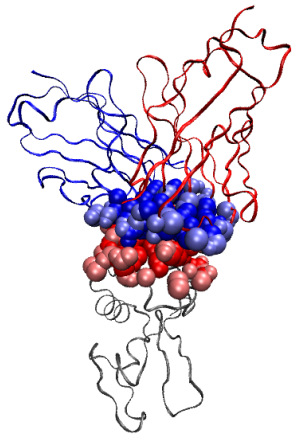
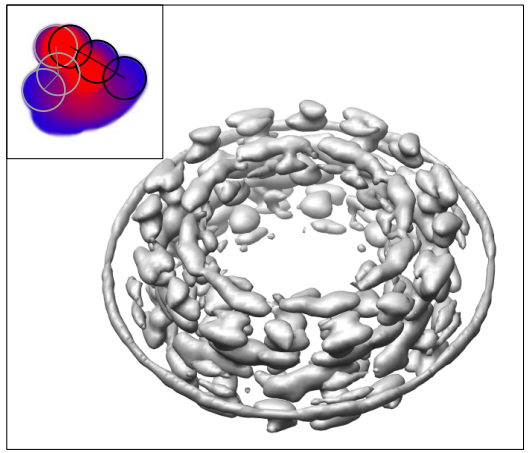
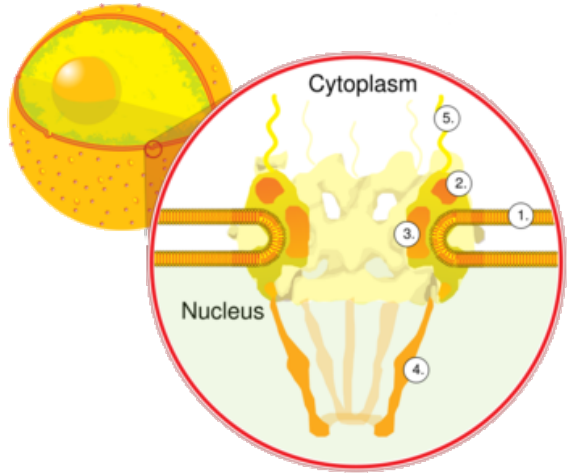
Modeling
Macro-Molecular Complexes and Assemblies

Frederic.Cazals@inria.fr
Algorithms - Biology - Structure
INRIA Sophia-Antipolis







informatics mathematics

Structural Bioinformatics: Early Advent and Positioning

▷ Bioinformatics vs high throughput experimental data

– **Structural data**
  - X ray structures, 1960, J. Kendrew and M. Perutz (Nobel 62)
  - NMR, 1982, K. Wüthrich (Nobel 02); electro-spray, J. Fenn (Nobel 02)
  - cryo-EM, 1975-93, R. Henderson

  ⇒ **Structural bioinformatics: Structure-to-Function paradigm**
– Genomic data
  - DNA sequences, 1977, F. Sanger, Nobel in 80 (and 58!)
  - Genomes on line: `http://www.genomesonline.org/`

  ⇒ Classical comp. biology (alignment, phylogeny, . . .)
– Omics: Transcriptomics, proteomics, Metabolomics etc
  - DNA chips, 1998, after P. Brown and I. Herskowitz

  ⇒ Systems biology

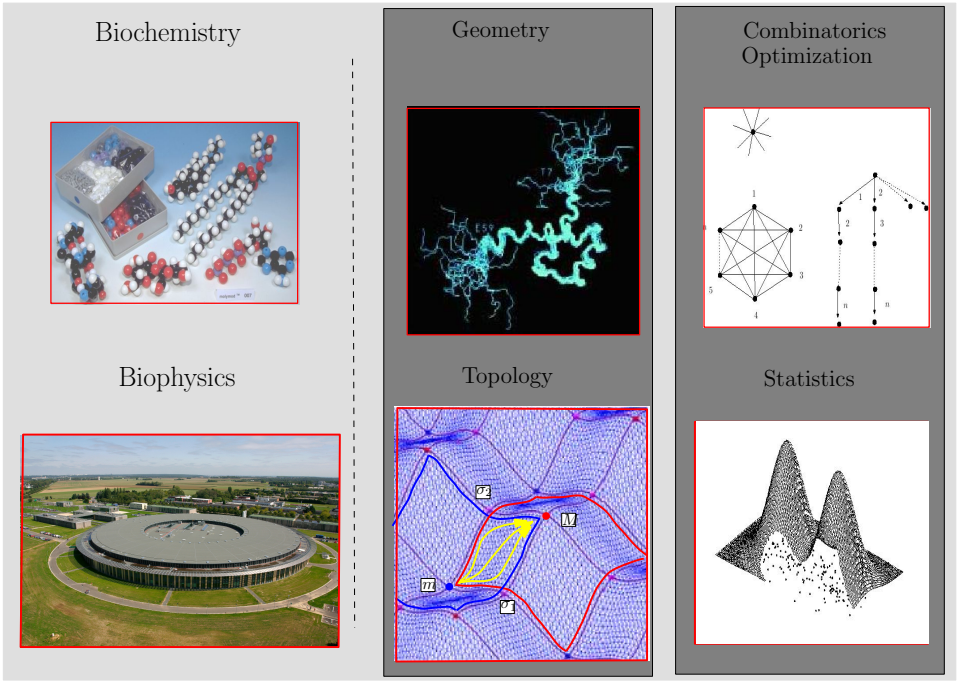
▷ Relative paucity of structures: incentive for modeling

$\sim 80k$ structures (PDB) (`http://www.rcsb.org`)
$\sim 10M$ non redundant sequences (`http://www.ncbi.nlm.nih.gov`)

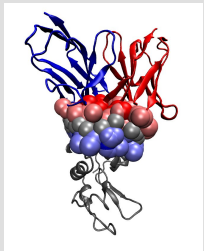
▷Ref: Levitt; PNAS 106; 2009

Our Vision

▷ Experiments and Modeling



Structure-to-Function



Docking (and Folding)

- Improved descriptions
- Improved predictions
  - atomic models (small complexes)
  - coarse models (PPI networks)

▷ Applied challenges

– Modeling protein complexes
  atomic models, large assemblies
– Modeling the flexibility of proteins
(– Systems biology)

▷ Mathematical - algorithmic foundations

– Geometric - topological modeling
  focus on stability analysis
– Graph theory, matching algorithms
– Optimization
– Machine learning
  dimensionality reduction
  statistical tests

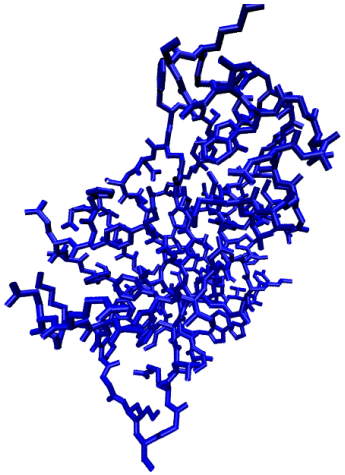
PART 1: Modeling High Resolution Protein Complexes
PART 2: Modeling Large Protein Assemblies
PART 3: Studies in Mass Spectrometry
PART 4: (Selected Algorithmic Details)
PART 5: Conclusion

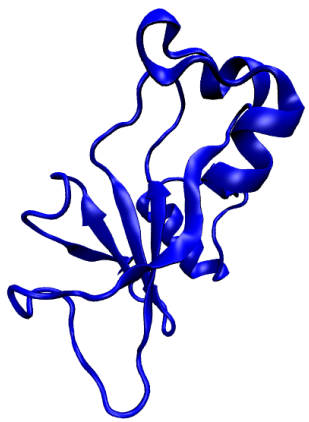
Part I

Modeling High Resolution Protein Complexes

Modeling High Resolution Protein Complexes

Geometry - Topology versus Biophysics: A Matter of Correlations





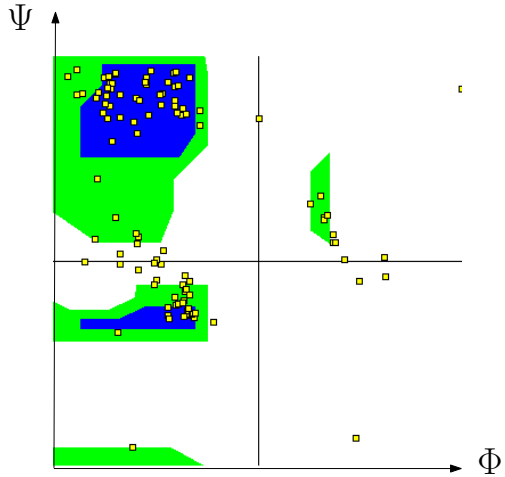


Fig. Hydrogen bonding in antiparallel b sheet

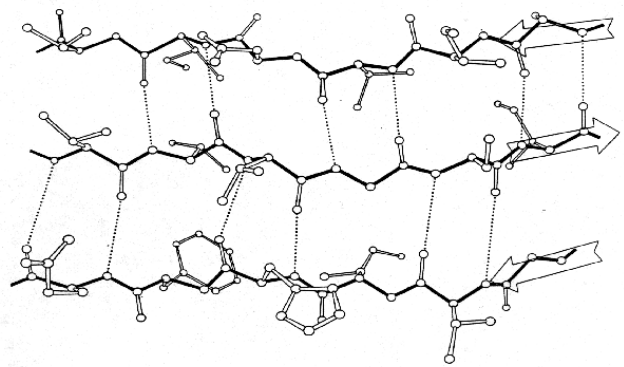
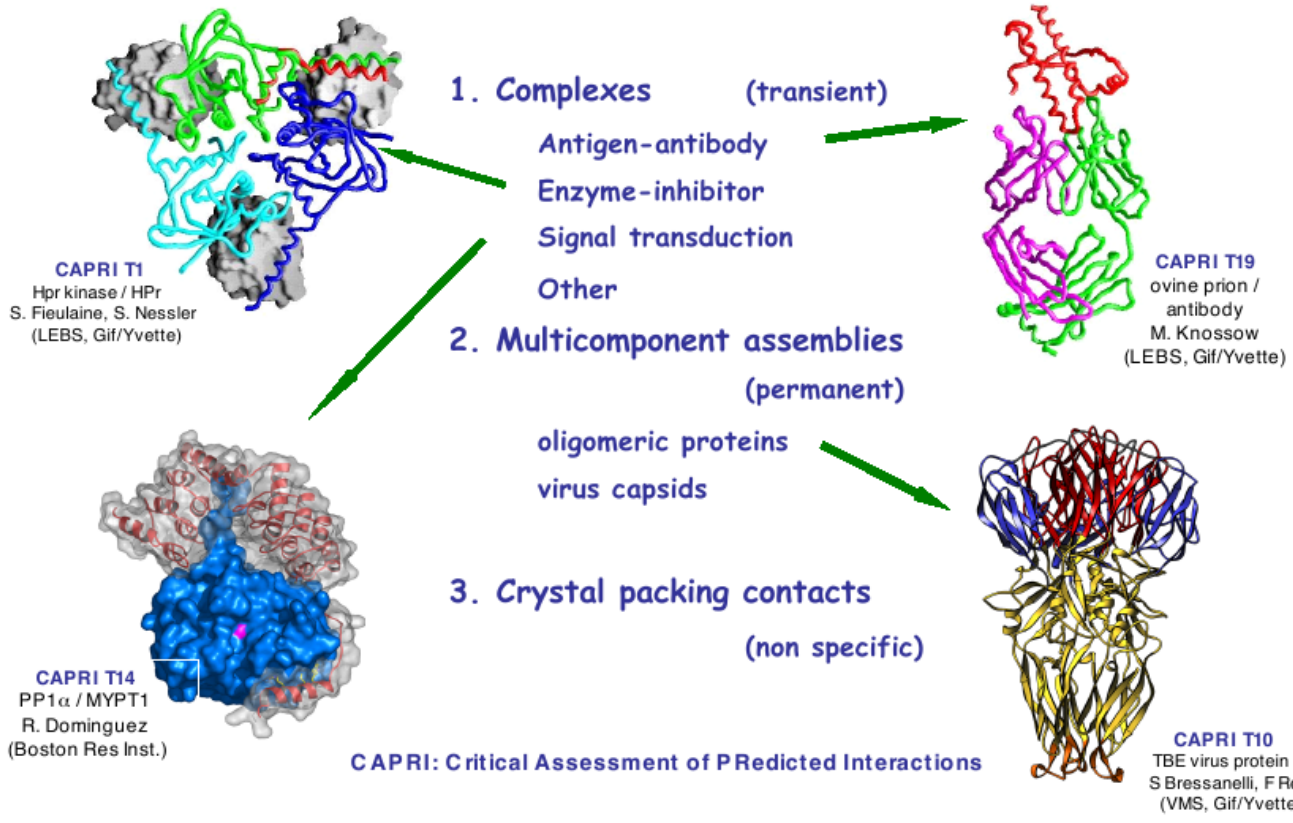


Figure 18. Drawing of a small piece of antiparallel β sheet (from SOD), illustrating the alternately narrow and wide pairing of H-bonds and the side-chain alternation above and below the plane of the sheet.

- ≪Geometry is not everything, but is is the most fundamental thing≫
  M. Connolly, 1982
- Building (phenomelogical) models: predicting, explaining

Diversity of Protein Assemblies: Quaternary Structure



[J. Janin]

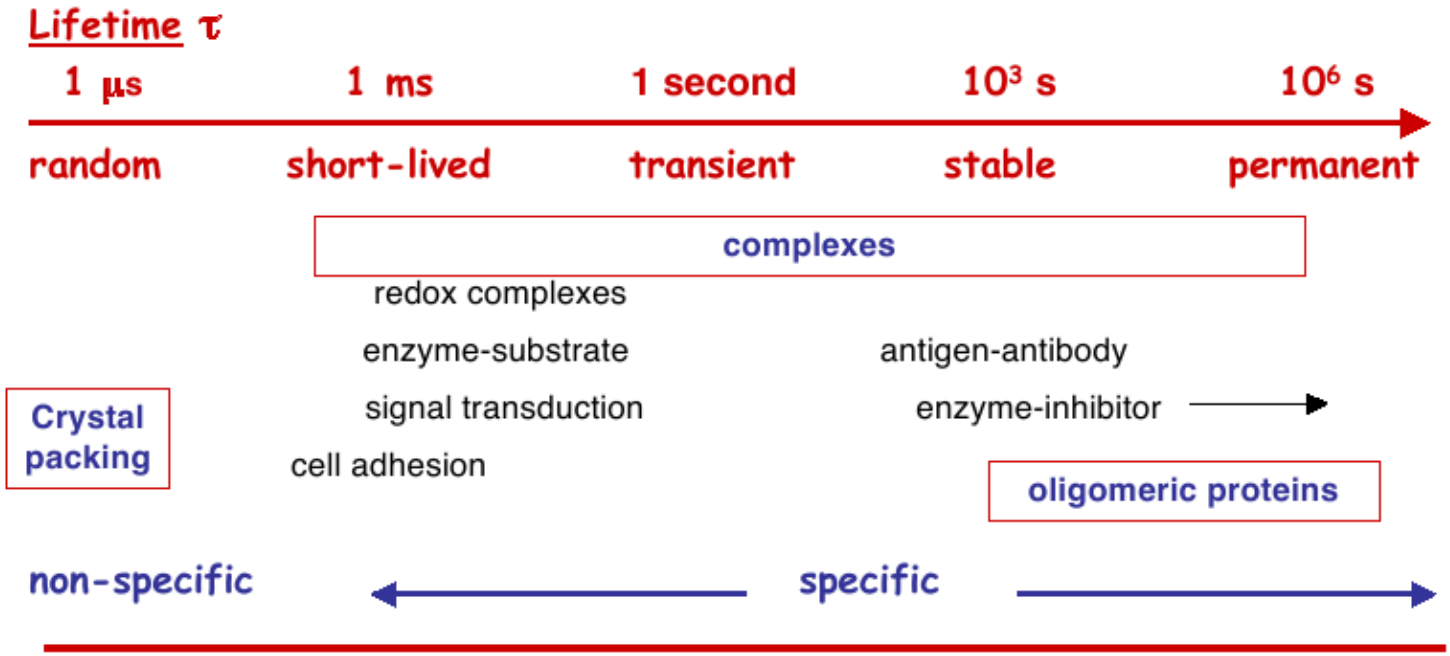
- Molecular mass: from $O(100\ kDa)$ up to 120 MDa (mammalian NPC)
- Structures vs sequences: 100,000 (PDB) versus 17,000,000 (NCBI RefSeq)

Ref: Janin et al; Quarterly reviews of biophysics; 2008

Ref: http://www.ncbi.nlm.nih.gov/RefSeq

Diversity of Protein Assemblies: Time Scales

- Biological time-scales



Short-lived complexes (τ<1 second) are relevant to many important biologically processes.
Only a few examples of these are present in the PDB (Nooren & Thornton, 2003).
These systems may resemble **crystal packing** more than permanent assemblies.

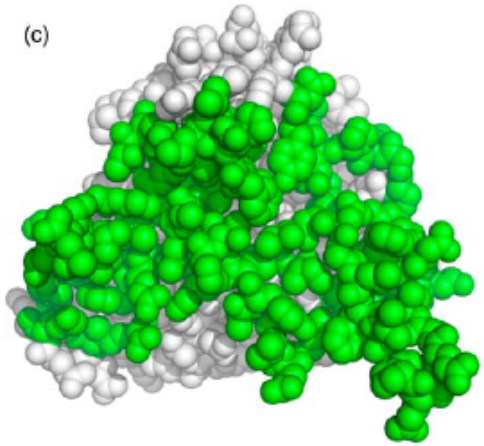
[J. Janin]

- Modeling: integration time step in MD ... femto-second

- Ref: Janin et al; Quarterly reviews of biophysics; 2008

Inferring Hot residues at Protein-Protein Interfaces

▷ Modeling protein complexes : core questions



- Stability of a complex (binding affinity): What are the key residues / atoms?
- Specificity of an interaction

▷ Strategies

Energy Experiments, directed mutagenesis: residues with high $\Delta\Delta G$; costly, incomplete

Modeling: free energy calculations (competition enthalpy/entropy (hydrophobic effect)); costly

Evolution Conserved residues: favored by evolution; hot residues tend to be conserved...

but may not apply; database dependent; conserved res. not at interface

Structure Shape, size, position of atoms; hot residues tend to be located in the interface *core*

Various interface models : core-rim, geometric footprint, Voronoi based

Modular Architecture of Protein-protein Interfaces

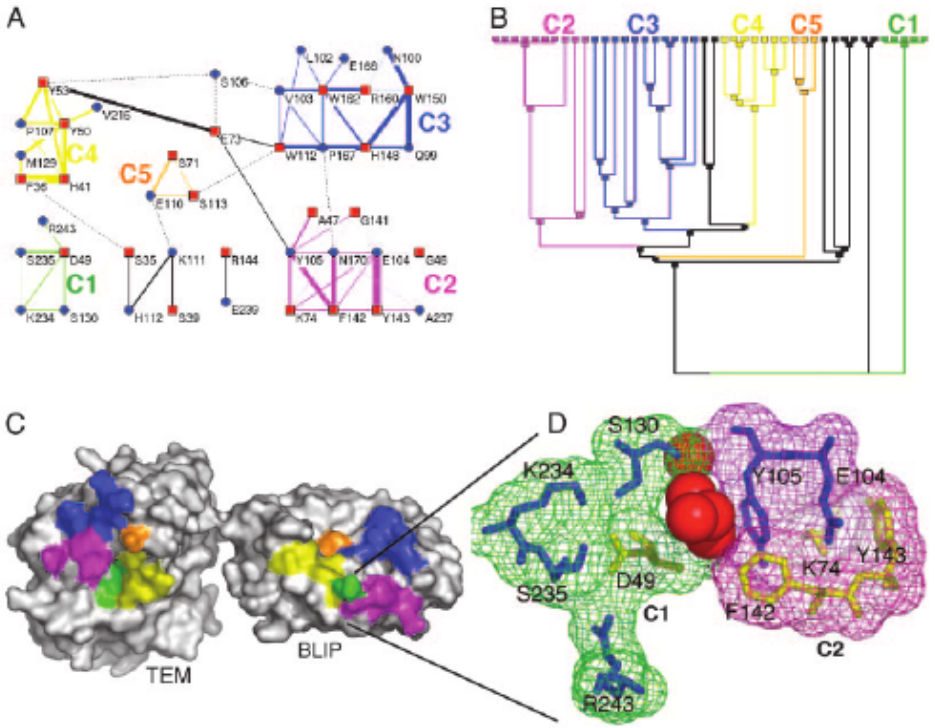


Fig. 1. Cluster analysis of the TEM1–BLIP interface. The interactions between residues located within the interface were extracted by using the CSU package (22) [for parameters, see Table 3; clustered with the SPC 3.3.13 tool (23)]. The interface was divided into five clusters of interactions, shown in *A* as a connectivity map, with the dendogram given in *B*, where the final nodes are the residues. A minimum of three residues is needed to form a cluster. The black lines indicate two residue interactions. (*C*) The location of the clusters is marked on the protein surfaces. An enlarged view of the two clusters (C1 and C2) is shown in *D* and includes the four water molecules separating the two clusters. The same color-coding is preserved throughout Fig. 1. In *A*, red squares mark BLIP residues, and blue circles mark TEM1 residues. In *D*, blue residues are for TEM1 and yellow for BLIP.

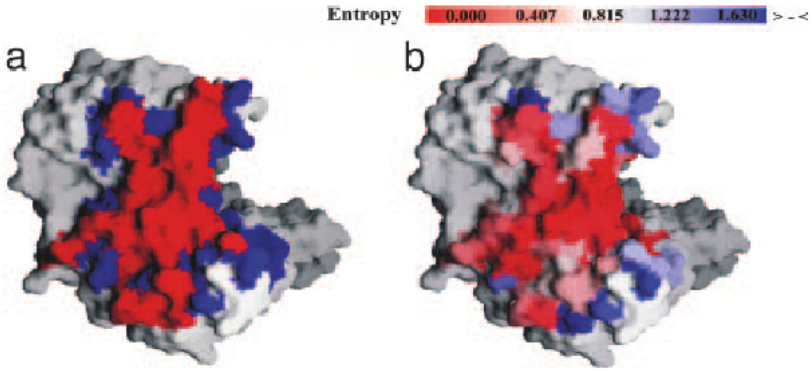
▷Schreiber et al, PNAS, 2005

- System: interface TEM1-β-lactamase – β-lactamase inhibitor protein (TEM1 - BLIP)
- Experiments: mutagenesis + ΔG through kinetics
- Modeling tools: clustering residus to define modules –based on atomic contacts
- Insights: Interface is modular; $\Delta\Delta G$: neg. NON additive in a module; (but add. between modules)

Inferring Hot residues at Protein-Protein Interfaces

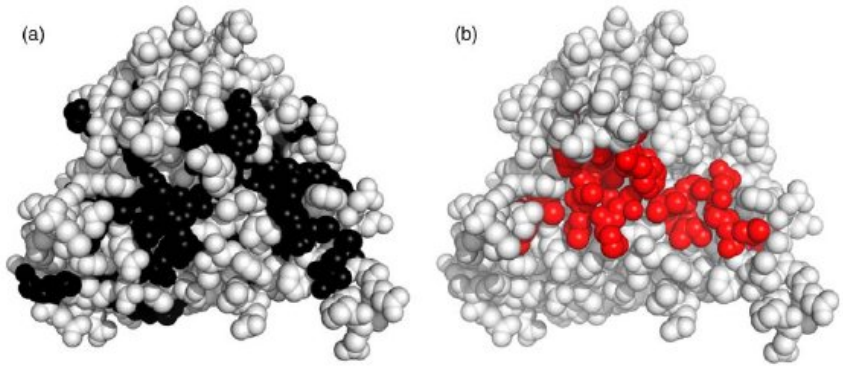
▷ Conservation vs geometry (core,rim)

▷Ref: Guharoy et al; PNAS, 2005



▷ Conservation vs dryness

▷Ref: Lichtarge et al; JMB; 2007



Protocol

Dissect interface core vs rim:

core: fully buried; rim: partly exposed

Conclusions

Core residues more conserved

Directed mutagenesis

Core residues : tend to exhibit higher $\Delta\Delta G$

Protocol

Run MD simulations

Measure Water residence times: dryness

Rationale for dryness :

interactions not perturbed by water fluxes

Conclusion

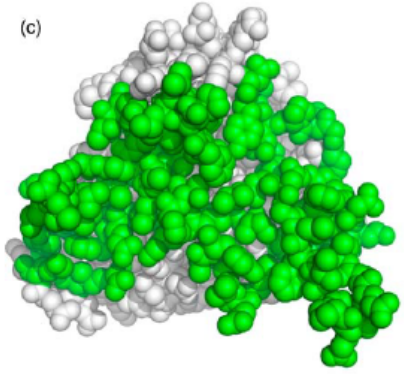
Conservation detects dry $\gg$ Conservation geom. footprint

▷ Rmk: statistics (P-values) are global: no assessment on a per-complex basis

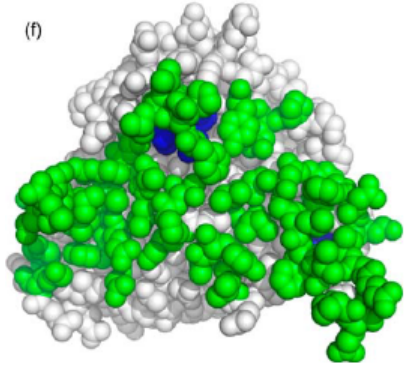
Predicting Important Residues: the Role of Dry Residues

- Important residues for P-P interactions
  - geometric footprint over/under predicts the hot residues
  - hot spot are known (in general) to be dehydrated (mutagenesis, dehydron, etc)
  - strong interactions : not perturbed by water fluxes (water might be quiet)

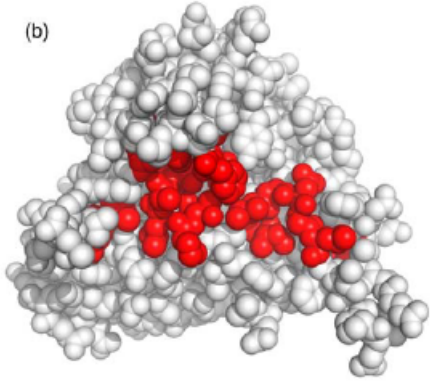
- 2DOR: interface residues within 7 Å



- 2DOR: interface residues: using ΔSAS



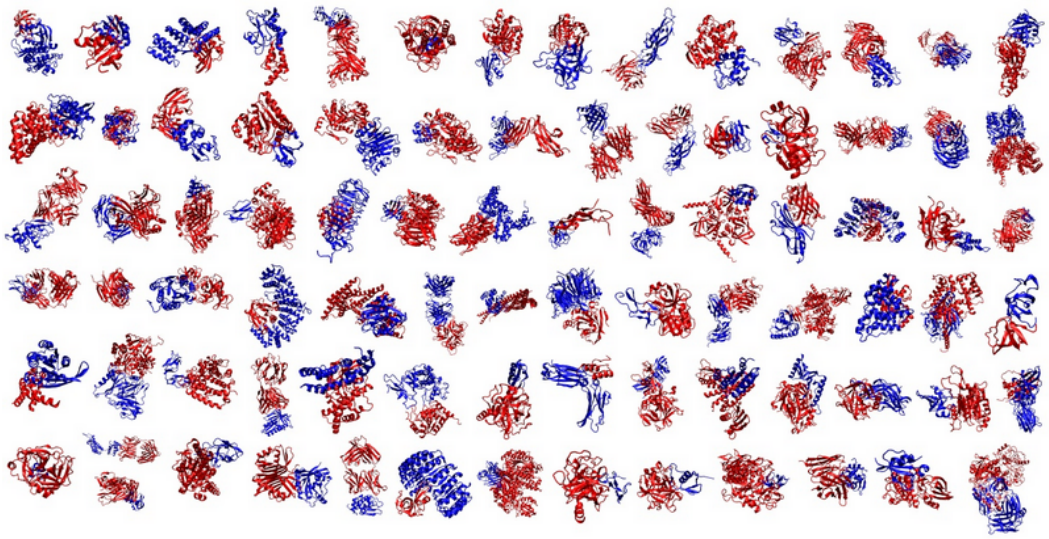
- 2DOR: dry residues



- Ref: Mihalek, Res, Lichtarge; JMB, 2007

Protein-Protein Interaction Affinity Database

http://bmm.cancerresearchuk.org/~bmmadmin/Affinity/



▷ Dissociation constant vs affinity

$$\Delta G = -RT \ln K_d / c^\circ$$

▷ NB: prediction based on unbound partners bound to mail for flexible cases

- ▷ 144 protein complexes
- ▷ Binding affinity known: ITC, SPR
  caveat: order of magnitude matter (pH, ion strength, . . .)
- ▷ Crystal structures known: bound complex, unbound partners
  induced flexibility upon docking

▷Ref: Kastritis et al; Protein Science (20), 2011

Scoring Functions versus Scoring at Random

- Testing
  two prototypical scoring functions vs a random permutation

- Decoys set: cf curve of expected number of successes $E(m)$
  either the scoring functions finds a near-native quickly
  or it is not any better than a random permutation

- CAPRI re-ranking:
  success in accordance with P-value (!)

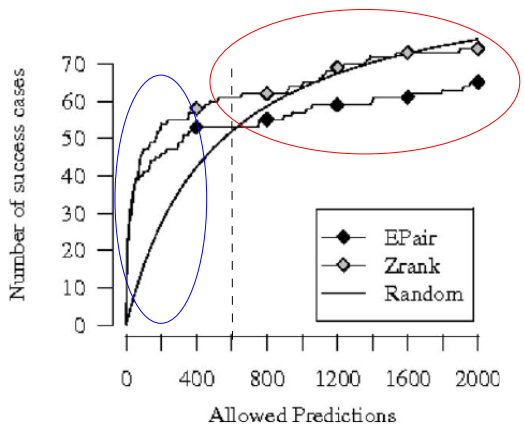


Figure 2
Success curves of ZRank and EPair compared with the random success curve for a small number of predictions (from 1 to 2000). See also Figure S1.

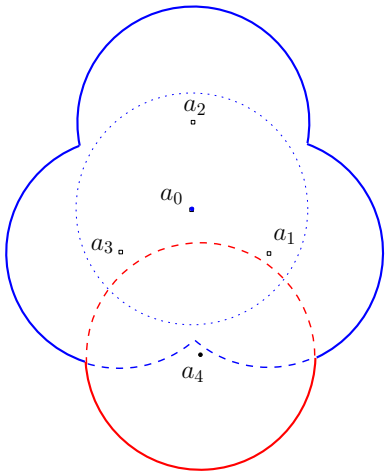
Table II
Statistical analysis of targets in the scoring section in CAPRI rounds 9–19

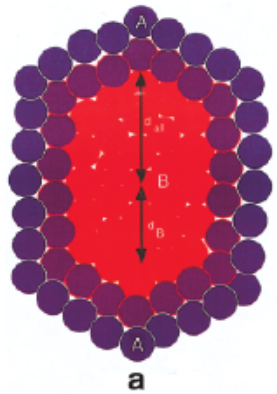
Target	Uploaded models	Accept or better models	*P* value (≥1 hit)	*P* value (≥2 hits)	Number of scorers	Successful scorers	Ratio of successful scorers	Optimal number of predictions
T25	700	36	0.41	0.09	6	6	1	2 (0.1)
T26	1171	60	0.41	0.09	8	4	0.5	2 (0.09)
T27	1093	123	0.7	0.31	12	11	0.92	1 (0.11)
T29	2192	167	0.54	0.17	10	7	0.7	1 (0.08)
T30	1346	2	0.015	~0	14	0	0	70 (0.1)
T32	599	15	0.226	0.023	5	2	0.4	4 (0.1)
T35	499	2	0.04	~0	11	1	0.1	26 (0.1)
T37	1700	76	0.37	0.07	11	10	0.9	2 (0.09)
T39	1400	4	0.03	~0	14	0	0	36 (0.09)
T40	2180	(354,134)	0.92	0.38[a]	14	11	0.78	1 (0.22)
T41	1208	299	0.94	0.75	13	11	0.85	1 (0.25)

▷Ref: Feliu et al; Proteins 78 (16); 2010

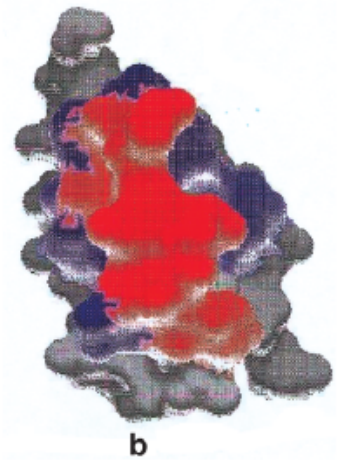
Classical Tools: Modeling Interfaces

- The core-rim model

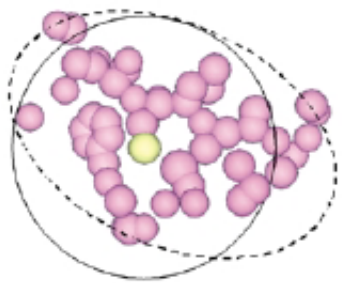


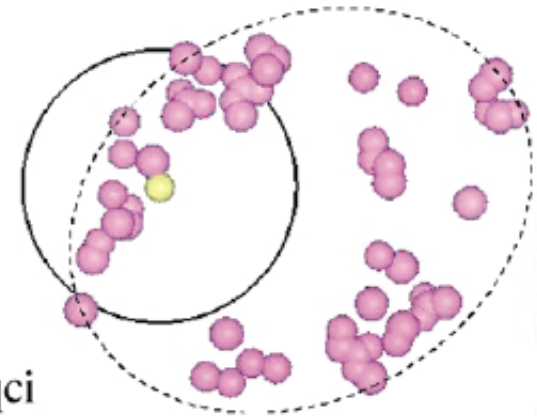


a


b

- Interface shape - (atom centric) packing density


1kba

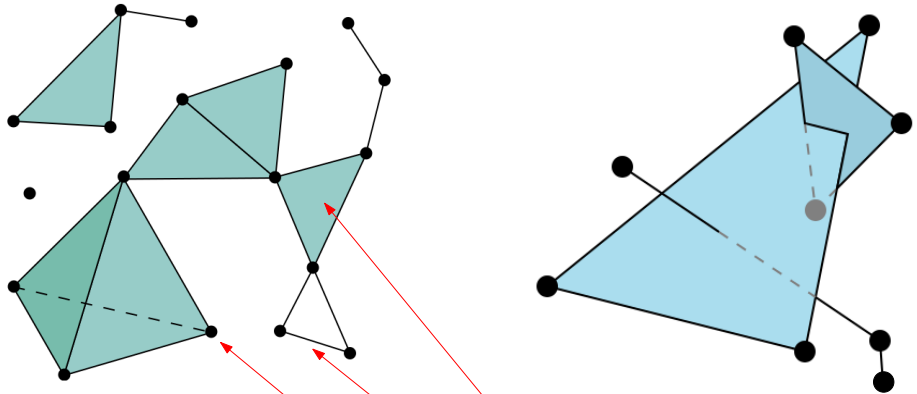

1qci

- Ref: Chakrabarti, Janin; Proteins; 2002
- Ref: Bahadur, Chakrabarti, Rodier, Janin; JMB; 2004

Modeling High Resolution Protein Complexes

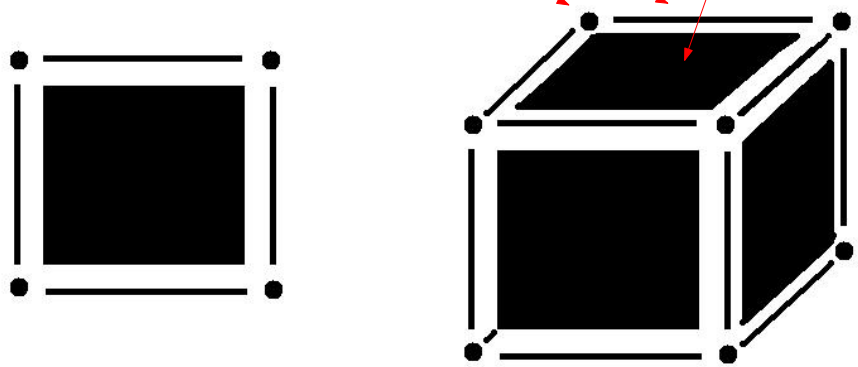
Linear Cell Complexes: Examples

Simplicial complex:lego of simplices (vertex, edge, triangle, tetrahedron,...)



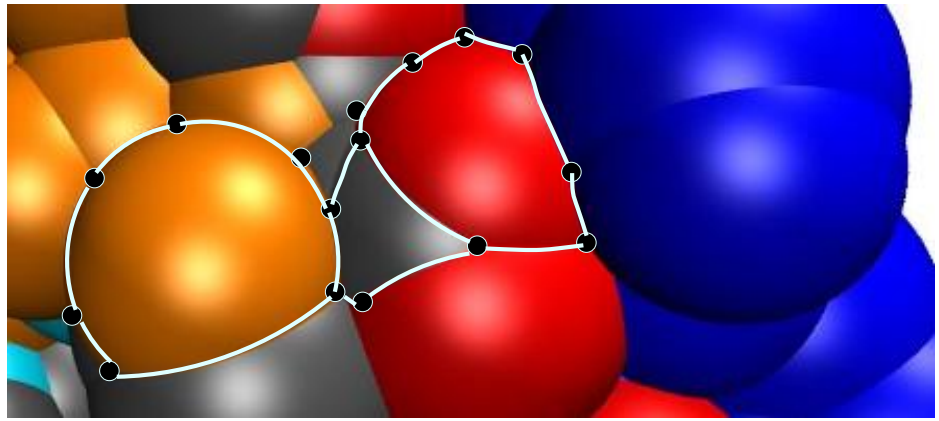
0-cell, 1-cell, 2-cell, 3-cell

Cubical complex:lego of hyper-cubes

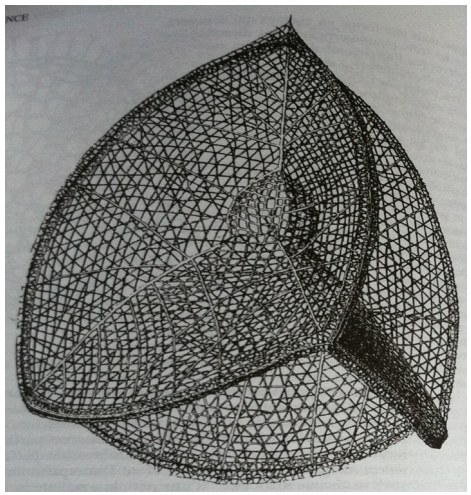


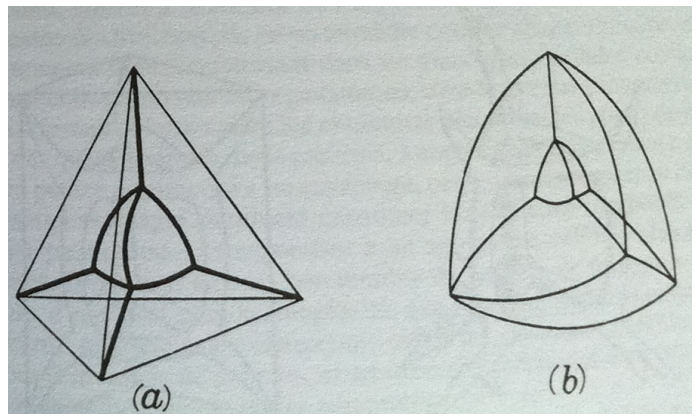
Curved Cell Complexes: Examples

Curved 2D cell complex: cells are points, circle arcs, spherical polygons

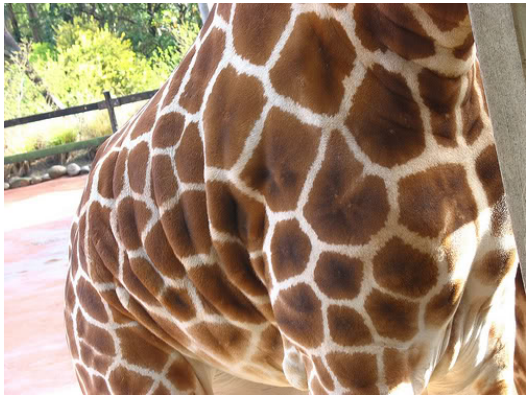


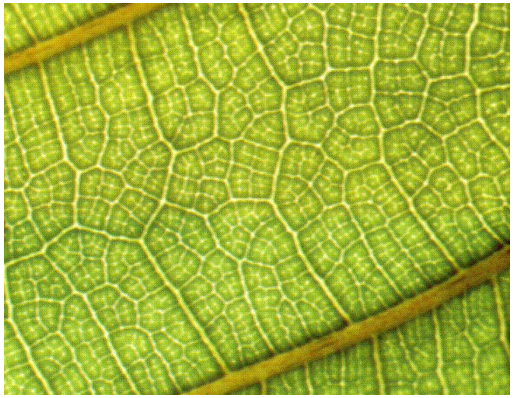
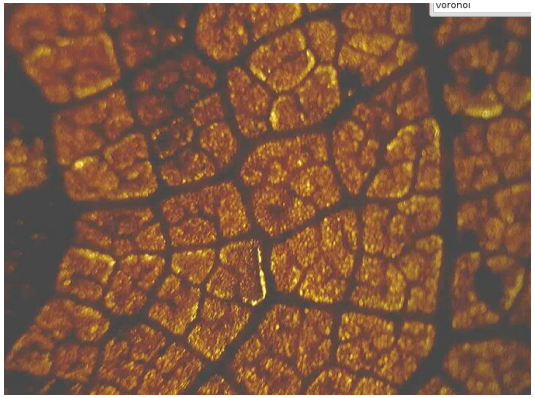
Curved 3D cell complex





Voronoi diagrams in Science and Growth Processes: Gallery



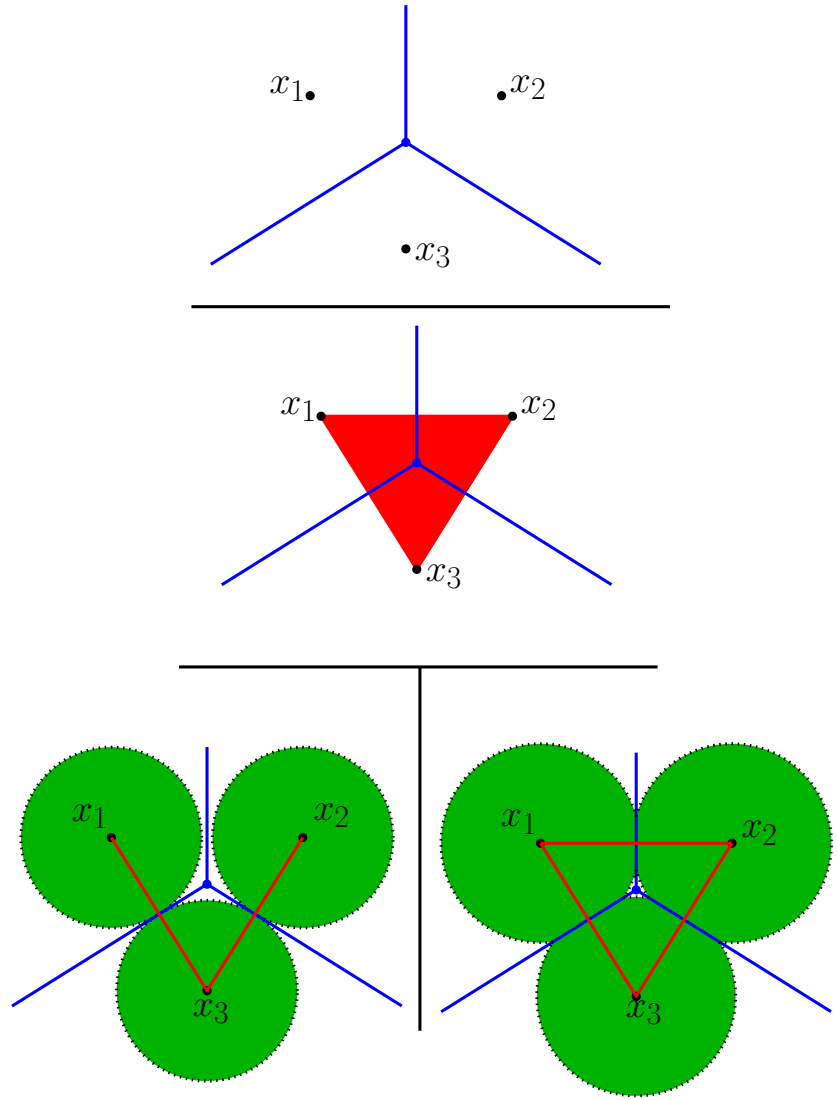



http://forum.woodenboat.com/showthread.php?112363-Voronoi-Diagrams-in-Nature

http://en.wikipedia.org/wiki/Giant's_Causeway

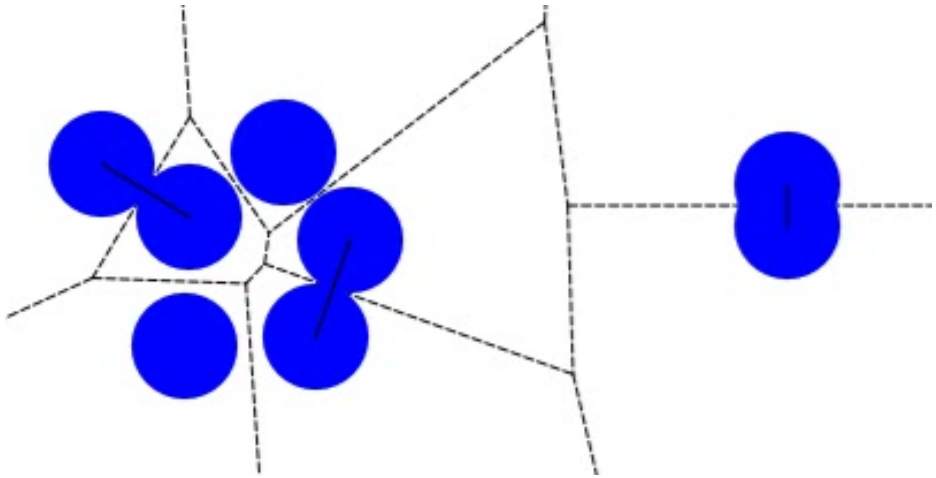
Euclidean Voronoi diagram and α-complex

- 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



α-shapes: Demo

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



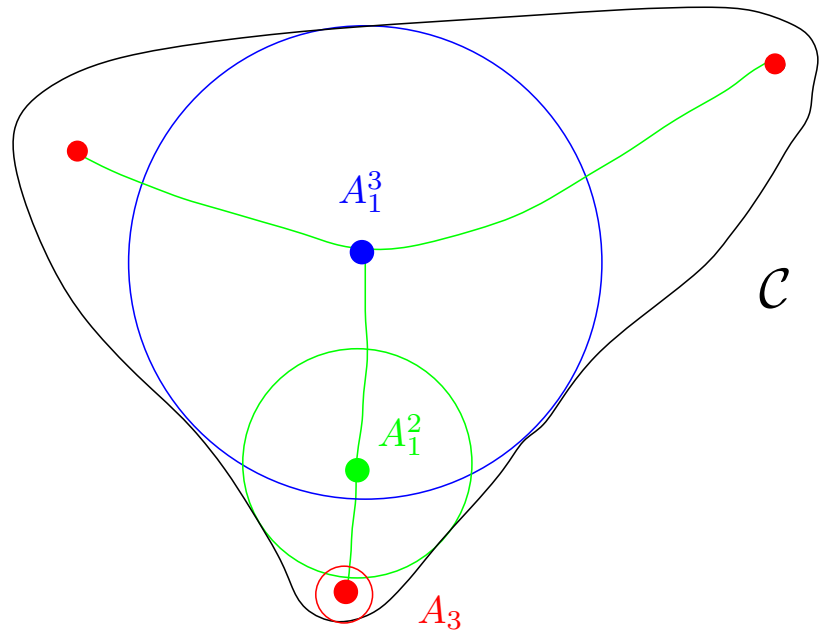
α-shapes : building a simplicial complex encoding the topology of the shape

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:

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

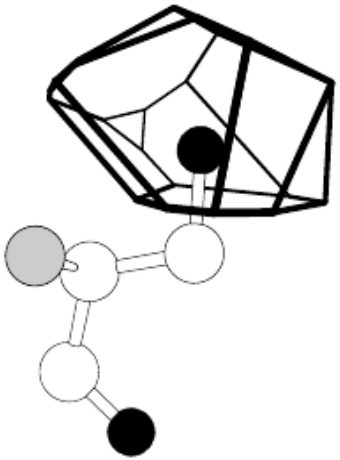
- Skeleton and local thickness:
  - Local: curvature properties
  - Global: related to bi/tri/tetra-tangent balls
- Medial axis transform: MAT



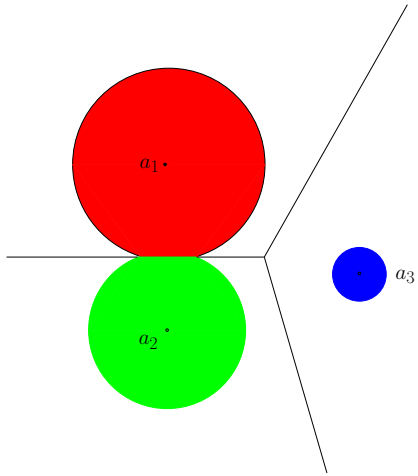
On the Volume of Union of Balls

- ▷ Context: discriminating native vs non-native states
  - – Describing the packing properties of atoms : surfaces and volumes
  - – Application: scoring functions

Voronoi region of atoms



Restricted Voronoi region



- ▷ STAR
  - – Monte Carlo estimates: slow
  - – Fixed precisions floating-point calculations: not robust

▷Ref: Gerstein, Richards; Crystallography Int'l Tables; 2002

▷Ref: McConkey, Sobolev, Edelman; Bioinformatics; 2002

▷Ref: McConkey, Sobolev, Edelman; PNAS 100; 2003

On the Volume of Union of Balls Cont'd

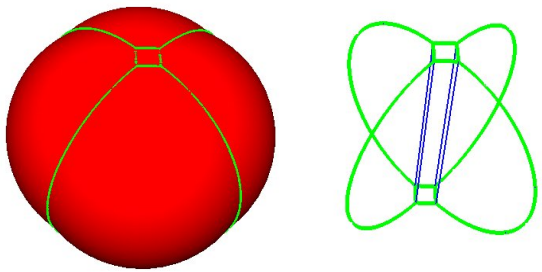
▷ Strategy developed: certified volume calculation

– Proved a simple formula for computing the volume of a restriction
– Analyzed the predicates and constructions involved
– Interval arithmetic implementation: certified range $[V_i^-, V_i^+] \ni V_i$

▷ Observation: Robustness requires mastering the sign of expressions

$$a + b\sqrt{\gamma_1} + c\sqrt{\gamma_2} + d\sqrt{\gamma_1\gamma_2}$$

with $\gamma_1 \neq \gamma_2$ algebraic extensions.



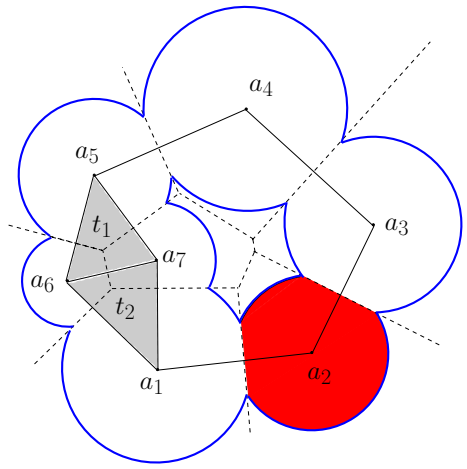
▷ Assessment

– 1st certified algorithm for volumes/surfaces of balls and restrictions
  – certified volume estimates (versus crude estimates)
  – (correct classification of atoms (exposed, buried; cf misclassification))
– 10x overhead w.r.t. to calculations using doubles

▷Ref: Cazals, Loriot, Machado, Teillaud; The 3dSK; CGAL 3.5; 2009

▷Ref: Cazals, Kanhere, Loriot; ACM Trans. Math. Software; 2011

Molecular Surfaces and Volumes: VORLUME and contenders



▷ Relative error computation r

$\tilde{t} = [t^-, t^+]$: VORLUME 's interval
e: estimate from contender
if $e < t^-$, then $r = (t^- - e)/t^-$
if $t^- \leq e \leq t^+$, then $r = 0$
if $e > t^+$, then $r = (t^+ - e)/t^+$

▷ Assessment: {S:surface, V:volume}× {G:global; R: per restriction } on a representative set from the PDB, of size 4405

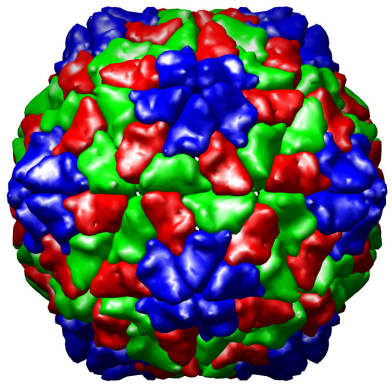
	$r = 0$	$r \in (0, 0.25]$	$r > 0.25$	r_{max}
Naccess, S_G	12.26	85.15	2.60	0.88
McC-et-al, S_G	27.33	72.67	0	0.10
Voidoo, V_G	9.58	90.42	0	3.43e-3
McC-et-al, V_G	0	99.98	0.02	0.29

▷Ref: Hubbard and Thornton; UCL Tech report; 1993 (Naccess)
▷Ref: Kleywegt and Jones; Acta Crystallographica D; 1994(Voidoo)
▷Ref: McConkey et al; Bioinformatics 18; 2002 (McC-et-al)

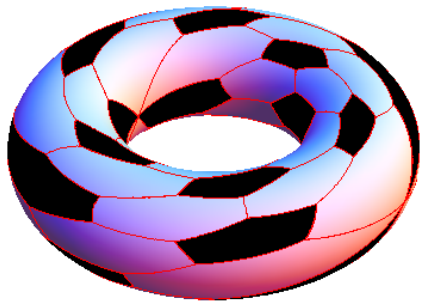
Geometry versus Topology:

The Theorem of Classification of Closed Surfaces in $\mathbb{R}^3$

A topological sphere: a genus 0 surface

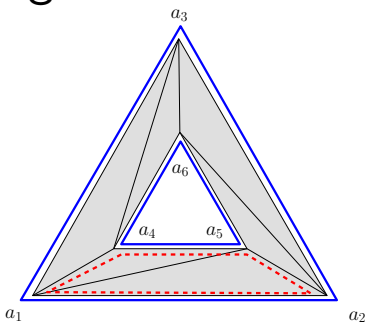


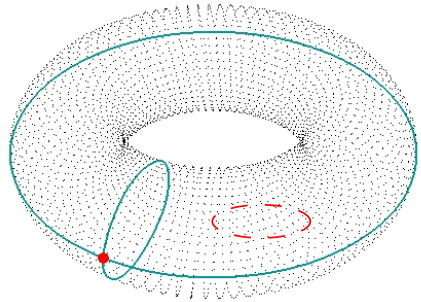
A topological torus: a genus 1 surface



Homology Theory

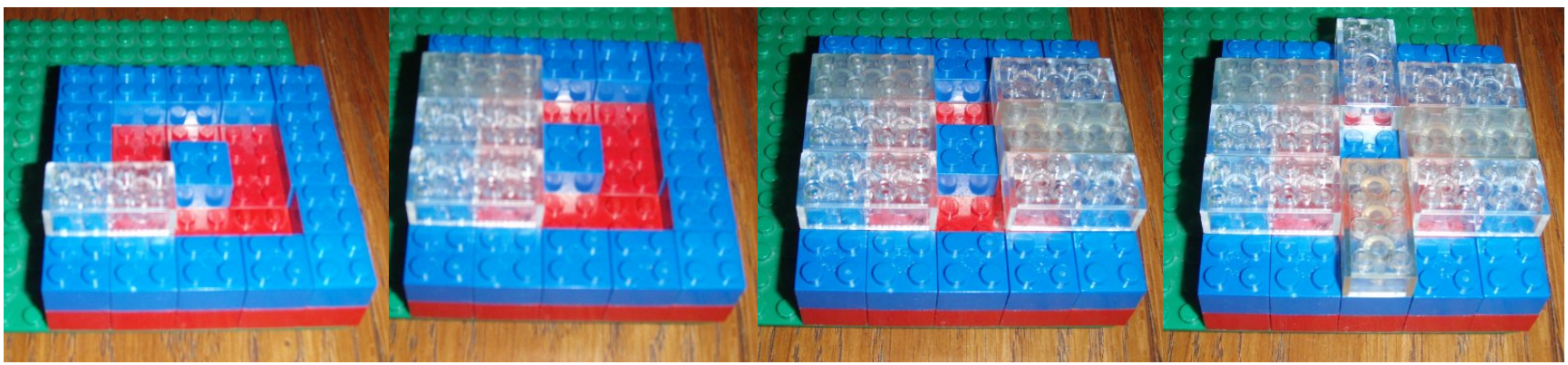
▷ Homology: counting k-dimensional cycles which do not bound (bound voids), regardless of their *thickness*





▷ Betti numbers count homology generators: examples in 3D

β_0: #cc β_1: # tunnels β_2: # voids



▷ Connexion to the Euler characteristic

$$\chi = \sum_{i=0,\ldots,d} (-1)^i \beta_i = \sum_{i=0,\ldots,d} (-1)^i (\#i - \text{dimensional cells})$$

Golf Courses and Energy Funnels

▷ Mr Euler playing golf

▷ Funnels on energy landscapes. . .

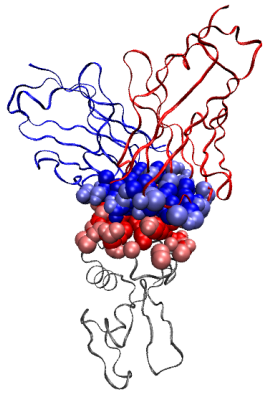
Euler characteristic?

Pitfall: index-1 saddles

Native state?

Modeling High Resolution Protein Complexes

Modeling the Interface of Macro-molecular Complexes



- Key questions: predicting the . . .
  stability of interfaces
  plasticity of complexes, dynamics of networks
  and their specificity

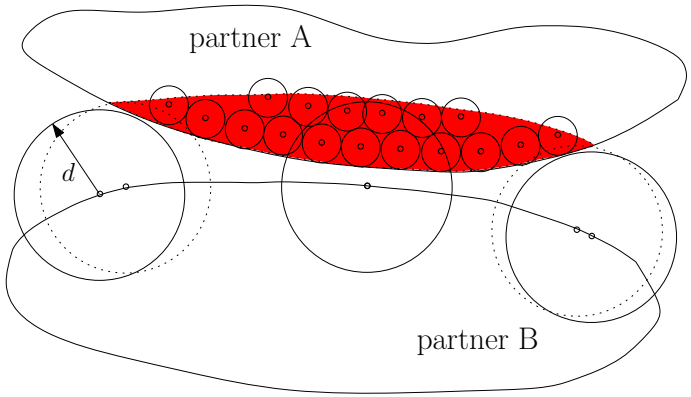
- Shape - topology:
  - # connected components, holes, voids / cavities [Homology]
  - morphology: *fat, skinny, dumbbell-like*

- Shape - geometry:
  - *privileged* contacts (pairs, triples, quadruples,...)
  - packing properties
  - accessibility (exposed vs buried atoms)
  - curvature information

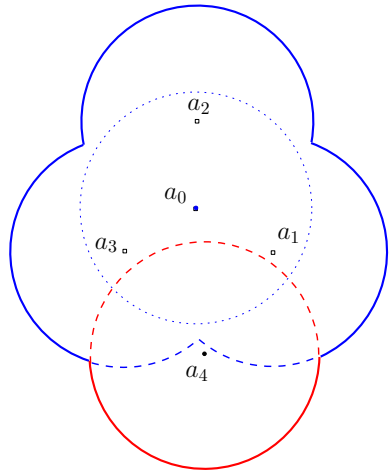
- Correlations with bio-physical quantities
  - conservation of amino-acids
  - biochemical properties

About Interface Models

▷ Distance threshold: geometric footprint



▷ Loss of solvent accessibility
(cf core and rim models)



▷ The Voronoi interface model

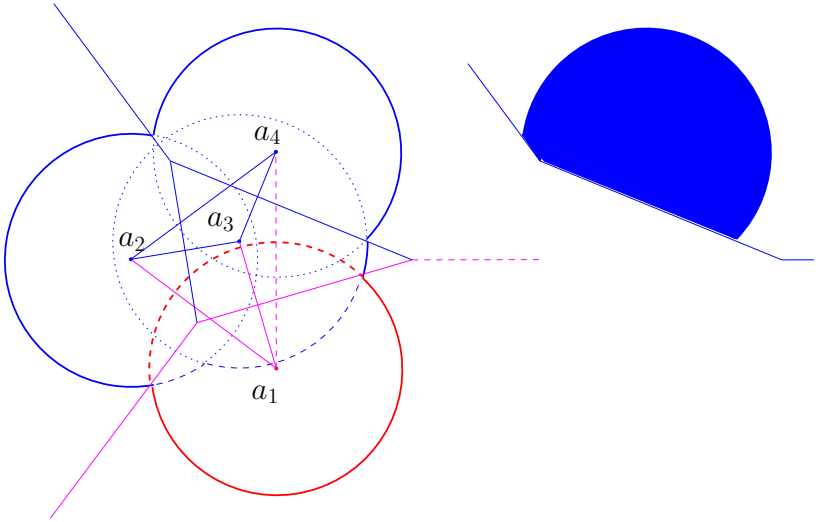
- A parameter free interface model
- Singles out a single layer of atoms
- Is amenable to geometric and topological calculations

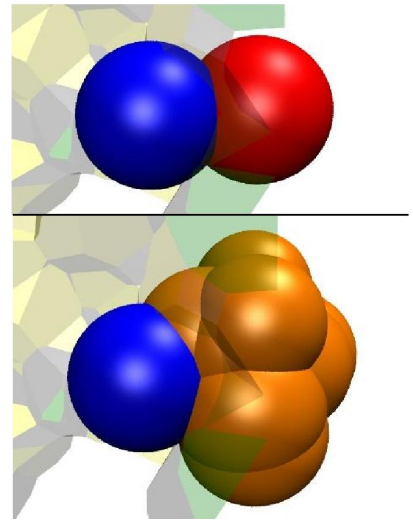
▷ Applications

- Wet biology: complex analysis and optimization — directed mutagenesis
- Structural modeling: scoring functions for docking
- Systems biology: mining contacts, mating orphan molecules, ...

Voronoi Interface : Definition

(Power Diagram Based Interface Definition)





- Interface : bicolor edges in 0-complex

Lemma. Any atom with $\Delta ASA > 0$ is an interface atom.

Attention. Converse is FALSE : cf 13% of interf. atoms missed by previous studies

Importance.
Such atoms are *nearest neighbors* (wrt to the power distance)

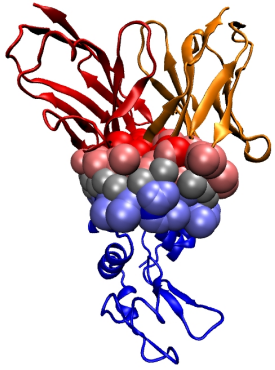
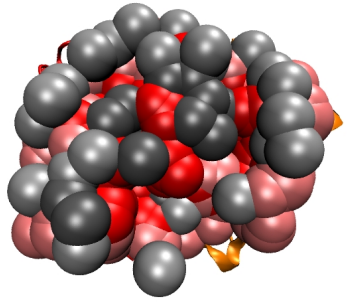
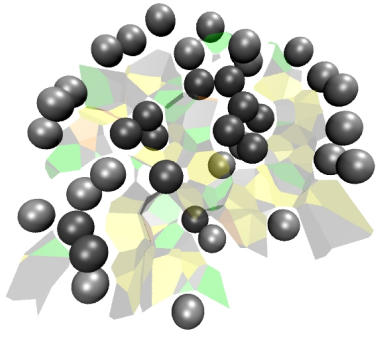
Voronoi interface: balance between geom. footprint and ΔASA

Ref: Cazals, Proust, Bahadur, Janin; Protein Science; 2006

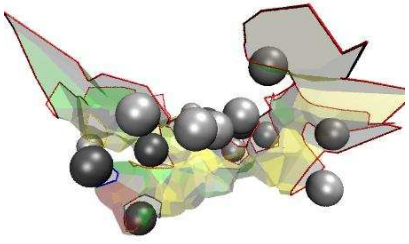
Voronoi Interfaces : Illustrations

(An integrated model from the atomic to the interface scale)

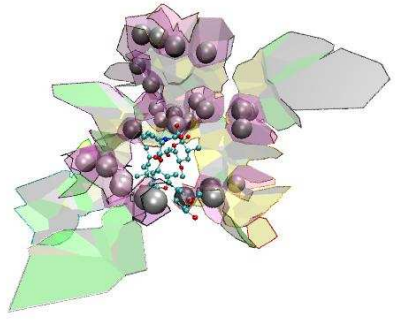
- Role of strutural water –antobody-antigen

- Curvature –protease-inhbitor



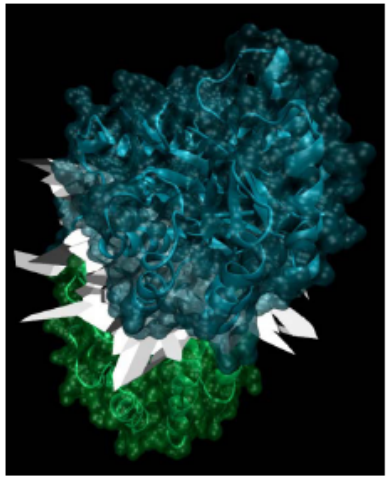
- Multi-patch structure –signal transduction



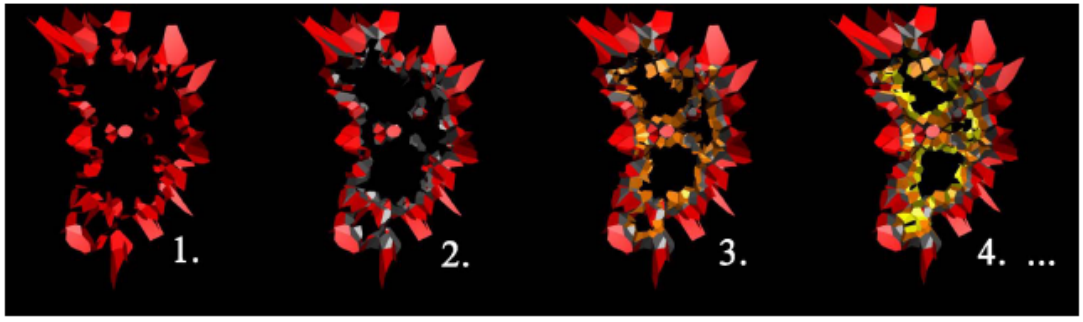
- Ref: Cazals, Proust, Bahadur, Janin; Protein Science; 2006

Modeling High Resolution Protein Complexes

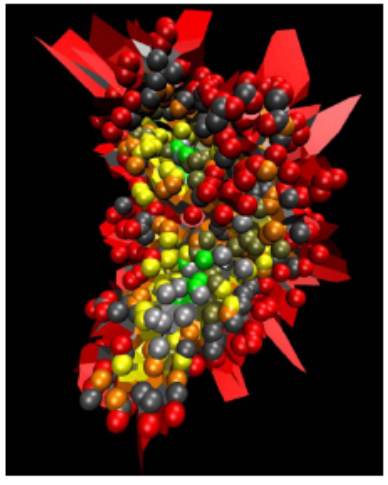
Shelling the Voronoi Interface: Illustration



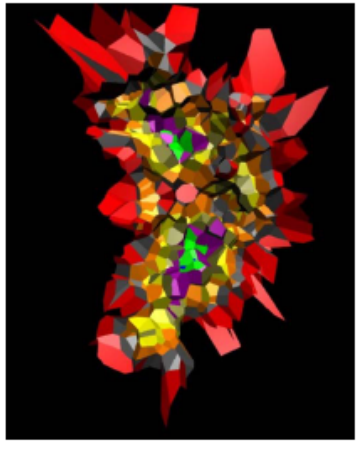
Dihydroorotate dehydrogenase (2DOR)



Shelling the Voronoi interface...



Projected on atoms

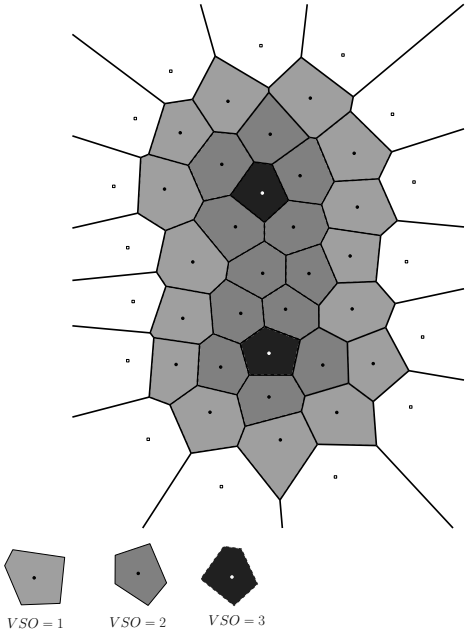


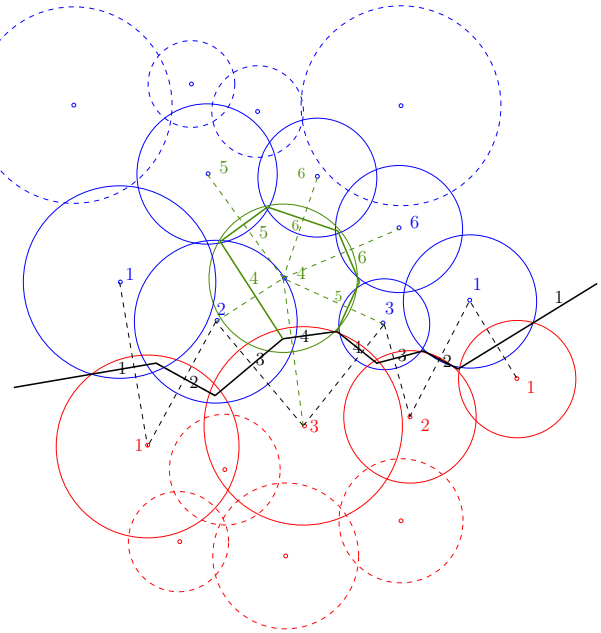
Shelled interface

- Properties?
- Evolution during an MD simulation?

▷Ref: Bouvier, Gruenberg, Nilges, Cazals; Proteins 76 2009

Voronoi Shelling Order: Definition





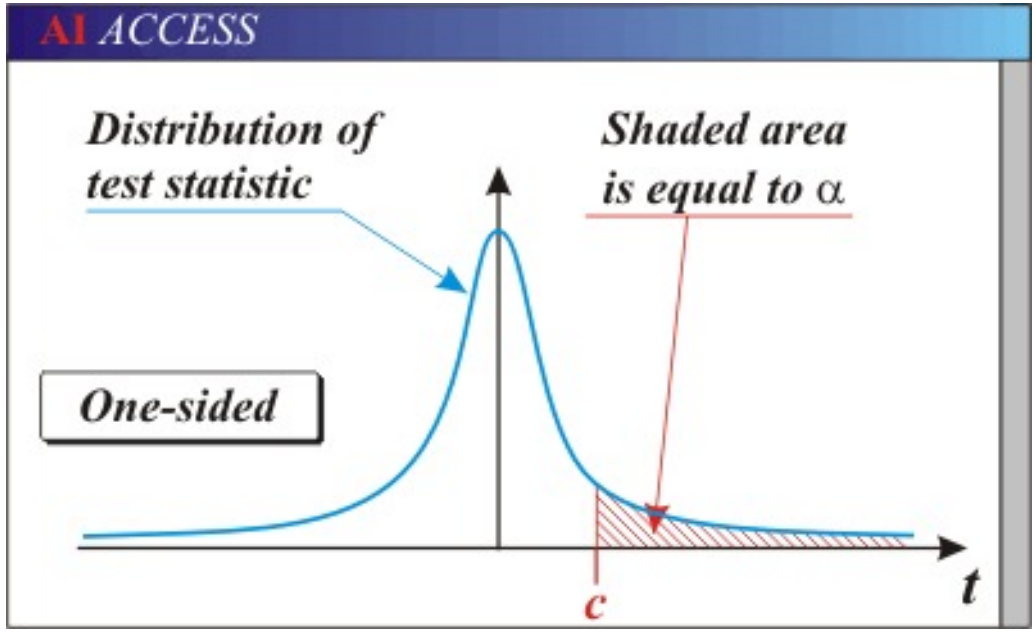
- Three stages
  - select bicolor Delaunay edges in the 0-complex
  - walking over the dual Voronoi facets/tiles
  - pulling back values onto the atoms

Testing Statistical Hypothesis: P-value and Errors are two probability distributions p and q identical?

- Null hypothesis and its alternative
  - H0 (the belief): $p = q$
  - Ha (alternative): $p \neq q$

- Testing H0
  - design a test statistic S
  - compute it from samples, say s_0
  - p-value for H0: $P(S > s_0)$
  - reject H0 if $P(S > s_0) < \alpha (= 0.05)$
    or if $s_o \notin$ acceptance region

- Type I error: H0 erroneously rejected
  α upper bounds the proba. of the type I error
- Type II error: H0 erroneously accepted
  - power of S for $p \neq q$: type II error
  - the statistic is called *consistent* if it
    the type II error converges to 0
    (when the # samples increases)



- Acceptance region:

  $1 - \alpha$ quantile of the null distributions
  hatched area

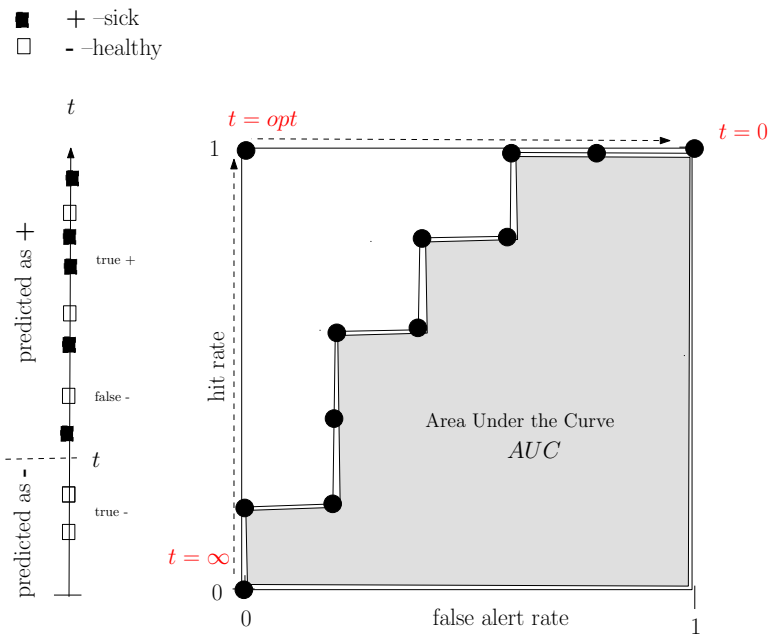
Receiver Operating Characteristic (ROC) curves

- Continuous variable t versus binary attribute $\{+, -\}$:

  prediction of $\{+, -\}$ based on position of t relative to a threshold t_0

$$\text{sensitivity=hit rate} = \frac{\textbf{true}+}{\textbf{true}+ + \textbf{false}-}, \quad \text{false alert rate} = \text{1-specificity} = \frac{\textbf{false}+}{\textbf{true}- + \textbf{false}+}$$

- Varying the threshold yields the ROC curve. Ideal situation:



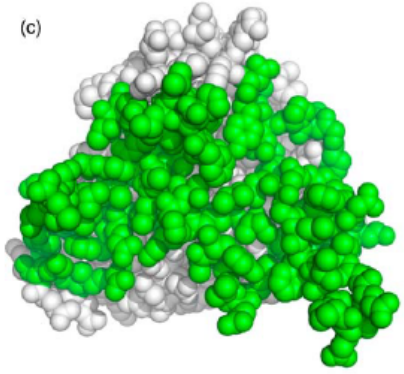
- p-value calculation for a particular value AUC_0:

  AUC_0 vs. distribution of areas over all permutations of + and -

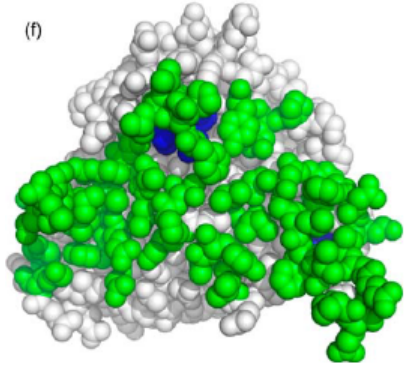
Predicting Important Residues: the Role of Dry Residues

- Important residues for P-P interactions
  - geometric footprint over/under predicts the hot residues
  - hot spot are known (in general) to be dehydrated (mutagenesis, dehydron, etc)
  - strong interactions : not perturbed by water fluxes (water might be quiet)

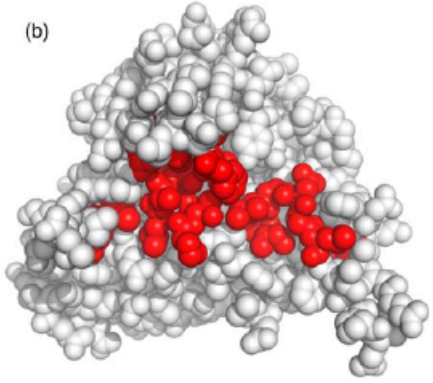
- 2DOR: interface residues within 7 Å



- 2DOR: interface residues: using ΔSAS



- 2DOR: dry residues

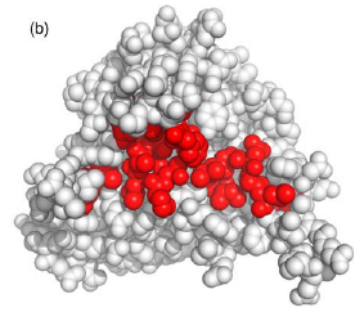


- Ref: Mihalek, Res, Lichtarge; JMB, 2007

Water Traffic and Conservation of Residues at Protein - Protein Interfaces

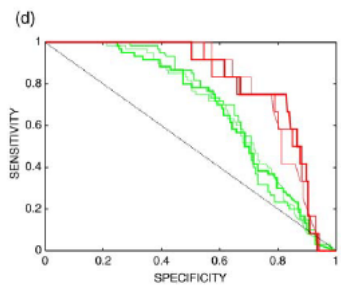
- ▷ Dry A.A. tend to be more *important*
- ▷ Protocol: MD simulation; A.A. s.t. $\Delta ASA > 0$
- ▷ Traffic intensity for A.A. i: $I_i = \frac{1}{T} \sum_w \frac{1}{\tau_w}$
- ▷ Dry residue w.r.t.traffic intensity:
  - $I_i \leq 0.005 ps^{-2}$ for homodimers
  - $I_i \leq 0.01 ps^{-2}$ for heterodimers
- ▷ Assessment with ROC curves:

▷ 2DOR: dry residues



conservation predicts dryness versus conservation predicts geom. footprint

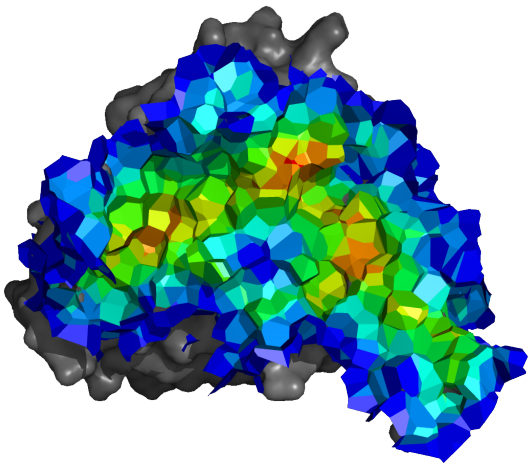
- ▷ Conclusions:
  - 3 conservations methods perform equally
  - **AUC(conserv. $\rightarrow$ dryness) $\gg$ AUC(conserv. $\rightarrow$ geom. footprint)**

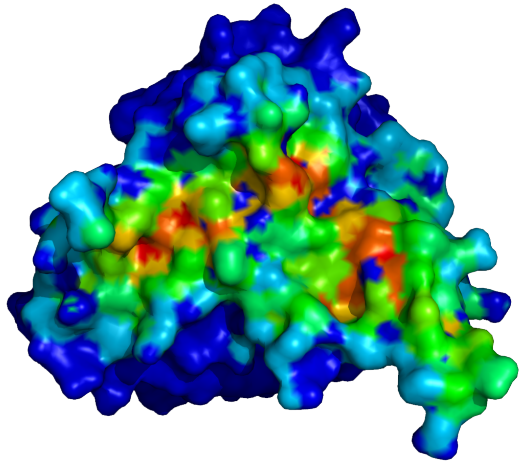


▷Ref: Mihalek, Res, Lichtarge; JMB, 2007

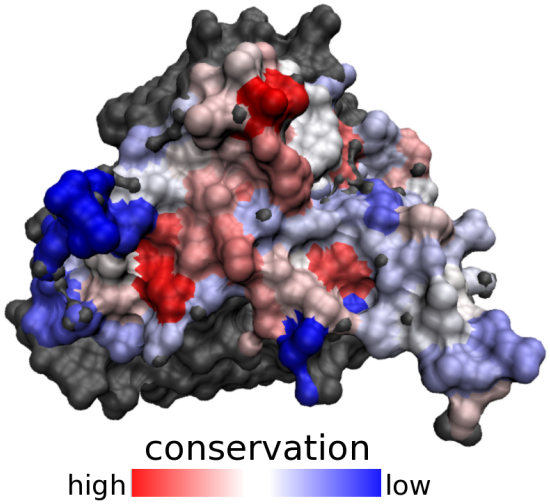
VSO versus Dryness – 2DOR

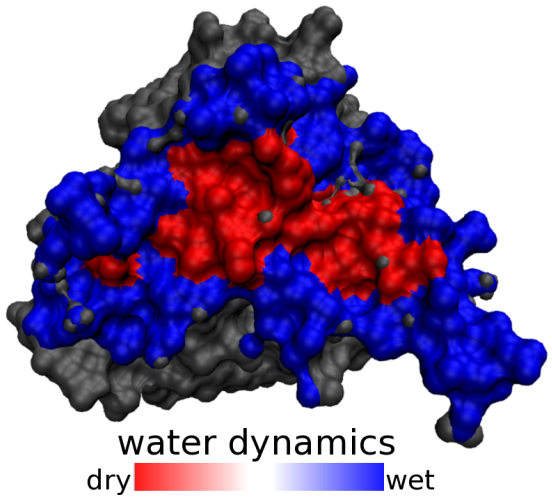
- VSO: facets and atoms

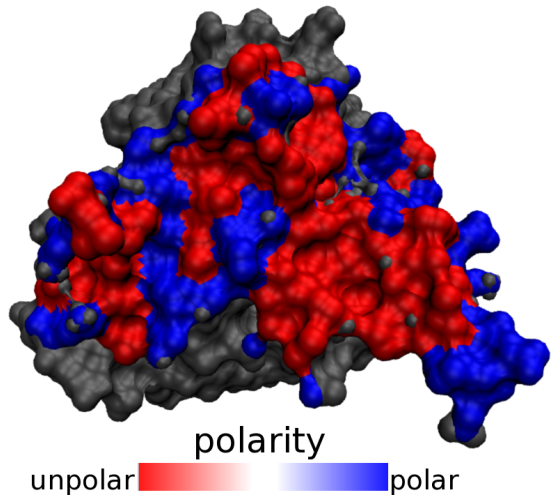




- Conservation, dryness, polarity







VSO, Dryness, Conservation: Statistical Significance of Predictions / Methodology

- Protocol for each set of complexes (36 homos, 18 heteros)

  ability of a continuous parameter to predict a binary attribute

- Four predictions for the two datasets:

VSO [cont.]$\longrightarrow$ dryness [threshold] conserv. [cont.] $\longrightarrow$ dryness [threshold]

conserv. [cont.]$\longrightarrow$ VSO [threshold] VSO [cont.] $\longrightarrow$ unpolar [bin.]

- Statistical assessment

  Per complex:

    AUC, p-value for null hypothesis

  Per dataset (homos, heteros):

    Combined p-value for k tests / Fisher's inverse Chi-square:
    $X^2 = -2 \sum_{i=1 \ldots k} \log p_i$ follows a chi-square with $2k$ dof

- Summary for a given prediction

– per complex: AUC + p-value

– per data set: average AUC + combined p-value

VSO, Dryness, Conservation: Statistical Significance of Predictions / Results

▷ 18 Heterodimers

PDB Id.	VSO→dryness		conserv.→dryness		conserv.→VSO		VSO→unpolar	
	AUC	P-value	AUC	P-value	AUC	P-value	AUC	P-value
...								
Reject H_0		18/18		8/18		8/18		11/18
Global	0.81	6e-74	0.64	3e-14	0.65	2e-09	0.63	1e-21

▷ 36 homodimers

PDB Id.	VSO→dryness		conserv.→dryness		conserv.→VSO		VSO→unpolar	
	AUC	P-value	AUC	P-value	AUC	P-value	AUC	P-value
...								
Reject H_0		36/36		25/36		14/36		27/36
Global	0.84	2e-265	0.63	2e-43	0.62	4e-20	0.64	2e-63

▷ Conclusions

VSO→dryness
universal correlation—valid on ALL individual cases

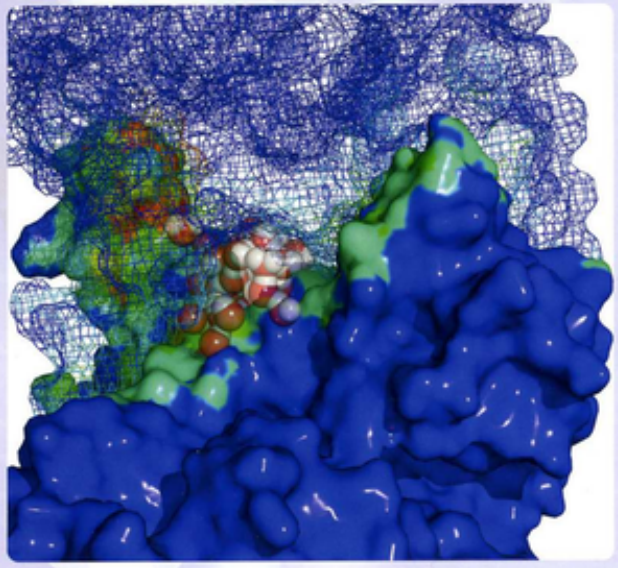
conserv.→dryness (cf Lichtarge et al, JMB 369, 2007) [no p-values]
conserv.→VSO (cf Chakrabarti et al, PNAS 102, 2005) [combined p-values only]
VSO→unpolar
global trend . . . but prediction often fails on an individual basis
binary core/rim interface models do not account for the subtlety of distributions of conservation/polarity
VSO provides a continuous parameterization of the interface

▷Ref: Bouvier, Gruenberg, Nilges, Cazals; Proteins, 2009

PROTEINS

STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

VOLUME 76, NUMBER 3, AUGUST 15, 2009



Shelling the Voronoi Interface of Protein–Protein Complexes

WILEY-BLACKWELL

ISSN 0887-3585

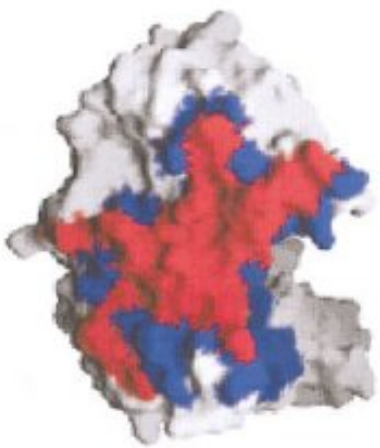
Articles published online in Wiley InterScience, 14 January 2009–8 April 2009

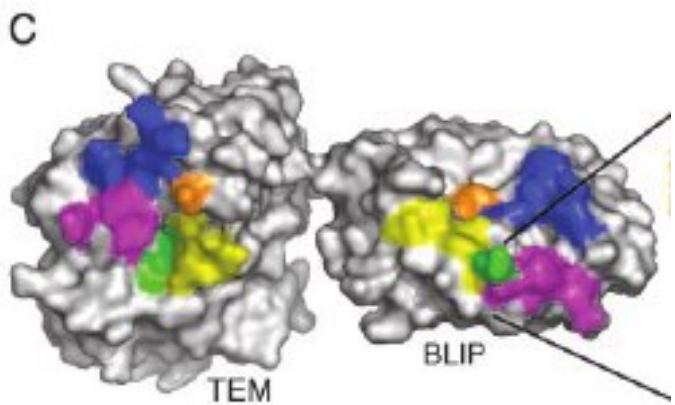
Modeling High Resolution Protein Complexes

On the Morphology of Binding Patches

- Current binding patch models : not designed for quantitative processing
  - Pro: used to mine correlations with biological - biophysical properties
  - Cons: core rim model : dissection based on solvent accessibility: binary model
- Global pairwise comparison for docking - clustering:
  - Pro: useful algorithms for rigid docking
  - Cons: not amenable to local comparisons
  - Cons: no *decomposability* of binding patches

- Understanding the morphology of binding patches
  $\longrightarrow$ simple geometric - topological model amenable to both types of studies





(a) Core-rim model [Janin et al, 2003-2009]
(b) Clustering into modules [Schreiber et al, PNAS, 2005]

Comparing Binding Patches: Quasi-isometric Subsets and Reduction to Max Clique

- Distance between two atoms i, j of M_1: $d^1_{i,j}$; likewise for M_2

- Root Mean Square Deviation of Internal Distances

  Given 2 sets of atoms S_1 and S_2 having the same size n

  and a one-to-one mapping m between them

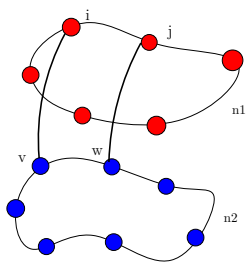
  $$RMSD_d(S_1, S_2) = \sqrt{\sum_{i<j} |d^1_{i,j} - d^2_{m(i),m(j)}|^2 / \binom{n}{2}}$$

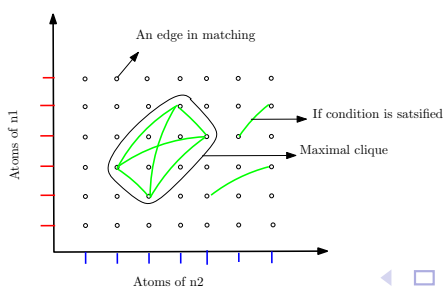
- Goal for two molecules M_1 and M_2: find the largest $S_1 \subset M_1$ and $S_2 \subset M_2$, and the corresponding mapping $m()$, such that $RMSD_d(S_1, S_2) \leq \epsilon$

- Reduction to Max Clique :

  Match atoms i, j of M_1 and k, l of M_2 iff $|d^1_{i.j} - d^2_{k.l}| \leq \epsilon$

  Therefore, $M_1 \bigcap M_2 =$ Size of maximum clique in product graph

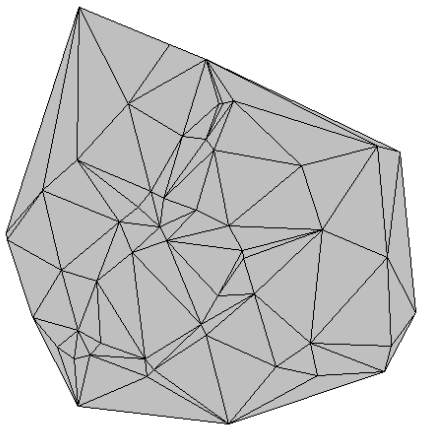




Shelling a Cell Complex

▷ Input

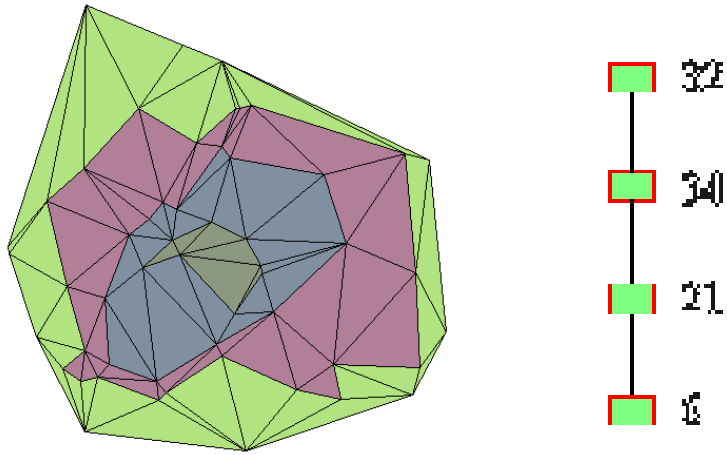
Cell complex say **D** dimensional
- Cells - dimension **D**
- Facets - dimension **D-1**
- Pivots - dimension **D-2**



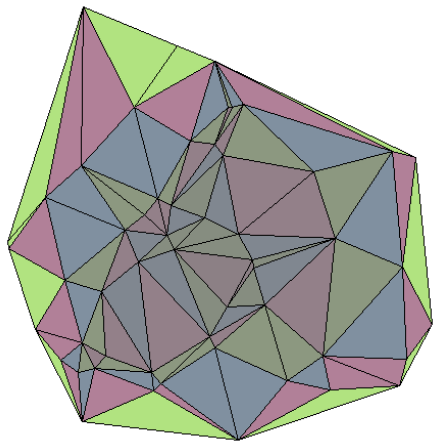
Example : 2D Alpha shape
Triangles are cells, edges are facets and vertices are pivots

▷ Output

▷ Shelling by pivoting

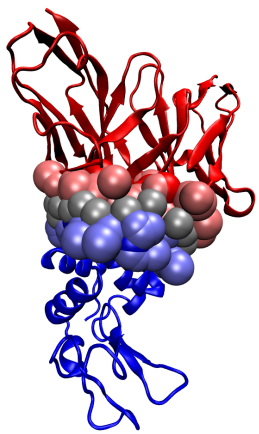


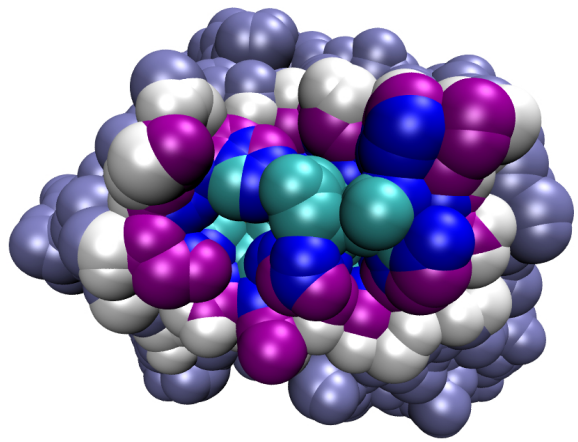
▷ Shelling by face connectivity

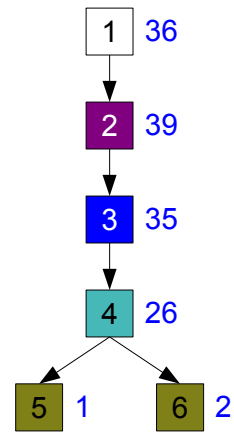


Shelling a Binding Patch yields a Topological Encoding

- From the complex: Voronoi-based identification of interface atom
- For each partner
  - compute the boundary of the union of balls into a Half-edge Data Structure:
    - spherical caps - circle arc - vertices
  - shell the HDS – as a cell complex
- Convert the output into an Atom Shelling Tree







Ordered Tree Edit Distance (TED)

- Editing T_1 into T_2 is based on 3 operations:
  node insertion | deletion | morphing
- Complexity, using dynamic programming: time: $O(n^3)$; space: $O(n^2)$
- Semantics of the 3 operations: problem dependent

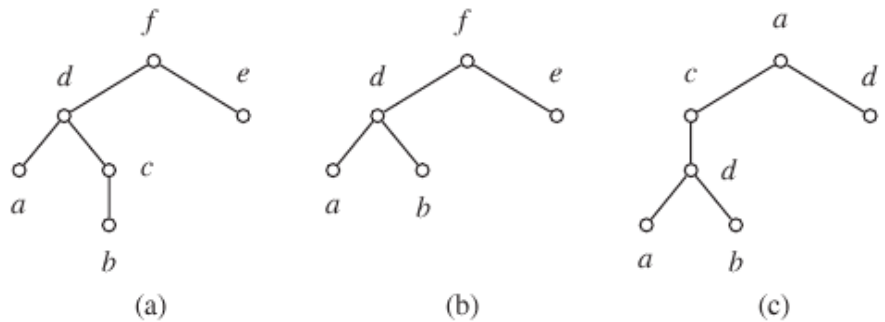
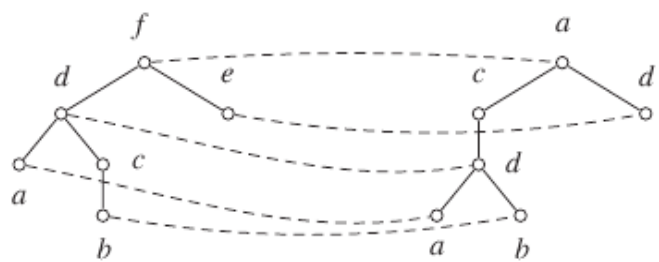


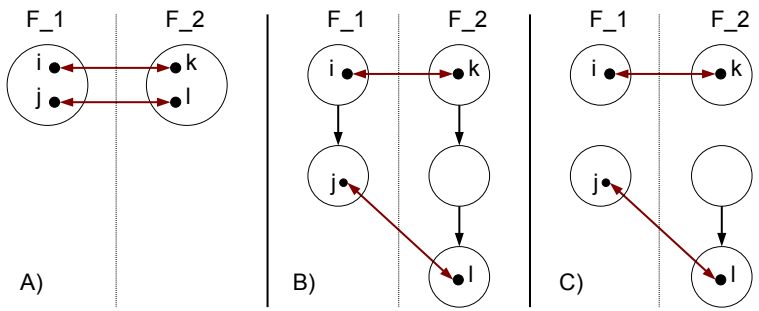
Fig. 2. Transforming (a) into (c) via editing operations. (a) A tree. (b) The tree after deleting the node labeled c. (c) The tree after inserting the node labeled c and relabeling f to a and e to d.



- Ref: Bille; TCS; 337 (205)

Application 1: Topological Comparison of Patches

- Input: the trees T_1 and T_2 encoding 2 binding patches
- Straight TED:
  cost of insertion - deletion: node size;
  cost of morphing shell s_1 into shell s_2: $\max(|\ s_1\ |, |\ s_2\ |) - \min(|\ s_1\ |, |\ s_2\ |)$
- The TED calculation delivers an Ordered Edit Distance Mapping:
  $M \subset Vertices(T_1) \times Vertices(T_2)$ s.t. $(v_1, v_2) \in M$ and $(w_1, w_2) \in M$, one has:
  (i) $v_1 = w_1$ iff $v_2 = w_2$, or
  (ii) v_1 is an ancestor of w_1 iff v_2 is an ancestor of w_2, or
  (iii) or v_1 is to the left of w_1 iff v_2 is to the left of w_2.
- Atoms matched meet (i,ii,iii): they are called isotopologic:
  $SIM_t(T_1, T_2)$: number of atoms matched
  $TED_t(T_1, T_2) = |\ T_1\ | + |\ T_2\ | - 2\ SIM_t(T_1, T_2)$
- Corresponding dissimilarity $\in 0..1$
  $DIS_t(T_1, T_2) = TED_t(T_1, T_2)/(|\ T_1\ | + |\ T_2\ |)$



Application 2: Geometric Comparison of Patches

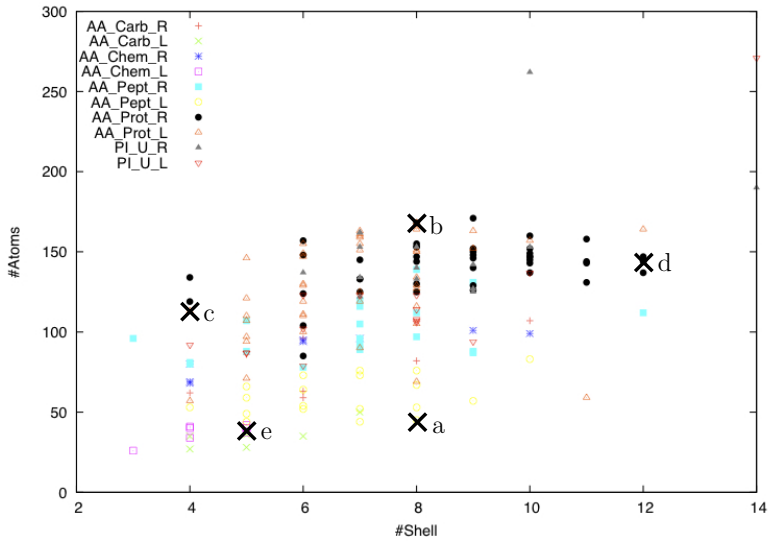
- Restrict the Max-Clique calculation to shells

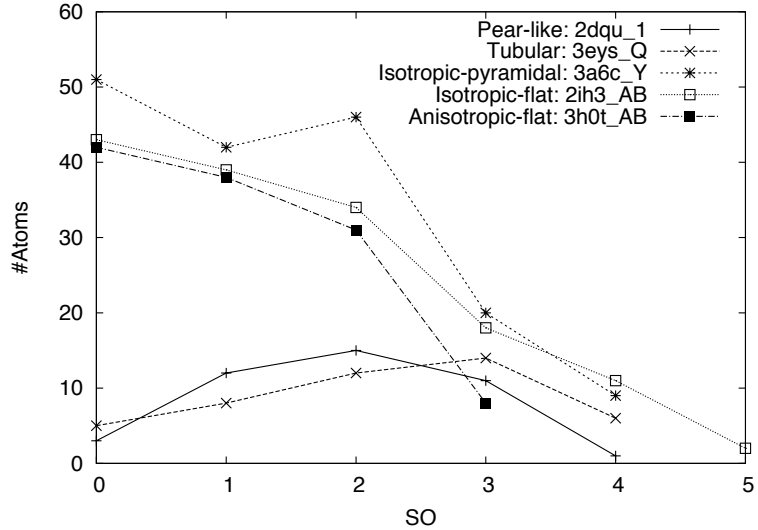
- Atoms matched are called isotopologic:
  $\mathrm{SIM}_g(T_1, T_2)$: number of atoms matched
  $\mathrm{TED}_g(T_1, T_2) = \mid T_1 \mid + \mid T_2 \mid -2\ \mathrm{SIM}_g(T_1, T_2)$
- Corresponding dissimilarity $\in 0..1$:
  $\mathrm{DIS}_g(T_1, T_2) = \mathrm{TED}_g(T_1, T_2)/(\mid T_1 \mid + \mid T_2 \mid)$

- Properties:
  $RMSD_d$ upper-bounded at the shell but NOT binding patch level
  Topology versus geometry
    similarity: $\mathrm{SIM}_g(T_1, T_2) \leq \mathrm{SIM}_t(T_1, T_2)$
    dissimilarity: $\mathrm{DIS}_g(T_1, T_2) \geq \mathrm{DIS}_t(T_1, T_2)$

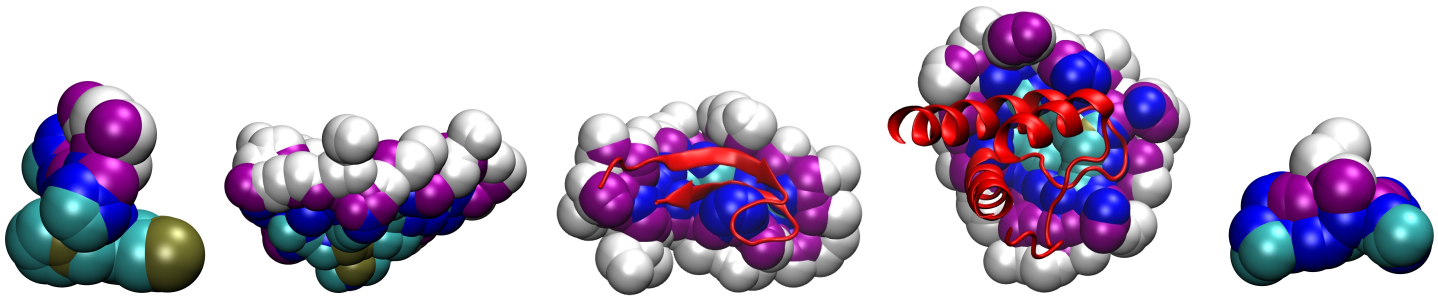
Binging Patches: Typical Morphologies

- Number of atoms as a function of the number of shells



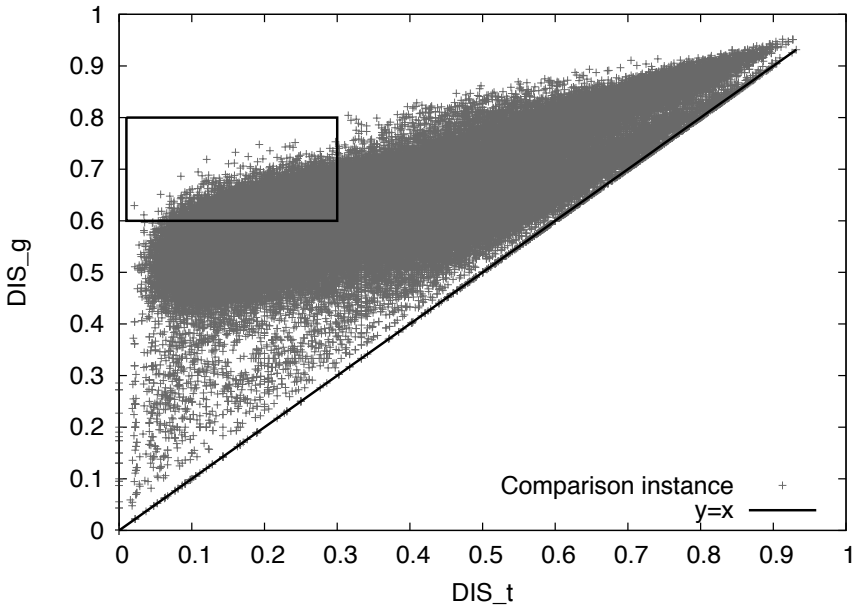


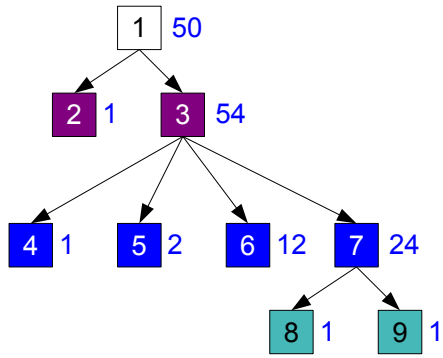
- Typical morphologies:

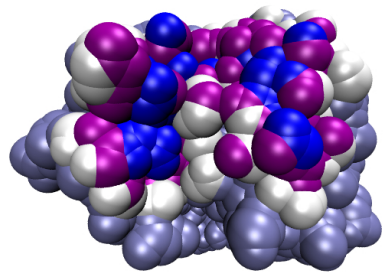


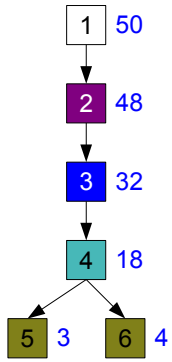
(a) tubular (b) isotropic-pyramidal (c) anisotropic-flat (d) isotropic-flat (e) pear-like

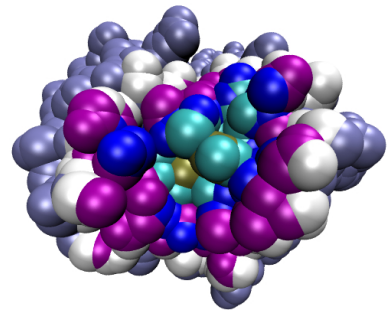
Similar Topology, Dissimilar Geometry





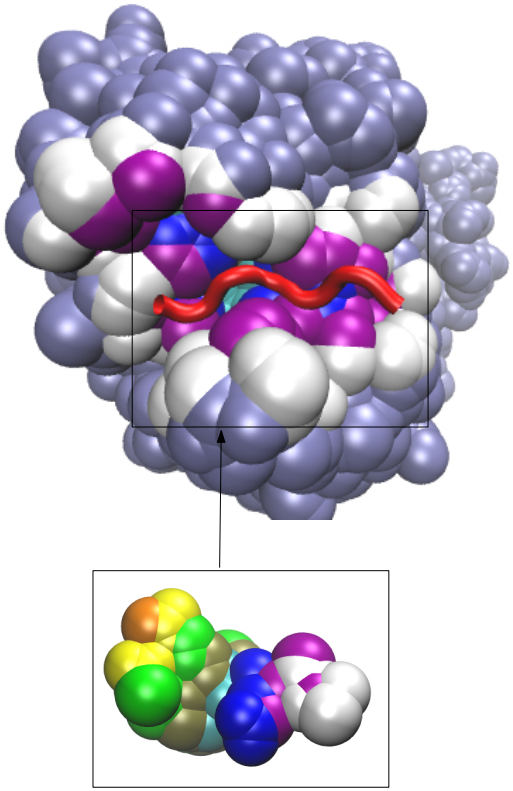






Symmetry of Patches and Homogeneity of Families

- Anisotropic vs tubular



- Identification favors the family rather than the complement

DB decomposition:

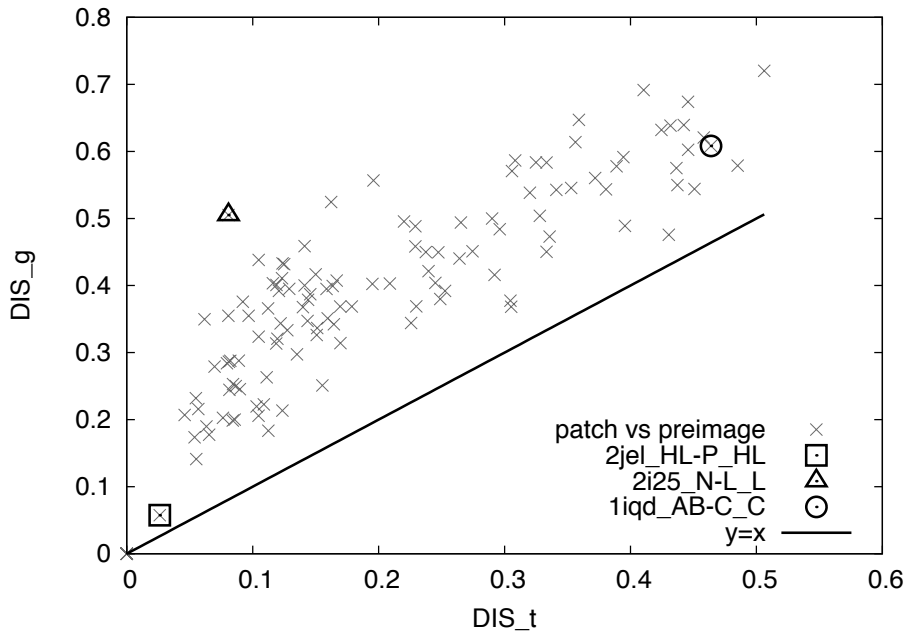
$$\mathcal{P} = p \cup P_{\setminus p} \cup \overline{P} \cup P^c$$

Family (=P)	(P, P) vs $(P, \overline{P})$	(P, P) vs (P, P^C)
AA_Carb_R	3.76e-06	3.02e-07
AA_Carb_L	5.15e-11	1.27e-13
AA_Chem_R	1.42e-08	1.30e-08
AA_Chem_L	3.44e-14	5.78e-17
AA_Pept_R	1.80e-17	1.31e-27
AA_Pept_L	9.47e-69	9.78e-70
AA_Prot_R	7.25e-04	3.93e-38
AA_Prot_L	2.86e-56	9.73e-49
PI_U_L	2.76e-23	6.25e-20
PI_U_R	7.10e-06	1.14e-14

Flexibility Upon Docking:

Rigid, Flexible,and Topo-rigid patches

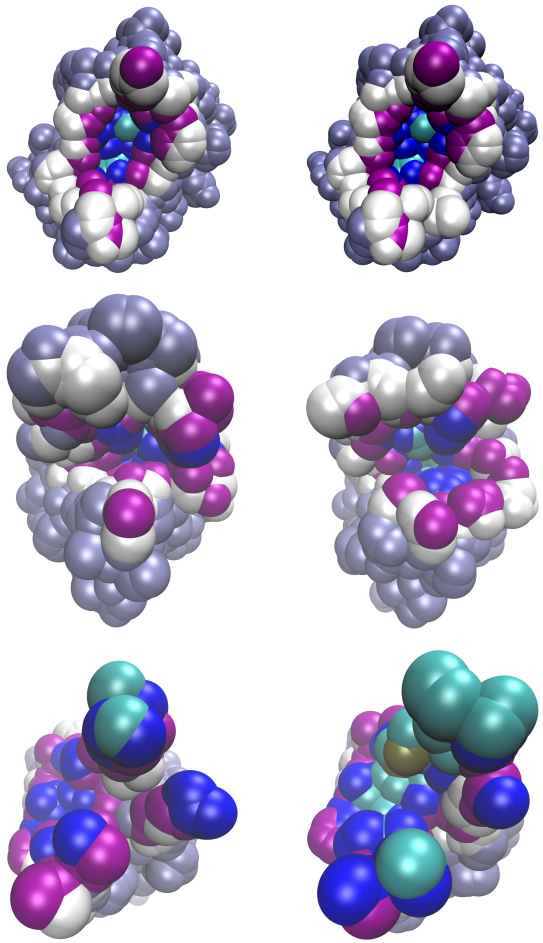
- Patch vs. prepatch on unbound partner



- Topologically rigid patches:

a third tier

- Rigid, topo-rigid, flexible



Affinity Benchmark: Predicting Binding Affinities

▷ Whole affinity benchmark:

Parameter	Pearson		Spearman		Maximal Information	
	C_{Pea}	p-value	C_{Spe}	p-value	C_{MIC}	p-value
IPL	0.31	$1.3e{-4}$	0.43	$7.6e^{-8}$	0.35	$7.6e^{-4}$
#Atoms	0.27	$1.2e{-3}$	0.37	$4.7e^{-6}$	0.24	
Depth	0.29	$4.8e{-4}$	0.35	$1.5e^{-5}$	0.26	
ΔASA	0.22	$8.9e{-3}$	0.33	$6.6e^{-5}$	0.25	
Firedock score	-0.17	$4.2e{-2}$	0.20	$1.8e^{-2}$	0.23	
I_RMSD	-0.11	$2.0e{-1}$	0.17	$4.3e^{-2}$	0.24	
#Shells	0.092	$2.7e{-1}$	-0.16	$5.4e^{-2}$	0.16	
DIS_g	0.16	$5.8e{-2}$	-0.14	$8.5e^{-2}$	0.24	
Assymetry	0.045	$5.9e{-1}$	-0.094	$2.6e^{-1}$	0.19	
DIS_t	0.029	$7.2e{-1}$	-0.089	$2.9e^{-1}$	0.20	

The Internal Path Length yields the best against $(-\ln K_d)$.

▷ As a function of the flexibility observed upon docking:

I-rmsd (Å)	ΔASA		#Atoms		Depth		IPL	
	C_{Spe}	p-value	C_{Spe}	p-value	C_{Spe}	p-value	C_{Spe}	p-value
$<$ 1 Å	0.52	$3.5e^{-6}$	0.58	$1.4e^{-7}$	0.54	$9.0e^{-7}$	0.59	$5.9e^{-8}$
in [1Å,1.5Å[	0.18	$2.7e^{-1}$	0.11	$5.0e^{-1}$	0.054	$7.5e^{-1}$	0.23	$1.7e^{-1}$
$\geq$ 1.5Å	0.26	$1.2e^{-1}$	0.34	$4.7e^{-2}$	0.34	$4.2e^{-2}$	0.41	$1.5e^{-2}$

Spearman's correlation coefficient as a function of the docking induced flexibility.

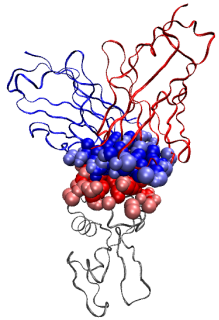
▷Ref: Kastritis et al; Protein Science (20), 2011

Modeling Protein Interfaces

- Voronoi models of protein interfaces
  F. Cazals and F. Proust and R. Bahadur and J. Janin
  Protein Science 15 (9), 2006
- Shelling Voronoi interfaces
  B. Bouvier and R. Grunberg and M. Nilges and F. Cazals
  Proteins 76 (3), 2009
- Voronoi interfaces: algorithms
  F. Cazals
  Int'l Conference on Pattern Recognition, 2010
- Modeling protein interfaces with `Intervor`
  S. Loriot and F. Cazals
  Bioinformatics 26 (7), 2010
- Shape Matching by Localized Calculations of Quasi-isometric Subsets
  F. Cazals and N. Malod-Dognin
  Int'l Conference on Pattern Recognition, 2011
- Characterizing the Morphology of Protein Binding Patches
  F. Cazals and and A. Bansal and N. Malod-Dognin
  Proteins 80 (12), 2012
- Computing the Volume of Union of Balls: a Certified Algorithm
  F. Cazals and H. Kanhere and S. Loriot
  ACM Trans. on Math. Software 38 (1), 2011

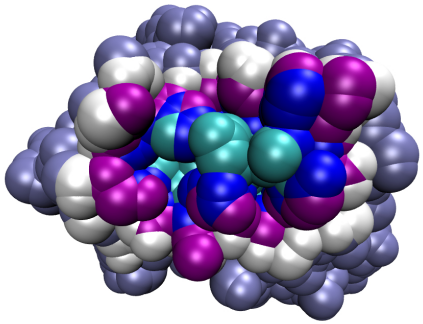
Sotware: Modeling Protein Interfaces

▷ intervor: modeling protein - protein interfaces

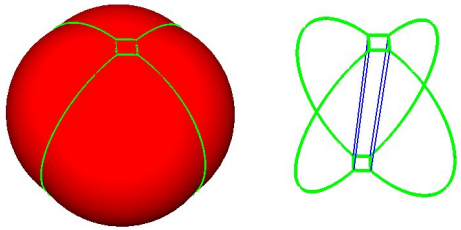


http://cgal.inria.fr/abs/Intervor;
Bioinformatics; 26 2010

▷ vorpatch: topological encoding of binding patches

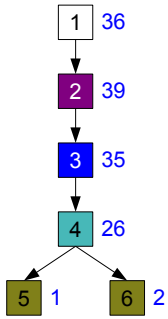


▷ vorlume: certified molecular surfaces and volumes



http://cgal.inria.fr/abs/Vorlume;
ACM Trans. Math Softw.; 2011

▷ compatch: comparing binding patches



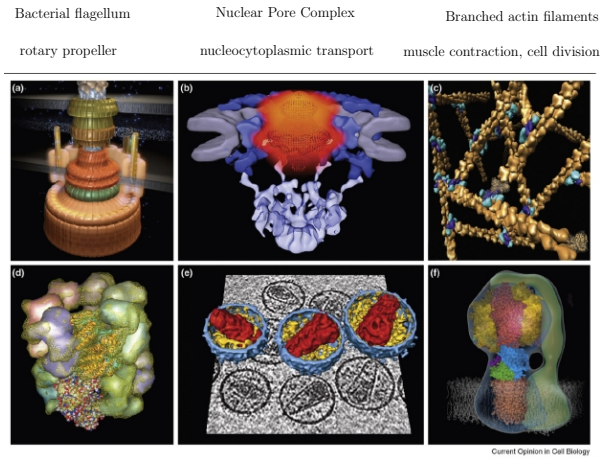
Part II

Modeling Large Protein Assemblies

Modeling High Resolution Protein Complexes

Structural Dynamics of Macromolecular Processes

Reconstructing Large Macro-molecular Assemblies



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

▷ Core questions

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

▷ Difficulties

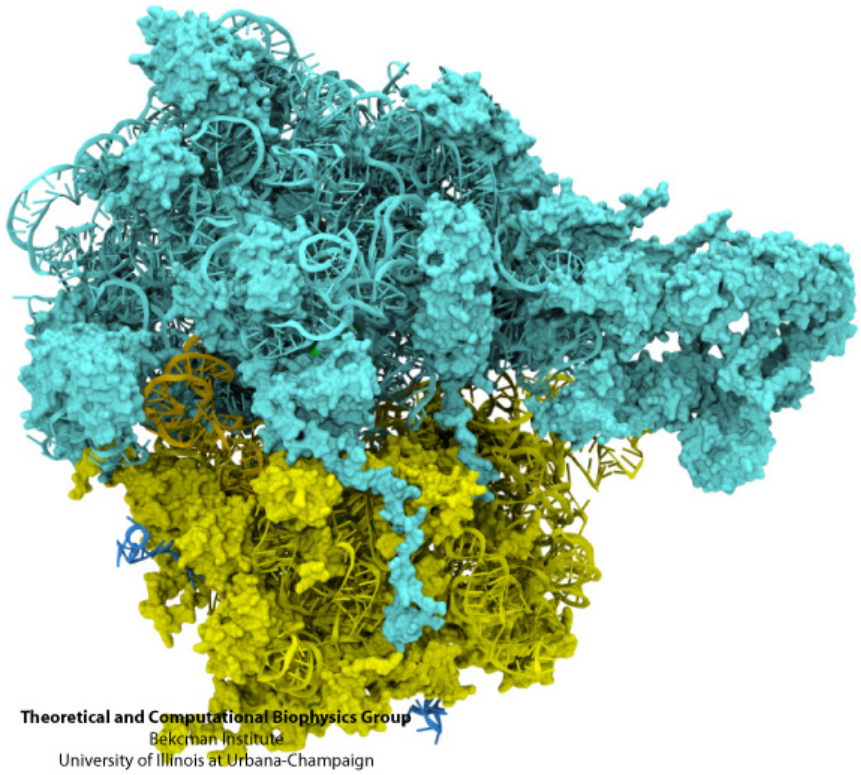
Modularity
Flexibility

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

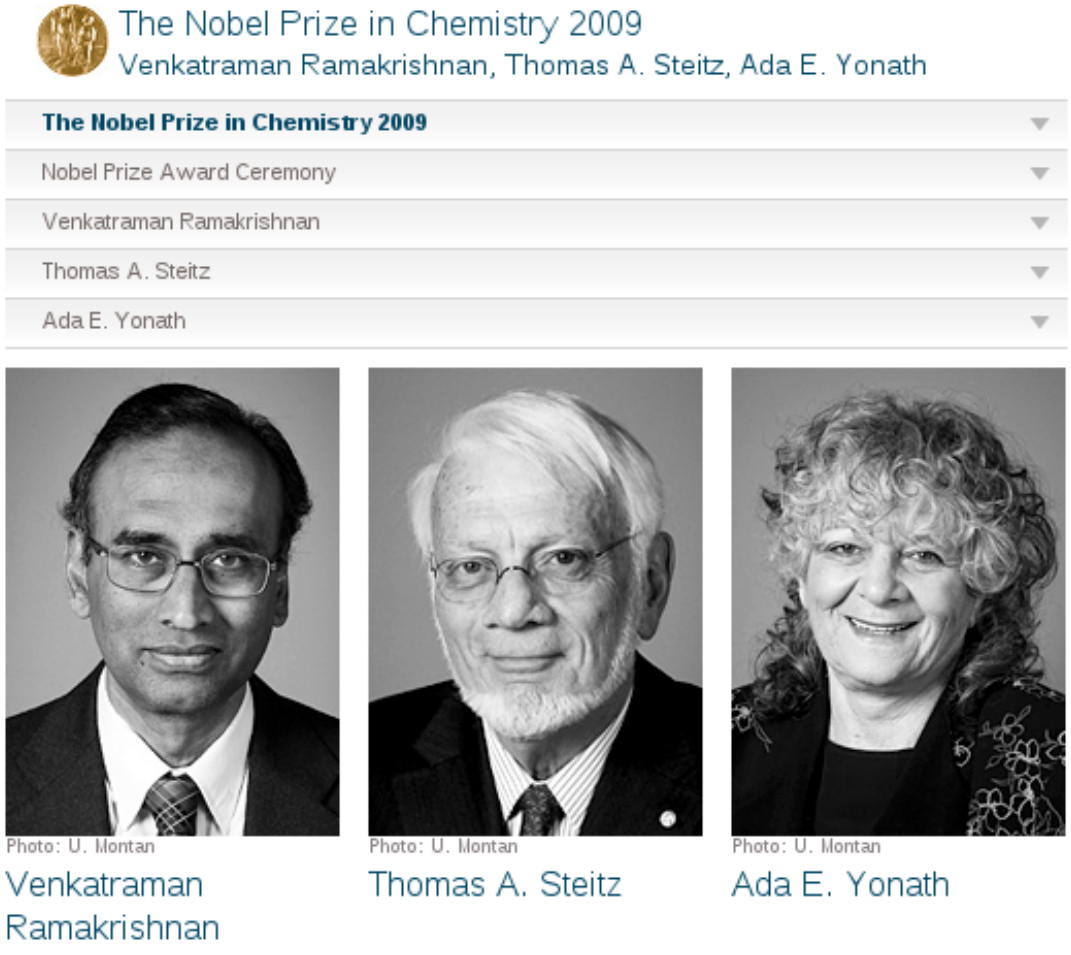
Modeling Large Protein Assemblies

The Biological Turing Machine: The Ribosome

videos-misc/video-ribosome-youtube



The 2009 Nobel Prize in Chemistry:
for studies of the structure and function of the ribosome



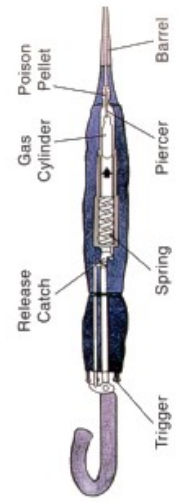
`http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2009/`

Cold War, the Ribosome, and Antibiotics

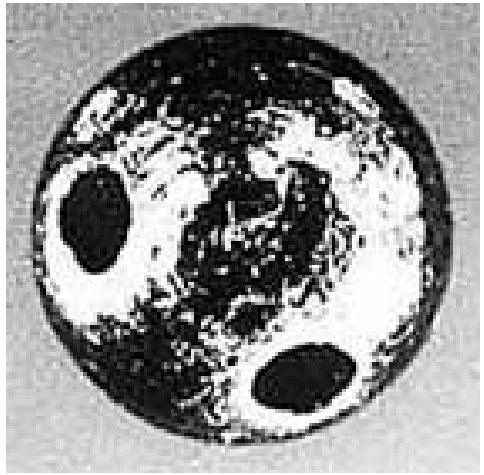
- Murder of Georgi Markov, September 1978
  - Bulgarian writer and broadcaster...
  - Communist defector working for the BBC World Service

- G. Markov

- Bulgarian umbrella



- Pinhead size pellet (1.52mm)



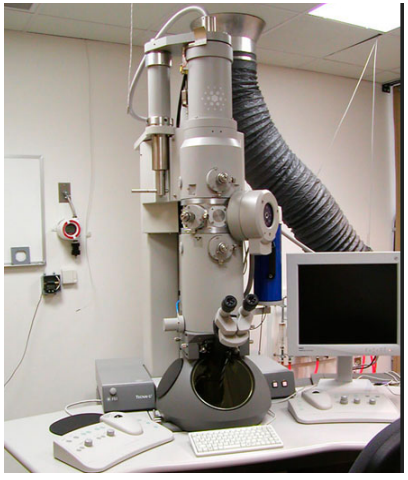
- Ricin: toxic glycoprotein inhibiting the ribosome
  - Two polypeptidic chains (30 kDa each) linked by one S-S bridge
  - Similar molecules used to block the ribosomes of bacteria (but not ours!)

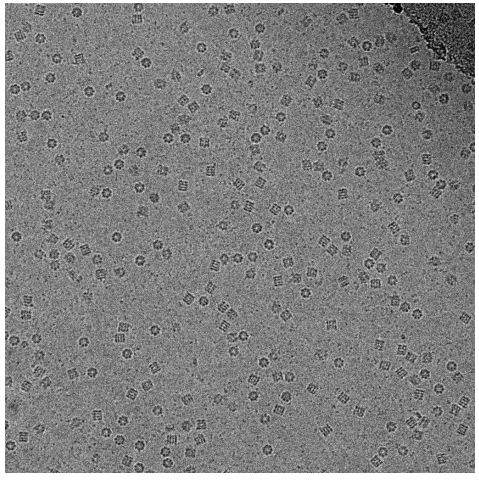
`http://fr.wikipedia.org/wiki/Georgi_Markov`

`http://www.efsa.europa.eu/fr/scdocs/doc/contam_op_ej726_ricin_`

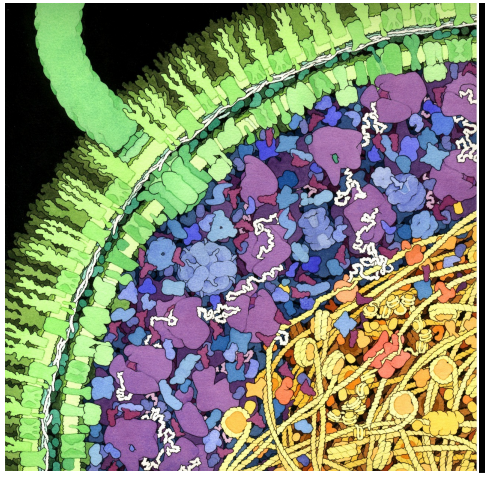
Modeling Large Protein Assemblies

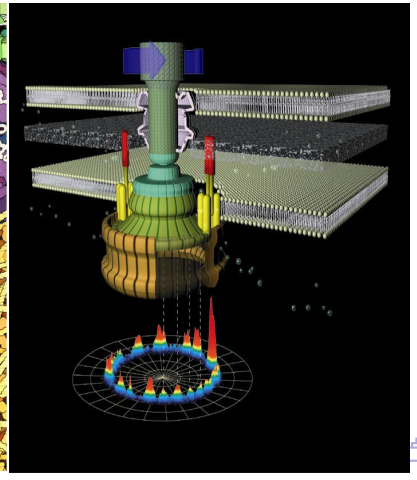
Molecular Motors, Cryo-electron Microscopy . . . and Fuzzy Density Maps





mols-archives/maps-EM-etc/EM-maps/bacteria/emd_5299-bacterial-flagel.mpeg

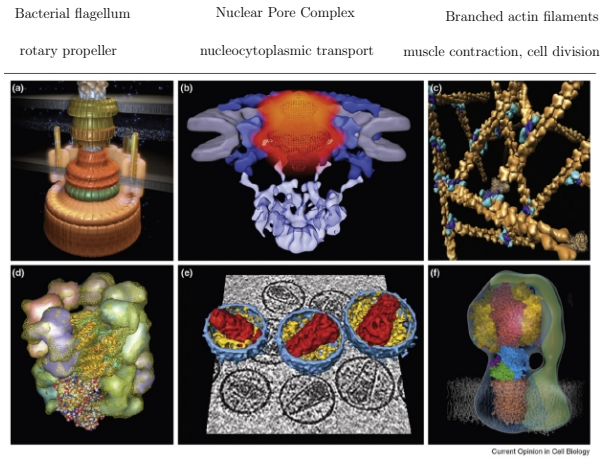




Modeling High Resolution Protein Complexes

Structural Dynamics of Macromolecular Processes

Reconstructing Large Macro-molecular Assemblies



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

▷ Core questions

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

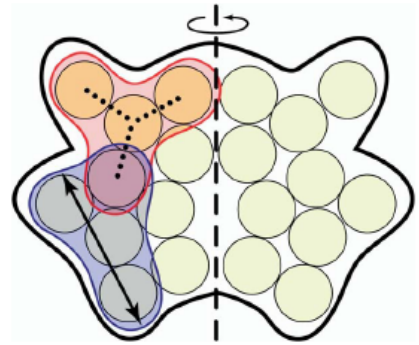
▷ Difficulties

Modularity
Flexibility

▷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,. . .)



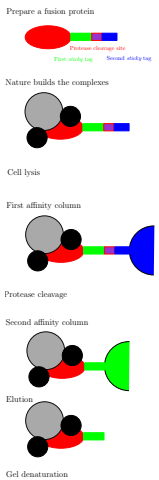
▷ Restraints, experimental data and . . . ambiguities:

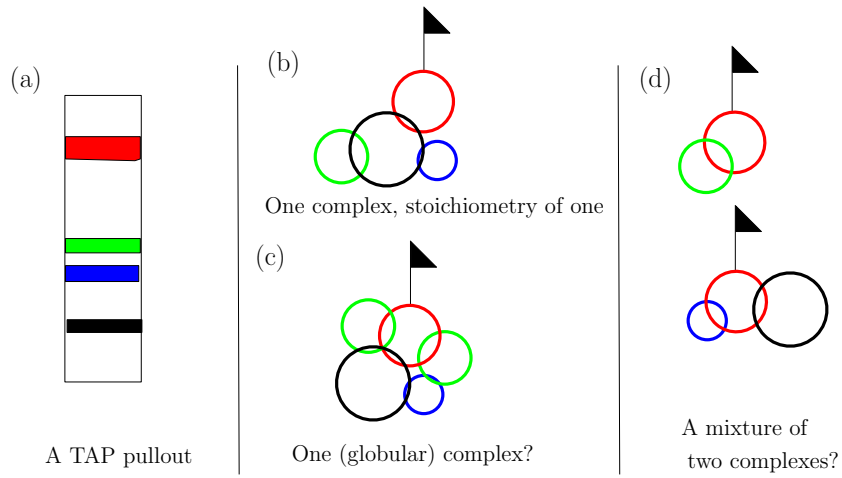
Assembly	: shape	cryo-EM	fuzzy envelopes
Assembly	: symmetry	cryo-EM	idem
Assembly	: sub-systems	mass spec.	stoichiometry
Complexes:	: interactions	TAP (Y2H, overlay assays)	stoichiometry
Instance:	: shape	Ultra-centrifugation	rough shape (ellipsoids)
Instances:	: locations	Immuno-EM	positional uncertainties

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

Tandem Affinity Purification Data

- TAP data: one composite = one pullout = list of protein types



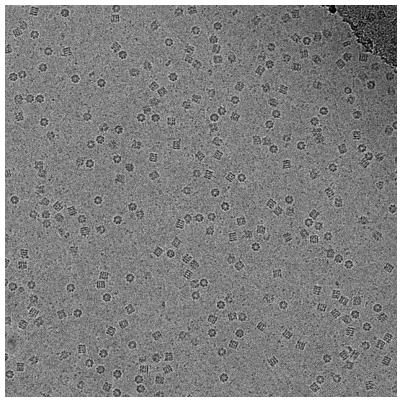


- Questions:
  - Is a given composite a complex or mixture of complexes?
  - Stoichiometry of a protein type within a complex?

- `Ref: Puig et al; Methods; 2001`

Reconstruction in Cryo-Electron Microscopy

- EM and cryo-EM



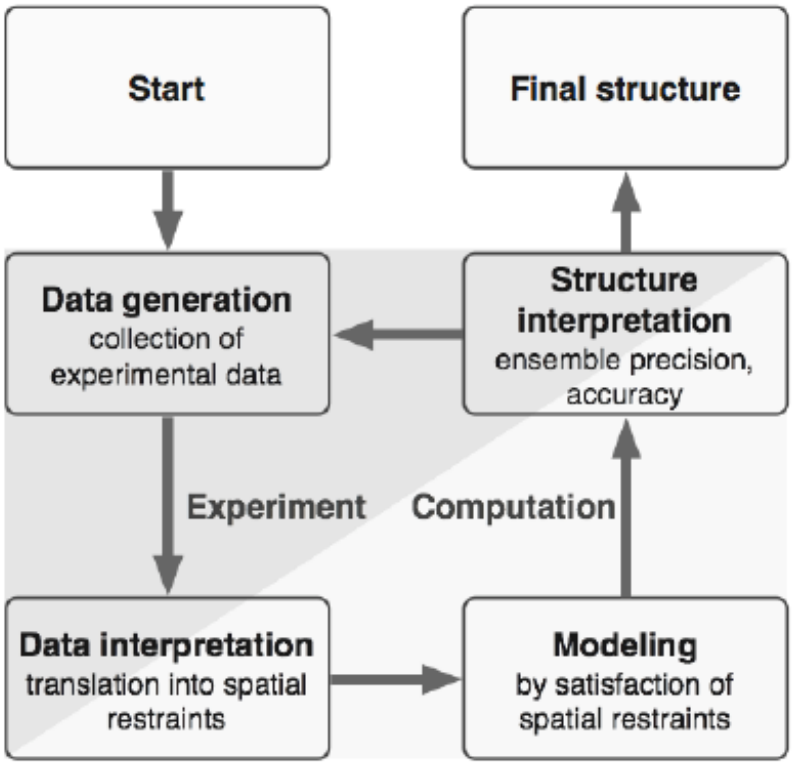
- Imaging protocol
  - electron beam in vacuum
  - sample destroyed in vacuum
    $\Rightarrow$ cryo-genized
- Large assemblies: difficulties
  - conformational diversity/stoichiometry
  - sample preparation : damaging

- Outpout: 3D density map
- Algorithms: 3D tomography

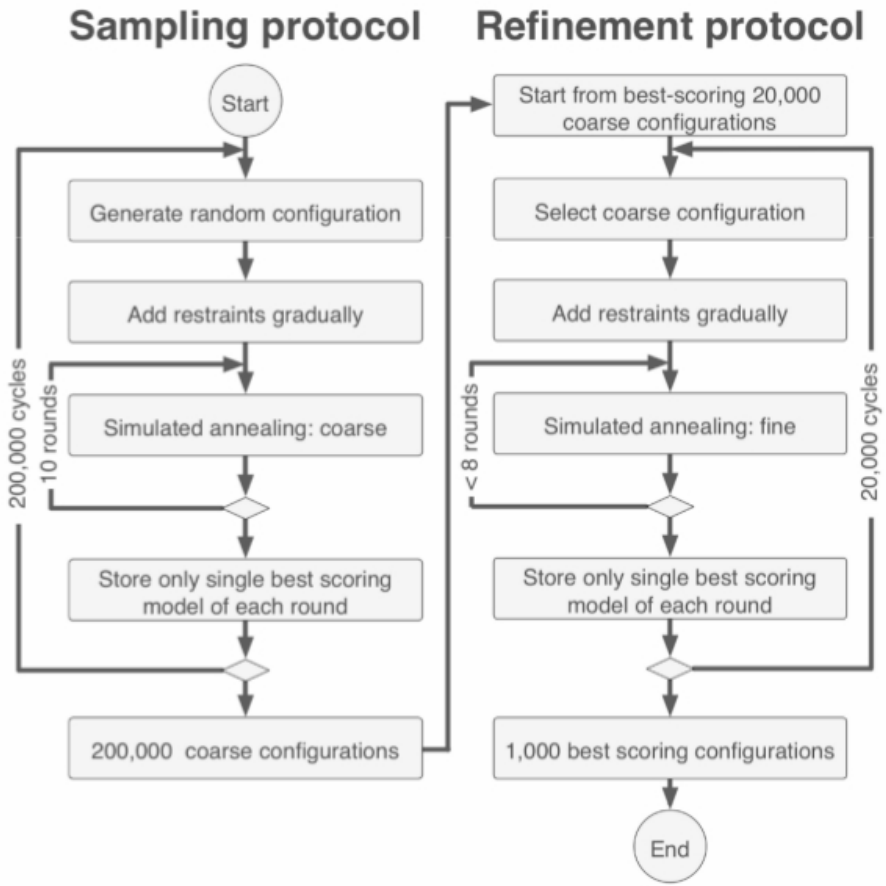
i.e. 3D recon. from 2D projections

Reconstruction by Data Integration: Protocols

▷ RDI: overall protocol



▷ RDI: optimization protocol

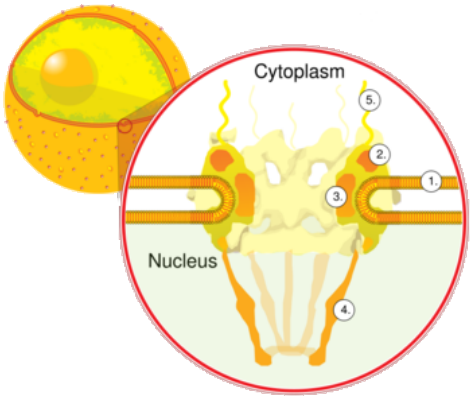


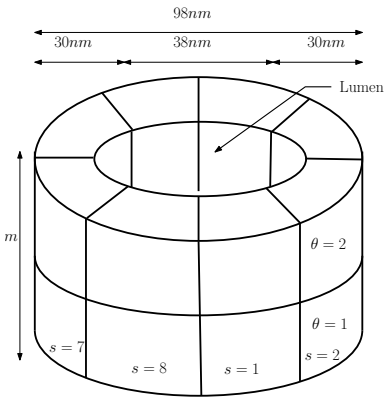
▷Ref: Alber et al; Ann. Rev. Biochem., 77, 2008

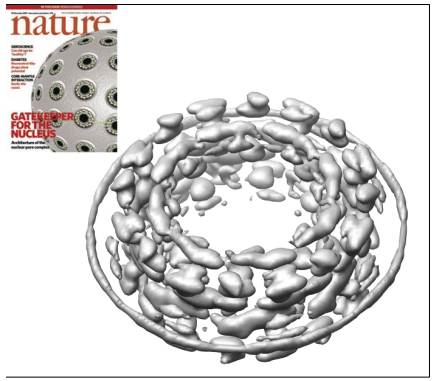
Modeling High Resolution Protein Complexes

The Nuclear Pore Complex: Structure and Reconstruction

▷ NPC: overview







– Eight-fold axial + planar symmetry
– 456 protein instances of 30 protein types ($456 = 8 \times (28 + 29)$)

▷ Reconstruction results: $N = 1000$ optimized structures (balls):
(i) blending the balls of all the instances of one type over the N structures: one 3D probability density map per protein type
(ii) superimposing these maps provides a global fuzzy model

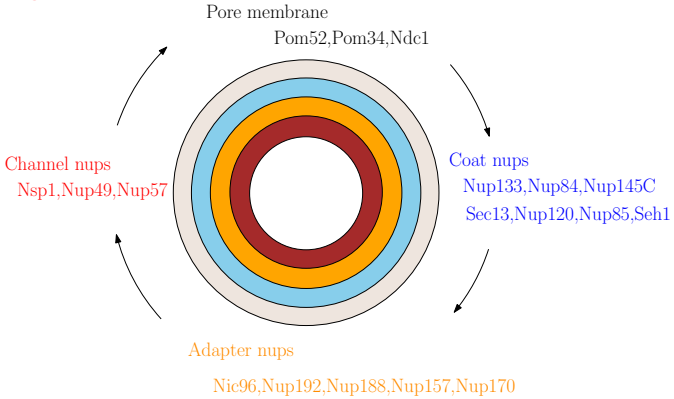
▷ Qualitative results:

Our map is sufficient to determine the relative positions within NPC
...limited precision; not to be mistaken with the density map from EM
The localization volumes . . . allow a visual interpretation of proximities

▷Ref: Alber et al; Nature; 450; 2007

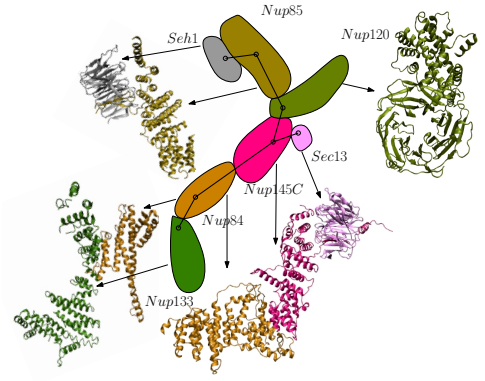
Putative Models of Sub-complexes: the Y-complex

▷ Symmetric core of the NPC



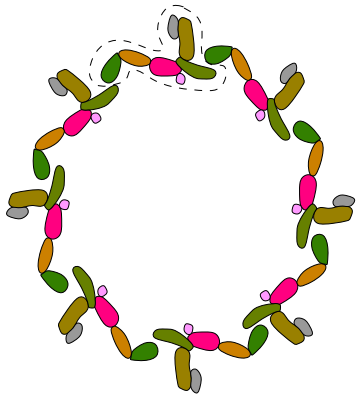
▷Ref: Blobel et al; Cell; 2007

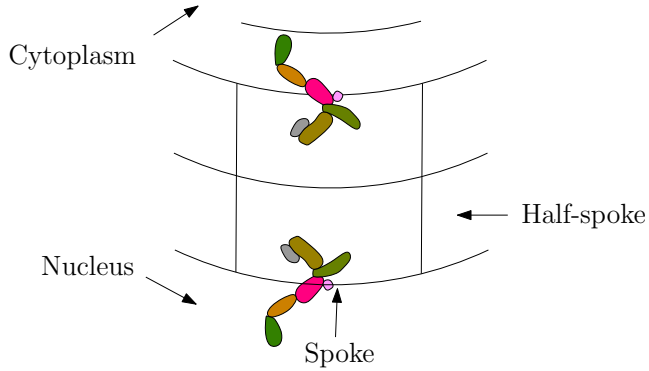
▷ The Y-complex: pairwise contacts



▷Ref: Blobel et al; Nature SMB; 2009

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





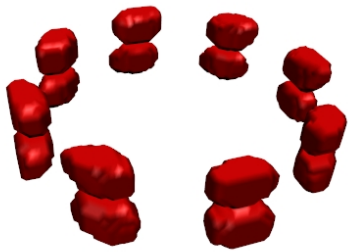
▷Ref: Seo et al; PNAS; 2009

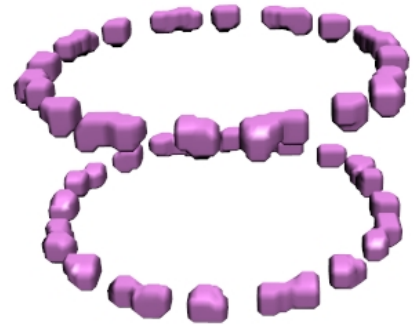
▷Ref: Brohawn, Schwarz; Nature MSB; 2009

NPC: Example Density Maps

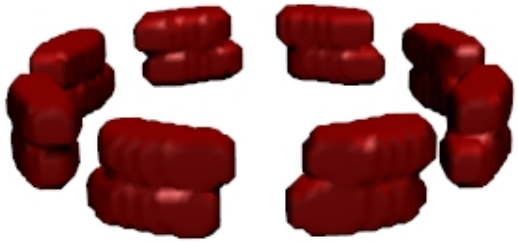
Stoichiometry vs number of connected components

- Two types of problems:
  - number of connected components vs stoichiometry
  - volume of each connected component vs. volume estimated from the sequence
- Cases: equal (Nup157); larger (Sec13)





- Cases: smaller (Nup170, Pom152)

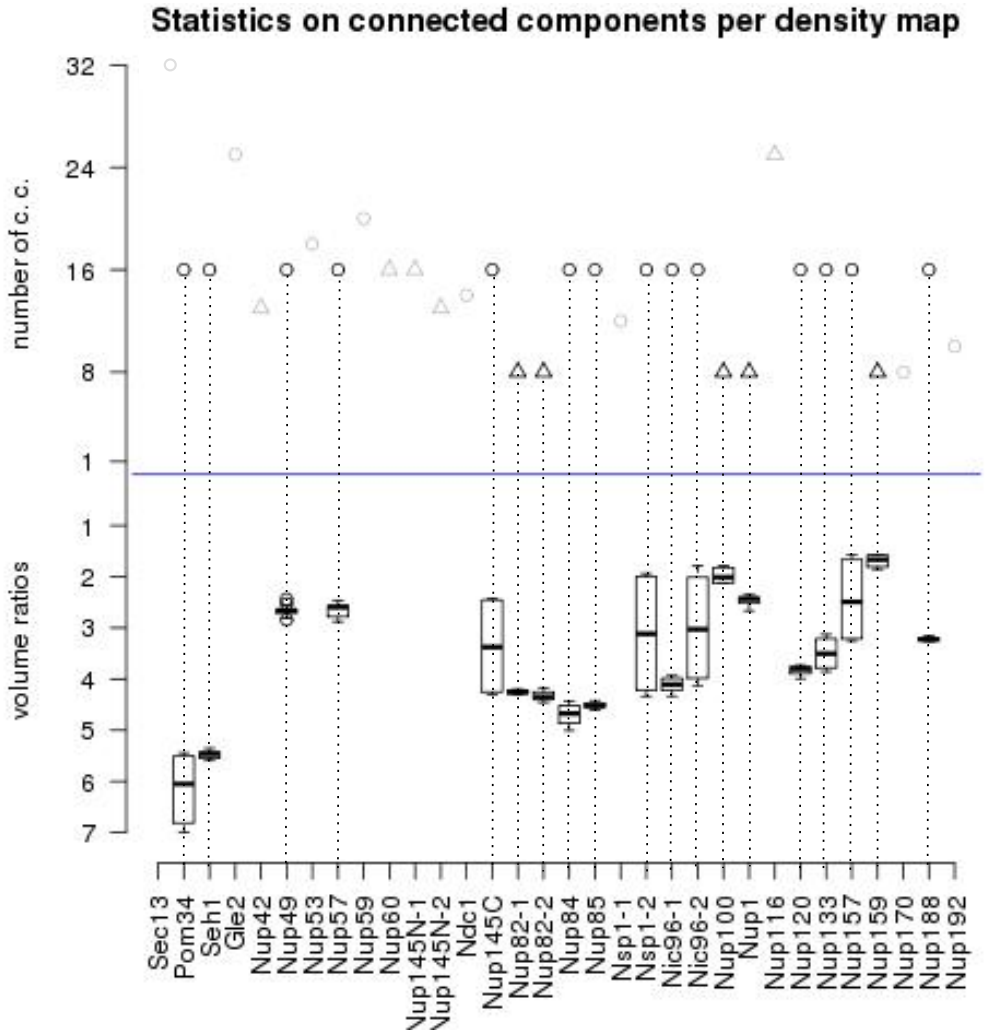


- Ref: Alber et al; Nature; 450; 2007

Uncertainties of the Density Maps

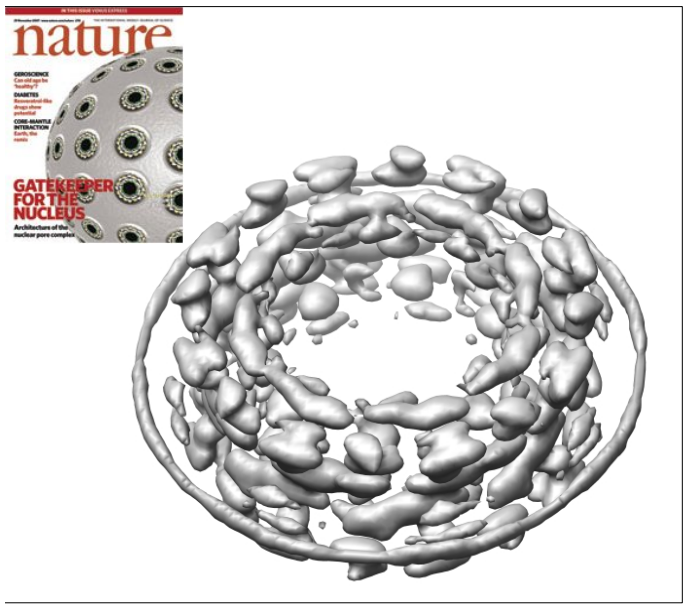
- Volume of connected components of non empty voxels vs. reference volume (estimated from the sequence)

$$\overline{V}(cc_i) = Vol(cc_i)/Vol_{ref}(P), \text{ for } i = 1, \ldots, p.$$

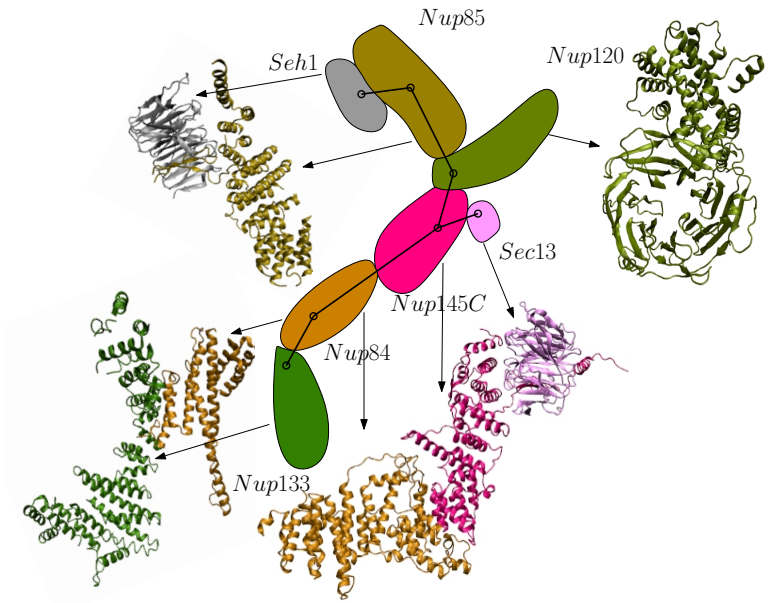


Large Assemblies: Bridging the Gap

- Global models: qualitative, low resolution
- Local models of sub-systems: atomic resolution (crystal stuctures)



Alber et al; Nature; 450; 2007

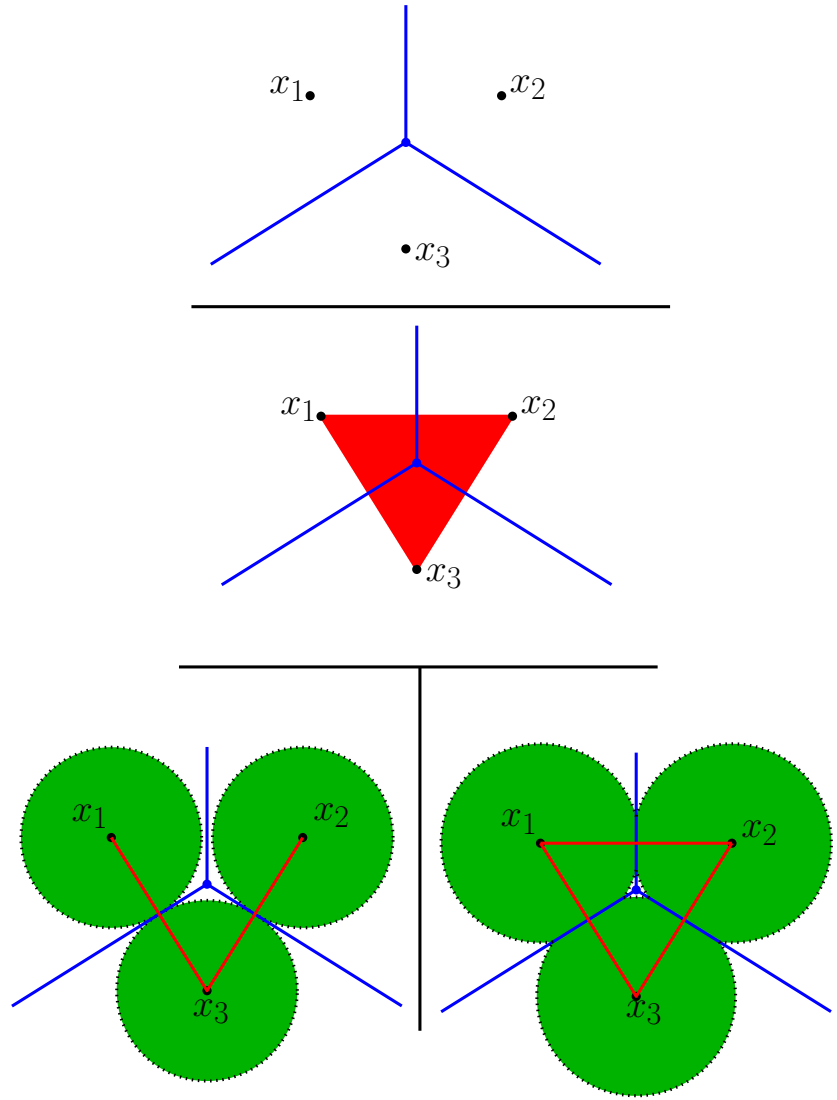


Blobel et al; Nature SMB; 2009

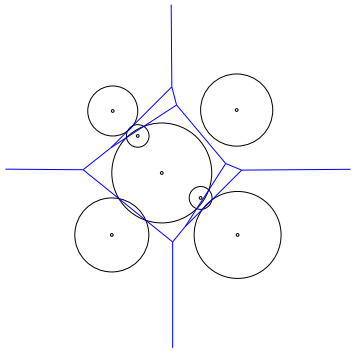
Modeling High Resolution Protein Complexes

Euclidean Voronoi diagram and α-complex

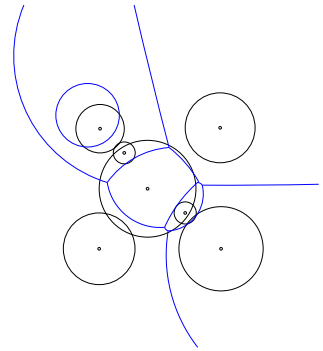
- Voronoi diagram of $\mathcal{S} = \{x_i\}$
  - Voronoi region $Vor(x_i)$:
    $\{p \mid d(p, x_i) < d(p, x_j), i \neq j\}$
- Dual complex $K(\mathcal{S})$
  - Delaunay triangulation (Euclidean case)
  - Simplex Δ: dual of $\bigcap_{x_i \in \Delta} Vor(x_i) \neq \emptyset$
- α-complex $K_\alpha(\mathcal{S})$
  - 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



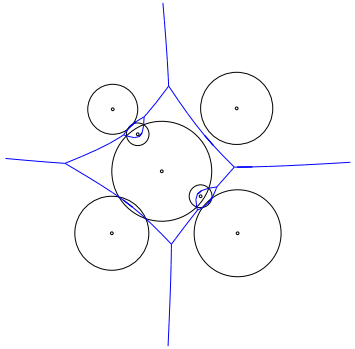
Growth Processes and Curved Voronoi diagrams



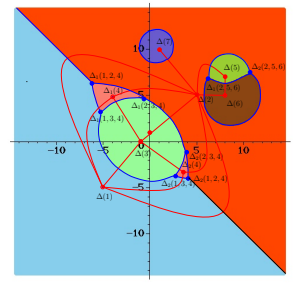
▷ Power diagram:
$d(S(c,r),p) = \|c-p\|^2 - r^2$



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



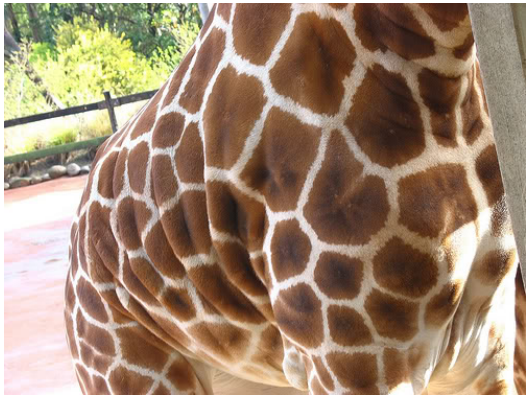
▷ Apollonius diagram:
$d(S(c,r),p) = \|c-p\| - r$

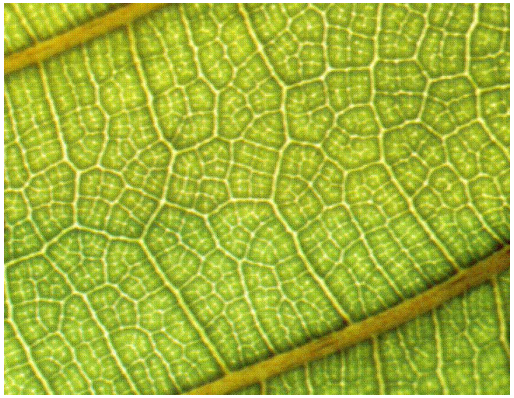
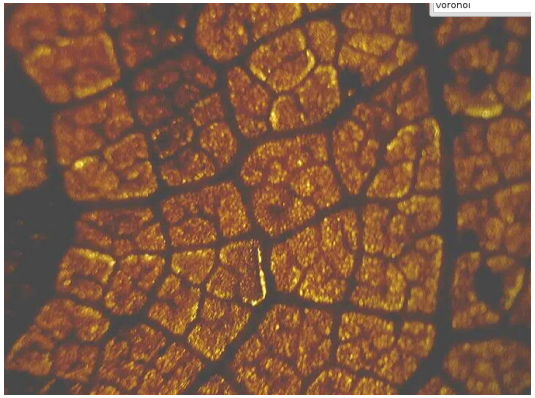


▷ **Compoundly Weighted Voronoi diagram**:
$d(S(c,\mu,\alpha),p) = \mu\|c-p\| - \alpha$

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

Voronoi diagrams in Science and Growth Processes: Gallery



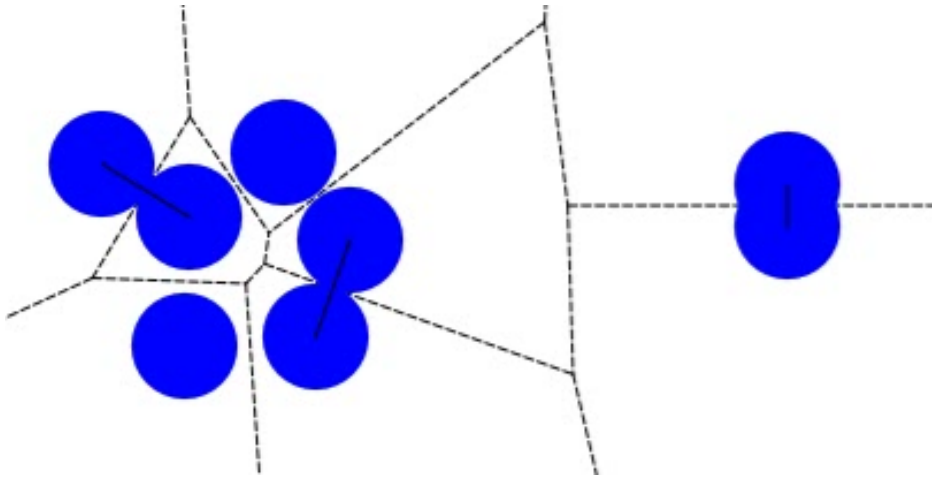



http://forum.woodenboat.com/showthread.php?112363-Voronoi-Diagrams-in-Nature

http://en.wikipedia.org/wiki/Giant's_Causeway

The α-complex: Demo

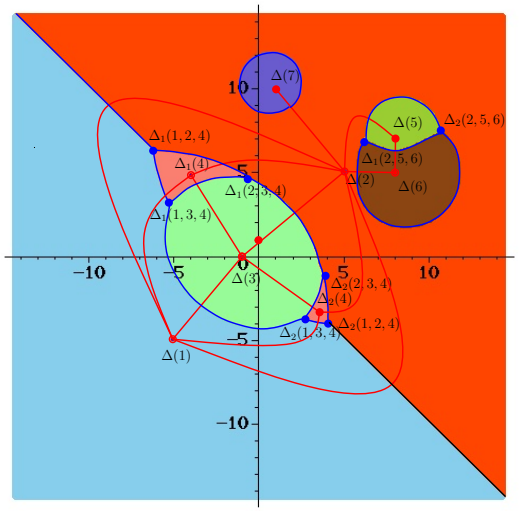
VIDEO/ashape-two-cc-cycle-video.mpeg

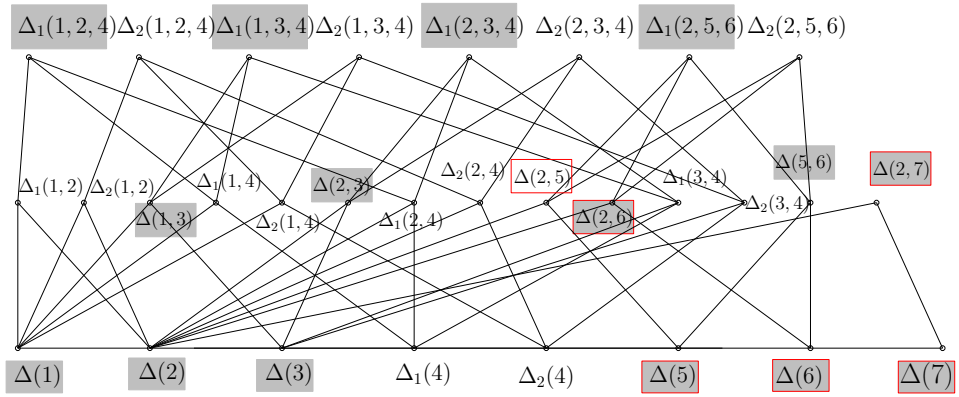


α-complex : building a simplicial complex encoding the topology of the growing balls

Compoundly-Weighted Voronoi Diagram

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



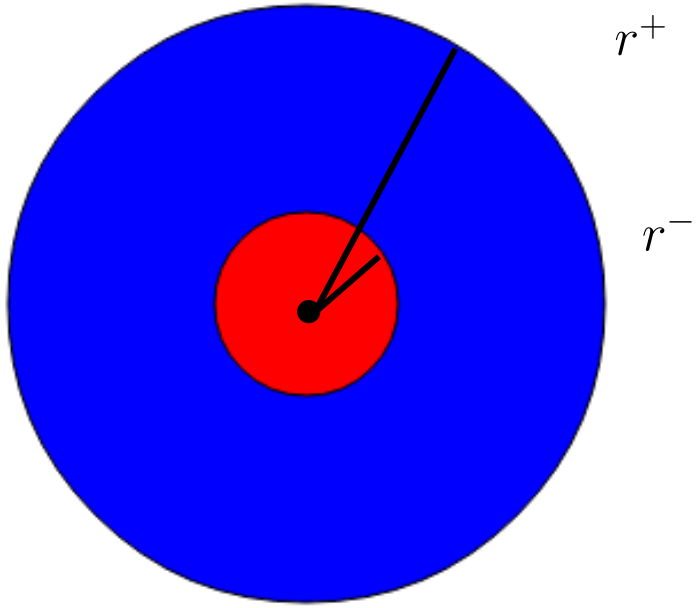


- About this diagram:
  - Bissectors are degree four algebraic surfaces
  - Voronoi regions: neither connected nor simply connected
  - Dual represented as an abstract simplicial complex
- Associated curved α-complex:
  - The affine filtration can be generalized
    - It allows computing the homological information

▷Cazals, Dreyfus; SGP 2010

Modeling High Resolution Protein Complexes

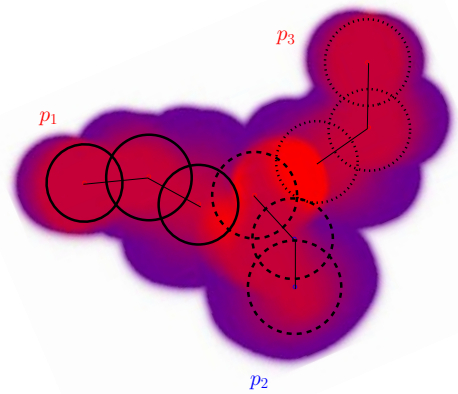
A Toleranced Ball



VIDEO/tol-ball-animation.html

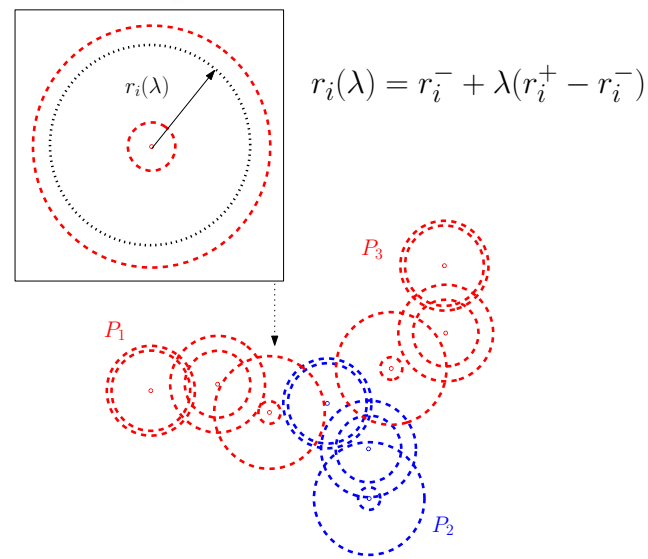
Uncertain Data and Toleranced Models: the Example of Molecular Probability Density Maps

- Probability Density Map of a Flexible Complex:
  - Each point of the probability density map: probability of being covered by a conformation



- Question:
  accommodating 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

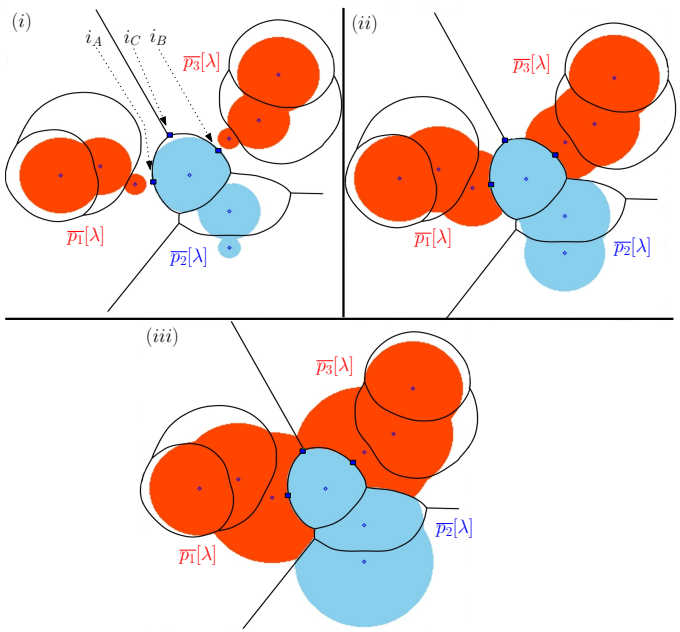


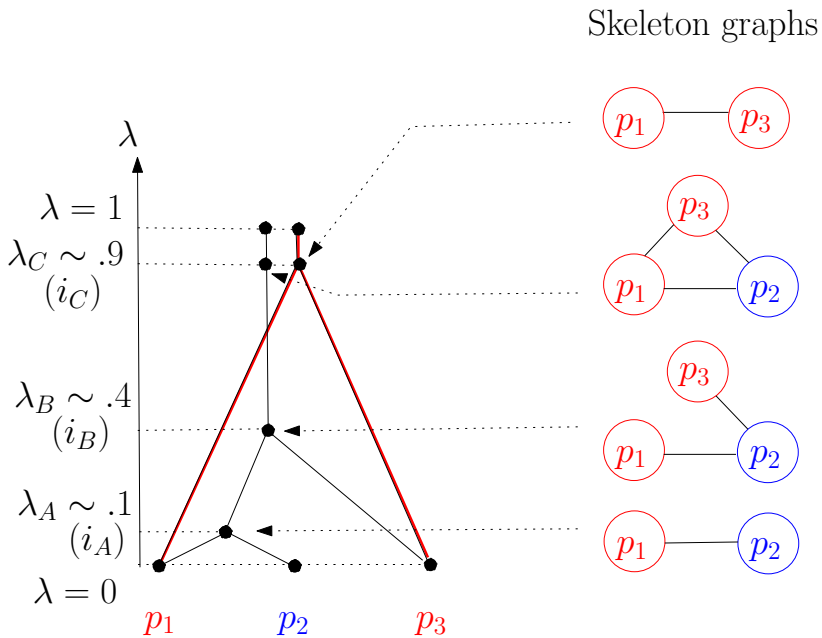
- Space-filling diagram $\mathcal{F}_\lambda$: a continuum of models
  - Radius interpolation: $r_i(\lambda) = r_i^- + \lambda(r_i^+ - r_i^-)$

- Multiplicative weights required

- Ref: Cazals, Dreyfus; Symp. Geom. Processing; 2010

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



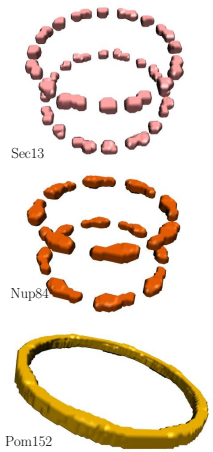


- ▷ Red-blue bicolor setting: red proteins are types singled out (e.g. TAP)
- ▷ Complexes and skeleton graphs: Hasse diagram
- ▷ Finite set of topologies: encoded into a Hasse diagram
  - – Birth and death of a complex
  - – Topological stability of a complex $s(c) = \lambda_d(C) - \lambda_b(C)$
- ▷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

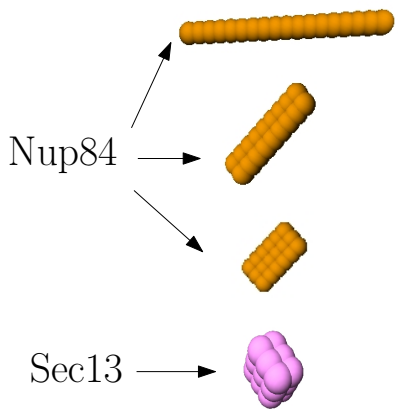
Toleranced Models for the NPC

- Input: 30 probability density maps from Sali et al.
- Output: 456 toleranced proteins
- Rationale:

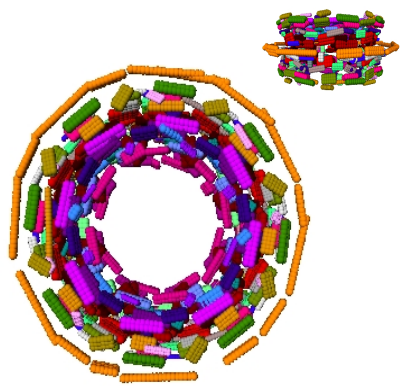
  $\longrightarrow$ assign protein instances to pronounced local maxima of the maps
- Geometry of instances:
  - four canonical shapes
  - controlling $r_i^+ - r_i^-$: w.r.t volume estimated from the sequence



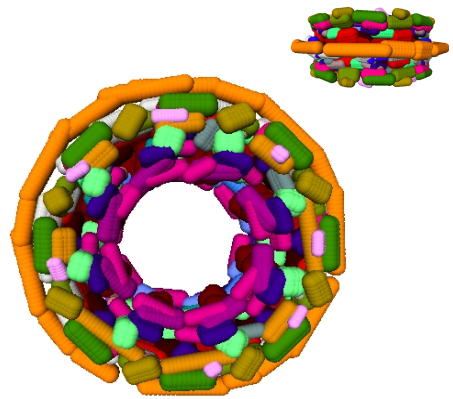
(i) Canonical shapes



(ii) NPC at $\lambda = 0$

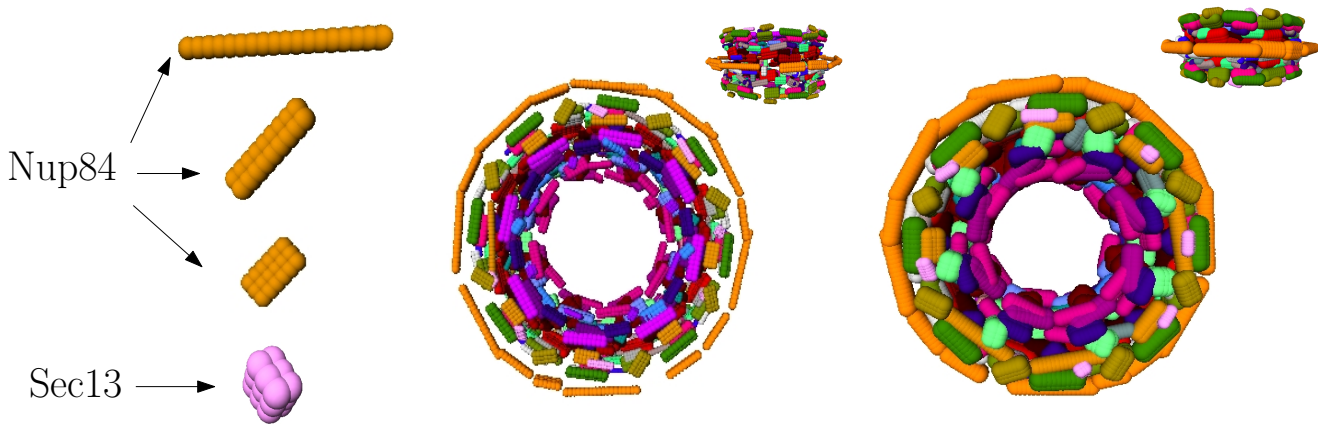


(iii) NPC at $\lambda = 1$



Density maps and local maxima
Building occupancy volumes
Building a Toleranced Model
Inferring the Hasse diagram encoding protein contacts

VIDEO/voratom-y-complex.mpeg



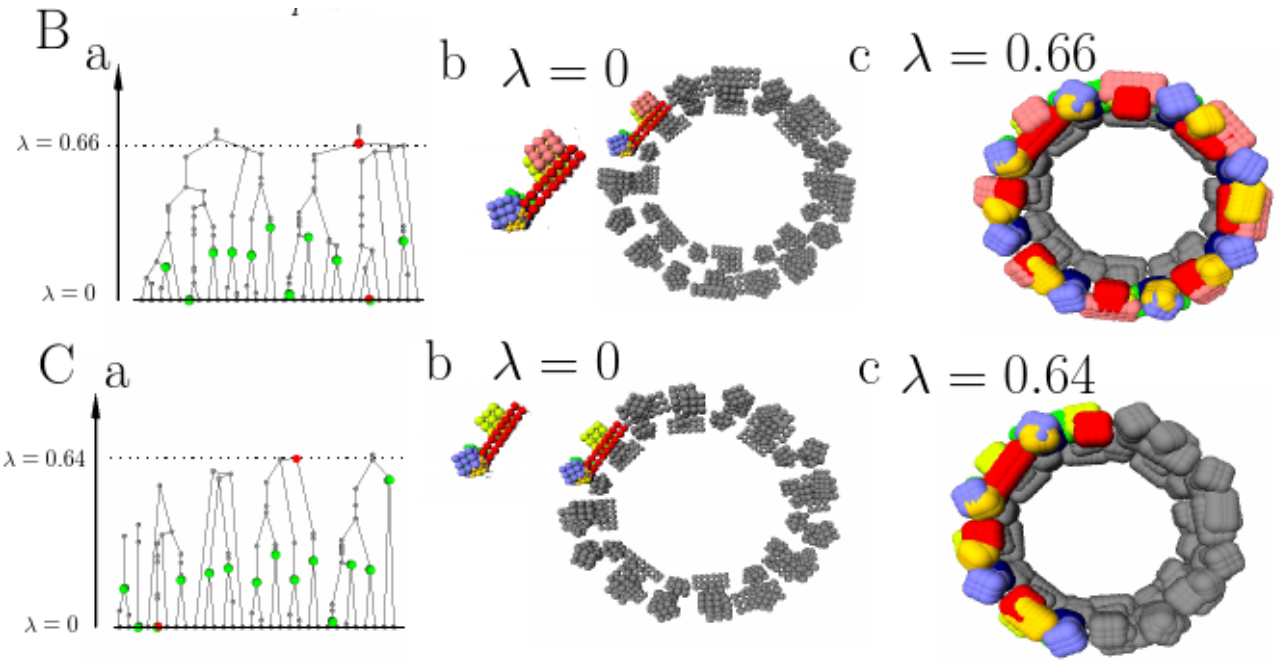
Modeling High Resolution Protein Complexes

Modeling Large Protein Assemblies

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

Geometry, Topology, Biochemistry

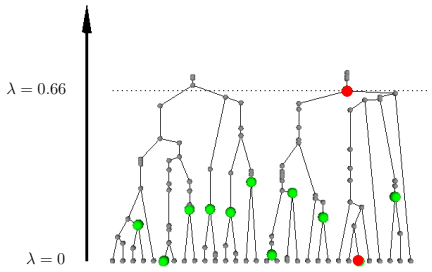
- Input:
  - Toleranced model
  - T: **set of proteins types**, the red proteins (types involved in a sub-complex)
- Output, overall assembly:
  - number of **isolated copies**: symmetry analysis
  - their topological stability: death date - birth date (cf α-shape demo)



- **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) $\lambda = 0.06$

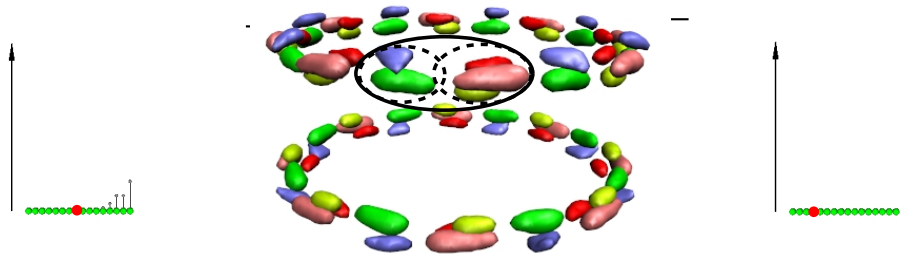
Nup85 involved in 14 / 16 contacts

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

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

(Nup120, Nup85, Seh1)

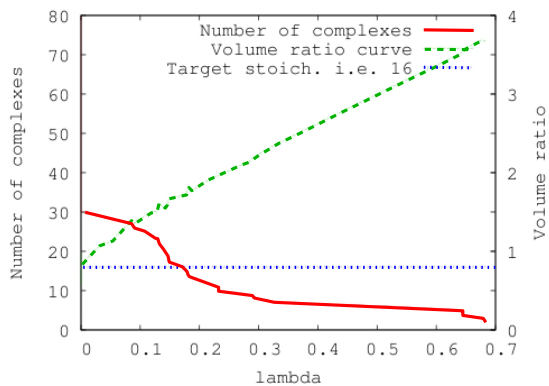
(Nup84, Nup145C, Nup133)



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



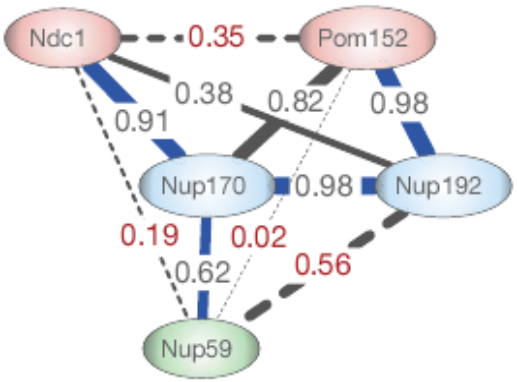
Complexes in the Hasse diagram: variation of the volume ratio as a function of λ

Ref: Harpaz, Gerstein, Chothia; Structure; 1994

Modeling Large Protein Assemblies

Contact Frequencies and their Limitations

- Proeminent contact frequencies out of the $\binom{30}{2} + 30 = 465$ pairs of protein types



– Contact frequency:
fraction of the 1000 models with $\geq$ one contact between instances of these types

– Freq. split into 3 classes, $a = 0.25, b = 0.65$:
$F_1 : f_{ij} \leq a$; $F_2 : a < f_{ij} < b$; $F_3 : b \leq f_{ij}$

– Limitations:
contact can be shallow
stoichiometry missing

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)

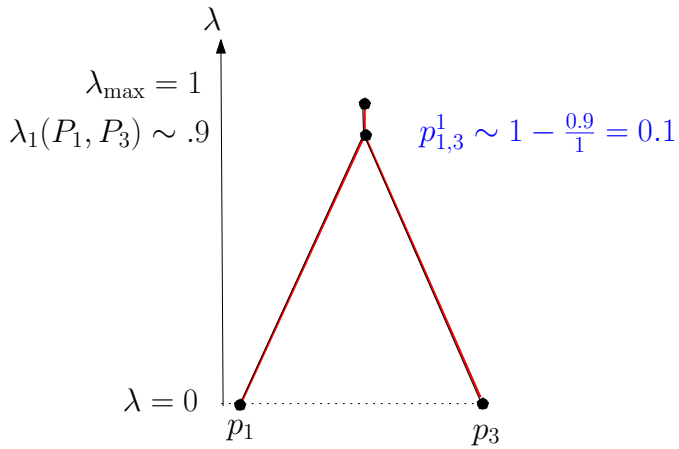
- **Contact probability** $p_{ij}^{(k)}$

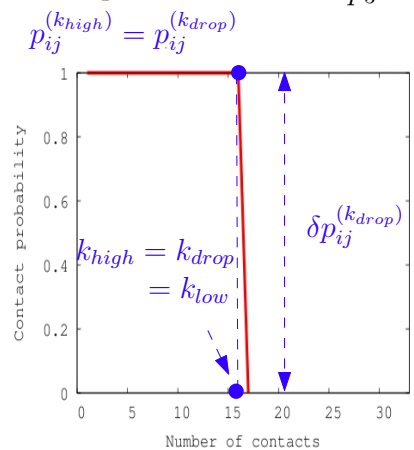
  - Consider:
    the Hasse diagram for $\lambda \in [0, \lambda_{\max}]$
    a stoichiometry $k \geq 1$

  - 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

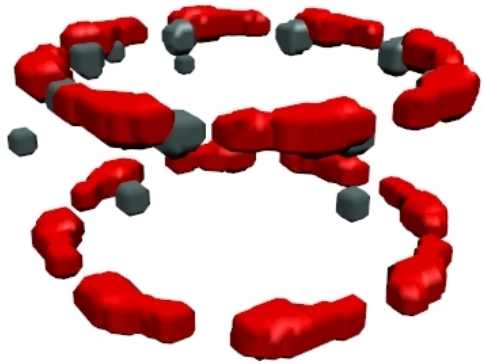




Contact Frequencies versus Contact Probabilities: Results

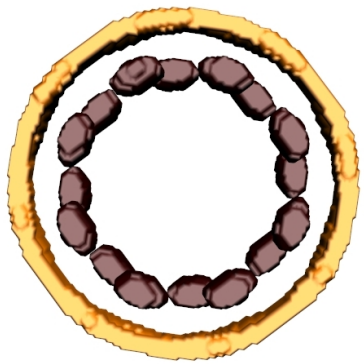
▷ **Under-represented contact in Sali et al:**

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



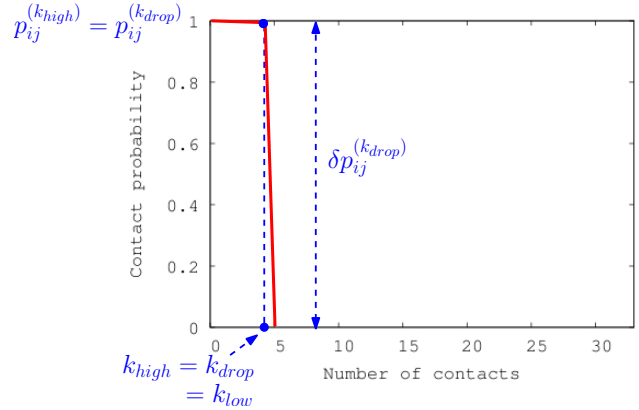
▷ **Over-represented contact in Sali et al:**

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



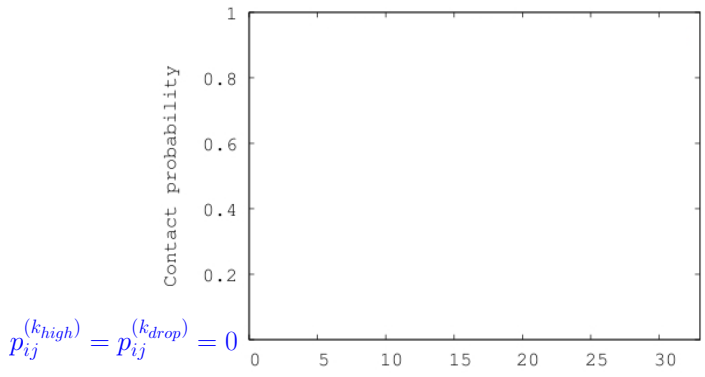
▷ **Corresponding contact curve:**

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



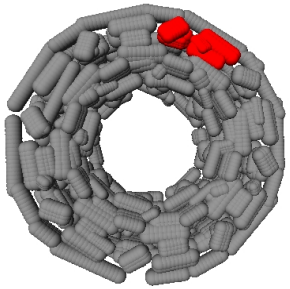
▷ **Corresponding contact curve:**

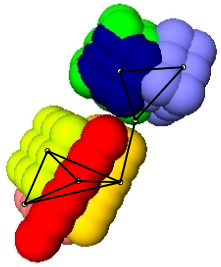
$Nup192 - Pom152 : p_{ij}^{(1)} = 0$

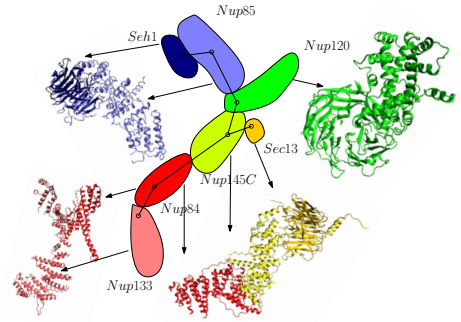


Modeling Large Protein Assemblies

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

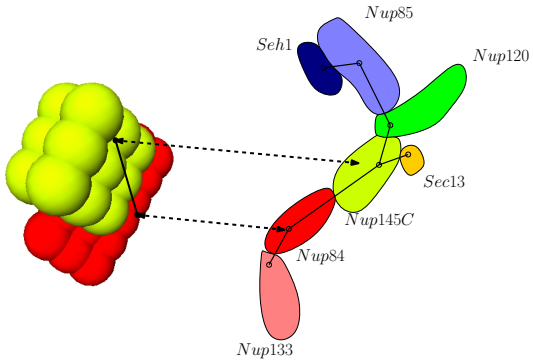


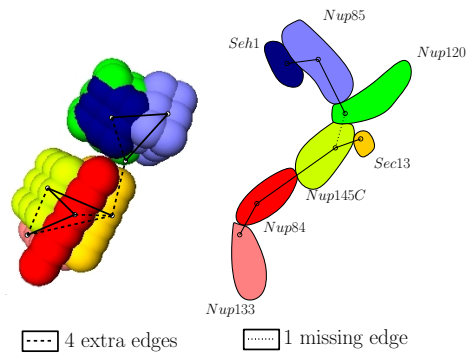




Assembly Complex: graph G_C Template: graph $G_{t|C}$

▷ Matching G_C against $G_{t|C}$: protein instance ↔ protein type; contact ↔ contact

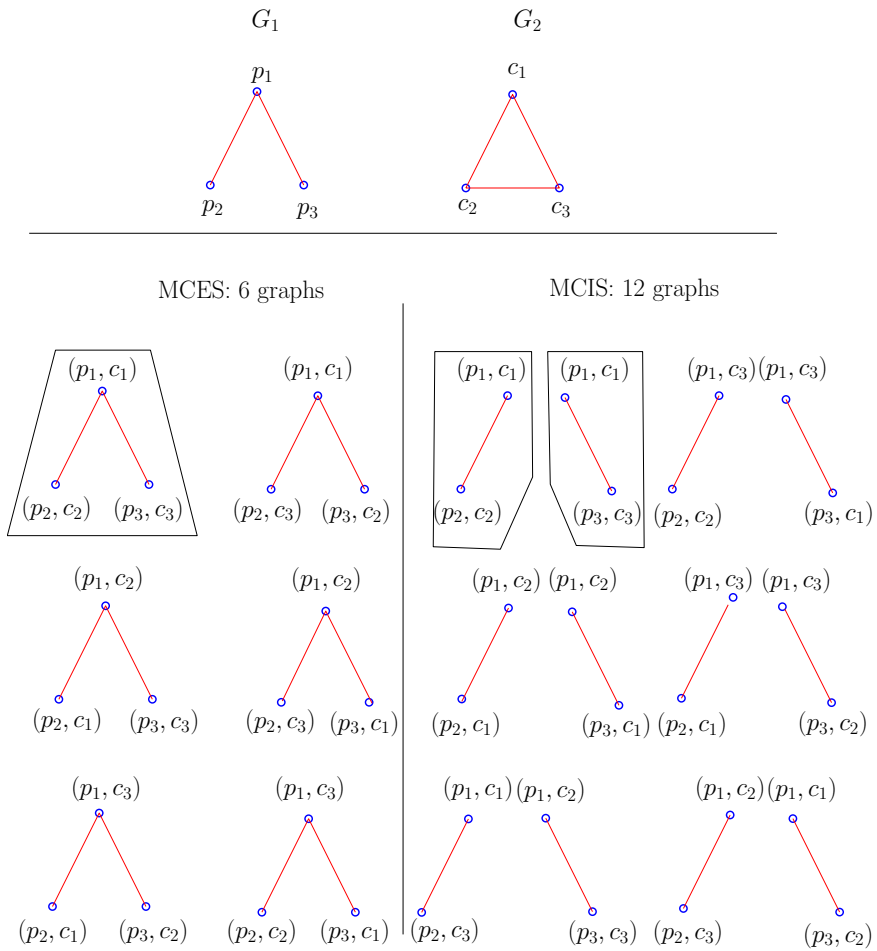




Exact superposition:
Perfect Matching

Approximate superposition:
Depleted/Complete/Exact Matching

MCES and MCIS: Illustrations



Top. Two labelled graphs G_1 and G_2
Bottom Left. The 6 MCES of G_1 and G_2
Bottom Right.: The 12 MCIS of G_1 and G_2.

▷Ref: Cazals, Karande; Theoretical Computer Science; 349 (3), 2005

▷Ref: Koch; Theoretical Computer Science; 250 (1-2), 2001

Matchings: Pre-requisites

▷ Comparing G_C and $G_{t|C}$

G_C: contact graph associated to a complex (node of Hasse diagram)
$G_{t|C}$: graph template (crystal structure, model, . . .)

▷ Matching: a mapping between vertices and edges of $G_{t|C}$ and G_C, defined either by a MCIS or a MCES.

▷ Qualifying matchings:

- $V^{\sim}$: *Matched protein type(s):* a protein type of $G_{t|C}$ with a corresponding instance in G_C.
- V^{-}: *Missing protein type(s):* a protein type of $G_{t|C}$ with no corresponding instance in G_C.
- $E^{\sim}$: *Matched contact(s):* a contact in $G_{t|C}$ with a counterpart in G_C.
- E^{-}: *Missing contact(s):* a contact in $G_{t|C}$ whose protein types match instances in G_C but with no corresponding contact in G_C.
- E^{+}: *Extra contact(s):* a contact in G_C whose protein instances match types in $G_{t|C}$ but with no corresponding contact in $G_{t|C}$.

▷ Signature of the matching M:

$$S(G_t; G_C; M) = \{V^{\sim}, V^{-}, E^{\sim}, E^{-}, E^{+}\}. \quad (1)$$

Matchings: The Four Classes

- Case study: $V^-\{=,\neq\}\emptyset$, $E^- \cup E^+\{=,\neq\}\emptyset$

Depleted matching:

$$V^- \neq \emptyset \text{ and } E^- \cup E^+ \neq \emptyset \tag{2}$$

Complete matching:

$$V^- = \emptyset \text{ and } E^- \cup E^+ \neq \emptyset \tag{3}$$

Exact matching:

$$V^- \neq \emptyset \text{ and } E^- \cup E^+ = \emptyset \tag{4}$$

Perfect matching:

$$V^- = \emptyset \text{ and } E^- \cup E^+ = \emptyset \tag{5}$$

- Calculations, depleted and complete matchings: a two-stage process

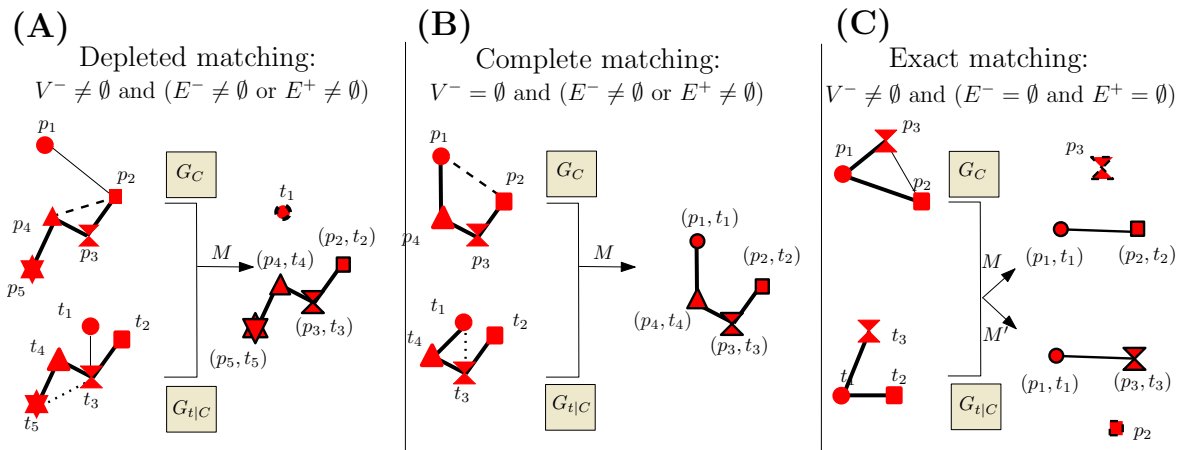
  MCES calculation between C of roots of the Hasse diagram and associated $G_{t|C}$
  search for ancestor of C yielding the same contacts, but minimizing $\mid E^+ \mid$

- Calculations: exact and perfect matchings

MCIS calculation between nodes G_C and $G_{t|C}$ maximality check
(matching not contained in larger matching for successor nodes)

Matchings: Illustrations

- 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)



Signature \ Matching	M
$V^\sim$ Matching Protein Types	(p_2,t_2) (p_3,t_3) (p_4,t_4) (p_5,t_5)
V^- Missing Protein Types	t_1
$E^\sim$ Matching Contacts	$(t_2,t_3)\leftrightarrow(p_2,p_3)$ $(t_3,t_4)\leftrightarrow(p_3,p_4)$ $(t_4,t_5)\leftrightarrow(p_4,p_5)$
E^- Missing Contacts	(t_3,t_5)
E^+ Extra Contacts	(p_2,p_4)

Signature \ Matching	M
$V^\sim$ Matching Protein Types	(p_1,t_1) (p_2,t_2) (p_3,t_3) (p_4,t_4)
V^- Missing Protein Types	$\emptyset$
$E^\sim$ Matching Contacts	$(t_2,t_3)\leftrightarrow(p_2,p_3)$ $(t_3,t_4)\leftrightarrow(p_3,p_4)$ $(t_1,t_4)\leftrightarrow(p_1,p_4)$
E^- Missing Contacts	(t_1,t_3)
E^+ Extra Contacts	(p_1,p_2)

Signature \ Matchings	M	M'
$V^\sim$ Matching Protein Types	(p_1,t_1) (p_2,t_2)	(p_1,t_1) (p_3,t_3)
V^- Missing Protein Types	t_3	t_2
$E^\sim$ Matching Contacts	$(t_1,t_2)\leftrightarrow(p_1,p_2)$	$(t_1,t_3)\leftrightarrow(p_1,p_3)$
E^- Missing Contacts	$\emptyset$	$\emptyset$
E^+ Extra Contacts	$\emptyset$	$\emptyset$

▷Ref: Cazals, Karande; Theoretical Computer Science; 349 (3), 2005

▷Ref: Koch; Theoretical Computer Science; 250 (1-2), 2001

Matchings for the Y complex: Illustrations

Depleted matchings:

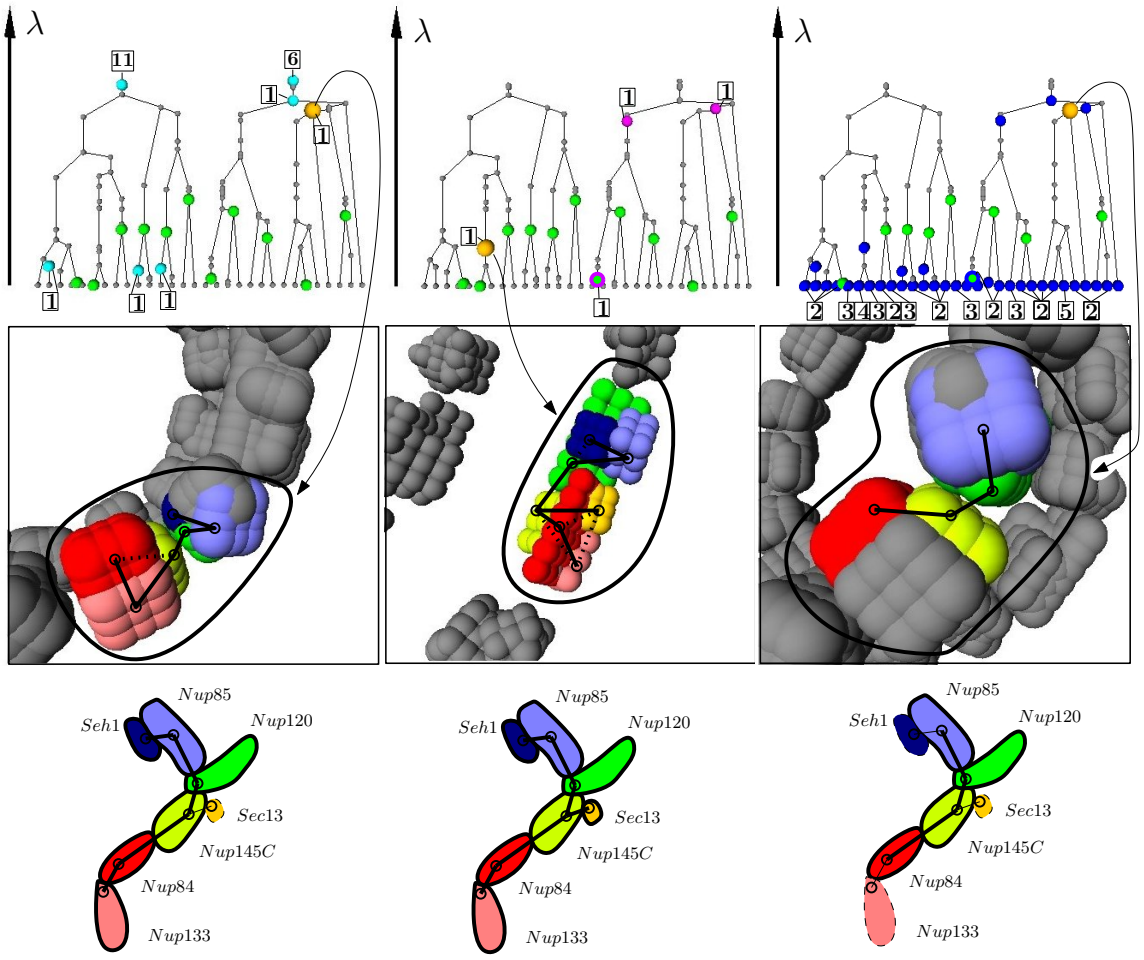
- missing nodes...
- problems on edges

Complete matchings:

- all nodes...
- problems on edges

Exact matchings:

- missing nodes...
- but edges perfect for these nodes



▷ Application: recovering the 16 copies of the Y: De. = 10+2; Co.: 4; Ex. : 0

▷Ref: Dreyfus, Doye, Cazals; Proteins, 2013

Matchings for the Y complex: Details

(A) Depleted matchings

Template; tag	#	$V^{\sim}$ (sub-complexes)	$\mid V^{-} \mid$	$\mid E^{\sim} \mid$	$\mid E^{-} \mid$	min $\mid E^{+} \mid$	max $\mid E^{+} \mid$	min $\overline{V}_{\lambda}$	max $\overline{V}_{\lambda}$
$G_t(Y);M_D(1)$	2		4	2	0	1/1	1/1	4.49	4.71
$G_t(Y);M_D(2)$	8		3	3	0	3/3	3/3	4.49	5.65
$G_t(Y);M_D(3)$	10		4	2	0	1/1	1/1	1.17	5.83
$G_t(Y);M_D(4)$	2		1	5	0	2/10	2/10	3.68	4.56

(B) Complete matchings

Template; tag	#	$V^{\sim}$ (sub-complexes)	$\mid E^{\sim} \mid$	$\mid E^{-} \mid$	min $\mid E^{+} \mid$	max $\mid E^{+} \mid$	min $\overline{V}_{\lambda}$	max $\overline{V}_{\lambda}$
$G_t(Y);M_C(1)$	4		6	0	4/15	7/15	1.00	4.95

(C) Exact matchings

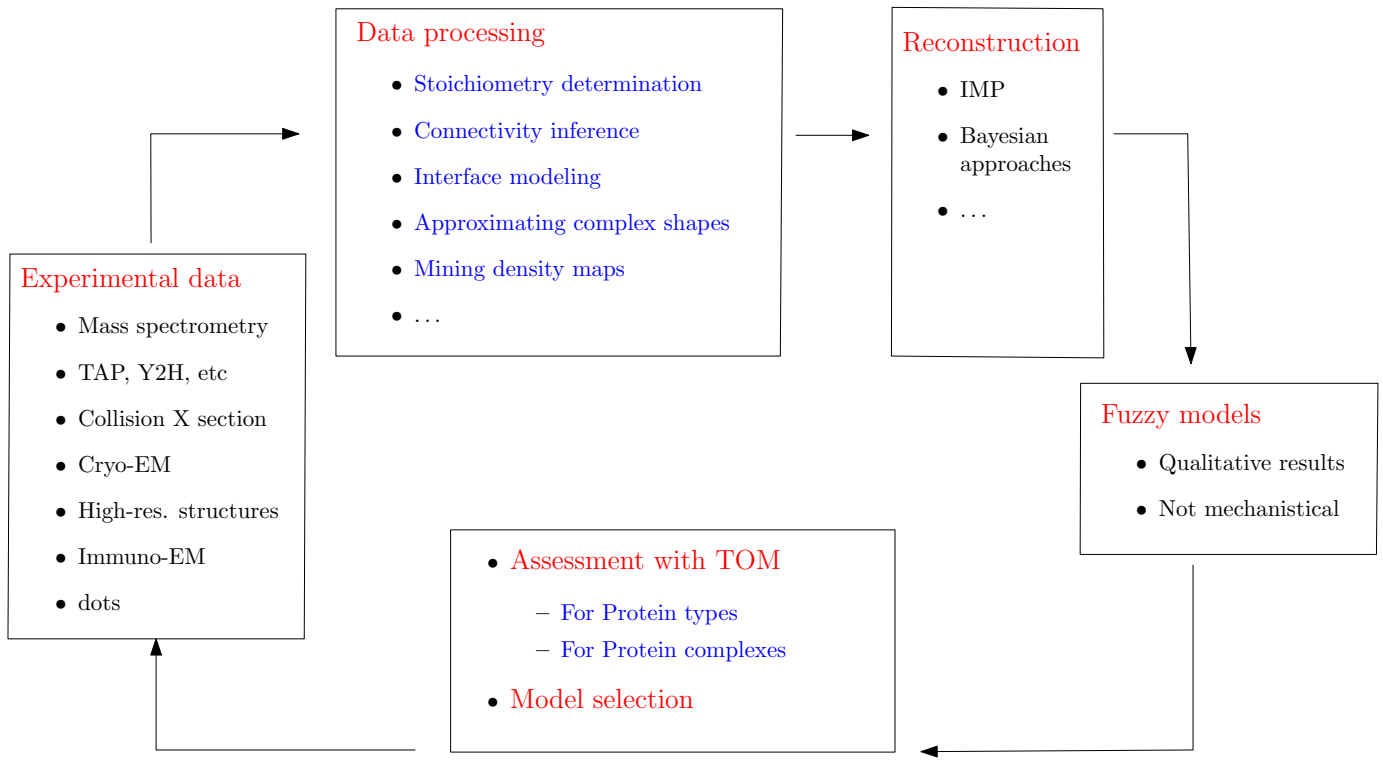
Template; tag	#	$V^{\sim}$ (sub-complexes)	$\mid V^{-} \mid$	min $\overline{V}_{\lambda}$	max $\overline{V}_{\lambda}$
$G_t(Y);M_E(1)$	5		3	1.02	3.42
$G_t(Y);M_E(2)$	15		5	0.77	0.89
$G_t(Y);M_E(3)$	1		4	0.86	0.86
$G_t(Y);M_E(4)$	5		4	0.78	0.86
$G_t(Y);M_E(5)$	7		5	0.81	0.86
$G_t(Y);M_E(6)$	6		5	0.80	0.84
$G_t(Y);M_E(7)$	1		3	1.05	1.05
$G_t(Y);M_E(8)$	4		6	0.54	0.63
$G_t(Y);M_E(9)$	16		5	0.77	1.27
$G_t(Y);M_E(10)$	10		5	0.88	0.91
$G_t(Y);M_E(11)$	1		4	2.15	2.15

Modeling High Resolution Protein Complexes

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



Outlook

- A new class of modeling problems
  - $O(1)$ chains: classical (pairwise) docking
  - $O(10)$ chains: docking crystal structures within cryo-EM envelopes
  - $O(100)$ chains: reconstruction by data integration
- Toleranced models: a modeling paradigm to incorporate uncertainties
  - Density maps in general: cryo-EM, probability density maps, etc
  - Positional uncertainties - soft docking
  - Atomic models: temperature factors
- A triple model assessment, local and global
  - Geometric : volume computation, symmetry analysis
  - Topological: stability, pairwise contacts
  - Biochemical: contacts and location of proteins

- Applications to coherence analysis and model selection
  $\longrightarrow$ getting the best out of global models obtained from data integration

- Compoundly weighted Voronoi diagram
  - Complicated ... yet encodes important features of the toleranced model
  - Incremental construction – in progress

Publications and Software

- Papers available from `http://team.inria.fr/abs/publications`

  Toleranced Models, theory
  : Symp. on Geometry Processing 2010

  Toleranced Models (isolated copies, contact probabilities)
  : Proteins 80(9), 2012

  Toleranced Models (graphical analysis)
  : Proteins 2013, in press

  Geometric optimization for collections of balls
  : ACM Trans. on Math. Soft. 2010, ACM IEEE Trans. CBB 2011
  : Submitted

  Graphs: algorithms for MCES, MICS, and maximal cliques
  : Theoretical Computer Science 2005 + 2008

  Mass spectrometry: stoichiometry determination
  : Submitted

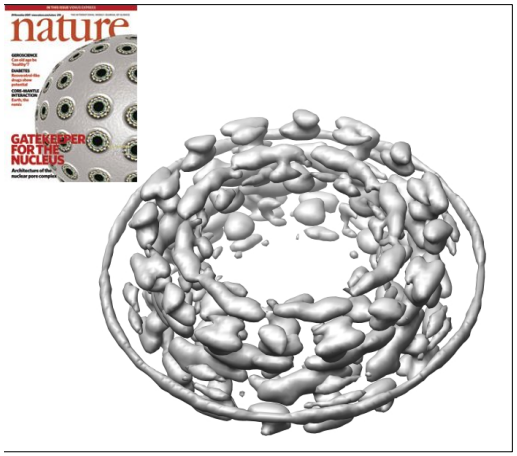
  Mass spectrometry: connectivity inference
  : Submitted

- Software available from `http://team.inria.fr/abs/software`

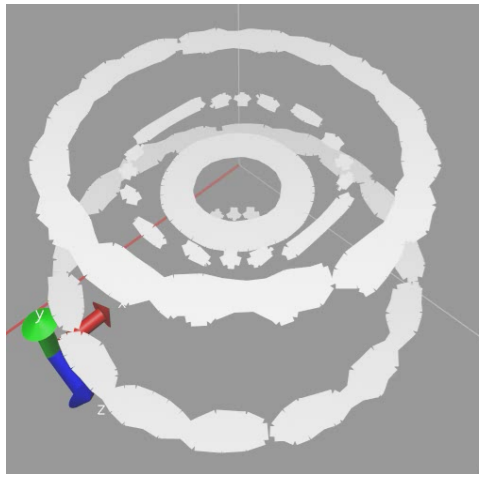
Modeling High Resolution Protein Complexes

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

- NPC: probability density maps



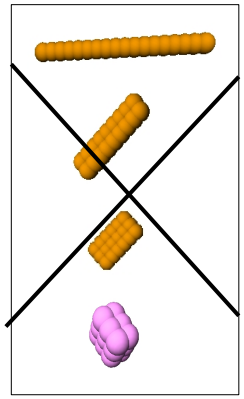
- Cryo-EM density maps

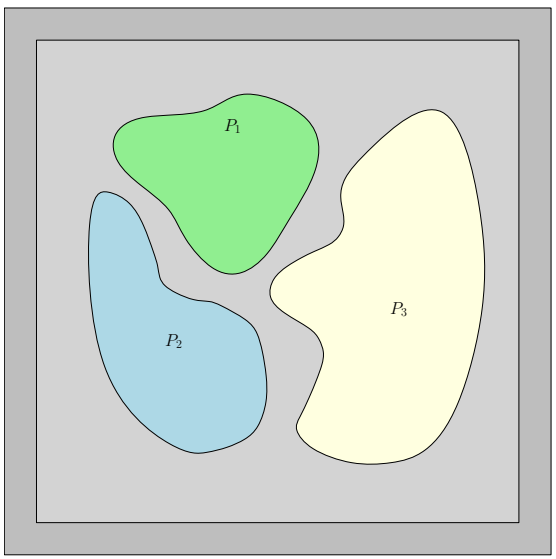


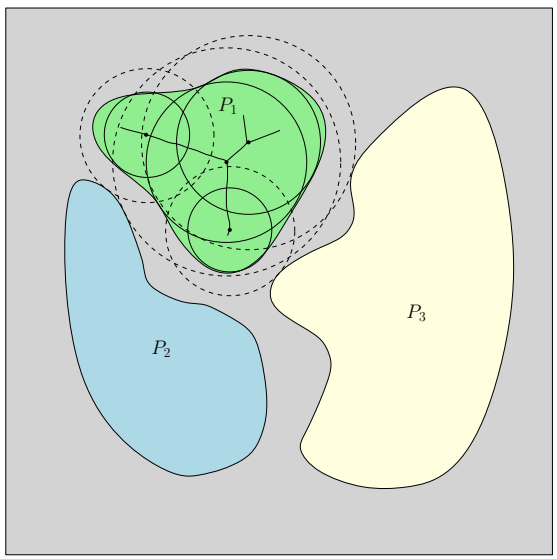
- Antelope canyon, AZ, USA

Building Complex Toleranced Models: Overview

- ▷ Step 1 : density map segmentation:
  Find a thick boundary and thick cracks such that
  the complement consists of k contractible regions
- ▷ Step 2: TOM construction for each protein
  Step 2a: cover the protein with inner balls
  Step 2b: cover the thick walls with outer balls





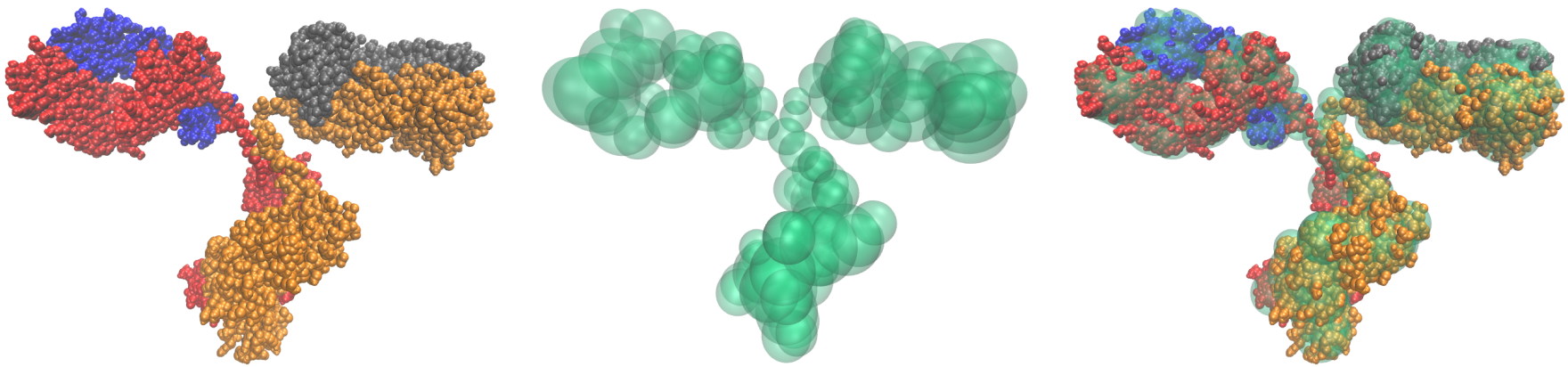


▷Ref: Chen, Freedman, Lampert; IEEE CVPR; 2011

▷Ref: Cazals, Dreyfus, Sachdeva, Shah; Preprint; 2012

Coarse Graining and Toleranced Model Building

- Coarse graining: the example a complete immunoglobulin
  - Atomic versus coarse grain model: 12533 atoms to 100 balls
  - Strategy: geometric version of max-k-cover, a NP-complete problem
- TOM building
  - Map segmentation
  - Geometric max-k-cover for the inner parts, expansion for the cracks



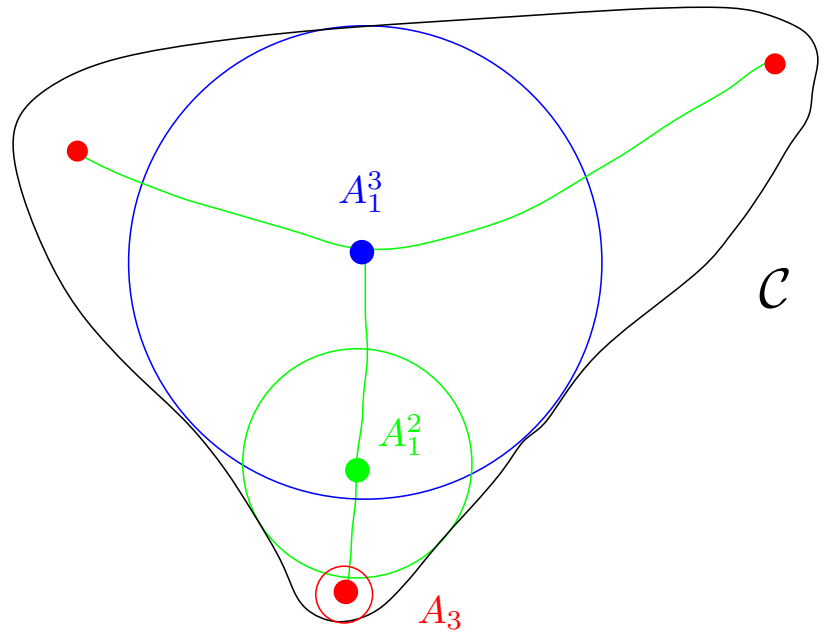
- Ref: Cazals, Dreyfus, Sachdeva, Shah; submitted

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:

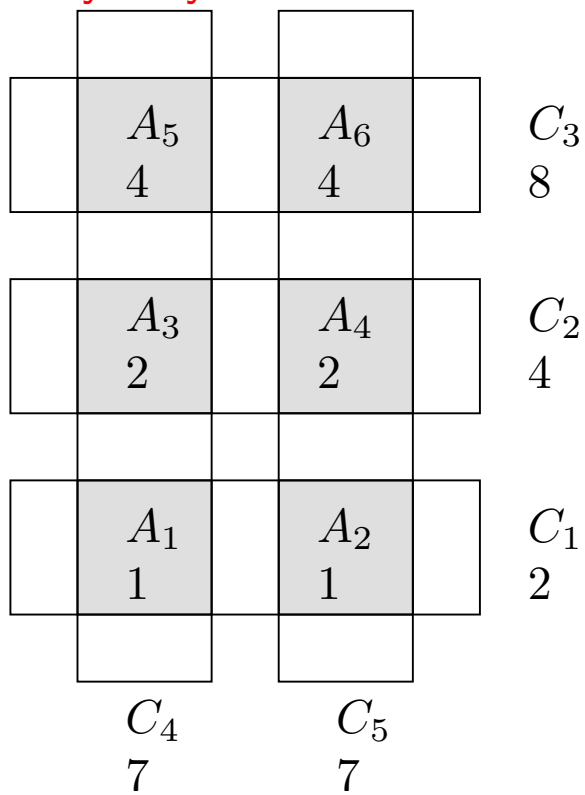
– problem is **NP**-complete

– 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:

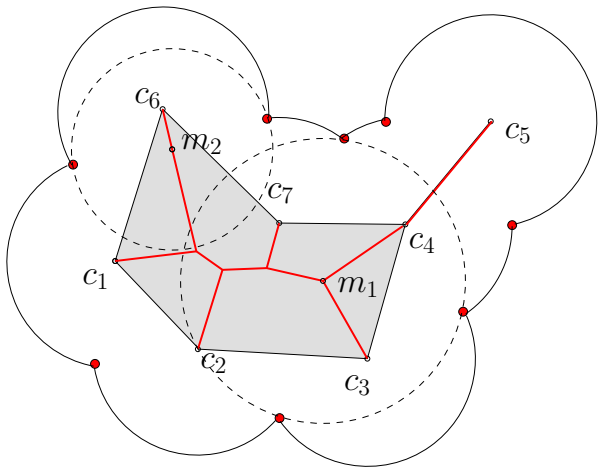


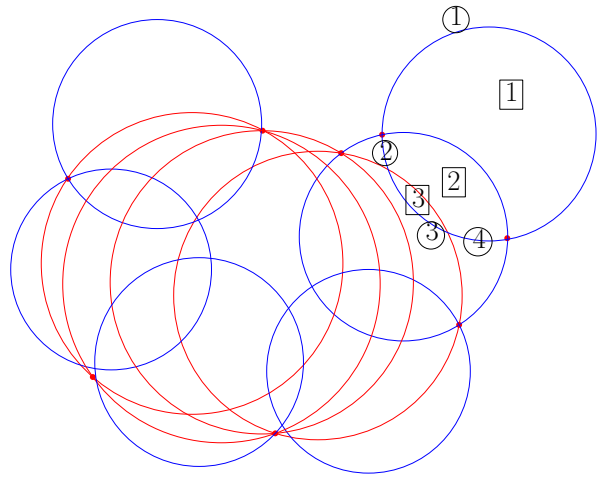
Greedy: $C_3 + C_2 = 12$

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

Our Results

▷ 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:

$$\mathcal{F}_{\mathcal{C}} = \bigcup_i B_i = \bigcup_{v \in \mathcal{V}} B_v^*. \tag{6}$$

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:

$$\begin{cases} 1-(1-1/k)^k > 1-1/e & \text{wrt to OPT weight (volume)} \\ 1-(1-1/n)^k & \text{wrt the total weight (volume)} \end{cases} \tag{7}$$

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

▷Ref: Cazals, Dreyfus, Sachdeva, Shah; Preprint; 2012

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

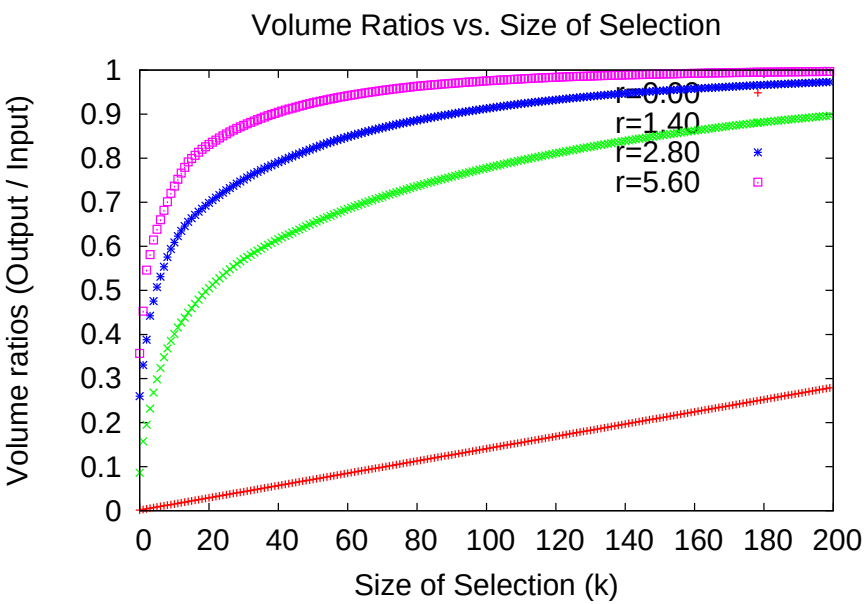
Greedy Assessment: Volume Covered

Incidence of the Topology

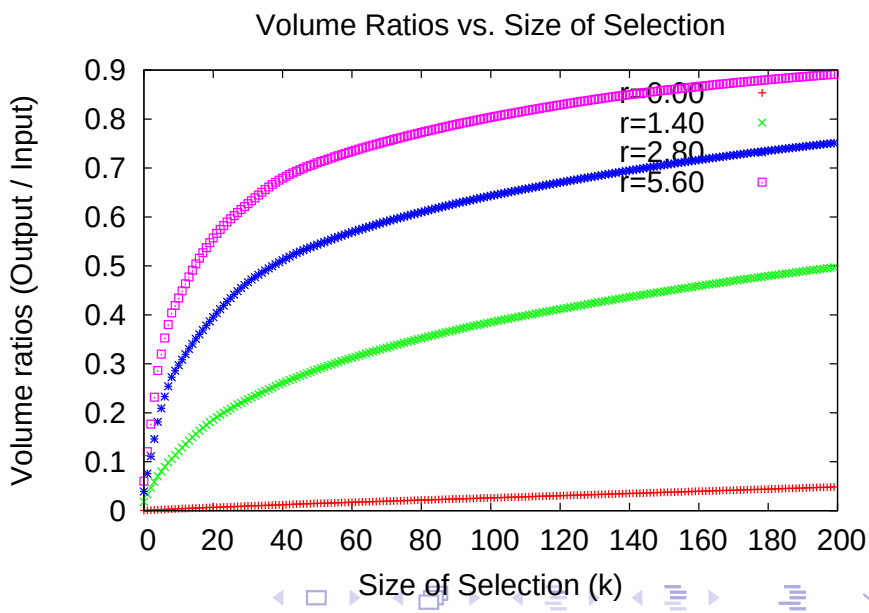
- Input domain versus domain of the selection: volume comparisons
  - $\mathcal{F}_{\mathcal{O}}^{r}$: input balls expanded by a quantity r
    - $\rightarrow r = 0$: input model
  - $\mathcal{F}_{\mathcal{S}}^{r}$: domain of the selection for the expanded model

  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: 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 $\partial\mathcal{F}_\mathcal{S}$:

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

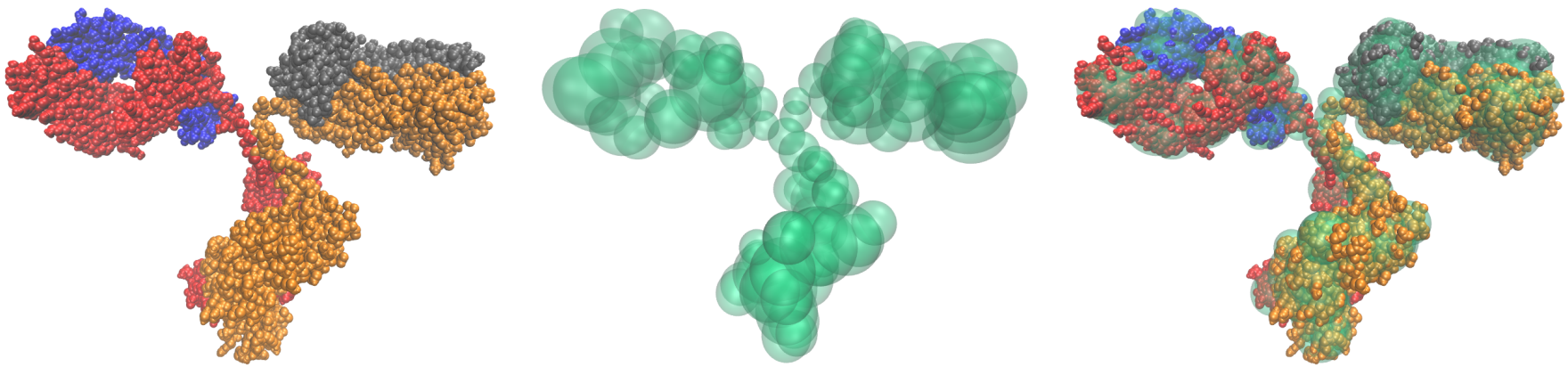
- Assessment on a set of 96 protein complexes (1008 -13214 atoms):

  volume ratio increases with the expansion radius r
  in particular with $r = 2.1$
    volume of the Solvent Accessible Model respected
    max distance of 5.58Å(about 3 atoms)

r	s/n	τ	τ_{SAS}	d_1	d_2	d_3	d_4
1.4	0.05	0.72 ± 0.04	0.72 ± 0.04	-10.39 ± 1.77	0.88 ± 0.53	-0.26 ± 0.00	4.57 ± 0.47
2.1	0.05	0.85 ± 0.03	0.99 ± 0.02	-4.44 ± 1.44	0.77 ± 0.00	-5.58 ± 1.01	2.60 ± 0.32
2.8	0.05	0.90 ± 0.02	1.19 ± 0.00	-2.79 ± 0.43	1.46 ± 0.01	-7.21 ± 1.02	1.80 ± 0.28

Coarse Graining and Toleranced Model Building

- Coarse graining: the example a complete immunoglobulin
  - Atomic versus coarse grain model: 12533 atoms to 100 balls
  - Strategy: geometric version of max-k-cover, a NP-complete problem
- TOM building
  - Map segmentation
  - Geometric max-k-cover for the inner parts, expansion for the cracks



- Ref: Cazals, Dreyfus, Sachdeva, Shah; submitted

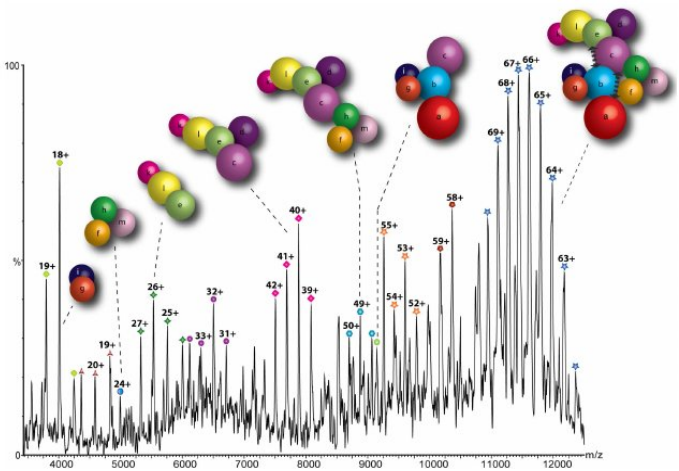
Part III

Studies in Mass Spectrometry

Stoichiometry Determination for Mass-spectrometry Data in Structural Proteomics

Frederic.Cazals@inria.fr
Joint work with D. Agarwal and N. Malod-Dognin

Algorithms-Biology-Structure, INRIA Sophia-Antipolis



Modeling High Resolution Protein Complexes

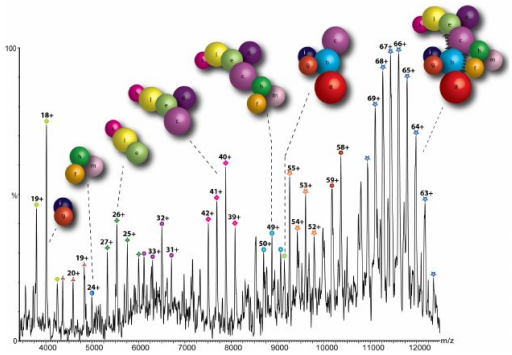
Stoichiometry Determination

Connectivity Inference

Mass Spectrometry and Structural Proteomics

▷ Electrospray: John Fenn

▷ Spectra and oligomers



▷ (Tandem) mass-spectrometry: principle

– Ionization of (complete | partially denaturated) protein complexes
  $\longrightarrow$ electrospray or MALDI
(selected) non-covalent interactions preserved

▷ PB 1: Stoichiometry determination: how many copies of each protein type?

– Post-translational modification of proteins
– Measurement errors, up to 2% — buffer molecules
– Incidence of isotopes

▷ PB 2: Connectivity inference: pairwise contacts?

▷Ref: Robinson et al; Chemical Reviews; 2007

▷Ref: Robinson et al; Trends in biochemical sciences; 2010

Stoichiometry Determination: the Exact and Interval Cases

▷ Notations:

- Weight vector for the p types: $\mathbf{W} = [w_1, w_2, \ldots, w_p]$
  integers . . . or floating point numbers
- Stoichiometry vector $\mathbf{S} = [s_1, s_2, \ldots, s_p]$
  non-negative solution: $s_i \geq 0$
  positive solution: $s_i > 0$ (relevant in biology)

▷ Exact vs interval problems:

report all vectors $\mathbf{S}$ such that

$$\sum_{i=1,\ldots,p} s_i \; w_i = M \qquad (8)$$

$$\mid \sum_{i=1,\ldots,p} s_i w_i - M \mid \leq \varepsilon. \qquad (9)$$

▷ Comments:

For integers masses: a.k.a. the *Money changing problem*
If isotopes taken into account: integral problem can be defined

Stoichiometry Determination: Context

▷ Denumerant on an integer M: for masses such that $\gcd(w_1, \ldots, w_p) = 1$, the # of non-negative solutions grows exponentially i.e.:

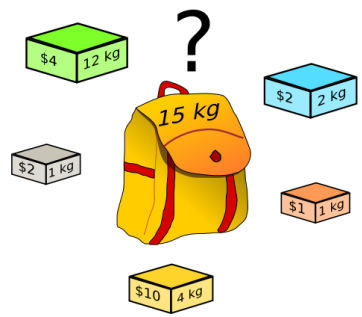
$$D(M) \sim \frac{M^{p-1}}{(p-1)! w_1 \ldots w_p}. \tag{10}$$

▷ Frobenius number $g_0(\mathbf{W})$:

- largest M which does not admit a non-negative representation
- computing $g_0(\mathbf{W})$: a NP-hard problem

▷ Unbounded knapsack (UKP): optimizing the value/utility of a knapsack of volume M

$$\begin{cases} \text{Maximize } \sum_{i=1}^{p} s_i \ v_i \\ \text{under the constraint } \sum_{i=1}^{p} s_i \ w_i \leq M. \end{cases}$$



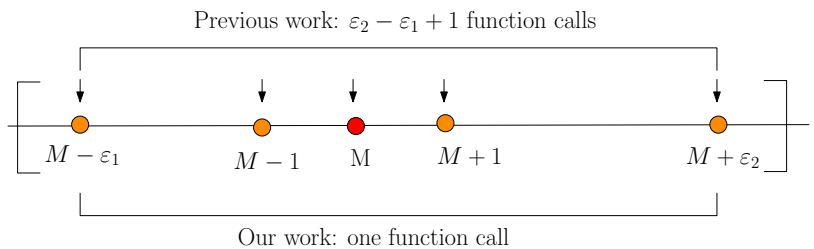
▷ Complexity-wise for UKP: decision problem (can a value $\geq V$ be achieved?) is NP-complete; the optimization problem (best value?) is NP-hard

▷ Subset-sum (SSP): UKP when the weight matches the value i.e. $w_i = v_i$.

SSP is NP-complete.

Contributions: Overview

- ▷ Context:
  - – Dynamic programming (DP): combining the solutions of sub-problems to solve (optimally) a given problem
  - – Time complexity: number of operations carried out
  - – Space complexity: number of memory cells required (#bytes)
  - – Output-sensitivity algo.: the time complexity depends on the # of solutions
- ▷ State-of-the-art for exact SD using DP:
  - – Exact case: pre-proc. time $O(pw_1)$; output sensitive $O(pw_1 D(M))$
  - – Interval case: iterating on all target masses



- ▷ Our contributions: two algorithms solving the interval case

1. Algorithm `DP++`

   Pre-processing: $O(p(M + \varepsilon))$
   Space complexity: $O(p(M + \varepsilon))$
   Output-sensitive

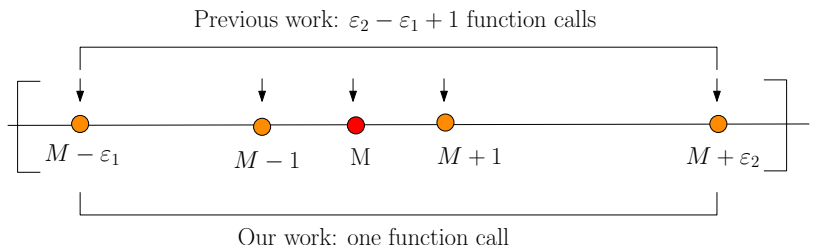
2. Algorithm `DIOPHANTINE`

   No-preprocessing
   Constant memory space $O(p)$
   Output-sensitive like

Noise level less than 1%: speed-up of 3-4 orders of magnitude!

Interval Stoichiometry Determination: Contributions

- Context: interval SD is an enumeration problem
  - Output-sensitivity solution?
  - Memory requirements as a function of p and $\mathbf{W}$?
- State-of-the-art for exact SD using dynamic programming:
  - Exact case: pre-proc. time $O(pw_1)$; output sensitive $O(pw_1 D(M))$
  - Interval case: iterating on all target masses



- Our contributions: two algorithms solving the interval case

1. Algorithm `DP++`

   Pre-processing: $O(p(M + \varepsilon))$
   Space complexity: $O(p(M + \varepsilon))$
   Output-sensitive

2. Algorithm `DIOPHANTINE`

   No-preprocessing
   Constant memory space $O(p)$
   Output-sensitive like

Noise level less than 1%: speed-up of 3-4 orders of magnitude!

Biological and Synthetic Datasets Studied

- Biological systems, with masses in the range 300 kDa to 50 MDa:

  Yeast 19S Proteasome lid
  COP9 Signalosome
  Eukaryotic Translation factor EIF3
  Yeast Exosome
  Rotary ATPases
  Nuclear Pore Complex (NPC): 1 spoke, 1 ring

- Synthetic systems:

  - Prime number instances
    p primes in the range $7,000 \ldots 70,000$

  - Random biological instances
    p protein types amidst the 6,700 yeast proteins

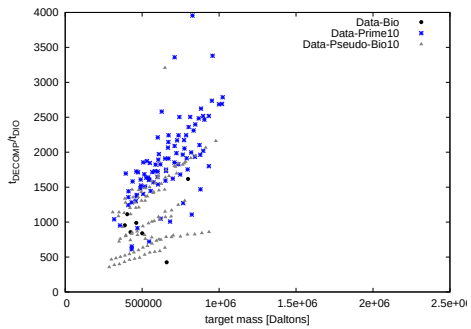
- Target masses: expressed in units of the Frobenius number

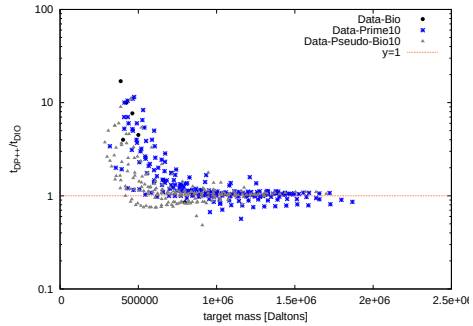
Observation: Enumeration Matters even at Null Noise Level

Complex	#sol, 0% noise	#sol, 0.1% noise	#sol, 1% noise	M (Da)	g_0^+	$\#sol(1\%\,noise)/2\varepsilon$	M/g_0^+
COP9	0	1	1	321,274 ± 35	961,855	3.12×10^{-4}	0.33
Y-19S-lid	0	0	1	376,151 ± 369	921,712	1.87×10^{-4}	0.41
EhATPase-sub-5	0	4	39	387,356 ± 230	682,901	5.03×10^{-3}	0.57
Y-exosome	0	13	149	402,708 ± 68	649,185	1.85×10^{-2}	0.62
EhATPase-sub-4	0	23	203	424,441 ± 148	682,901	2.38×10^{-2}	0.62
EhATPase-sub-3	0	77	751	461,674 ± 324	682,901	8.10×10^{-2}	0.68
EhATPase-sub-2	0	234	2,333	500,178 ± 294	682,901	2.32×10^{-1}	0.73
TtATPase	21	24,487	246,242	659,202 ± 131	607,304	18.7	1.08
EIF3	0	0	1	797,999 ± 180	1,257,629	8.05×10^{-5}	0.63
NPC-Y-ring	788	6,900,664	69,042,257	4,603,280	2,282,543	750	2.02
NPC-1-spoke	$[1.72 \times 10^{16}$, $1.73 \times 10^{16}]$	$[1.72 \times 10^{16}$, $4.77 \times 10^{16}]$	$[2.71 \times 10^{17}$, $3.44 \times 10^{17}]$	5,276,467	3,169,210	$\geq 2.57 \times 10^{12}$	1.66

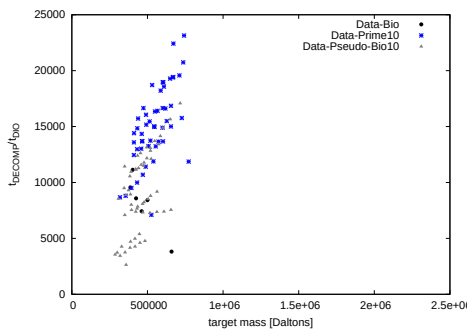
Timing Results as a Function the Noise Level

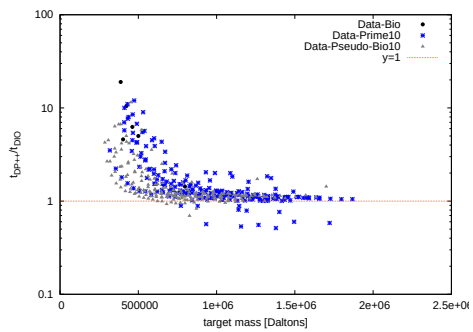
noise level 0% : DP++ and DIOPHANTINE: slow-down of ≤ 10
noise level 0.1% : DP++ and DIOPHANTINE: speed-up of factor $O(1000)$
noise level 1% : DP++ and DIOPHANTINE: speed-up of factor $O(10,000)$





Noise Level: 0.1%





Noise Level: 1%

DECOMP vs DIOPHANTINE DECOMP vs DP++

▷ Key idea: our algorithms avoid redundant calculations when solving interval SD by repeatedly calling the exact OS SD algorithm

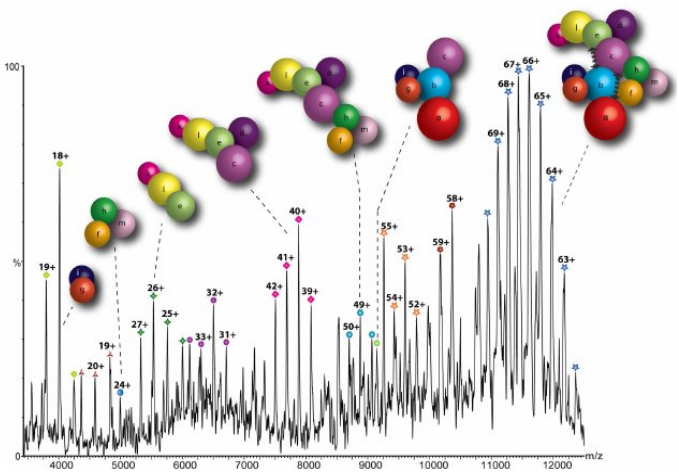
More Results

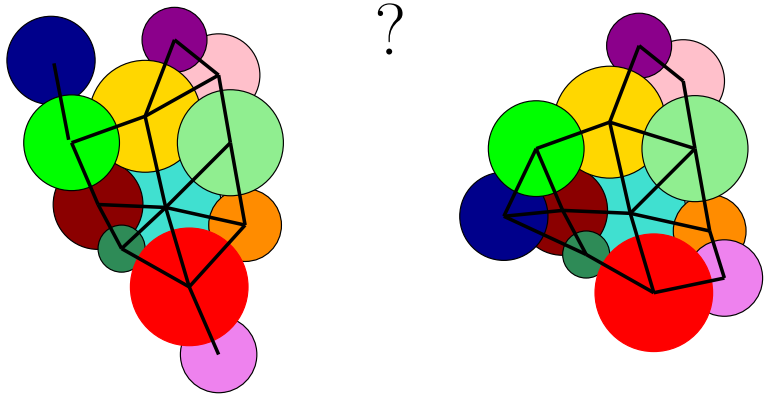
- Estimating the number of solutions
  - Upper and lower bounds based on combinatorial identities
  - Sharpness of estimates for practical purposes
- Exact case: convergence of the number of solutions to the denumerant
- Noisy case: output sensitivity of `DIOPHANTINE`

Outlook

- Reference algorithms to solve the interval stoichiometry determination problem
  - Paper: `http://team.inria.fr/abs/publications/`
  - Software: `http://team.inria.fr/abs/software/`

- Input for the connectivity inference problem





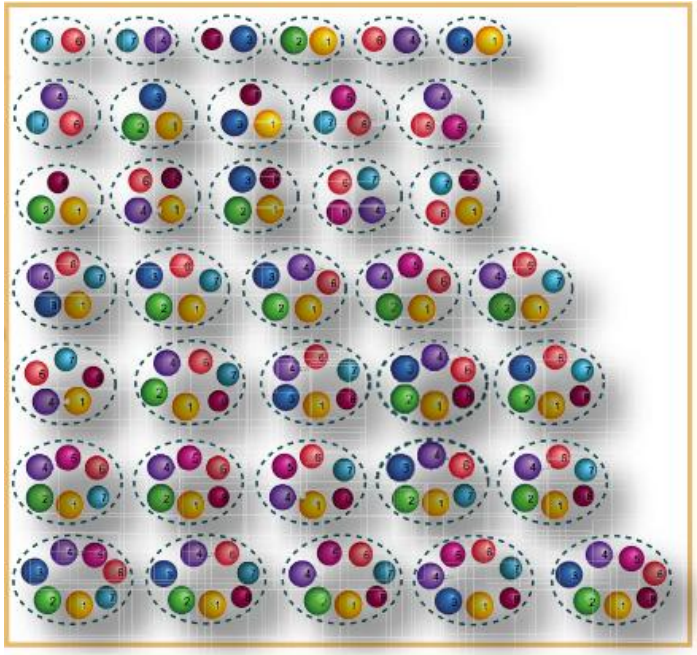
- Ref: Taverner, Robinson et al; Accounts of chemical research; 2008

Modeling High Resolution Protein Complexes

Stoichiometry Determination

Connectivity Inference

Connectivity Inference: Problem Specification



▷ A two steps process:

– using various chemical conditions
generate various oligomers

– using MS or MS/MS:
stoichiometry determination
oligomer composition

▷ Formal specification:

– A set V of proteins

– A set of k subsets $V_i \subset V, i = 1, \ldots, k$

– Find a graph $G = (P, E)$ such that:
$| E |$ minimal and
the sub-graph of G
induced by each V_i is connected

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

Connectivity Inference: Overview of Contributions

- Hardness: problem is NP-complete and APX-hard (bad news!)
  - reduction to set cover

- Algorithm, exact:
  - based on Mixed Integer Linear Programming (MILP)
    several options to ensure connectivity of induced subgraphs
  - allows generating all solutions matching $OPT + k$

- Algorithm, approximate: greedy
  - next edge: maximizes the change in number of connected components
    only requires Union-Find data structures
  - approximation factor: $2(\log n + \kappa)$, with κ max. # oligomers of a vertex

- Experiments on the biggest systems know to date:
  - our solutions more parsimonious than those of contenders
  - edges reported: (almost) perfect agreement with known contacts

▷Ref: `Inria ABS + Inria COATI, submitted`

Assessment: Protocol

Comparing predicted versus experimentally observed contacts

- Reference contacts set C, defined from:
  - C_{Xtal}: from a high resolution crystal structure
  - C_{Dim}: from dimers (TAP, etc)
  - C_{XL}: from cross-linking
  - $C_{Exp} = C_{Xtal} \cup C_{Dim} \cup C_{XL}$

- Assessing an ensemble of solutions $\mathcal{S}$, returned by say algo. `MILP`

– *precision* of solution $S \in \mathcal{S}$ wrt C: $P_{MILP;C}(S) = \mid S \cap C \mid$
  $\rightarrow$ precision is maximum if $S \subset C$ i.e. no false positive

– *precision* $P_{MILP;C}(\mathcal{S})$ of an ensemble of solutions $\mathcal{S}$, assessed by:
  (min, median,max) of the precisions of the solutions $S \in \mathcal{S}$

– *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

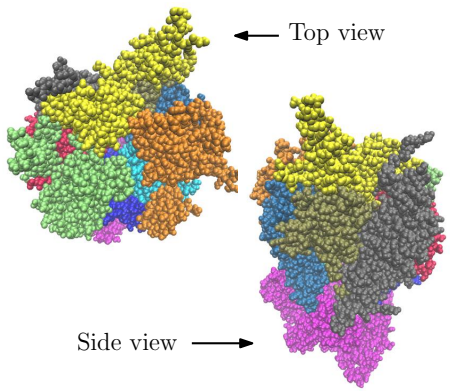
– *score of a solution* $S \in \mathcal{S}$: the sum of the scores of its contacts

– *consensus solutions* $\mathcal{S}^{cons_\cdot}$: achieve the maximum score in $\mathcal{S}$

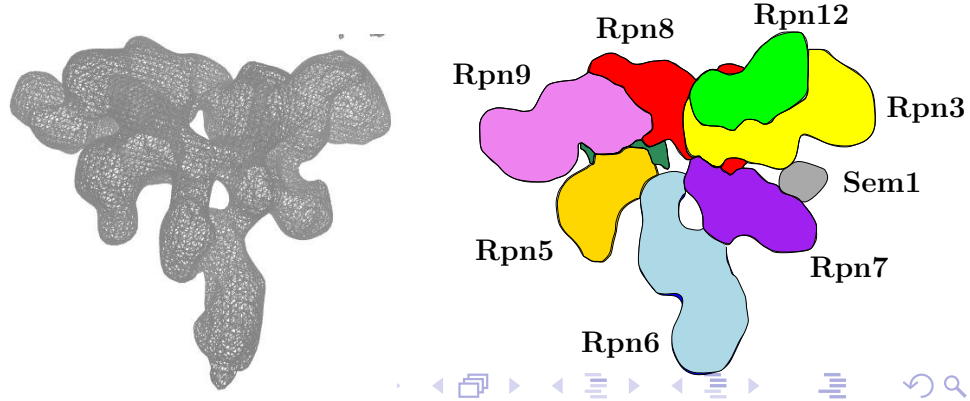
Assemblies Under Scrutiny, with exhaustive lists of contacts

- Yeast Exosome:
  - assembly involved in RNA processing and degradation
  - 10 different protein types with unit stoichiometry
  - input from MS: 21 complexes
- Yeast 19S Proteasome lid
  - proteasomes: elimination/degradation of damaged or misfolded proteins
  - proteasome lid: 9 distinct protein types each with unit stoichiometry
  - input from MS: 14 complexes
- Eukaryotic Translation factor eIF3
  - eIF are proteins involved in the initiation phase of the eukaryotic translation
  - eIF3: complex of 13 different protein types each with unit stoichiometry
  - input from MS: 27 complexes

- Yeast exosome: crystal structure



- Proteasome lid: cryo EM map

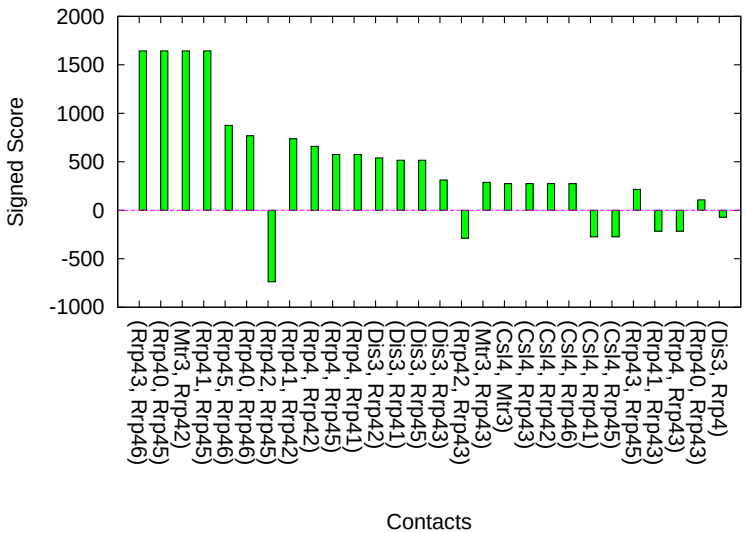


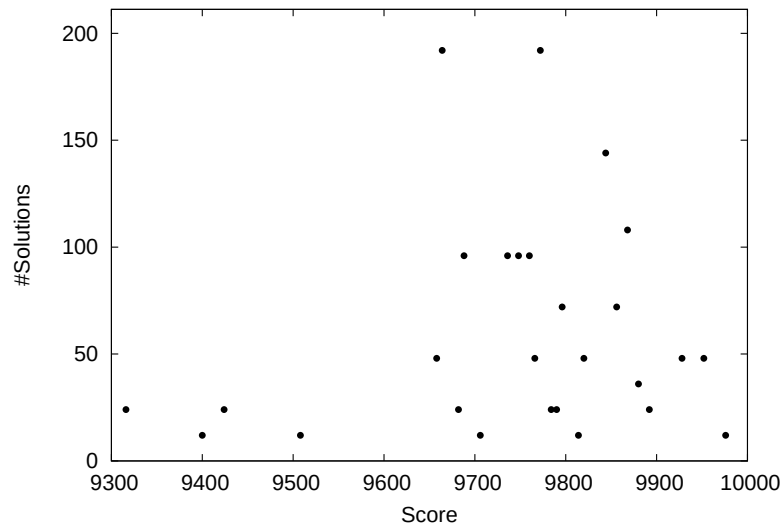
Results — with a focus on the exosome

▷ Parcimony of solutions

Complex	#types	Ref. set C	$\lvert C\rvert$	$\lvert S_{NI}\rvert$	$P_{MILP;C}(S_{NI})$	$\lvert S_G\rvert$	$P_{MILP;C}(S_G)$	$\lvert S_{ILP}\rvert$	$\lvert \mathcal{S}_{MILP}\rvert$	$P_{MILP;C}(\mathcal{S}_{MILP})$	$\lvert \mathcal{S}^{cons.}_{MILP}\rvert$	$P_{MILP;C}(\mathcal{S}^{cons.}_{MILP})$
Exosome	10	C_{Xtal}	26	12	12(100%)	10	10 (100%)	10	1644	(7, 9, 10)	12	(8, 9, 10)
19*S Lid*	9	C_{Exp}	16	9 (*NC*)*	7(77.8%)	10	8 (80%)	10	324	(6, 7, 10)	18	(8, 8, 10)
eIF3	13	C_{Exp}	19	17**	14 (82.3%)	14	9 (64.2%)	14	2160	(8, 9, 11)	432	(8, 9, 10)

▷ Scores and false positives





(Left) Contact scores for solutions in $\mathcal{S}_{MILP}$, w.r.t C_{Xtal}
(Right)Distribution of scores for solutions in $\mathcal{S}_{MILP}$

Part IV

(Selected Algorithmic Details)

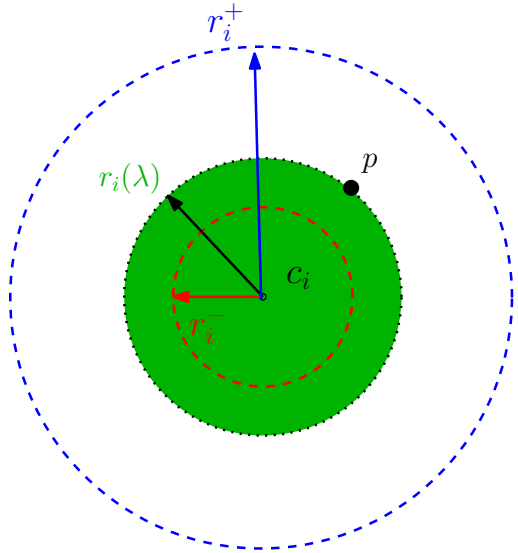
Modeling High Resolution Protein Complexes

Compoundly Weighted Voronoi Diagrams and their λ-Complex

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 $\overline{S_i}[\lambda]$
    $\Leftrightarrow r_i(\lambda) = \| c_i p \|$
    $\Leftrightarrow \lambda = \frac{\|c_i p\| - r_i^-}{\delta_i}$



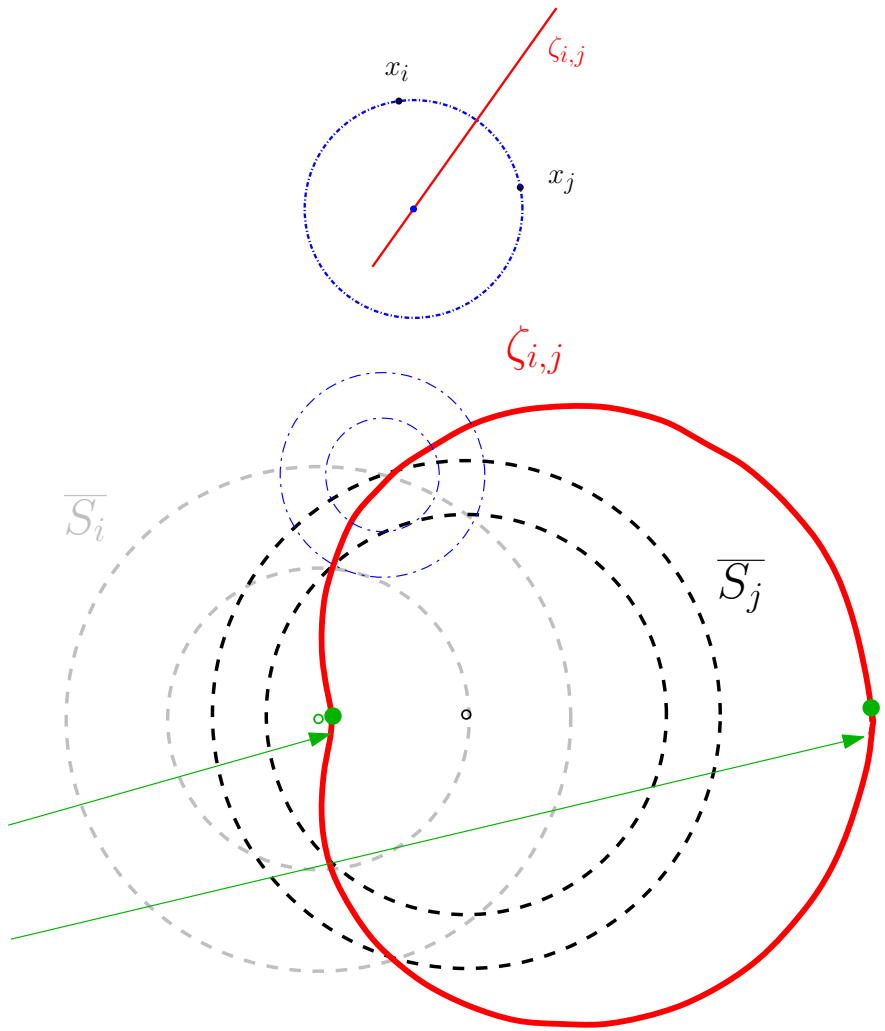
- **From Toleranced Ball to Compoundly Weighted Point:**
  - $S_i(c_i; \mu_i = \frac{1}{\delta_i}, \alpha_i = \frac{r_i^-}{\delta_i})$
  - $\lambda(S_i, p) = \frac{1}{\delta_i} \| c_i p \| - \frac{r_i^-}{\delta_i}$

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

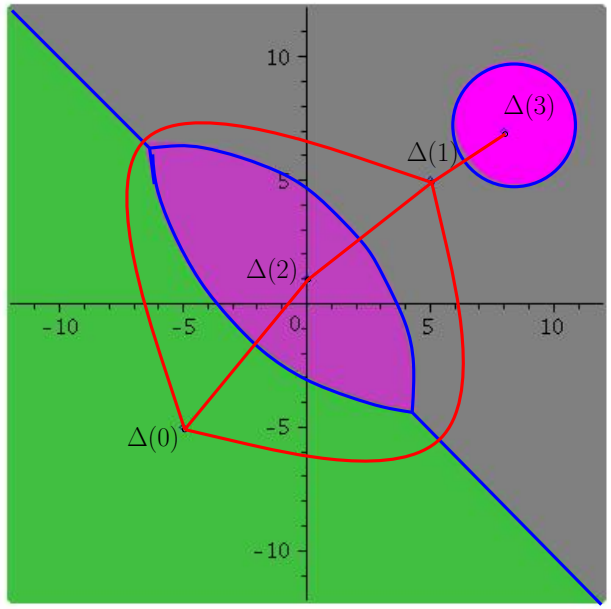
Bisectors

- 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 $\overline{S_{i_0}}[\overline{\rho}], \ldots, \overline{S_{i_k}}[\overline{\rho}]$



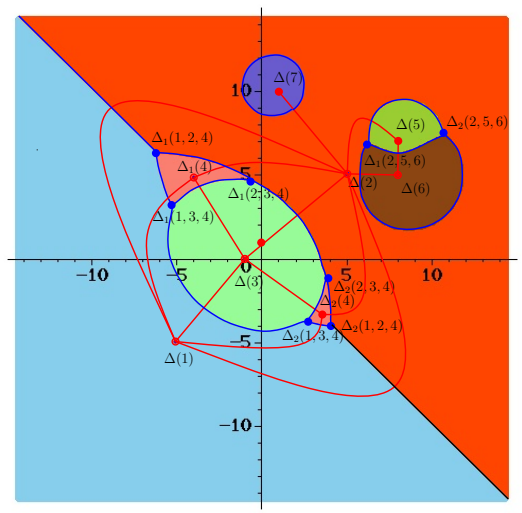
Voronoi Diagram and its Dual Complex: Topological Complications



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

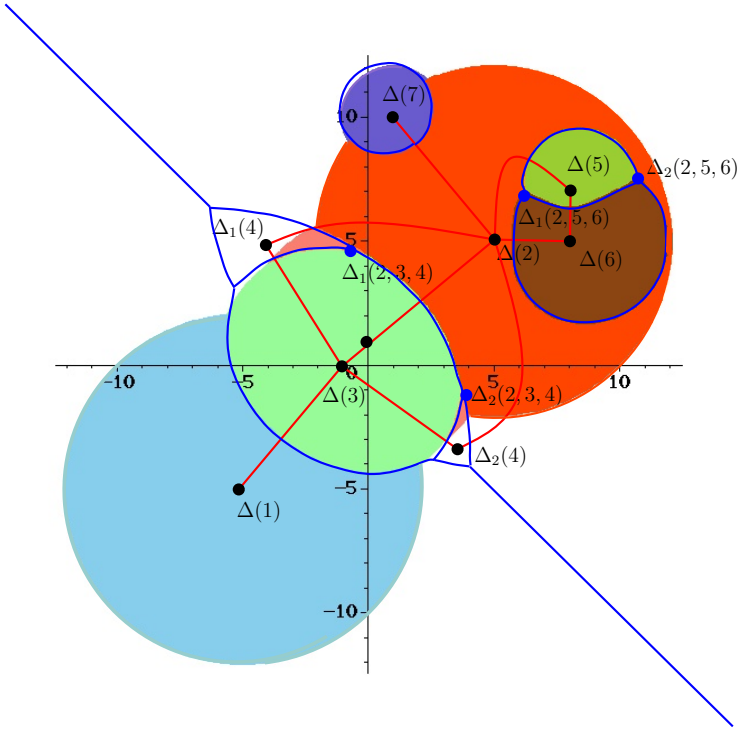
▷ Voronoi region – in all generality:
- Neither connected : collection of **faces**
- Nor simply connected

▷ Dual complex:
- Not a **triangulation**
  $\longrightarrow$ 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		$(\underline{\mu}_{\Delta(T)}, \gamma_{\Delta(T)}]$	$(\gamma_{\Delta(T)}, +\infty]$

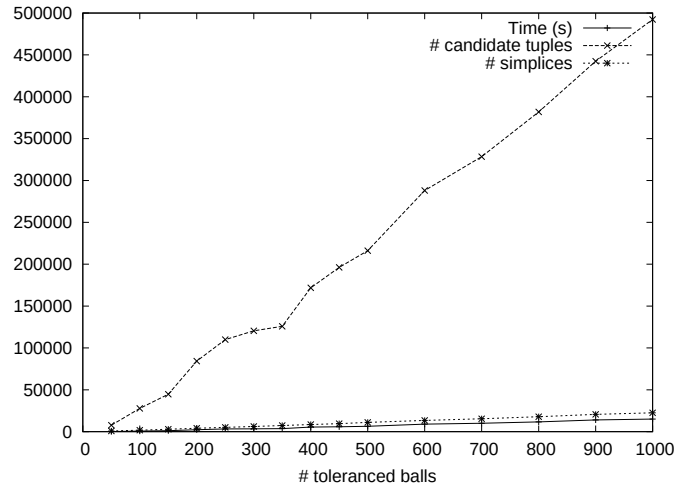
Algorithms

- **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



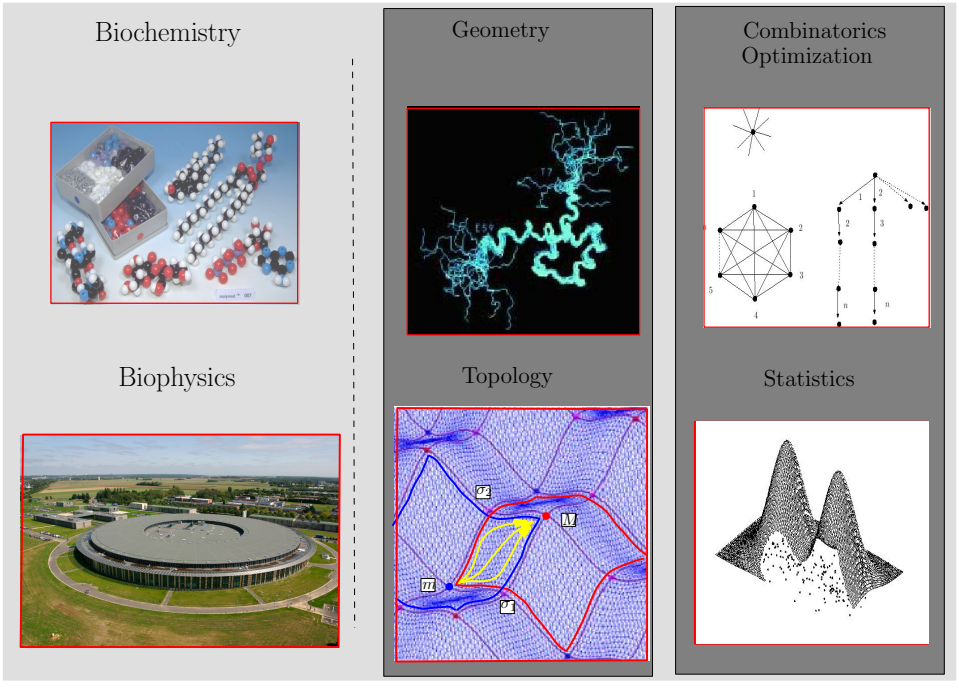
(**Random** Toleranced balls)

Part V

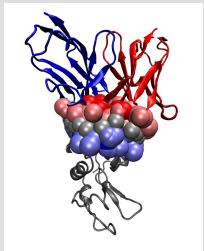
Conclusion

Our Vision

▷ Experiments and Modeling



Structure-to-Function



Docking (and Folding)

- Improved descriptions
- Improved predictions
  - atomic models (small complexes)
  - coarse models (PPI networks)

▷ Applied challenges

– Modeling protein complexes
  atomic models, large assemblies
– Modeling the flexibility of proteins
(– Systems biology)

▷ Mathematical - algorithmic foundations

– Geometric - topological modeling
  focus on stability analysis
– Graph theory, matching algorithms
– Optimization
– Machine learning
  dimensionality reduction
  statistical tests