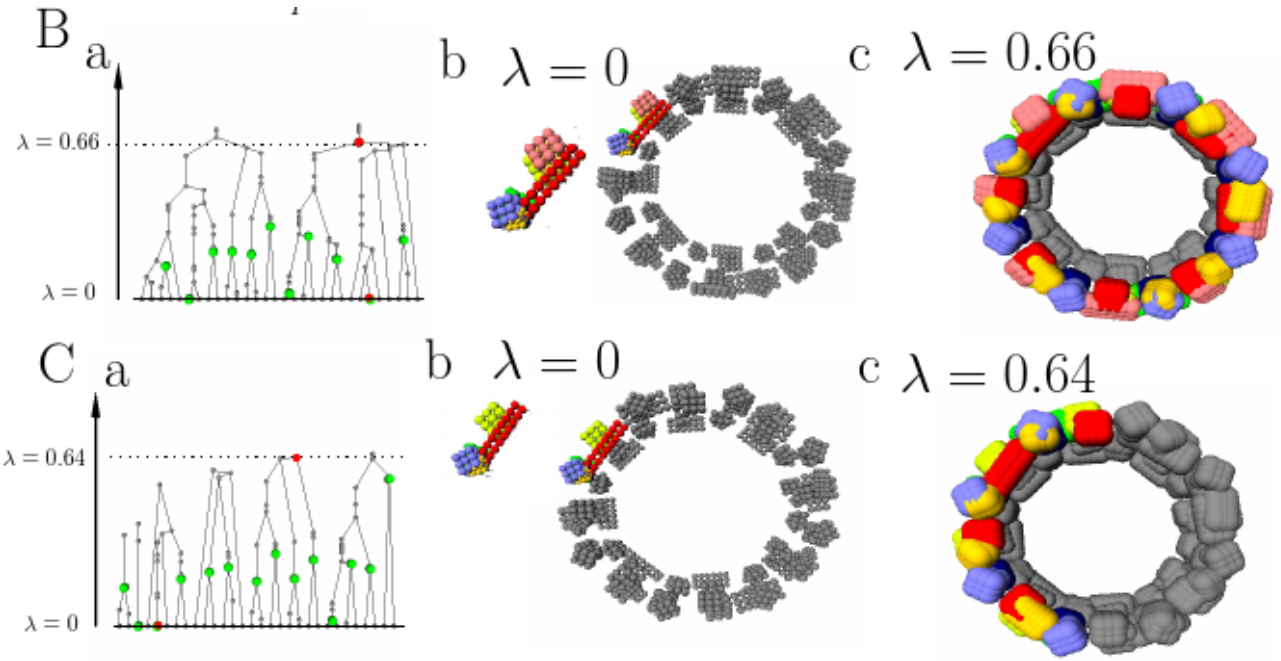
▷Ref: Brohawn, Schwarz; Nature MSB; 2009

⇒ Bridging the gap between both classes of models?

Assessment w.r.t. a Set of Protein Types: Isolated Copies

Geometry, Topology, Biochemistry

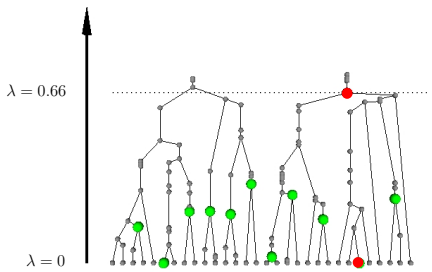
- Input:
  - Toleranced model
  - T: **set of proteins types**, the red proteins (types involved in a sub-complex)
- Output, overall assembly:
  - number of **isolated copies**: symmetry analysis
  - their topological stability: death date - birth date (cf α-shape demo)



- **B**: closure of the 2 rings; **C**: painting Nup133 in blue

Closure of the Two Rings Involving Y-complexes: Pairwise Contacts

- **The TOM supports Blobel's hypothesis**



Events accounting for the closure

9 (Nup133, Nup85) $\lambda \in [0.09, 0.70]$
5 (Nup84, Nup85) $\lambda \in [0.52, 0.69]$
1 (Nup133, Nup120) $\lambda = 0$
1 (Nup84, Nup120) $\lambda = 0.06$

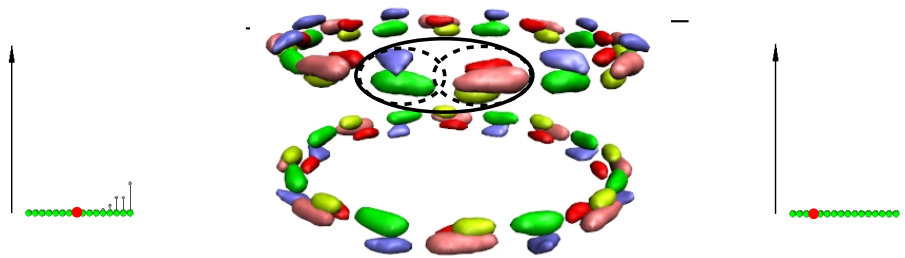
Nup85 involved in 14 / 16 contacts

- Inner structure of the **Y-complexes into two sub-units**

Density maps: contour plot; Hasse diagram per sub-unit

(Nup120, Nup85, Seh1)

(Nup84, Nup145C, Nup133)



Three Analysis of the Toleranced Model of an Assembly

- **Local:**
  - Tracking copies of sub-complexes in the assembly
    $\rightarrow$ **Hasse diagram**
- **Global:**
  - Inspecting pairwise protein contacts
    $\rightarrow$ **Contact probabilities**
  - Controlling the volume of merging complexes
    $\rightarrow$ **Volume ratio**

Contact Frequencies versus Contact Probabilities: Definitions

- **Contact frequency** f_{ij} from Sali et al
  - Given N optimized bead models of the NPC:
    f_{ij} : fraction of the N models with at least one contact (P_i, P_j)

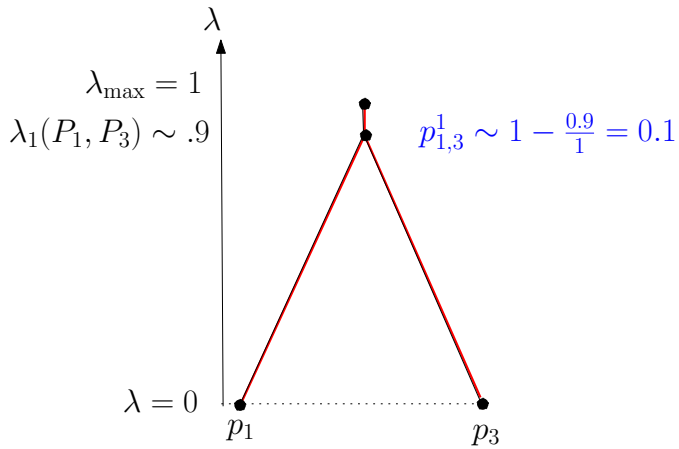
- **Contact probability** $p_{ij}^{(k)}$

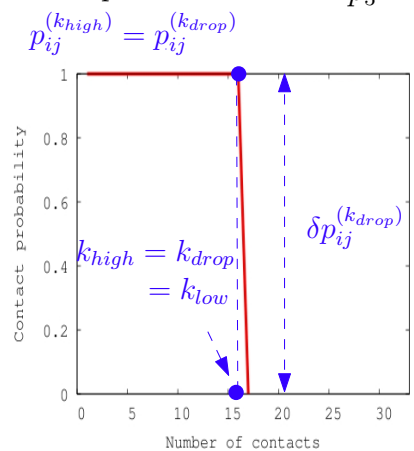
  - Consider:
    the Hasse diagram for $\lambda \in [0, \lambda_{\max}]$
    a stoichiometry $k \geq 1$

  - Define: $\lambda_k(P_i, P_j)$: smallest λ
    $\exists$ k contacts between P_i and P_j

  - **Contact proba.**: $p_{ij}^{(1)} = \lambda_{\max} - \lambda_1(P_i, P_j)/\lambda_{\max}$

  - **Contact curve**: $p_{ij}^{(k)}$ as a function of k

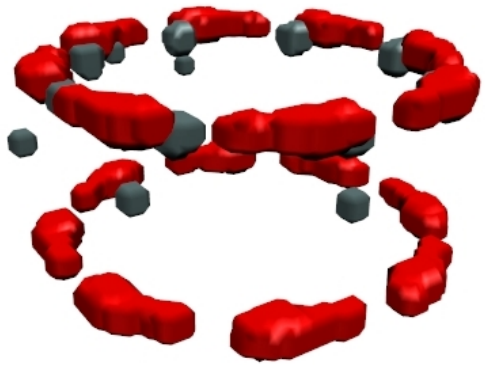




Contact Frequencies versus Contact Probabilities: Results

▷ **Under-represented contact in Sali et al:**

$Nup84 - Nup60 : f_{ij} = 0.07$

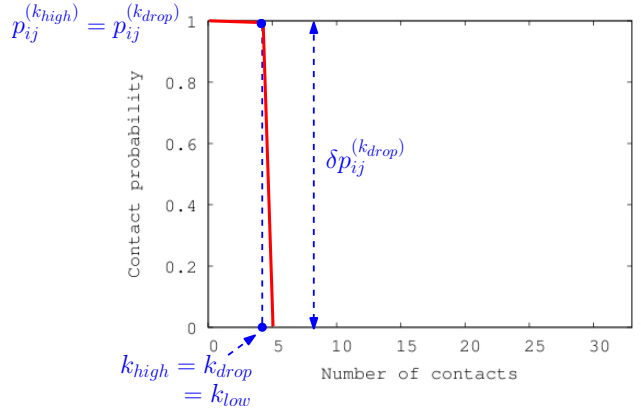


▷ **Over-represented contact in Sali et al:**

$Nup192 - Pom152 : f_{ij} = 0.98$

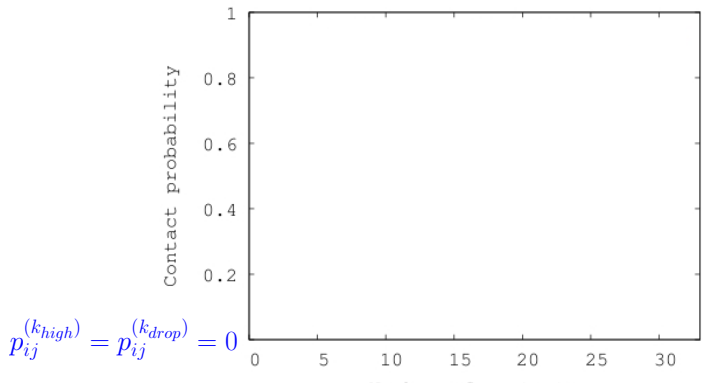
▷ **Corresponding contact curve:**

$Nup84 - Nup60 : p_{ij}^{(4)} = 1$



▷ **Corresponding contact curve:**

$Nup192 - Pom152 : p_{ij}^{(1)} = 0$



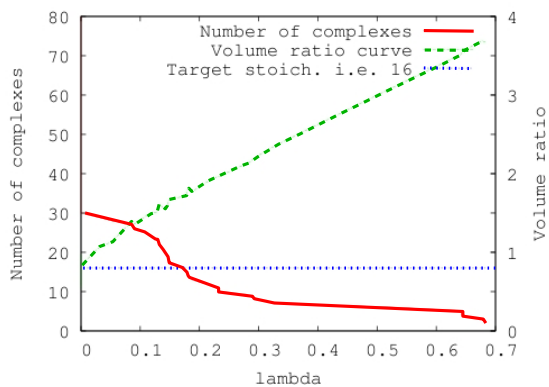
Three Analysis of the Toleranced Model of an Assembly

- **Local:**
  - Tracking copies of sub-complexes in the assembly
    $\rightarrow$ **Hasse diagram**
- **Global:**
  - Inspecting pairwise protein contacts
    $\rightarrow$ **Contact probabilities**
  - Controlling the volume of merging complexes
    $\rightarrow$ **Volume ratio**

Assessment w.r.t. a Set of Protein Types: Volume Ratios

- Definition:
  - **Reference volume** of
    - **a protein:** volume estimated from its sequence of amino-acids
    - **a complex:** sum of reference volumes of its constituting proteins

- Output, per complex:
  - **volume ratio**: volume occupied vs. expected volume
- Output, in conjunction with the Hasse diagram:
  - **curve**: evolution of volume ratio of evolving complexes

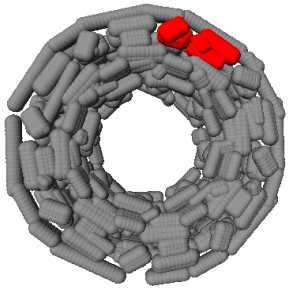


Complexes in the Hasse diagram: variation of the volume ratio as a function of λ

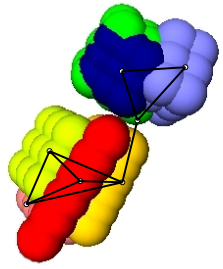
Ref: Harpaz, Gerstein, Chothia; Structure; 1994

Modeling Contacts in Macro-molecular Assemblies

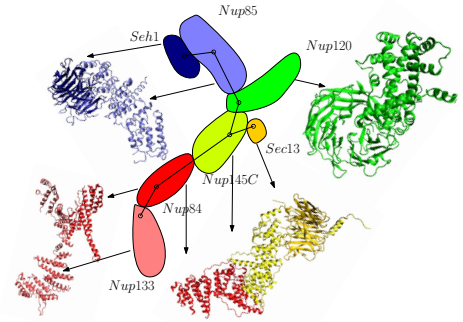
Assessing a Toleranced Model with Respect to a High-resolution Structural Model



Assembly



Complex: skeleton graph

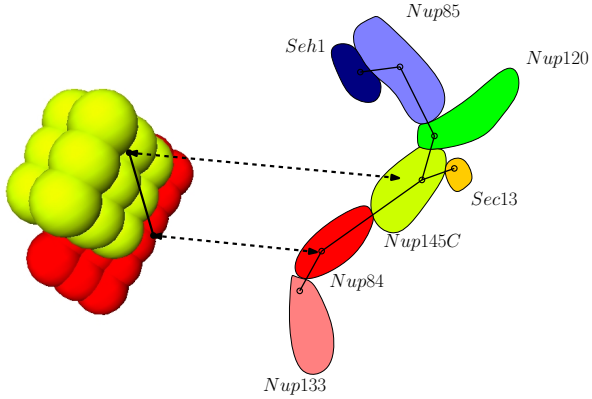


Template: skeleton graph

Matching between a Complex and a Template:

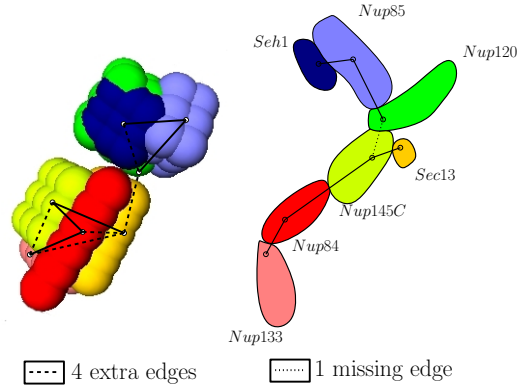
Protein instance $\leftrightarrow$ Protein type

Contact $\leftrightarrow$ Contact



Exact superposition:

Perfect Matching



Approximate superposition:

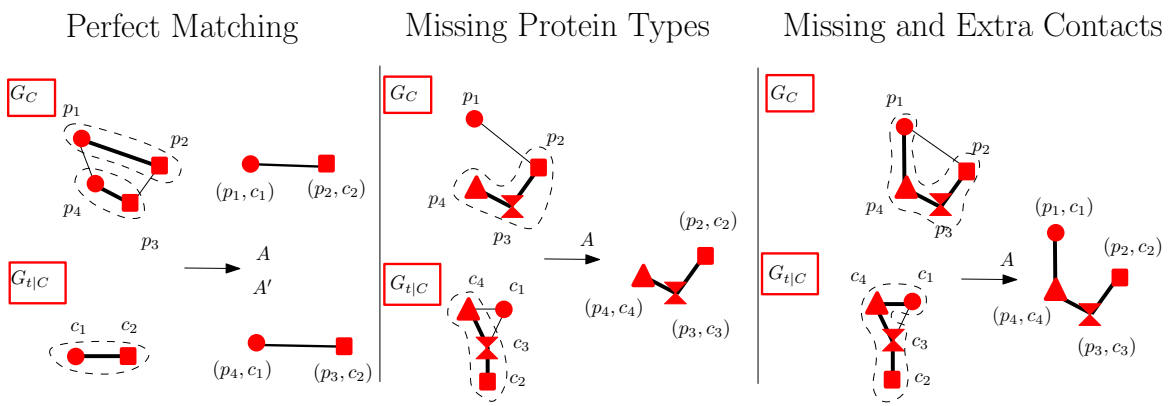
Alternate Matching

Assessment w.r.t. a High-resolution Structural Model: Contact Analysis

- Input: **two skeleton graphs**
  - template G_t, the red proteins : contacts within an atomic resolution model
  - complex G_C: skeleton graph of a complex of a node of the Hasse diagram

- Output: graph comparison, complex G_C versus template G_t:
  (common/missing/extra) $\times$ (proteins/contacts)

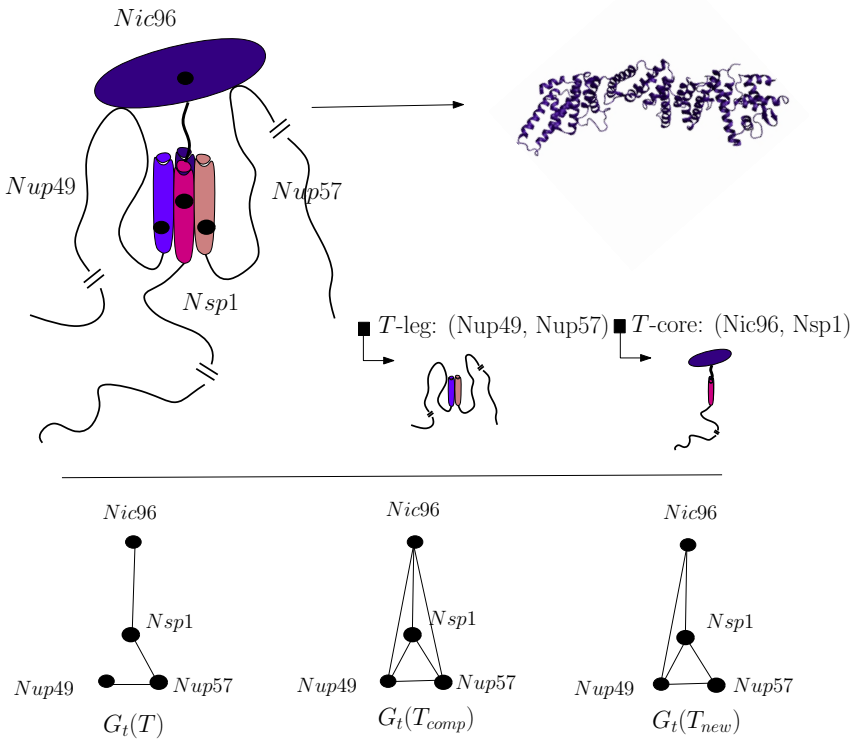
- Graph theory problems:
  Perfect Matching: All Maximal Common Induced Sub-graphs **(MCIS)**
  Alternate Matching: All Maximal Common Edge Sub-graphs **(MCES)**



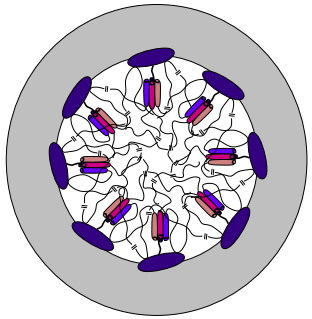
▷Ref: Cazals, Karande; Theoretical Computer Science; 349 (3), 2005
▷Ref: Koch; Theoretical Computer Science; 250 (1-2), 2001

A New Template for the T-complex

- **T-complex and its skeletons**
  Note the filaments



- **Putative positions**
  wrt the inner ring of the NPC



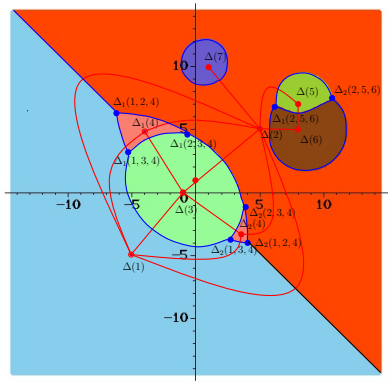
- **Perfect Matching:**
  - $G_t(T)$: 0 matching with T-complex
    $\rightarrow$ **Extra contacts** (Nup49, Nsp1)
  - $G_t(T_{comp})$: 2 matching with T-complex
    $\rightarrow$ **Missing contacts** (Nup57, Nic96)
  - $G_t(T_{new})$: 10 matching with T-complex
    $\rightarrow$ **Best coherence** with toleranced model

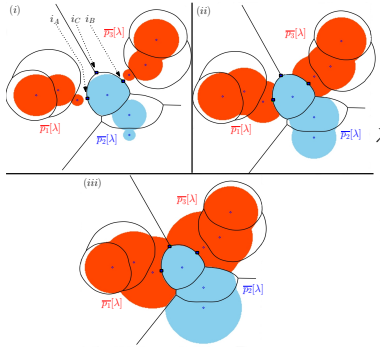
- **Contact analysis**: asymmetric role of Nup49 and Nup57; new template

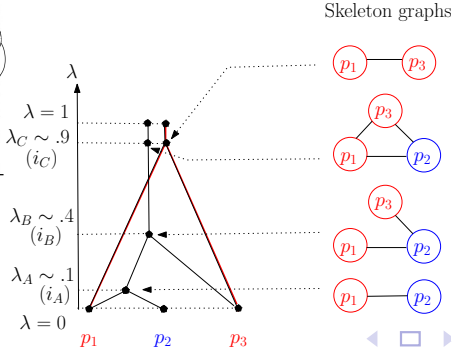
Modeling Contacts in Macro-molecular Assemblies

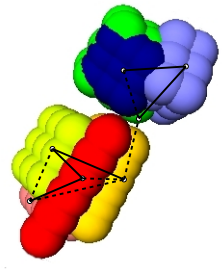
Conclusion and Outlook

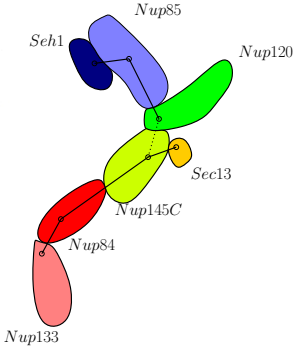
- **Compoundly Weighted Voronoi diagram**
  - Geometric and topological analysis
  - Output sensitive algorithm
  - λ-complex and its computation
- **Toleranced models and their applications**
  - Representing models with uncertainties
  - Bridging the gap *global - fuzzy versus local - atomic resolution* models
- **Reconstruction assessment**
  - A panoply of tools to perform the assessment of large protein assembly models
  - . . . of interest in a virtuous loop reconstruction – assessment
- **Software**
  - Algorithms to compute the CW diagram and the λ-complex (CGAL-style)
  - A generic C++ library for modeling and assessing large assemblies











4 extra edges 1 missing edge

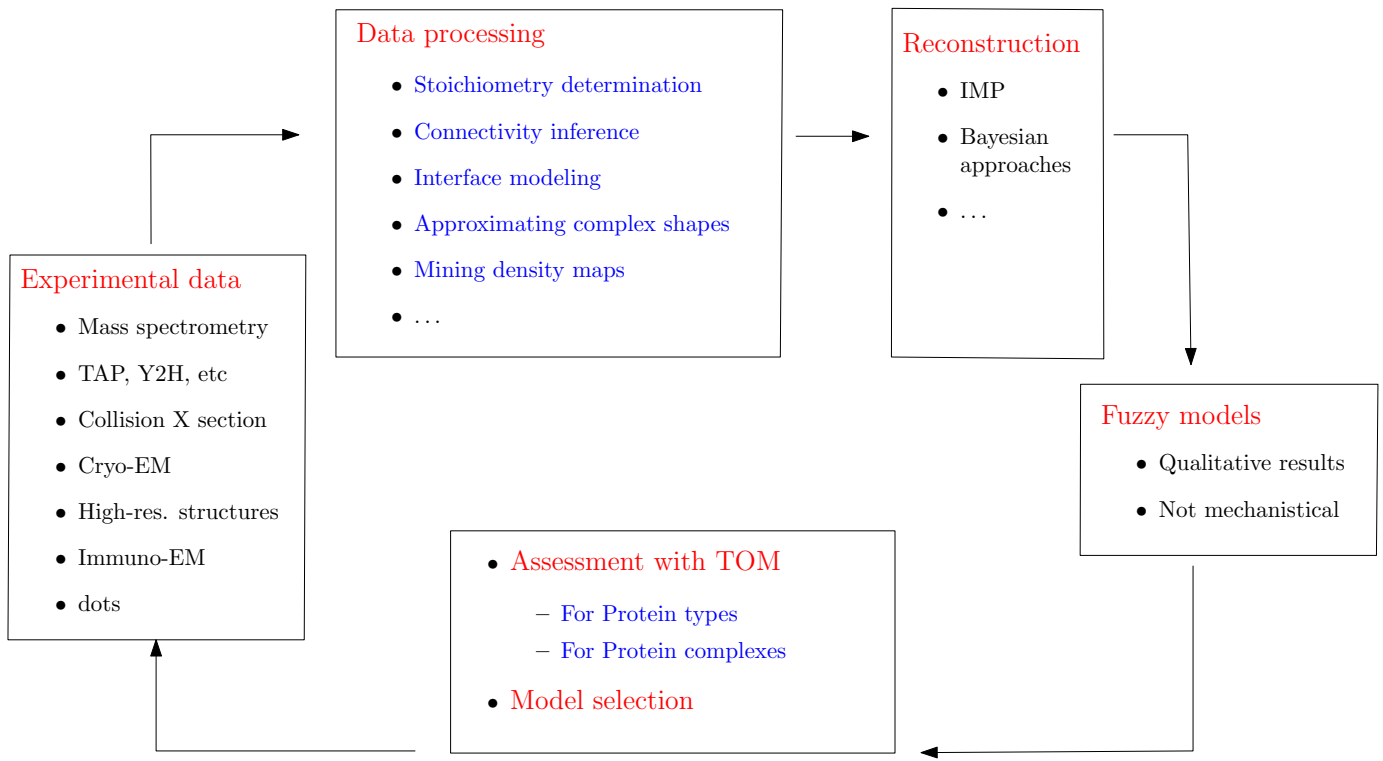
Perspectives

- **Compoundly Weighted Voronoi diagram**
  - Study of homological features (Euler characteristic)
  - Faster computation (Incremental algorithm)
- **Toleranced models**
  - Enhanced approximation of protein shapes
  - Interest of other non linear growth models (e.g Mobius)
- **Applications**
  - Toleranced models in a different context (e.g, cryoEM or crystal structures)
  - Reconstruction by data integration and model selection

Toleranced Models for Large Assemblies: Positioning

- ▷ Methodology: modeling with uncertainties
  - Toleranced models: continuum of shapes vs fixed shapes
  - Topological and geometric stability assessment (curved α-shapes)
- ▷ Applications to toleranced complexes
  - Protein types (contact probabilities)
  - Protein complexes (morphology, contacts)

http://team.inria.fr/abs



References

- Modeling Macro-molecular Complexes : a Journey Across Scales, in *Modeling in Computational Biology and Biomedicine: a Multi-disciplinary Endeavor*, F. Cazals and P. Kornprost Editors, Springer, 2012.
- Multi-scale Geometric Modeling of Ambiguous Shapes with Toleranced Balls and Compoundly Weighted alpha-shapes, F. Cazals, Tom Dreyfus, Computer Graphics Forum (SGP) 2010 29(5): 1713–1722.
- Probing a Continuum of Macro-molecular Assembly Models with Graph Templates of Sub-complexes T. Dreyfus, and V. Doye, and F. Cazals Proteins: structure, function, and bioinformatics, 81 (11), 2013.
- Assessing the Reconstruction of Macro-molecular Assemblies with Toleranced Models T. Dreyfus, and V. Doye, and F. Cazals Proteins: structure, function, and bioinformatics, 80 (9), 2012.
- A note on the problem of reporting maximal cliques F. Cazals, and C. Karande Theoretical Computer Science, 407 (1–3), 2008.

Overview

PART 1:Connectivity Inference from Native Mass Spectrometry Data

PART 2:Building Coarse Grain Models

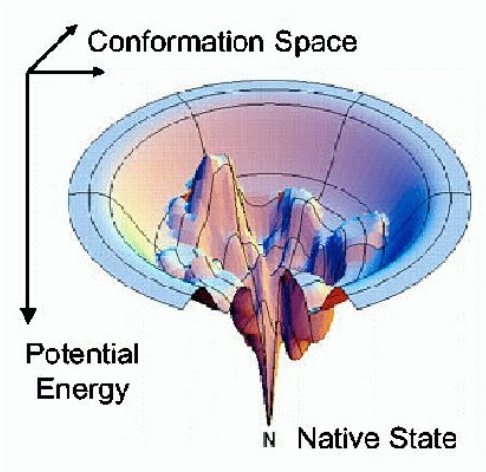
PART 3:Handling uncertainties in Macro-molecular Assembly Models

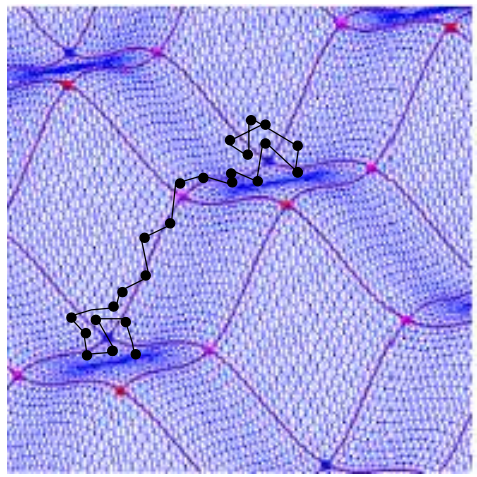
PART 4:Conformational Ensembles and Energy Landscapes: Analysis

PART 5:Conformational Ensembles and Energy Landscapes: Comparison

Conformational Ensembles and Energy Landscapes: Analysis

F. Cazals, A. Roth, T. Dreyfus
C. Robert, IBPC / CNRS





informatics mathematics

Modeling Contacts in Macro-molecular Assemblies

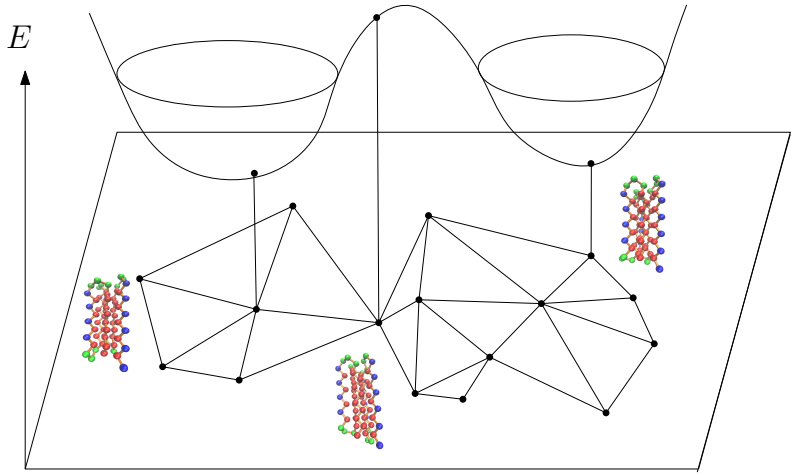
Landscapes: Intuitions

Example Test System: BLN69

Landscapes: Multiscale Topographical Analysis

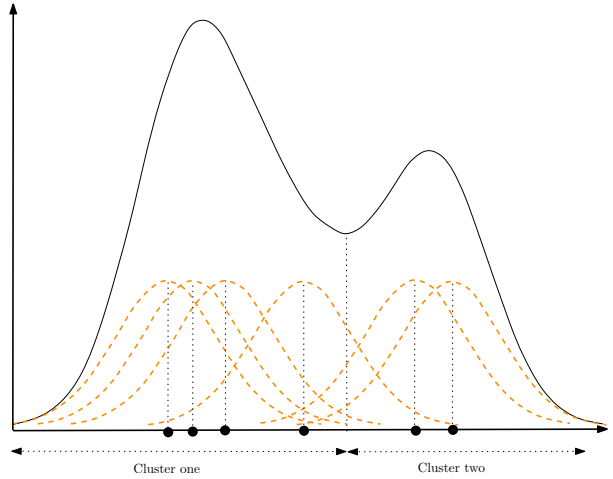
Analyzing Landscapes

▷ Energy landscape



- Input: point set + energies
- Output: minima, saddles, attraction basins

▷ Density estimates



- Input: point set
- Output: one cluster per significant local maximum

▷ Common points:

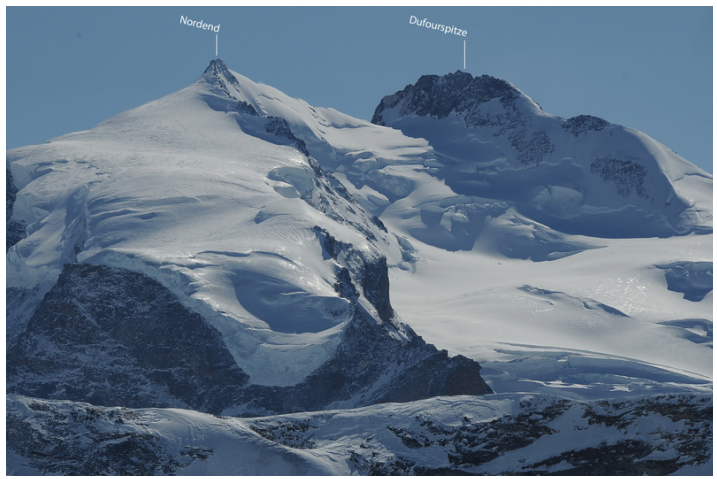
- Input consists of a set of points / conformations
- The elevation defines a landscape
- Neighbors used to define a graph / estimate a density

Landscapes and Peaks: What is a Peak !?

- 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`

Modeling Contacts in Macro-molecular Assemblies

Landscapes: Intuitions

Example Test System: BLN69

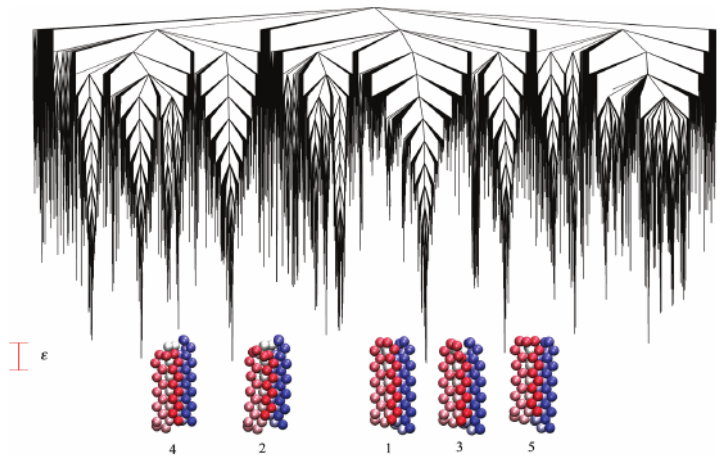
Landscapes: Multiscale Topographical Analysis

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

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

- Disconnectivity graph describing merge events between basins

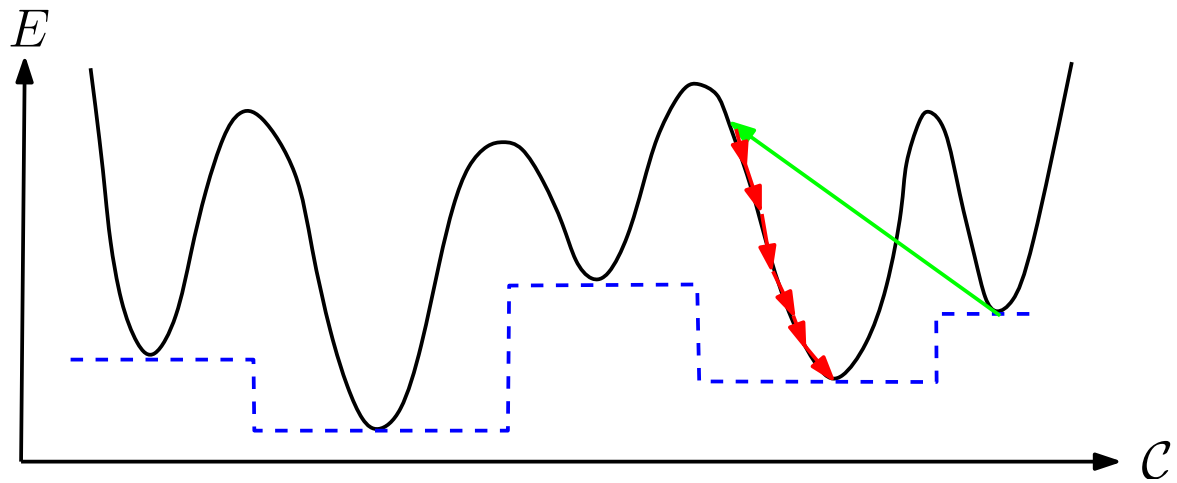


Ref: Oakley, Wales, Johnston, J. Phys. Chem., 2011

Sampling the PEL using Numerical Methods

The Example of Basin-Hoppinp

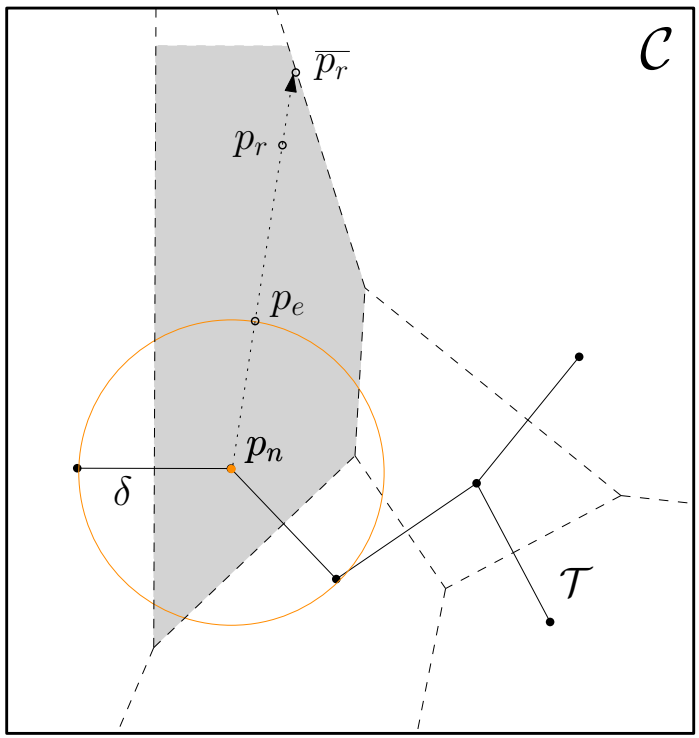
- Basin-hopping and the basin hopping transform
  - Random walk in the space of local minima
  - Requires a *move set* and an *acceptance test* (cf Metropolis) and the ability to descend the gradient

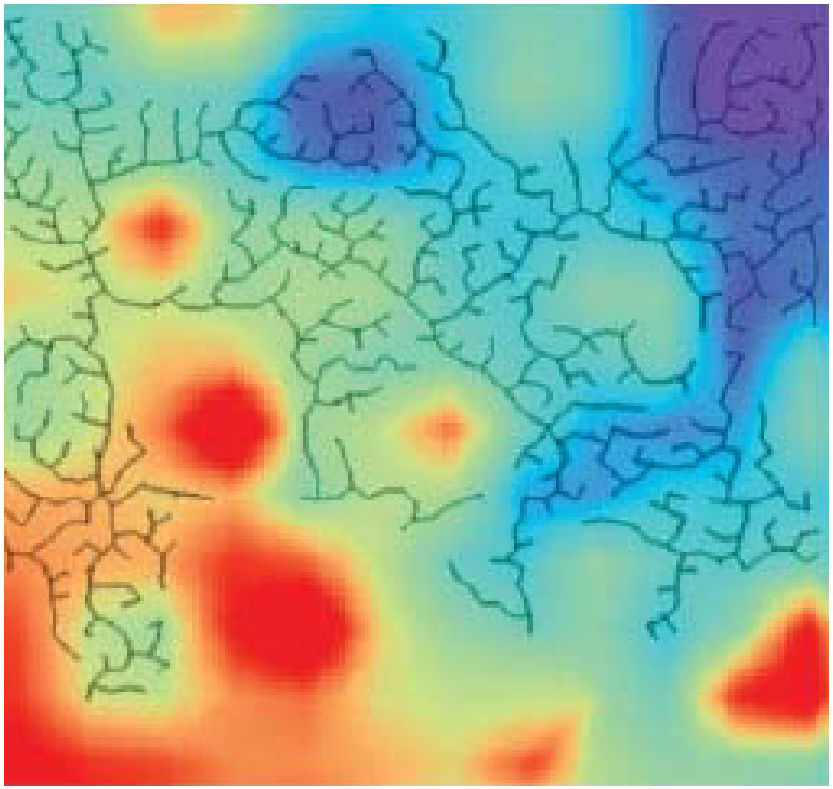


Ref: Schön and Jansen, Prediction, determination and validation of phase diagrams via the global study of energy landscapes, Int' J. of Materials Research, 2009

Landscape Exploration: Transition based Rapidly Growing Random Tree (T-RRT)

- 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

Modeling Contacts in Macro-molecular Assemblies

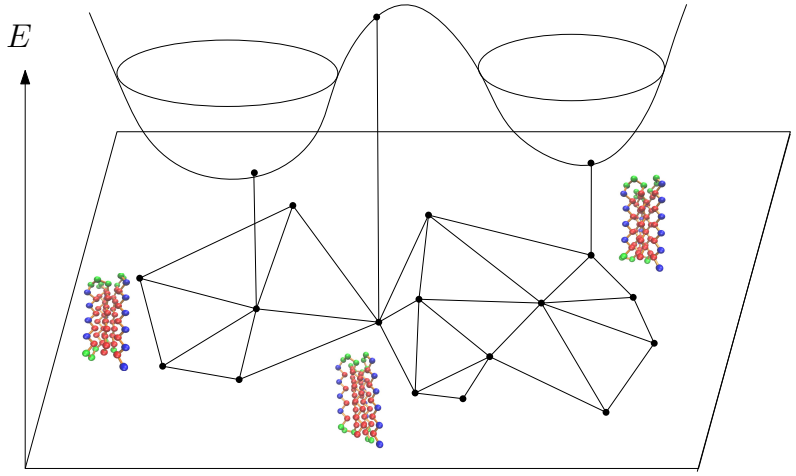
Landscapes: Intuitions

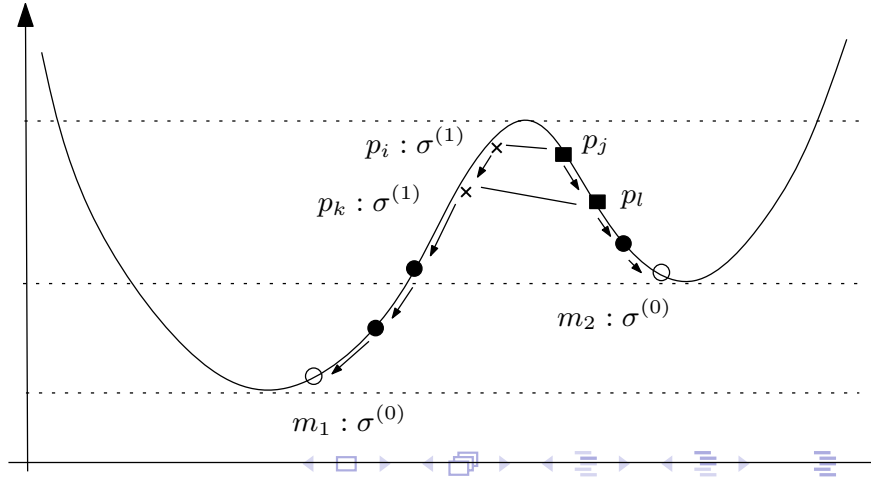
Example Test System: BLN69

Landscapes: Multiscale Topographical Analysis

Representing Sampled Landscapes

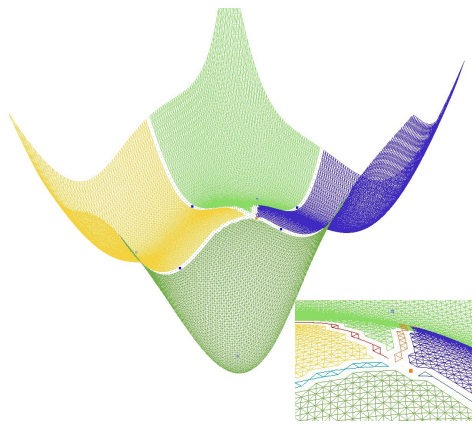
- ▷ Ground space: conformational space
- ▷ Elevation: potential energy / score
- ▷ Nearest neighbor graph (NNG)
  - connect each sample to its k-nearest neighbors (l-RMSD)
  - faces the curse of dimensionality . . . yet, strategies to fudge around data structures to handle NN queries in metric spaces
- ▷ Pseudo-gradient vector field: oriented NNG i.e. connect each sample to its highest neighbor





Energy Landscape Analysis: Morse Sketching

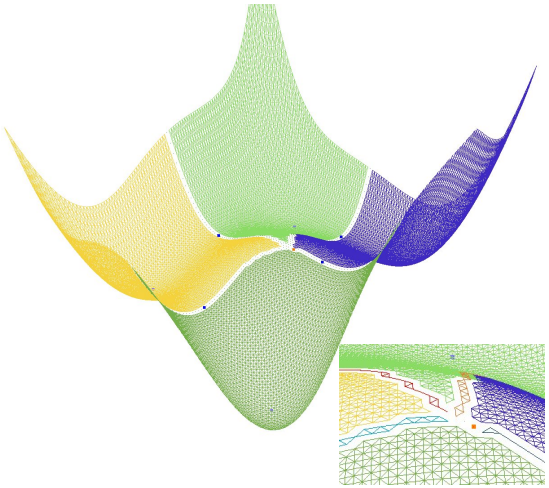
- Input:
  - a collection of conformations $\{c_i\}$
  - or better: samples and the associated local minima. But ...
    - requires the gradient of the energy / score
    - or derivative free optimization methods (CMA-ES)
- Output:
  - Transition graph connecting minima and saddles
  - Basins associated with local minima
- Method:
  - Simulate a gradient descent from each point
  - Identify *ridges across* basins, aka bifurcations



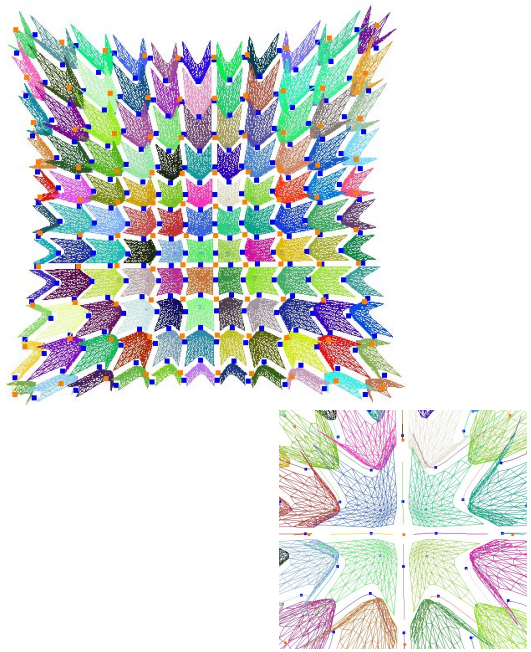
Critical Points and Stable Manifolds
Illustrations for functions $z = f(x, y)$

- Following the pseudo-gradient yields:
  - Local minima
  - Stable manifold of local minima: points flowing to local minima
  - Index one saddles

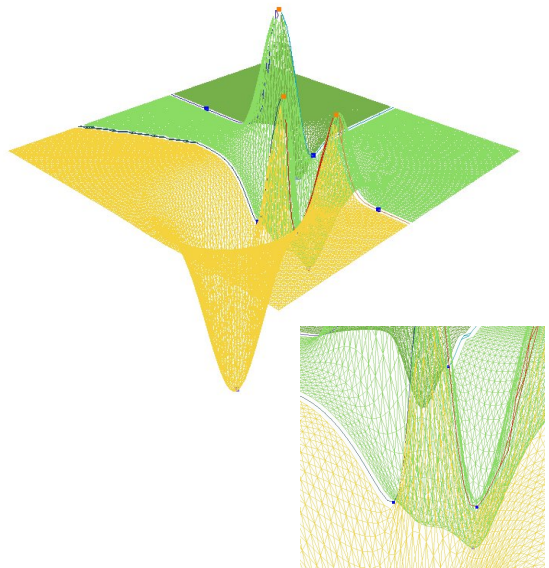
- Himmelblau
  (4,4,1)



- Rastrigin
  (121,220,100)

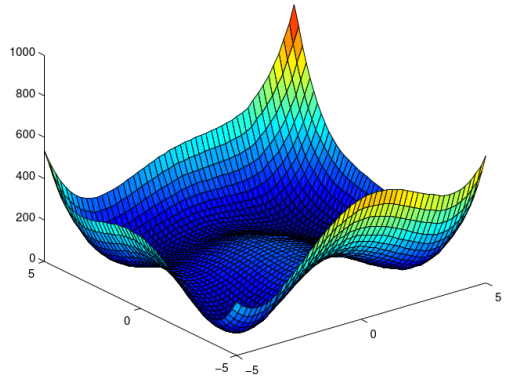


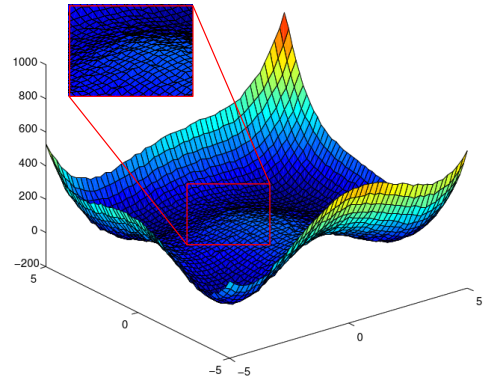
- Gauss6a
  (3,5,3)

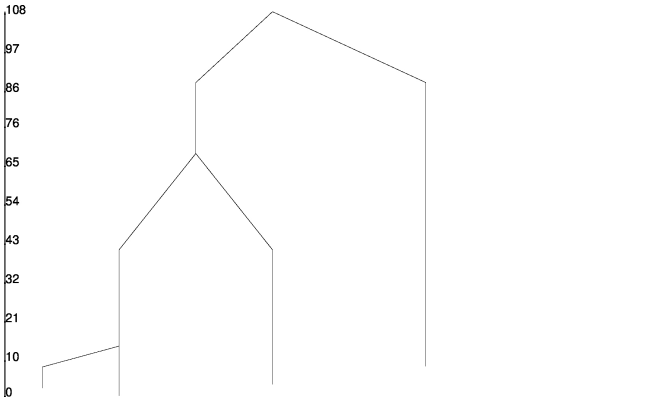


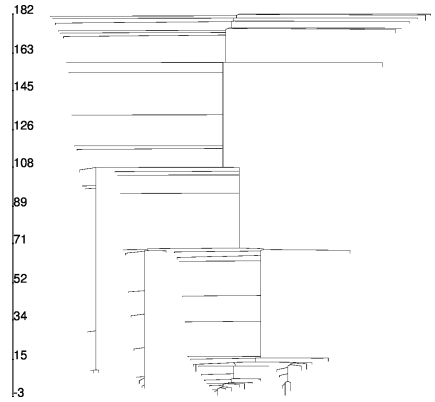
Landscape Analysis at a Glimpse:

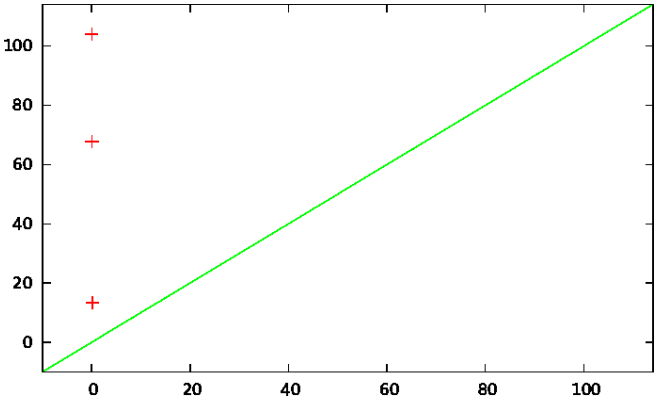
The Himmelblau function: $f(x, y) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2$

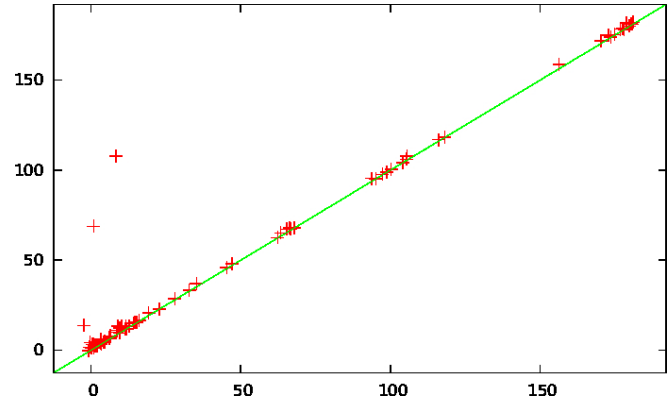






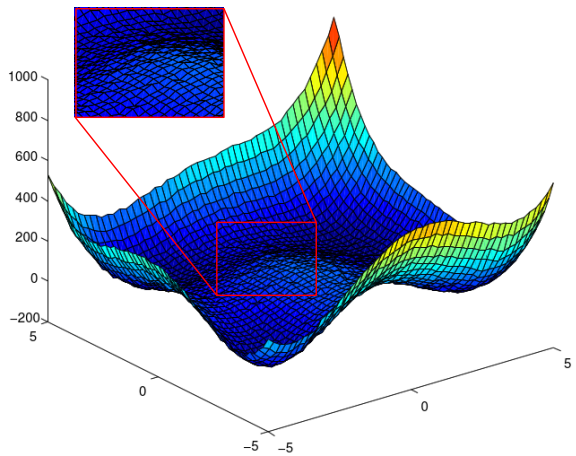




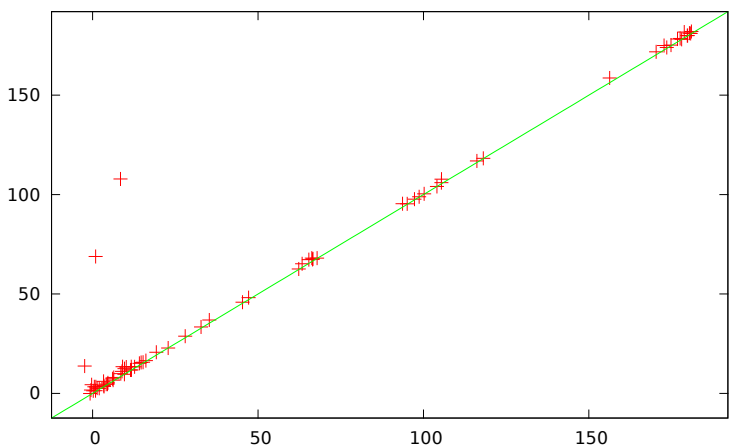


Sweeping a landscape yields: Persistence Diagram and the Disconnectivity Graph

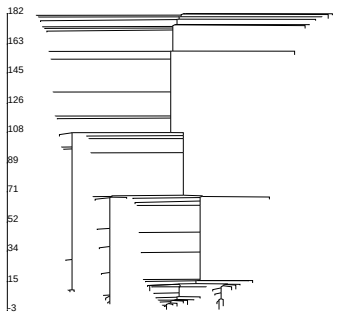
▷ Toy noisy landscape

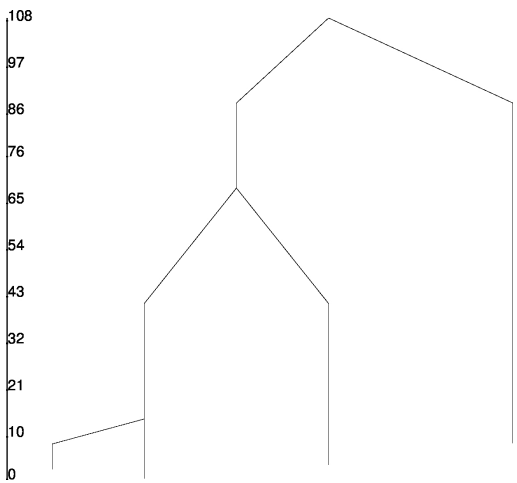


▷ Persistence diagram for sub-level sets



▷ Disconnectivity graph:
noisy and simplified

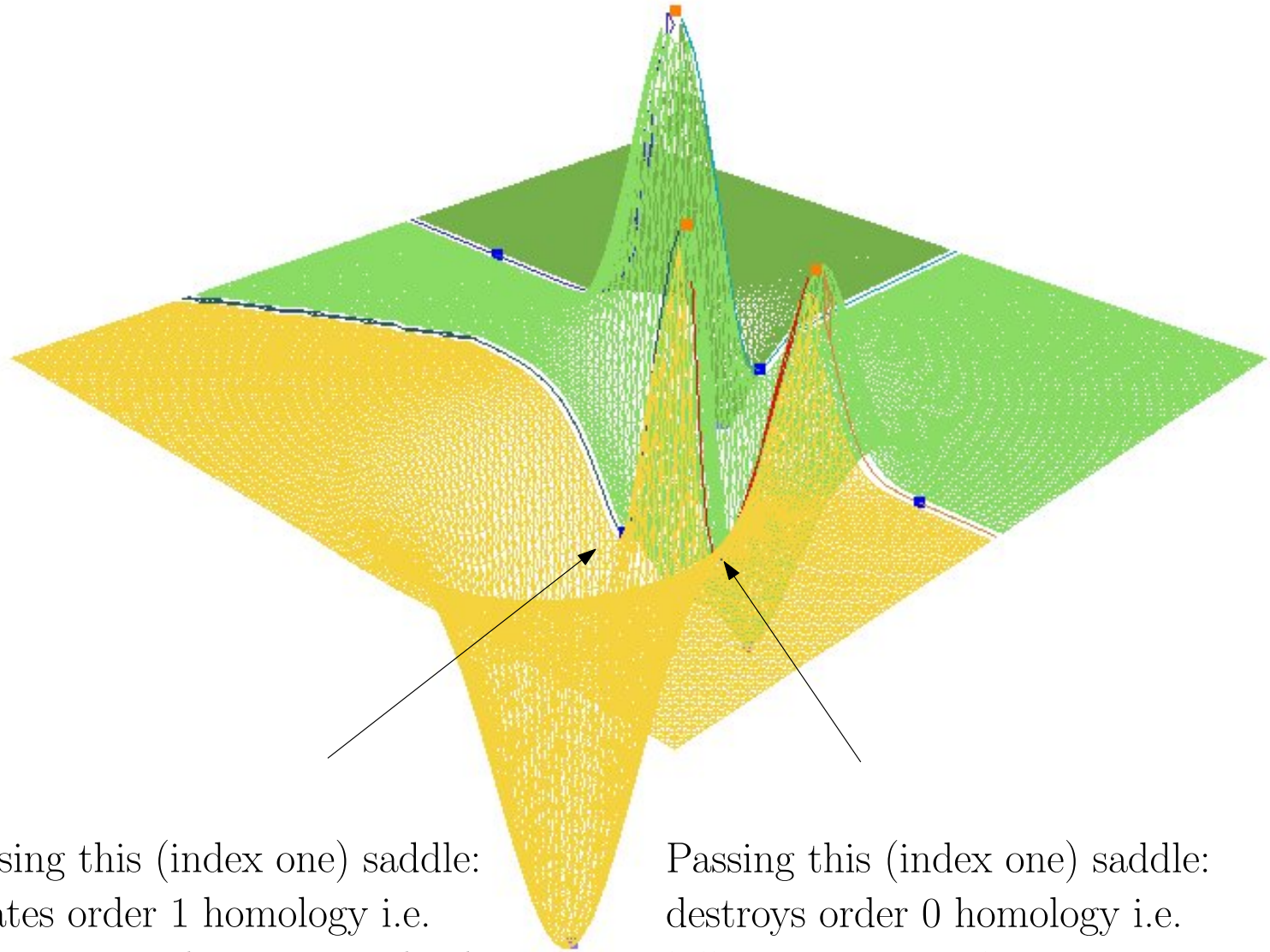




▷Ref: Chazal et al, ACM SoCG; 2011

▷Ref: Cazals, Cohen-Steiner; Comput. Geometry Th. & Appl.; 2011

Morse Theory:
Destruction and Creation of Homology Generators



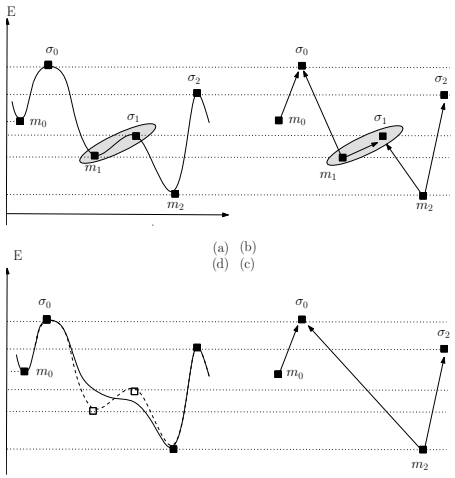
Passing this (index one) saddle: creates order 1 homology i.e. creates one loop around the mountain

Passing this (index one) saddle: destroys order 0 homology i.e. kills one connected component

Persistence, Simplification and Transition Paths $(\min, \sigma, \min)$

a.k.a. the re-routing algorithm

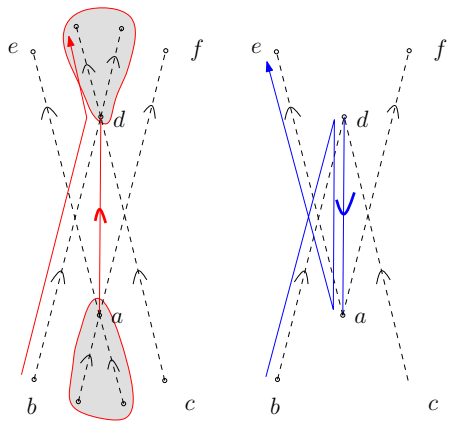
▷ Landscape simplification from the Morse-Smale chain complex



– The cc of a min dies upon encountering the *nearest* saddle

– News paths upon simplif: $(\mathit{min}, \sigma, \mathit{min})$
min: minima accessible from dead saddle

▷ Key operations: multiplexing and redistribution of stable manifolds



▷ Simplifying: reverting the flow
→ re-routing paths (in codimension one)

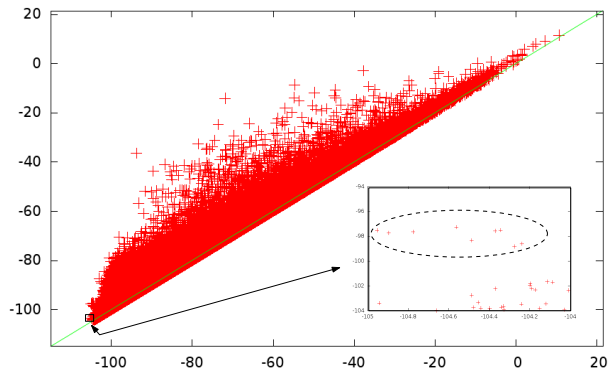
▷ Output:
– simplified Hasse diagram / persistence diagram
– stable basins partitioning the samples
– transition paths across stable basins

▷Ref: Cazals, Cohen-Steiner; Comput. Geometry Th. & Appl.; 2011

BLN69: Persistence reveals Novel Local Minima

- Selection of local minima m_i of interest by energy and persistence:
  - Range on energy: $m_i \in$ sub-level set $E \leq h$
    NB: High energies unlikely at room temperature
  - Upper bound on persistence: barriers of max. height δh

- Persistence of the 458,082 local minima in BLN69-all

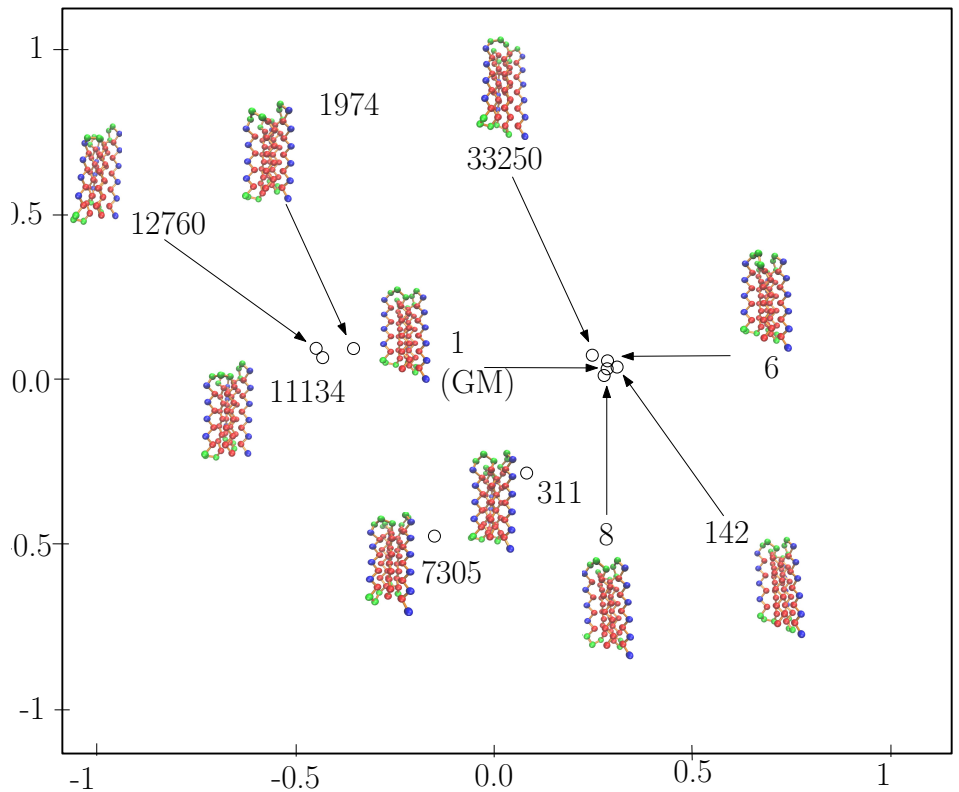


  - Inset: range query on energy and persistence
    40 minima in BLN69-all with energy $E < -104\epsilon$
    The 10 most persistent minima: 6 known + 4 new ones

- Ref: Cazals et al, under revision

BLN69: Dimensionality Reduction Reveals the Relative Positions of Low Handing Minima

- A three step process:
  - Step 0: select local minima of interest
  - Step 1: compute pairwise distances (lRMSD in ambiant space, or cumulative lRMSD on the graph of nearest neighbors
  - Step 2: apply dimensionality reduction, say Multidimensional Scaling



References

- Persistence-based clustering in Riemannian manifolds, F. Chazal and L. Guibas and S. Oudot and P. Skraba, ACM SoCG 2011.
- Reconstructing 3D compact sets, F. Cazals and D. Cohen-Steiner, CGTA, 2011.
- Conformational Ensembles and Sampled Energy Landscapes: Analysis and Comparison, F. Cazals and T. Dreyfus and D. Mazauric and A. Roth and C. Robert. Under revision, `https://hal.archives-ouvertes.fr/hal-01076317`

Overview

PART 1:Connectivity Inference from Native Mass Spectrometry Data

PART 2:Building Coarse Grain Models

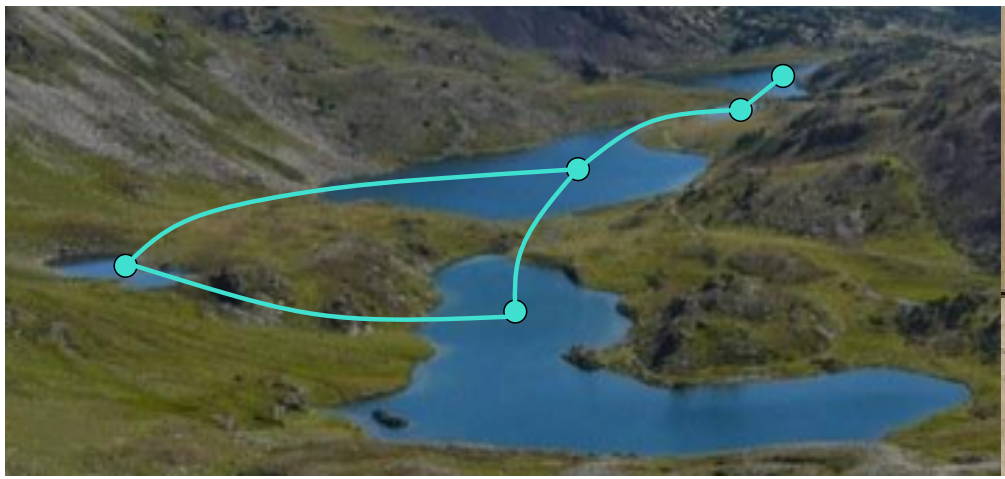
PART 3:Handling uncertainties in Macro-molecular Assembly Models

PART 4:Conformational Ensembles and Energy Landscapes: Analysis

PART 5:Conformational Ensembles and Energy Landscapes: Comparison

Sampled Energy Landscapes: Comparison

F. Cazals, D. Mazauric



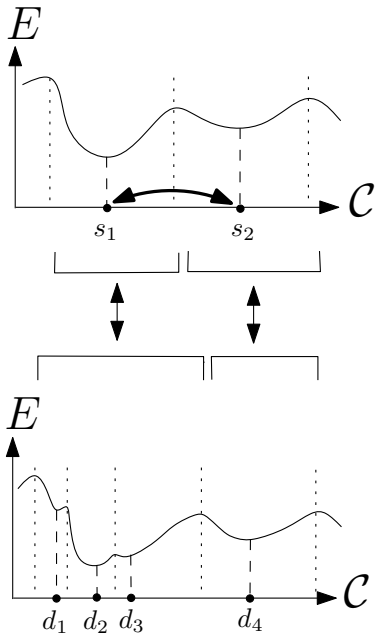
Modeling Contacts in Macro-molecular Assemblies

Algorithms

Results

Comparing (Sampled) Energy Landscapes: Motivation

- Comparing (sampled) landscapes:
  - Assessing the coherence of two force fields for a given system (atomic,CG)
  - Comparing two related systems: protein wild type/mutated
  - Comparing two simulations: different initial conditions, algorithms



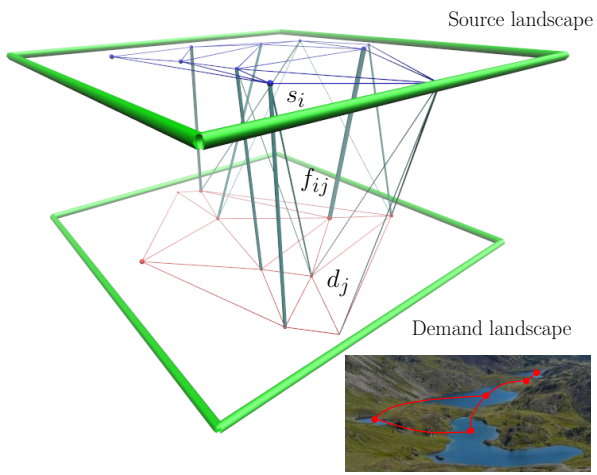
- Idea: find a mapping between basins considering
  - the similarity between the *native states* (one per basin)
  - the coherence between the *volumes* of the basins (their probabilities)

- NB: Terminology: sampled potential energy landscape:
vertex weighted transition graph associated with a simulation,
i.e. the subgraph of the whole transition graph *revealed* by the simulation.

Comparing (Sampled) Energy Landscapes via Their Transition Graphs

- Input: given a source landscape PEL_s and a demand landscape PEL_d
- Sampled landscape modeled as a transition graph:
  - One conformation per basin: $s_i \in \mathrm{PEL}_s$, $d_j \in \mathrm{PEL}_d$
    $+$ a metric $d_{\mathcal{C}}$ between conformations
  - One probability per basin
    $$w_i^{(s)} = \int_{B_i} (\exp \frac{-V(c)}{k_B T} dc)/Z, \quad \sum_i w_{i=1,\ldots,n_s}^{(s)} = 1$$
  - Transitions between basins

- Output: *transport plan* i.e. *flow quantities* f_{ij}
  f_{ij}: amount (of probability) flowing from basin $i \in \mathrm{PEL}_s$ to basin $j \in \mathrm{PEL}_d$



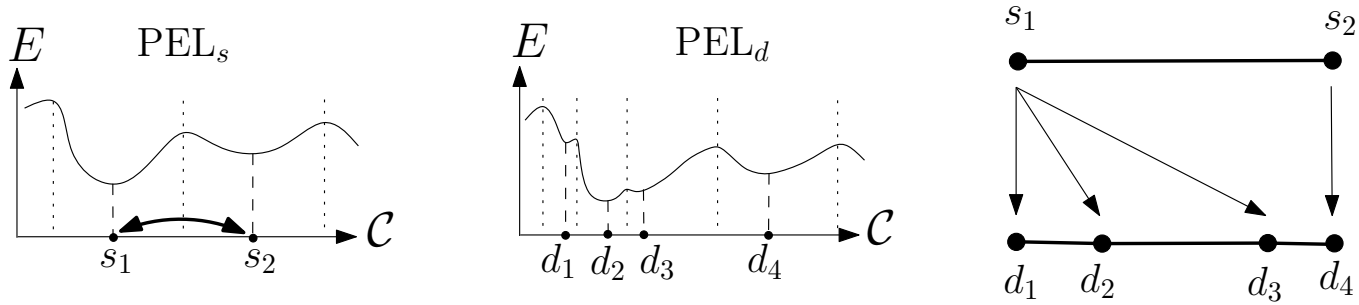
NB: the transport plan is a *mapping* between basins; it induces a *transport cost* (a distance) between landscapes.

Coding a Sampled Landscape into a Transition Graph

- Step 1: Morse sketching yields a *transition graph*:
  - Basins and their weights
  - Transitions between these basins
- Step 2: landscape simplification with topological persistence:
merge basins with non-significant *barrier heights* into more stable basins
- Step 3: assign masses to the remaining minima: yields a vertex weighted transition graph

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

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

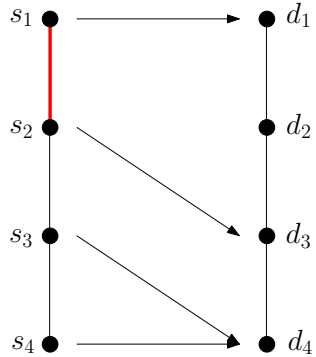
- Pros and cons:
  - Information used: location of minima, weight of basins
  - Linear program: solved in polynomial time
  - Connectivity information not used

- Ref: Rubner, Tomasi, Guibas, IJCV, 2000

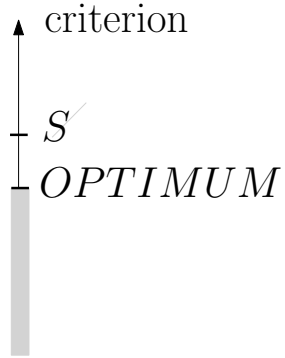
Checkpoint

Comparisons involving Connectivity Constraints

▷ EMD: may violate the connectivity constraints



▷ Hardness



▷ Problem Earth-Mover-Distance with connectivity constraints (`EMD-CC`):

Find the least cost transport plan such that every connected subgraph of PEL_s exports towards a connected subgraph of PEL_d

▷ Our results

- Decision problem is NP-complete (reduction: 3-partition problem)
- Optimization problem is not in APX
  If **P** $\neq$ **NP**: no polynomial algorithm with constant approx factor
- Yet: greedy polynomial algorithm producing admissible solutions

▷ Algorithms `Alg-EMD-LP` versus `Alg-EMD-CCC-G`:

`Alg-EMD-LP`: fast, but may violate connectivity constraints

`Alg-EMD-CCC-G`: slower, but respects connectivity constraints

▷`Ref: Cazals, Mazauric; submitted`

Modeling Contacts in Macro-molecular Assemblies

Algorithms

Results

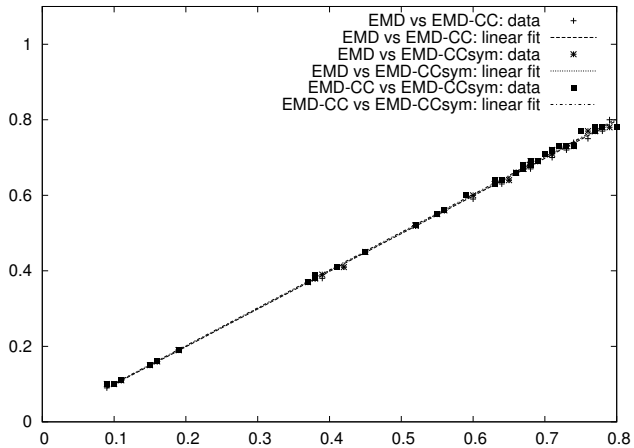
BLN69: `Alg-EMD-LP` and `Alg-EMD-CCC-G` Connectivity versus Demand Satisfaction

- Protocol:
  - for each of the 10 lowest local minima: one simulation of 10^4 samples
  - data processing yields transition graphs of varying size
    $\#V \in [27, 439], \#E \in [439, 1672]$
  - for each pair of landscapes (A, B) out of the 45 pairs:
    computation of Alg-EMD-LP(A, B),
    Alg-EMD-CCC-G(A, B), Alg-EMD-CCC-G(B, A)

- Connectivity and demand satisfaction:
  - `Alg-EMD-LP` violates the connectivity constraints: worst-cases are
    constraint satisfied for 41% of the source vertices (100% : perfect)
    constraint satisfied for 24% of the source edges (100% : perfect)
  - `Alg-EMD-CCC-G` almost saturates the demand
    worst-case is 99.23% of the demand

BLN69: `Alg-EMD-LP` and `Alg-EMD-CCC-G` Costs

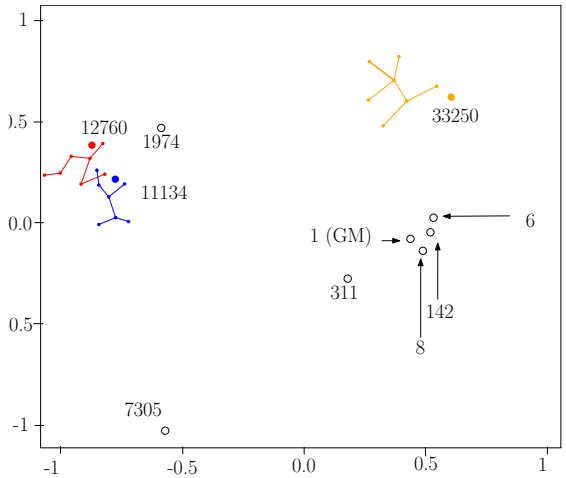
▷ `Alg-EMD-LP` and the two `Alg-EMD-CCC-G` yield identical costs:



- Three comparisons:
  Alg-EMD-LP(A, B)
  Alg-EMD-CCC-G(A, B), Alg-EMD-CCC-G(B, A)
- Linear correlations coeffs ~ 0.99
- `Alg-EMD-CCC-G` does not exhibit significant asymmetry on these cases

▷ Consistence with the relative positions of the local minima

Min distance:
  0.09 for (12760, 1134)
Max distance:
  0.79 for (12760, 33250)
But:
0.19 for (6, 142)



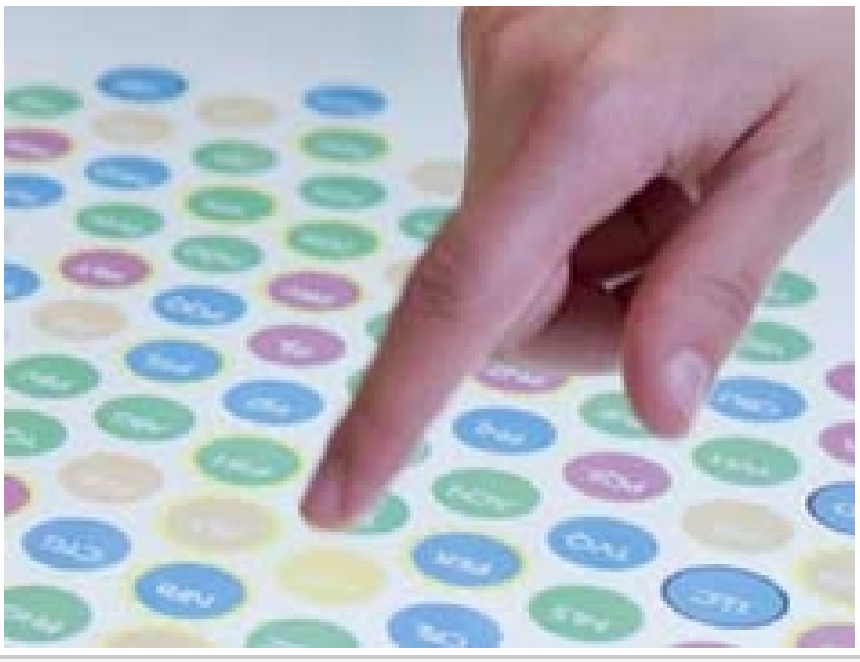
References

- Conformational Ensembles and Sampled Energy Landscapes: Analysis and Comparison, F. Cazals and T. Dreyfus and D. Mazauric and A. Roth and C. Robert. Under revision, `https://hal.archives-ouvertes.fr/hal-01076317`
- Mass Transportation Problems with Connectivity Constraints, with Applications to Energy Landscape Comparison, F. Cazals and D. Mazauric. Submitted.
- A new mallows distance based metric for comparing clusterings, Zhou, Ding and Li, Jia and Zha, Hongyuan, 22nd international conference on Machine learning, 2005.

Positions

Post-doctoral research fellowships

Campaign 2014: post-doctoral positions



In 2014, Inria is offering many post-doctoral positions, each lasting about 12 or 24 months, for holders of a PhD or other doctorate.
The list of subjects, which may exceed the number of vacant positions, will be updated regularly.