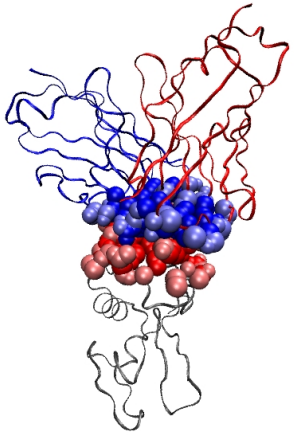
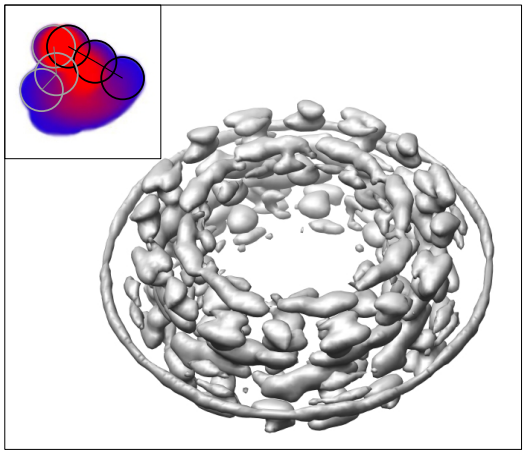
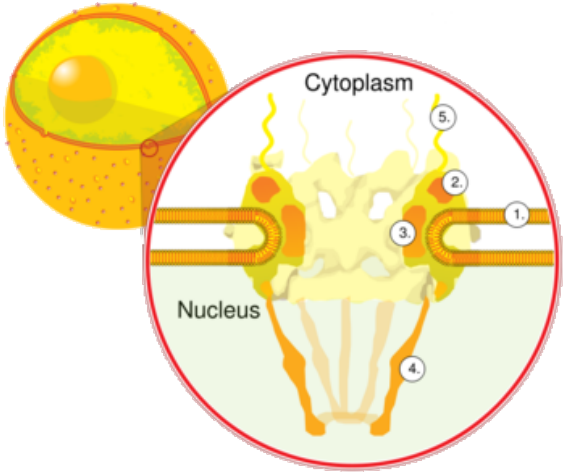
Modeling Protein Complexes and Assemblies with Voronoï Diagrams

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







Outline

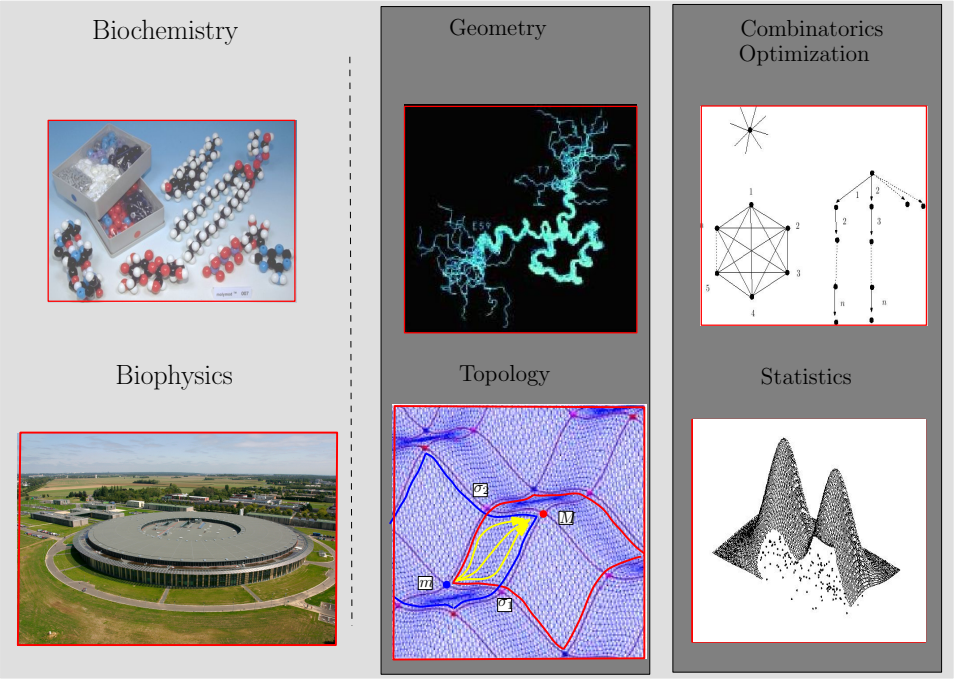
- Modeling high resolution protein complexes
  - Protein - protein interface?
  - Mining the stability - specificity of interactions
    predicting binding affinities
  - Understanding solvation properties
  - Template based docking

- Modeling large protein assemblies
  - Reconstruction by data integration
  - Handling uncertainties on protein shapes and positions
  - Assessing the reconstruction of fuzzy models

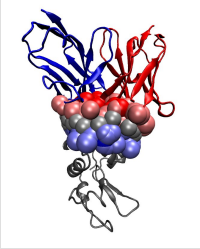
- Algorithms
  - Notions on cell complexes
  - Delaunay - Voronoi diagrams, α-shapes
  - Elementary notions in statistics
  - Comparing trees - the Tree Edit Distance

Our Vision

▷ Experiments and Modeling



Structure-to-Function



Docking (and Folding)

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

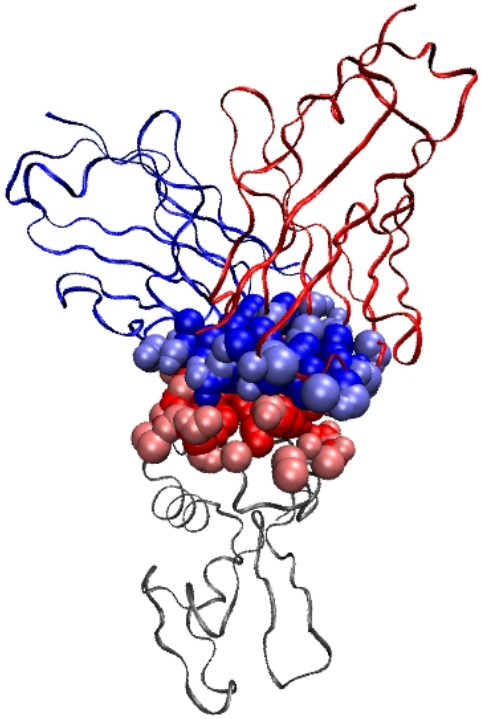
▷ Questions

- Modeling protein complexes
- Modeling the flexibility of proteins
- Bridging the gap to systems biology

▷ Partial answers from

- Geometric - topological modeling stability analysis
- Graph theory matching algorithms
- Statistical testing
- Dimensionality reduction investigating correlations

Part I: Modeling High Resolution Protein Complexes



Protein Interfaces: Key Questions

Geometric Intermezzo: Voronoi Diagrams and Relatives

Describing Protein Interfaces

Mining Biophysical Properties at Protein Interfaces

From Protein Interfaces to Protein Binding Patches

Modeling High Resolution Protein Complexes

Protein Interfaces: Key Questions

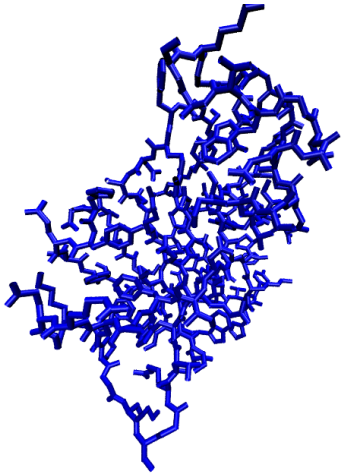
Geometric Intermezzo: Voronoi Diagrams and Relatives

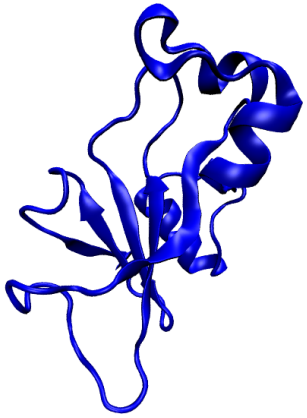
Describing Protein Interfaces

Mining Biophysical Properties at Protein Interfaces

From Protein Interfaces to Protein Binding Patches

Geometry - Topology versus Biophysics: A Matter of Correlations





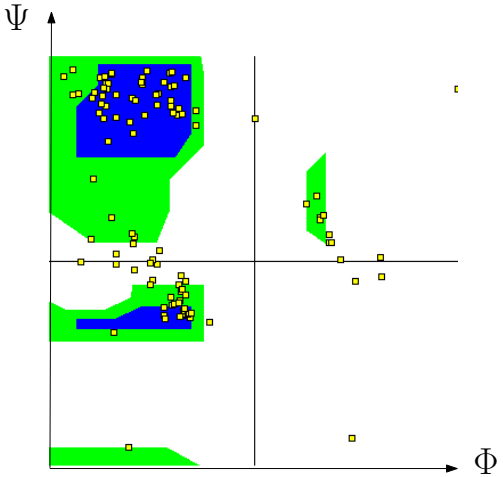


Fig. Hydrogen bonding in antiparallel b sheet

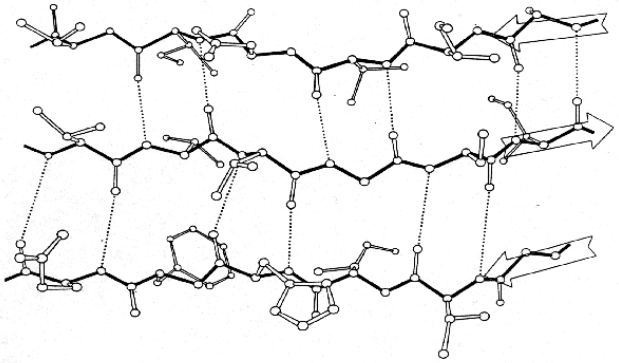
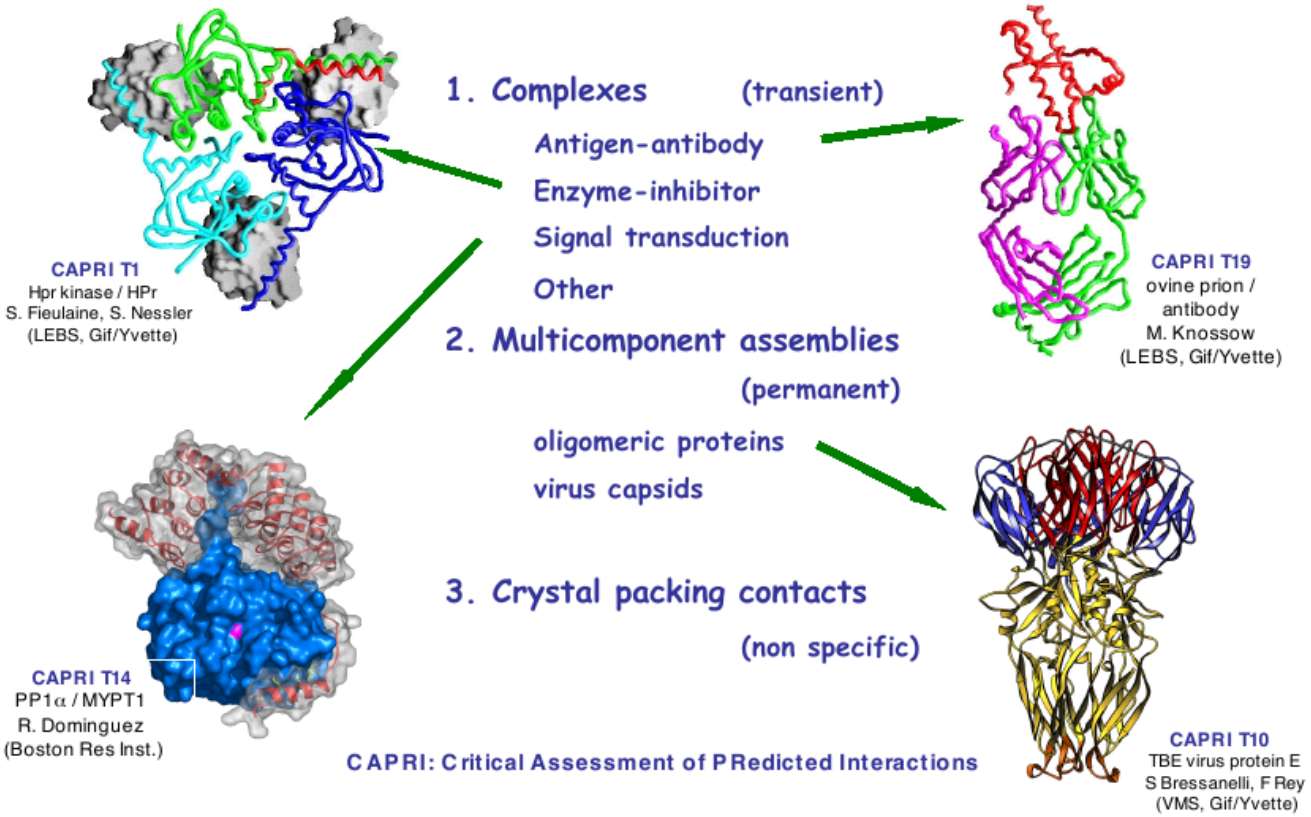


Figure 18. Drawing of a small piece of antiparallel β sheet (from SOD), illustrating the alternately narrow and wide pairing of H-bonds and the side-chain alternation above and below the plane of the sheet.

- ≪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]

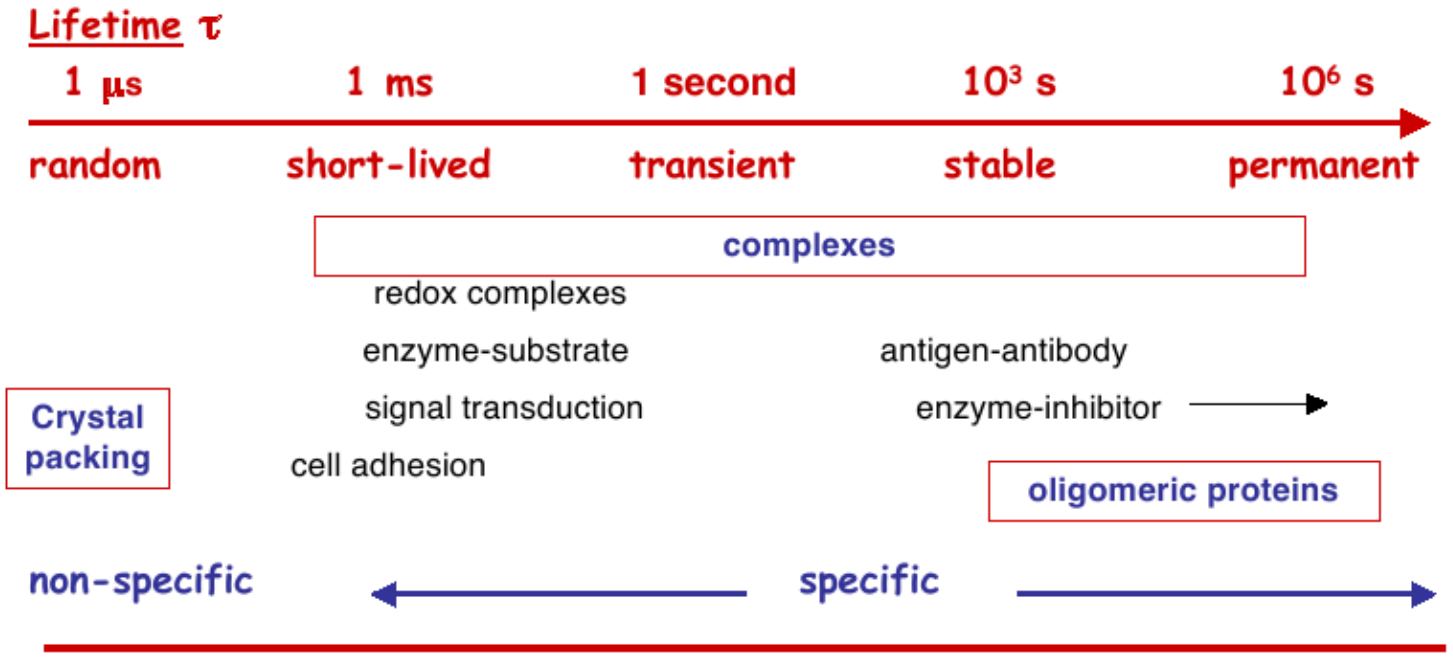
- ▷ Molecular mass: from $O(100\ kDa)$ up to 120 MDa (mammalian NPC)
- ▷ Structures vs sequences: 100,000 (PDB) versus 17,000,000 (NCBI RefSeq)

▷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.

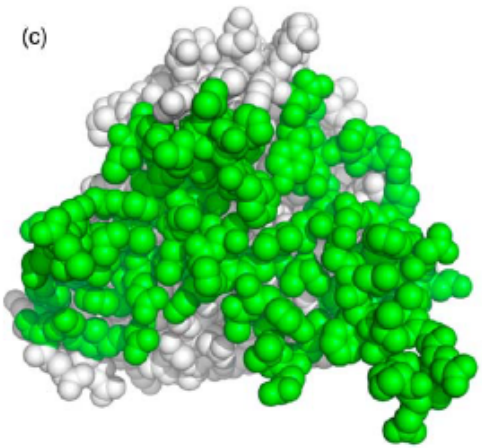
[J. Janin]

- 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

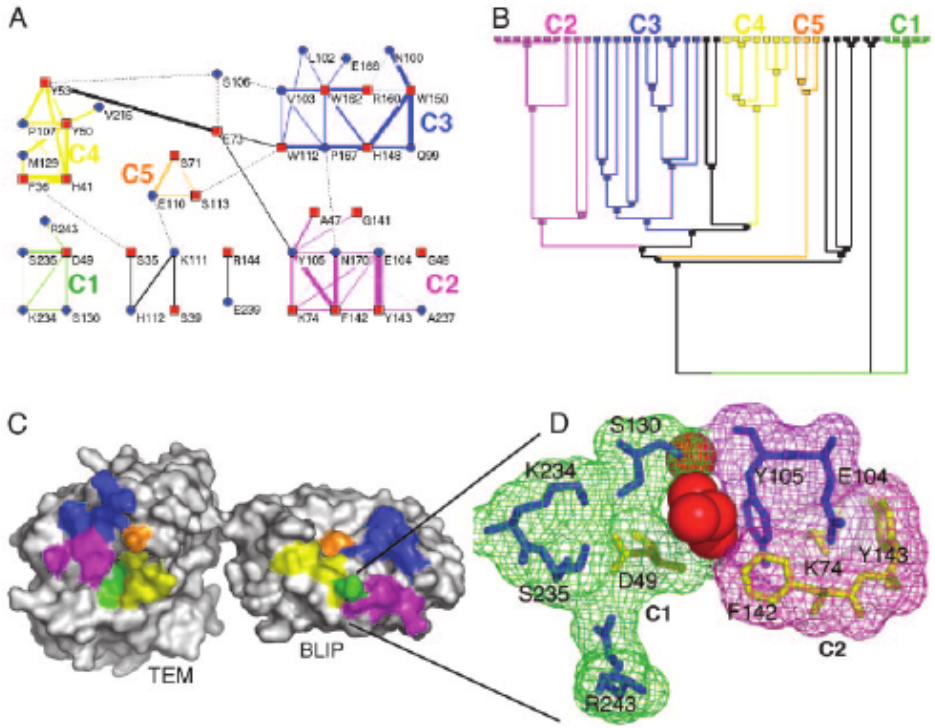


Fig. 1. Cluster analysis of the TEM1–BLIP interface. The interactions between residues located within the interface were extracted by using the CSU package (22) [for parameters, see Table 3; clustered with the SPC 3.3.13 tool (23)]. The interface was divided into five clusters of interactions, shown 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.

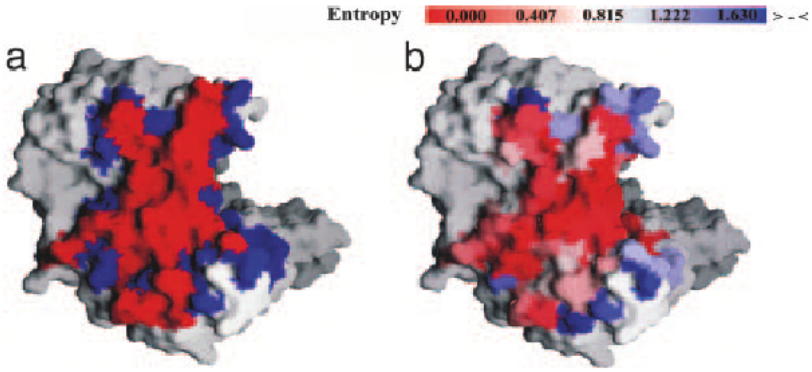
▷Schreiber et al, PNAS, 2005

- 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



Protocol

Dissect interface core vs rim:

core: fully buried; rim: partly exposed

Conclusions

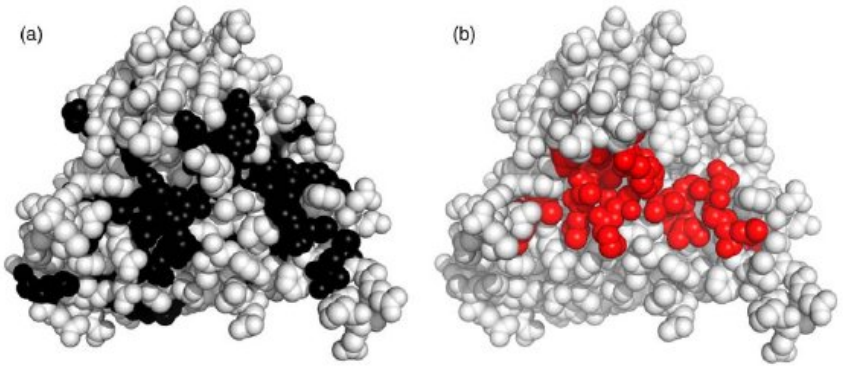
Core residues more conserved

Directed mutagenesis

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

▷ Conservation vs dryness

▷Ref: Lichtarge et al; JMB; 2007



Protocol

Run MD simulations

Measure Water residence times: dryness

Rationale for dryness :

interactions not perturbed by water fluxes

Conclusion

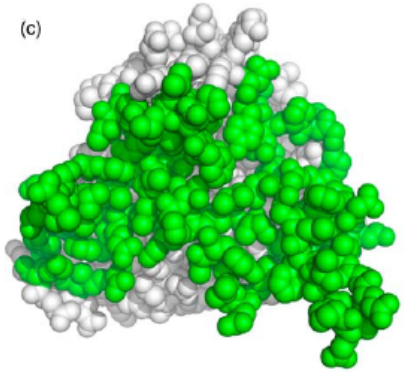
Conservation detects dry $\gg$ Conservation geom. footprint

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

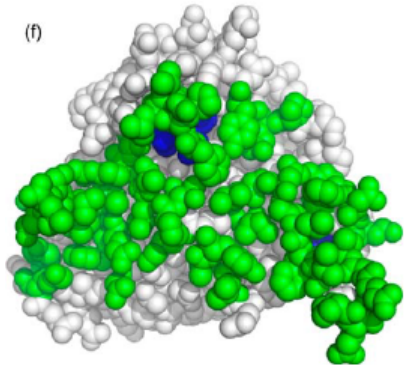
Predicting Important Residues: the Role of Dry Residues

▷ 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)

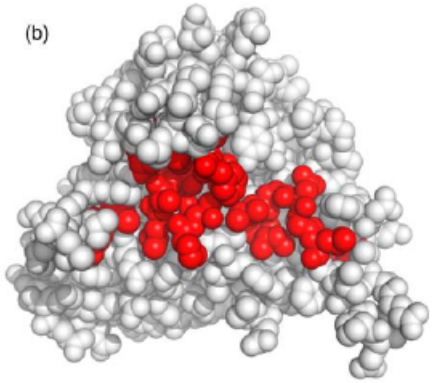
▷ 2DOR: interface residues within 7 Å



▷ 2DOR: interface residues: using ΔSAS



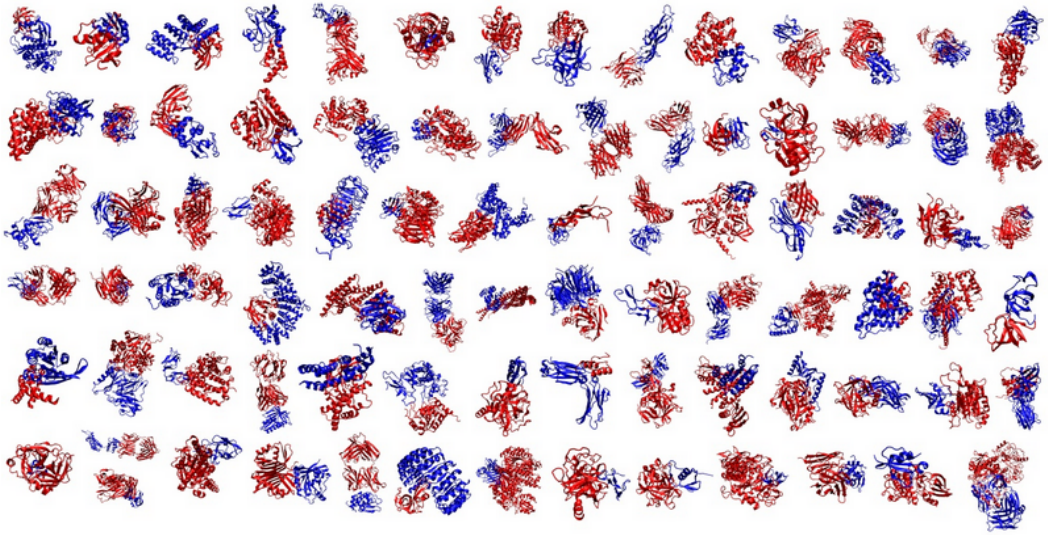
▷ 2DOR: dry residues



▷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 (!)

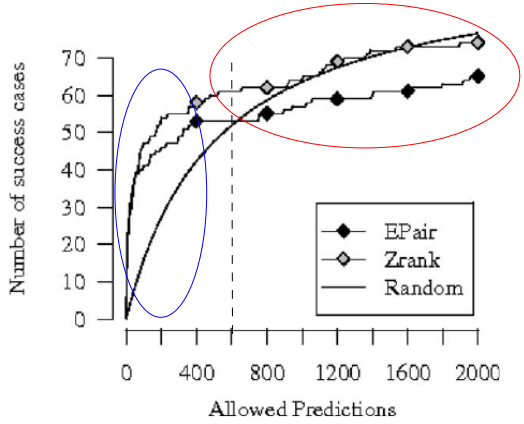


Figure 2
Success curves of ZRank and EPair compared with the random success curve for a small number of predictions (from 1 to 2000). See also Figure S1.

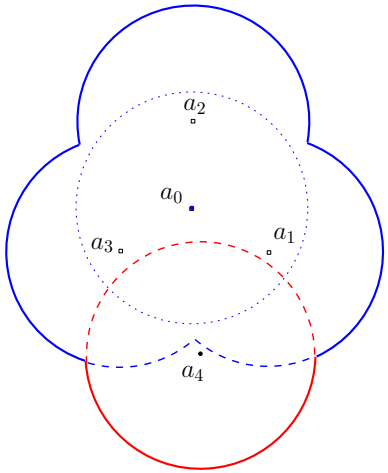
Table II
Statistical analysis of targets in the scoring section in CAPRI rounds 9–19

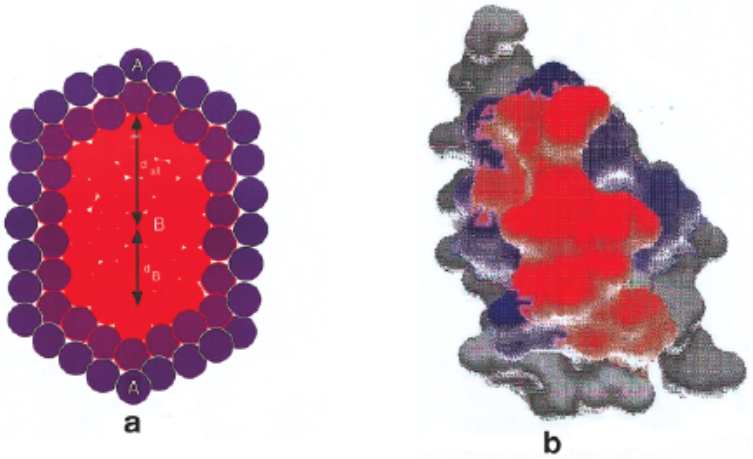
Target	Uploaded models	Accept or better models	*P* value (≥1 hit)	*P* value (≥2 hits)	Number of scorers	Successful scorers	Ratio of successful scorers	Optimal number of predictions
T25	700	36	0.41	0.09	6	6	1	2 (0.1)
T26	1171	60	0.41	0.09	8	4	0.5	2 (0.09)
T27	1093	123	0.7	0.31	12	11	0.92	1 (0.11)
T29	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

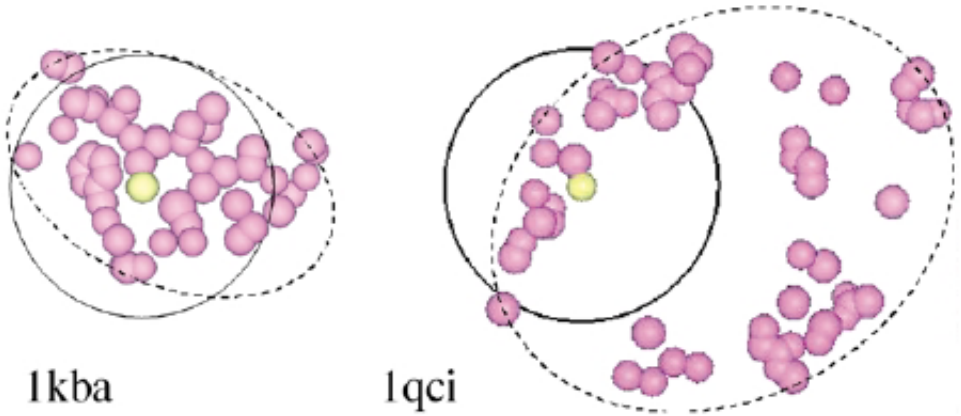
Classical Tools: Modeling Interfaces

- The core-rim model





- Interface shape - (atom centric) packing density



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

Modeling High Resolution Protein Complexes

Protein Interfaces: Key Questions

Geometric Intermezzo: Voronoi Diagrams and Relatives

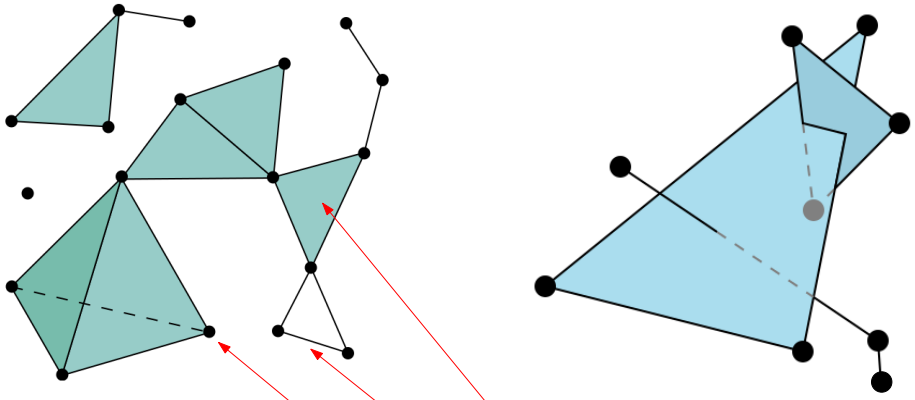
Describing Protein Interfaces

Mining Biophysical Properties at Protein Interfaces

From Protein Interfaces to Protein Binding Patches

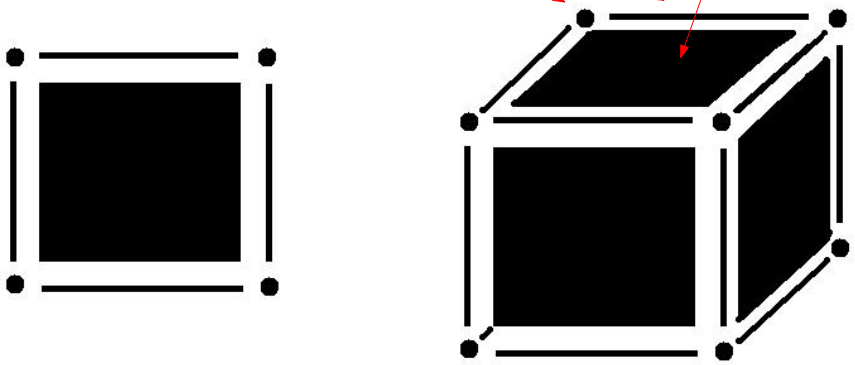
Linear Cell Complexes: Examples

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



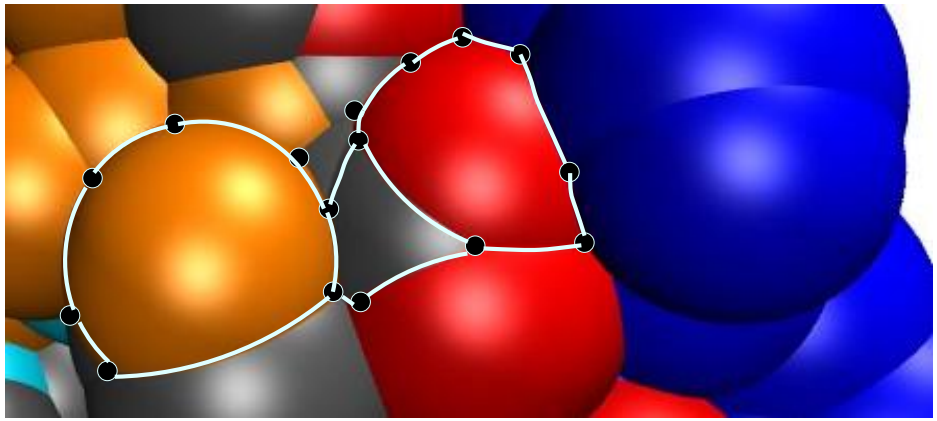
0-cell, 1-cell, 2-cell, 3-cell

Cubical complex:lego of hyper-cubes

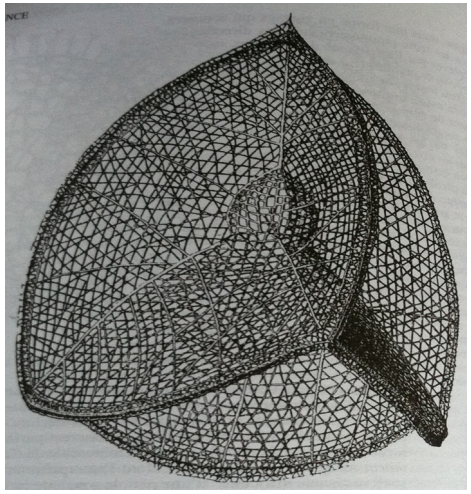


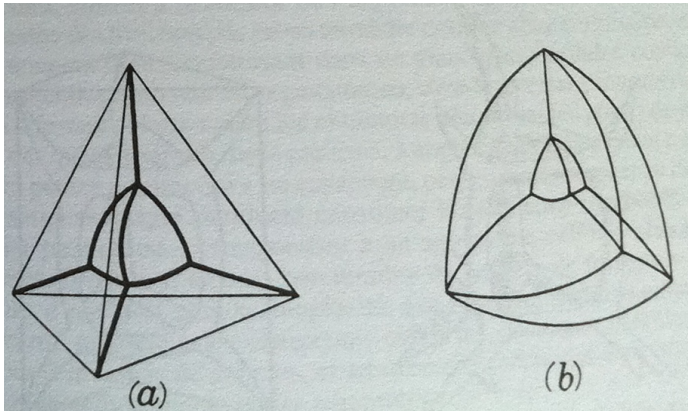
Curved Cell Complexes: Examples

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

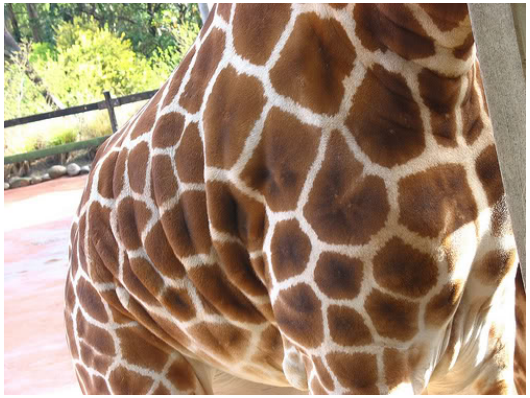


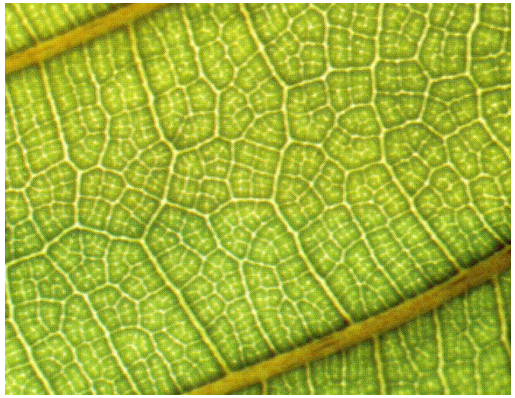
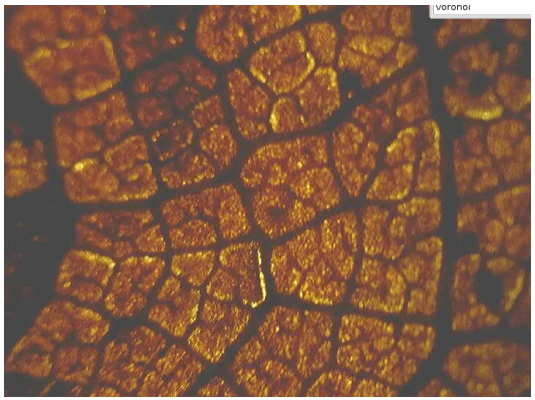
Curved 3D cell complex





Voronoi diagrams in Science and Growth Processes: Gallery



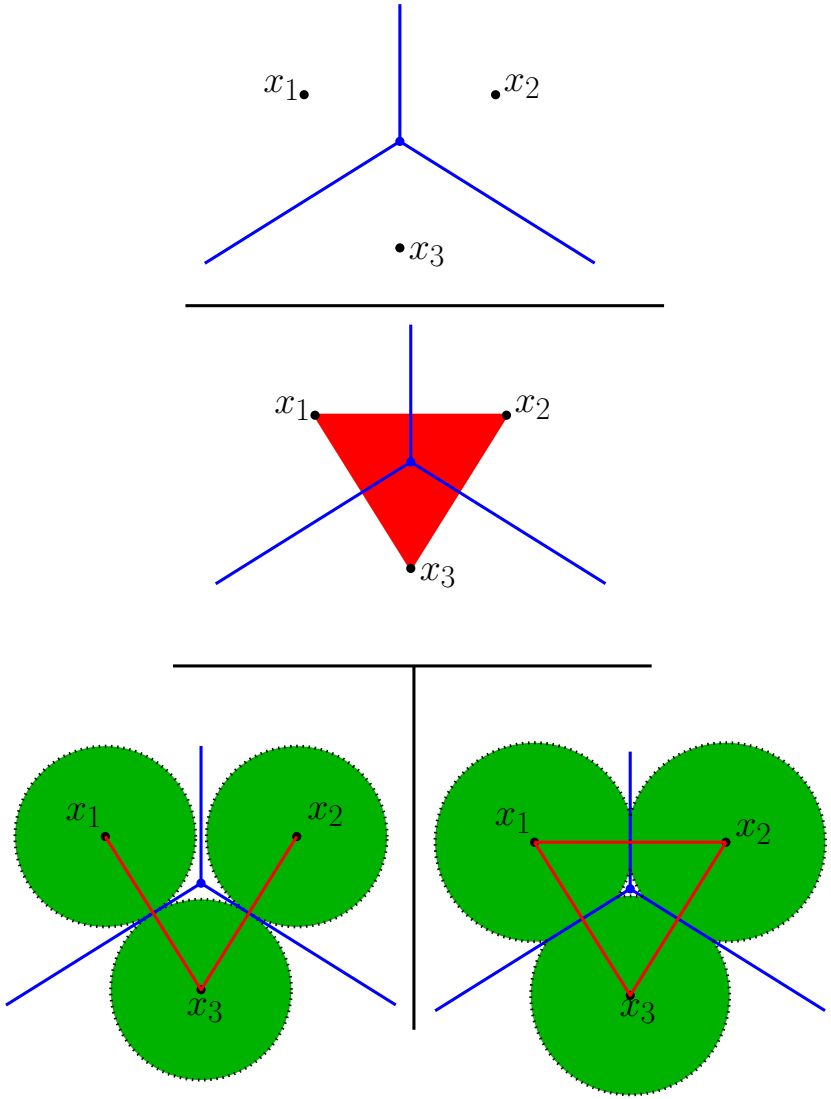



http://forum.woodenboat.com/showthread.php?112363-Voronoi-Diagrams-in-Nature

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

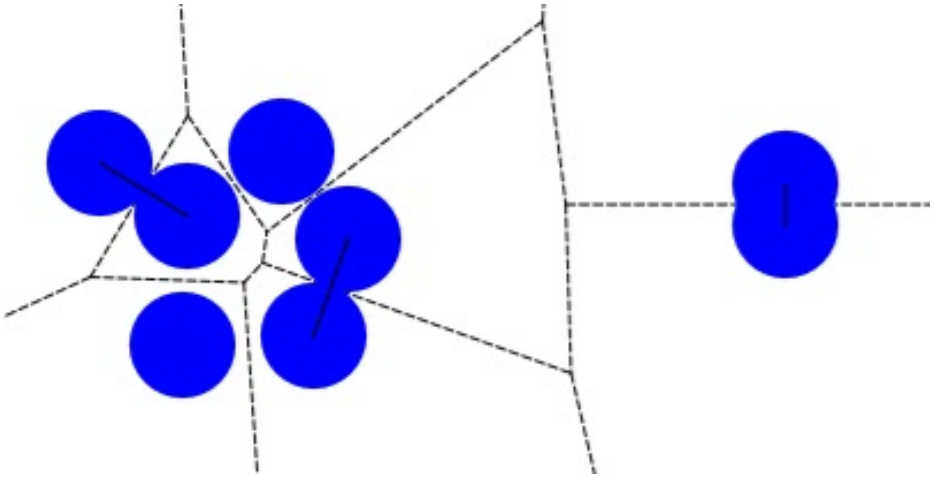
Euclidean Voronoi diagram and α-complex

- 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

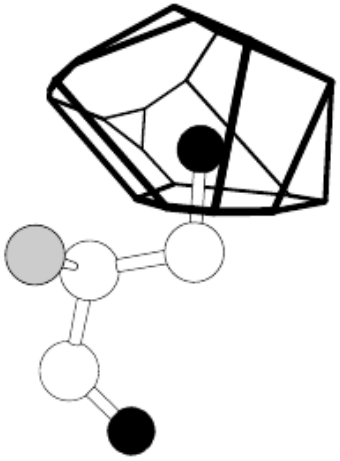


α-shapes : building a simplicial complex encoding the topology of the shape

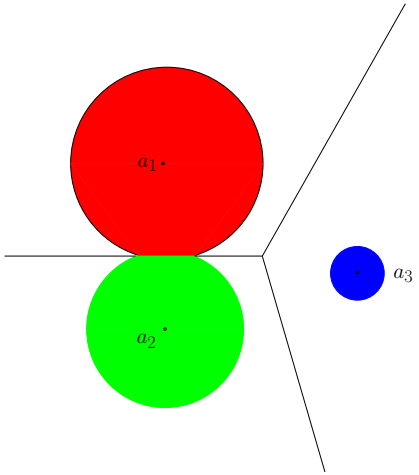
On the Volume of Union of Balls

▷ 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

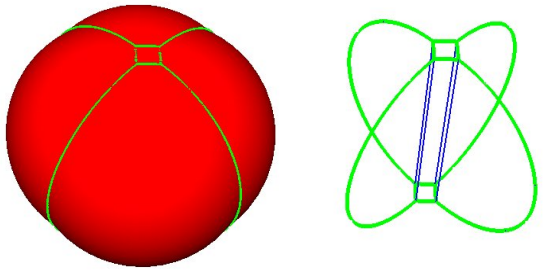
▷ Strategy developed: certified volume calculation

– 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.



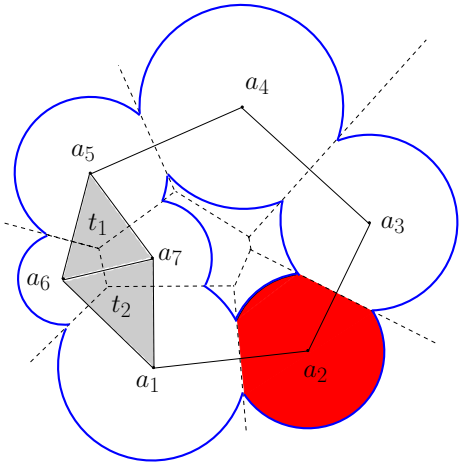
▷ Assessment

– 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)

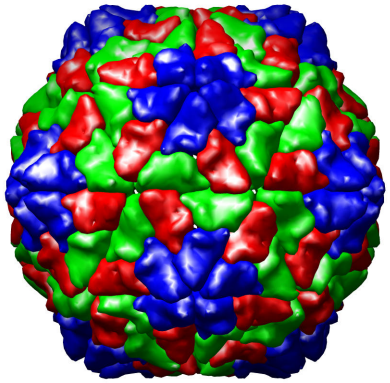
▷Ref: McConkey et al; Bioinformatics 18; 2002 (McC-et-al)

Geometry versus Topology:

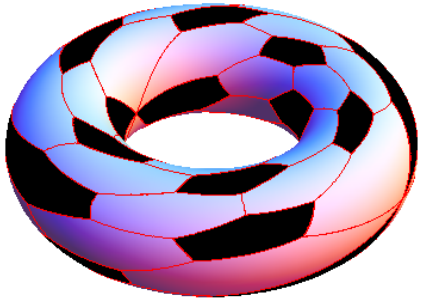
The Theorem of Classification of Closed Surfaces in $\mathbb{R}^3$

A topological sphere: a genus 0 surface



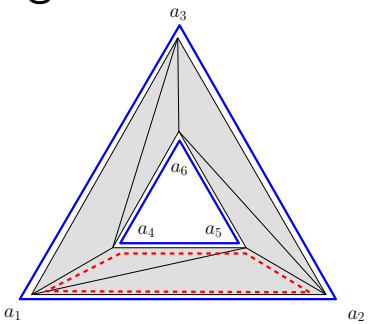


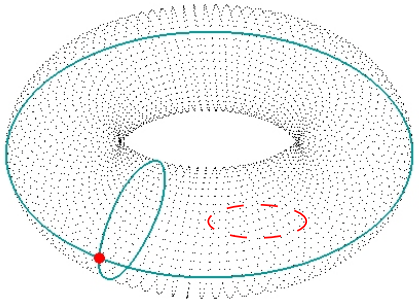
A topological torus: a genus 1 surface


Homology Theory

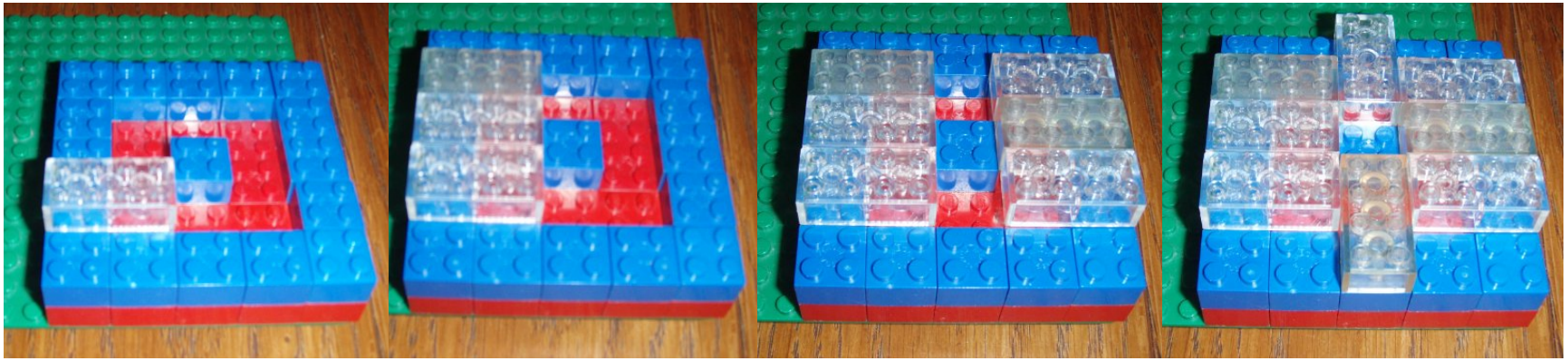
▷ Homology: counting k-dimensional cycles which do not bound (bound voids), regardless of their *thickness*





▷ Betti numbers count homology generators: examples in 3D

β_0: #cc $\qquad$ β_1: # tunnels $\qquad$ 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 Again

▷ Mr Euler playing golf

▷ Funnels on energy landscapes. . .

Euler characteristic?

Pitfall: index-1 saddles

Native state?

Modeling High Resolution Protein Complexes

Protein Interfaces: Key Questions

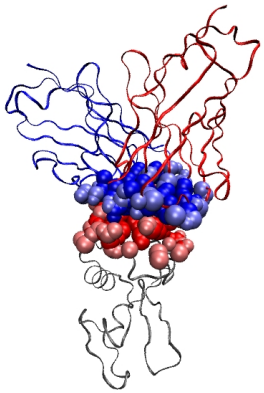
Geometric Intermezzo: Voronoi Diagrams and Relatives

Describing Protein Interfaces

Mining Biophysical Properties at Protein Interfaces

From Protein Interfaces to Protein Binding Patches

Modeling the Interface of Macro-molecular Complexes



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

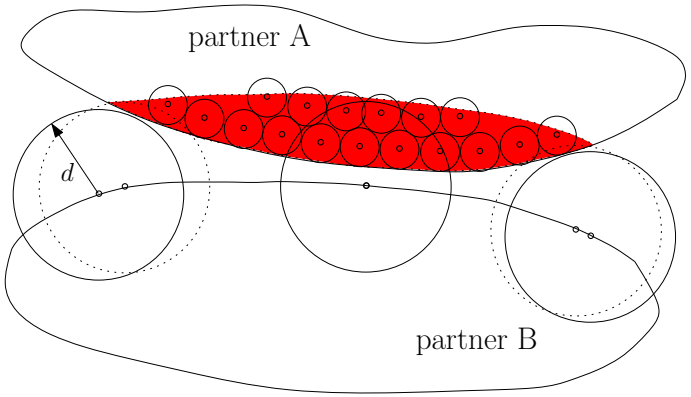
- Shape - topology:
  - # connected components, holes, voids / cavities [Homology]
  - morphology: *fat, skinny, dumbbell-like*

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

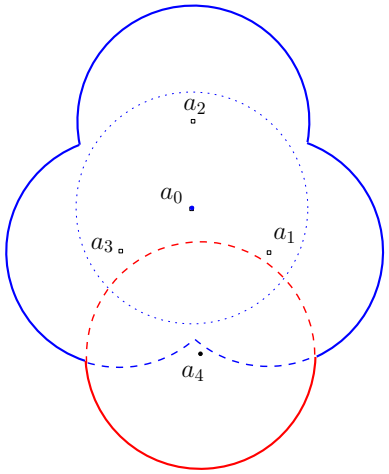
- Correlations with bio-physical quantities
  - conservation of amino-acids
  - biochemical properties

About Interface Models

▷ Distance threshold
(geometric footprint)



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



▷ The Voronoi interface model

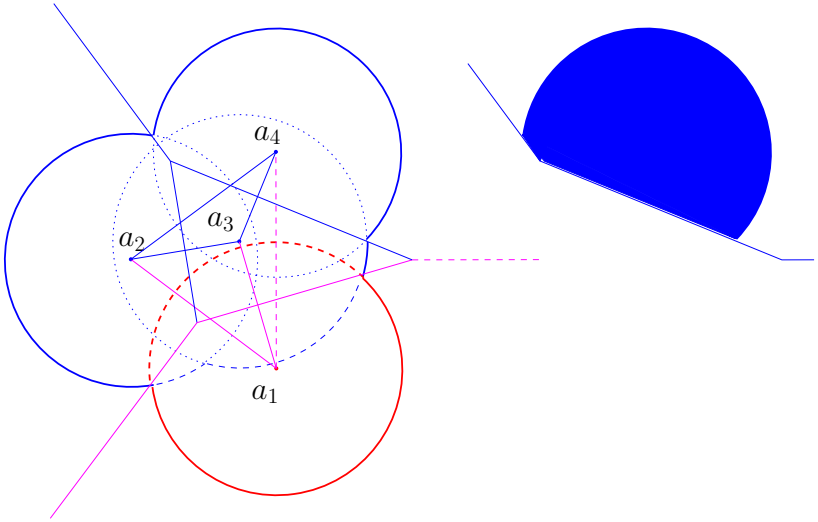
- A parameter free interface model
- Singles out a single layer of atoms
- Is amenable to geometric and topological calculations

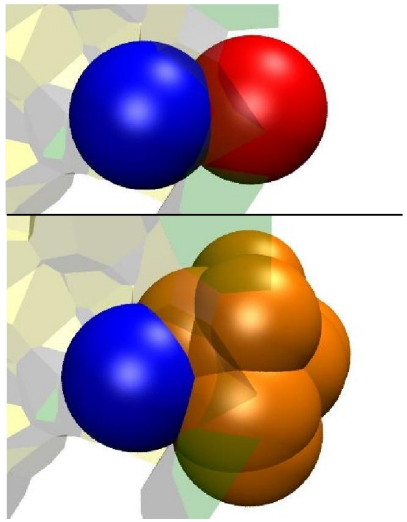
▷ Applications

- 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)

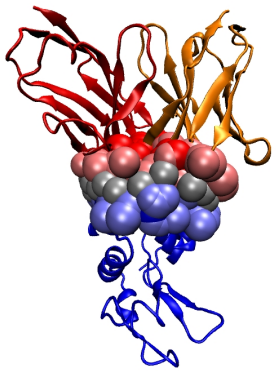
Voronoi interface: balance between geom. footprint and ΔASA

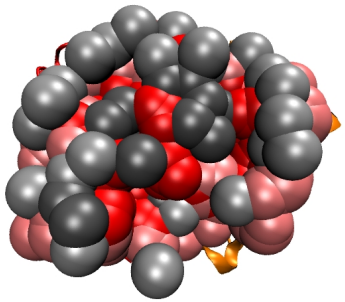
▷Ref: `Cazals, Proust, Bahadur, Janin; Protein Science; 2006`

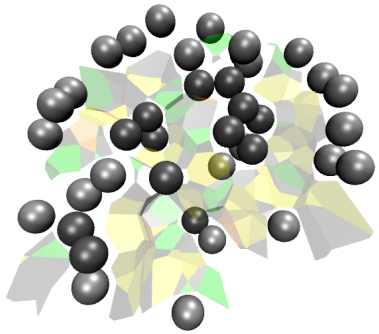
Voronoi Interfaces : Illustrations

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

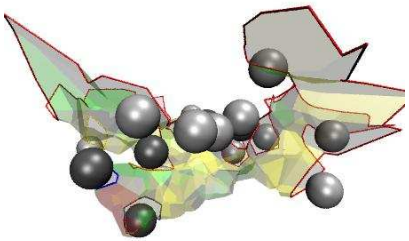
- Role of strutural water –antobody-antigen



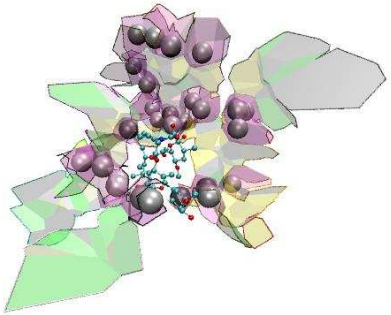




- Curvature –protease-inhbitor



- Multi-patch structure –signal transduction



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

Modeling High Resolution Protein Complexes

Protein Interfaces: Key Questions

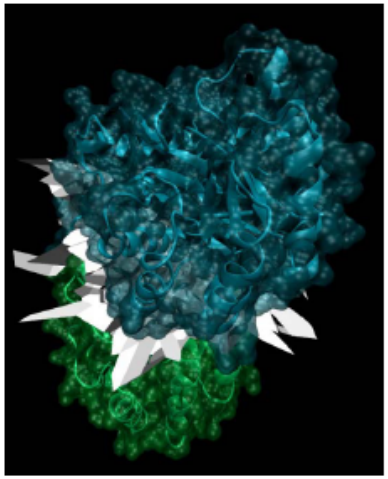
Geometric Intermezzo: Voronoi Diagrams and Relatives

Describing Protein Interfaces

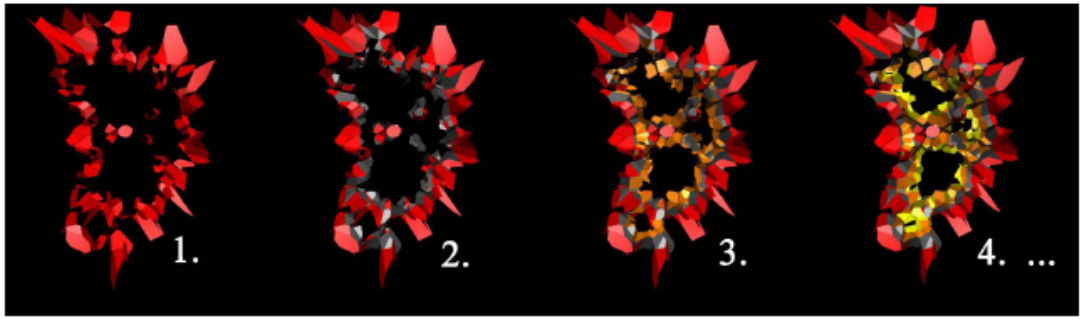
Mining Biophysical Properties at Protein Interfaces

From Protein Interfaces to Protein Binding Patches

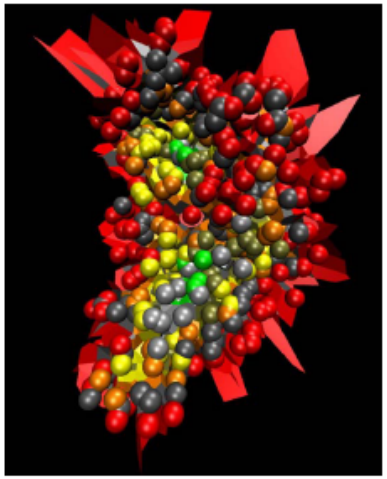
Shelling the Voronoi Interface: Illustration



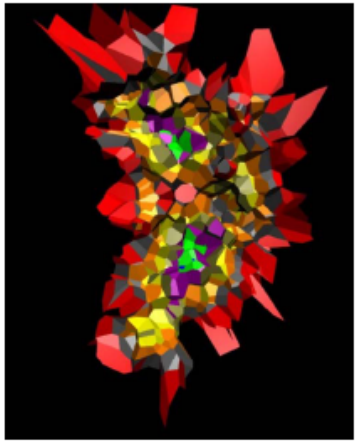
Dihydroorotate dehydrogenase (2DOR)



Shelling the Voronoi interface...



Projected on atoms

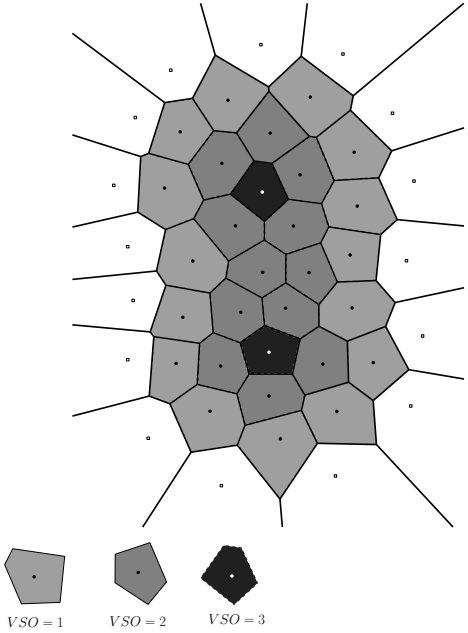


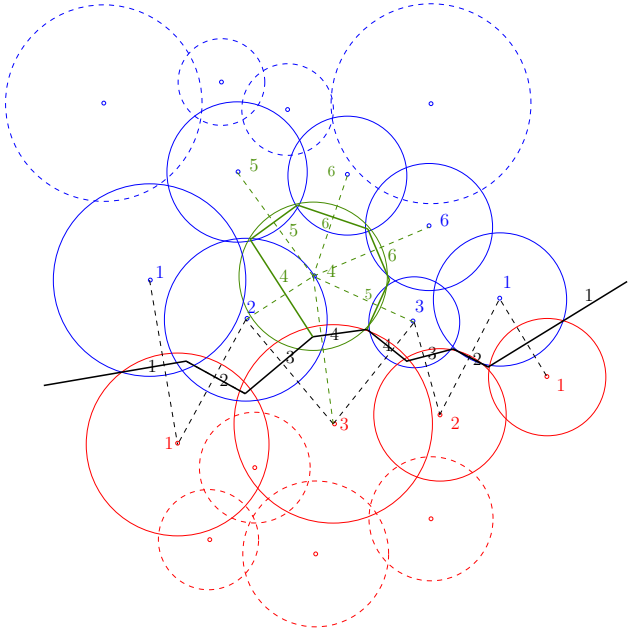
Shelled interface

- Properties?
- Evolution during an MD simulation?

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

Voronoi Shelling Order: Definition



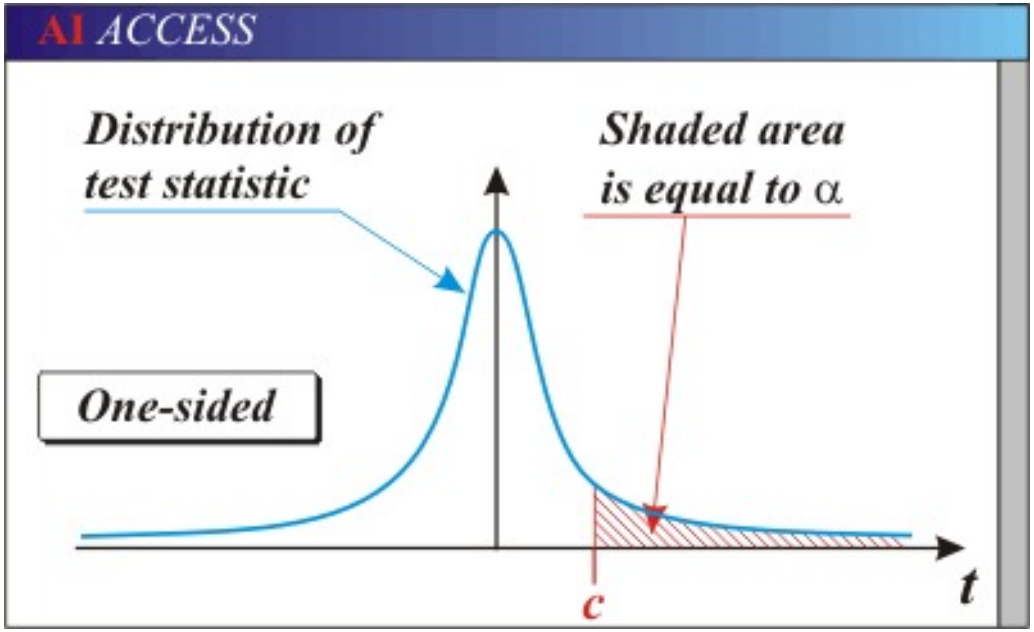


- Three stages
  - select bicolor Delaunay edges in the 0-complex
  - walking over the dual Voronoi facets/tiles
  - pulling back values onto the atoms

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

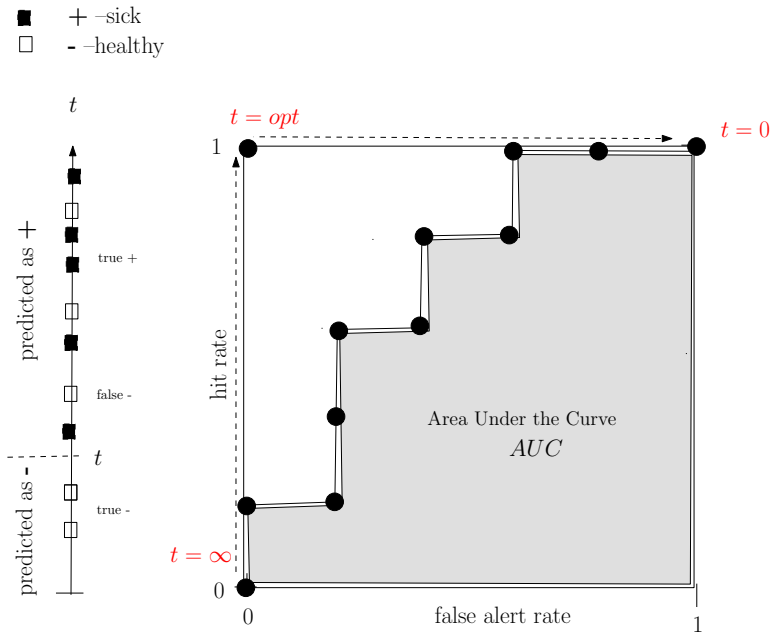
Receiver Operating Characteristic (ROC) curves

- 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:



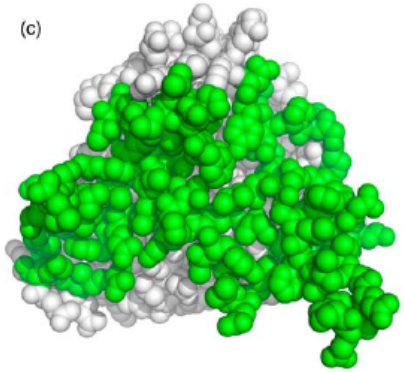
- p-value calculation for a particular value AUC_0:

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

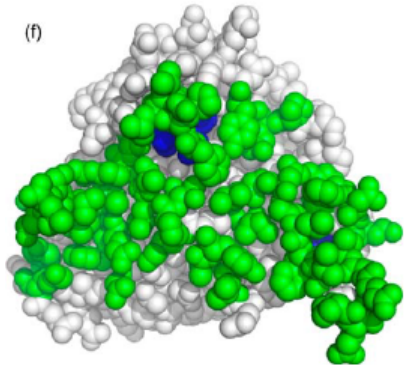
Predicting Important Residues: the Role of Dry Residues

▷ 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)

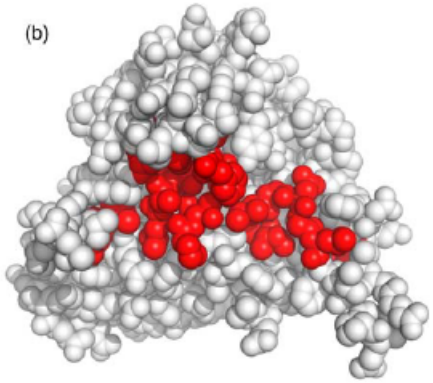
▷ 2DOR: interface residues within 7 Å



▷ 2DOR: interface residues: using ΔSAS



▷ 2DOR: dry residues

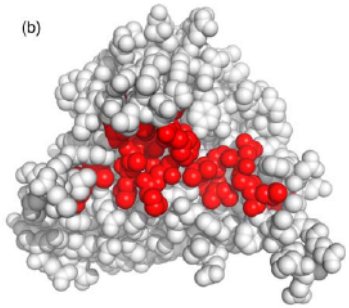


▷Ref: Mihalek, Res, Lichtarge; JMB, 2007

Water Traffic and Conservation of Residues at Protein - Protein Interfaces

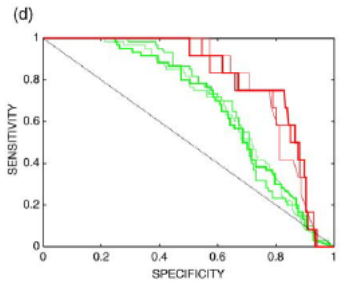
- ▷ Dry A.A. tend to be more *important*
- ▷ Protocol: MD simulation; A.A. s.t. $\Delta ASA > 0$
- ▷ Traffic intensity for A.A. i: $I_i = \frac{1}{T}\sum_w \frac{1}{\tau_w}$
- ▷ Dry residue w.r.t.traffic intensity:
  - $I_i \leq 0.005 ps^{-2}$ for homodimers
  - $I_i \leq 0.01 ps^{-2}$ for heterodimers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

▷ 2DOR: dry residues



conservation predicts dryness versus conservation predicts geom. footprint

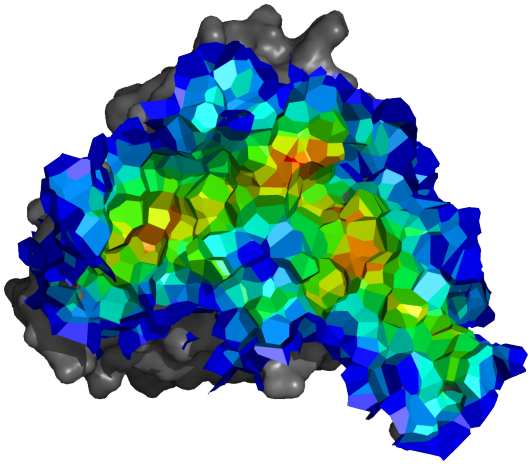
- ▷ Conclusions:
  - 3 conservations methods perform equally
  - **AUC(conserv. $\rightarrow$ dryness) $\gg$ AUC(conserv. $\rightarrow$ geom. footprint)**

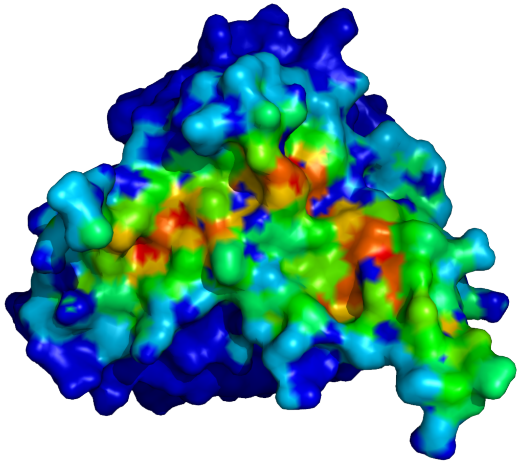


▷Ref: Mihalek, Res, Lichtarge; JMB, 2007

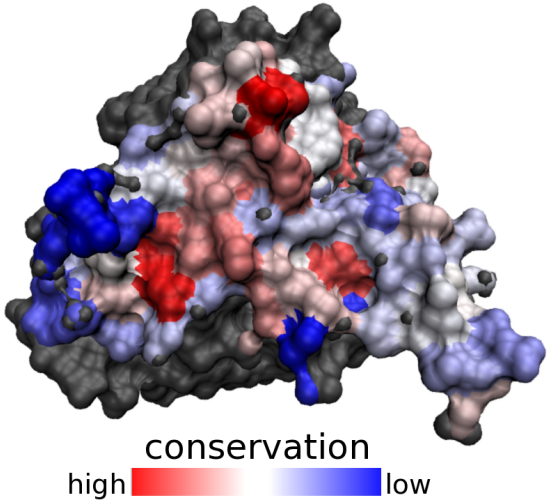
VSO versus Dryness – 2DOR

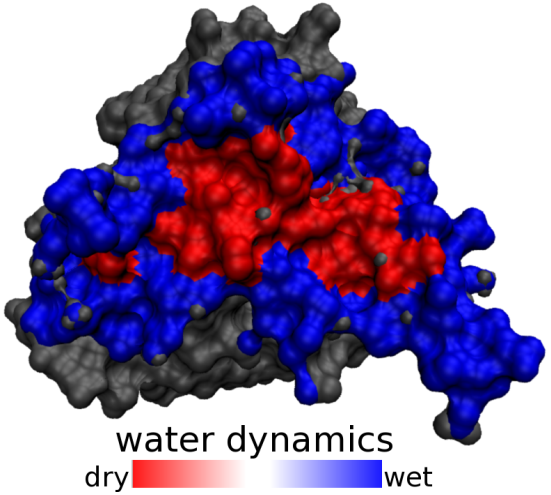
- VSO: facets and atoms

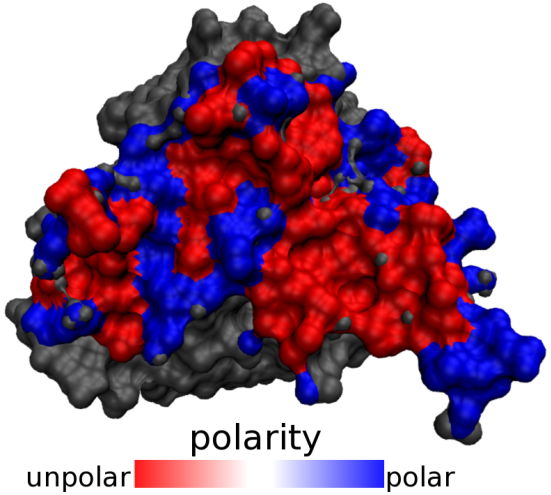




- Conservation, dryness, polarity







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.]$\rightarrow$ dryness [threshold] conserv. [cont.] $\rightarrow$ dryness [threshold]

conserv. [cont.]$\rightarrow$ VSO [threshold] VSO [cont.] $\rightarrow$ 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

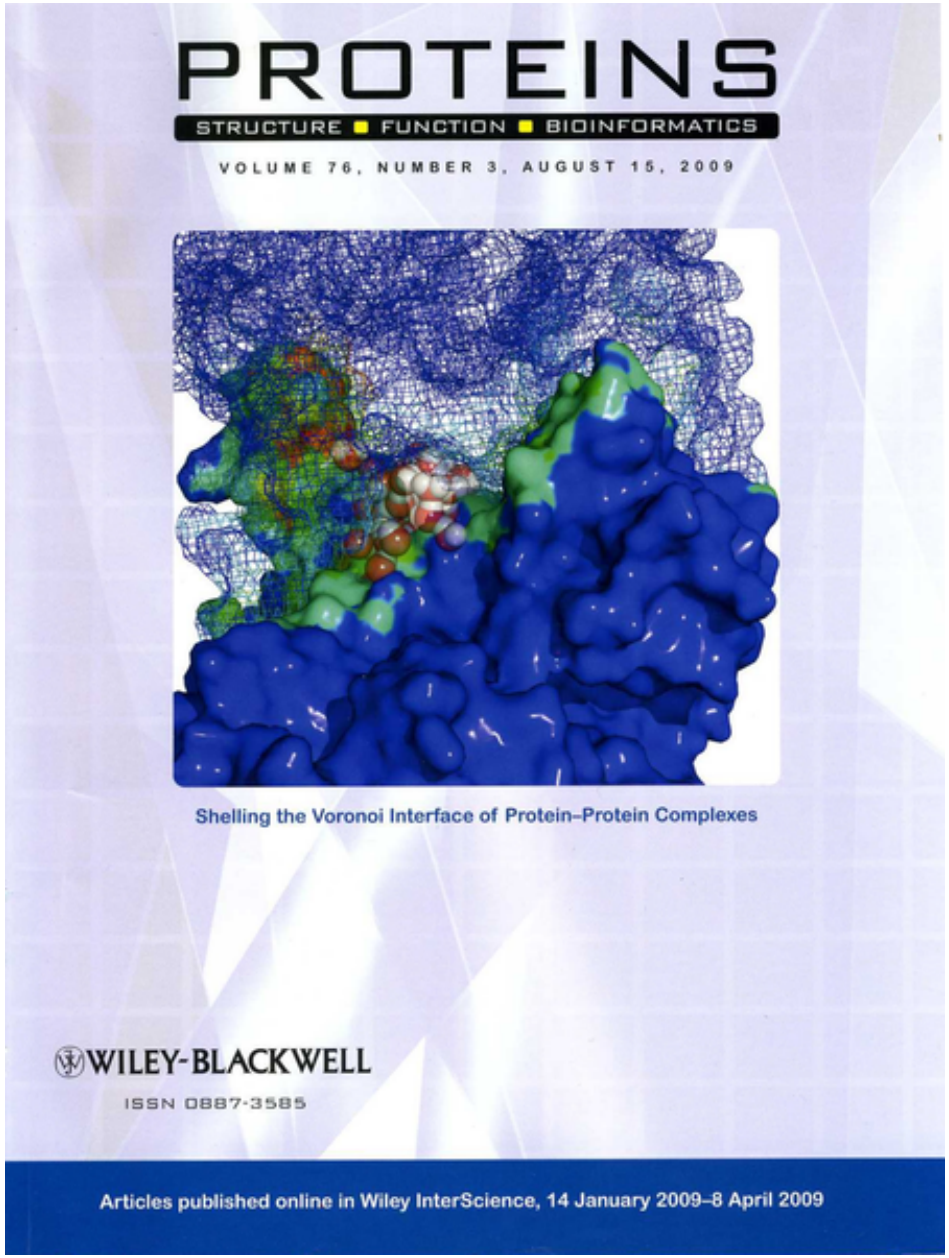
PDB Id.	VSO→dryness		conserv.→dryness		conserv.→VSO		VSO→unpolar	
	AUC	P-value	AUC	P-value	AUC	P-value	AUC	P-value
...								
Reject H_0		36/36		25/36		14/36		27/36
Global	0.84	2e-265	0.63	2e-43	0.62	4e-20	0.64	2e-63

▷ 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

Protein Interfaces: Key Questions

Geometric Intermezzo: Voronoi Diagrams and Relatives

Describing Protein Interfaces

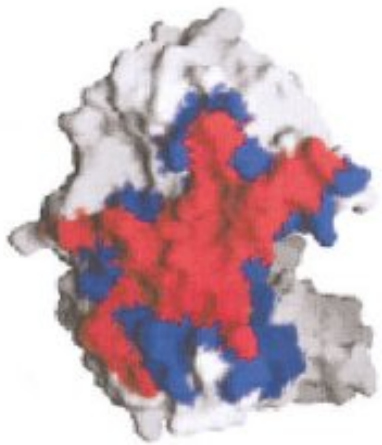
Mining Biophysical Properties at Protein Interfaces

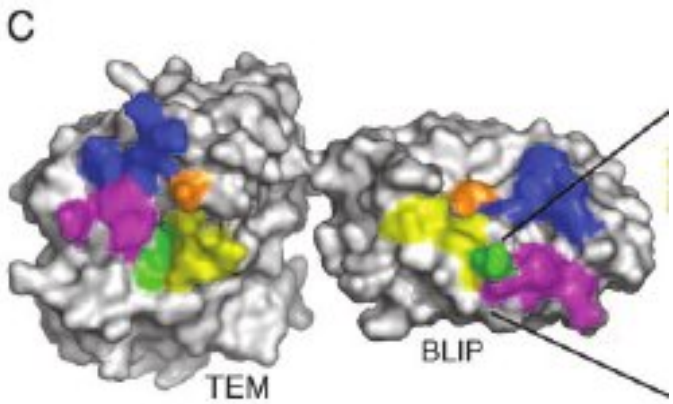
From Protein Interfaces to Protein Binding Patches

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

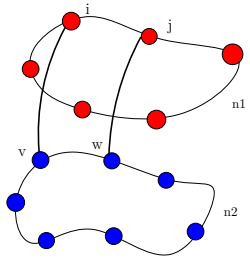
  $$RMSD_d(S_1, S_2) = \sqrt{\sum_{i<j} |d^1_{i,j} - d^2_{m(i),m(j)}|^2 / \binom{n}{2}}$$

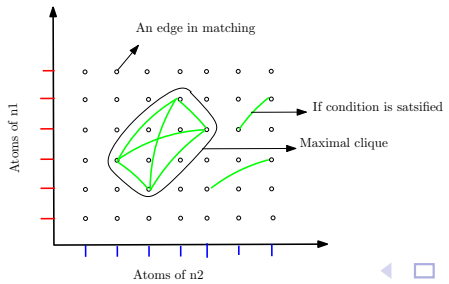
- Goal for two molecules M_1 and M_2: find the largest $S_1 \subset M_1$ and $S_2 \subset M_2$, and the corresponding mapping $m()$, such that $RMSD_d(S_1, S_2) \leq \epsilon$

- 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

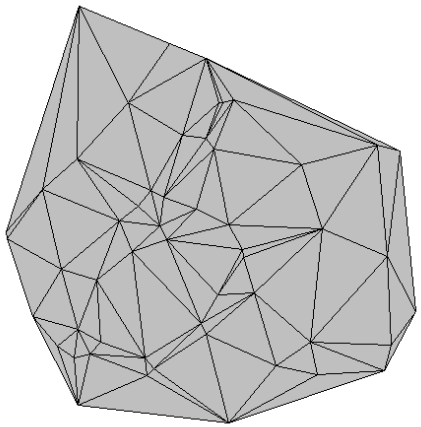




Shelling a Cell Complex

▷ Input

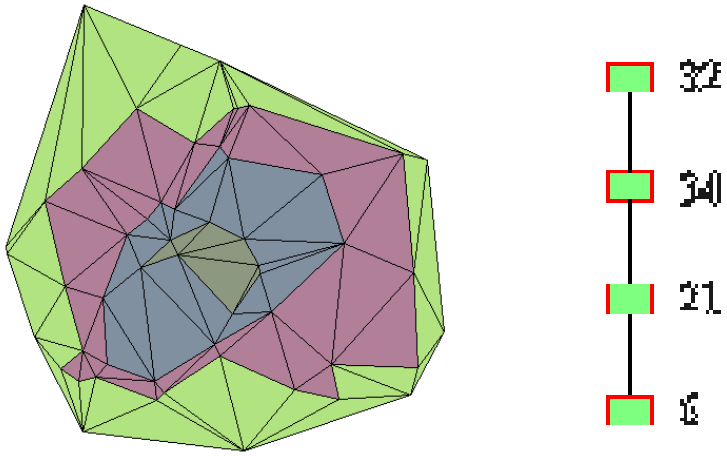
Cell complex say **D** dimensional
- Cells - dimension **D**
- Facets - dimension **D-1**
- Pivots - dimension **D-2**



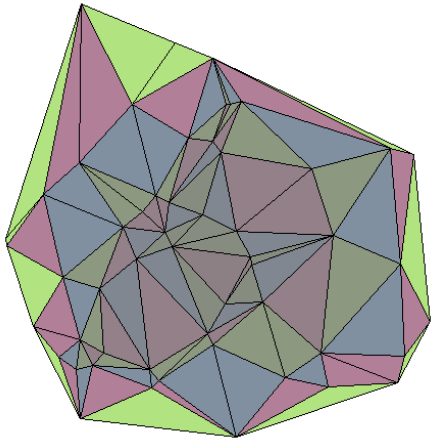
Example : 2D Alpha shape
Triangles are cells, edges are facets and vertices are pivots

▷ Output

▷ Shelling by pivoting

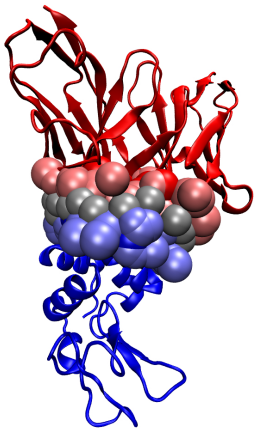


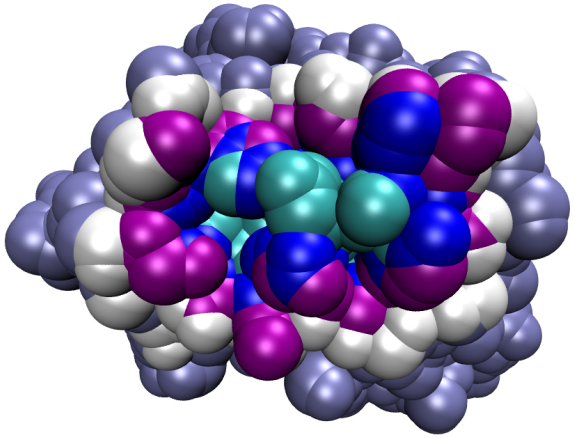
▷ Shelling by face connectivity

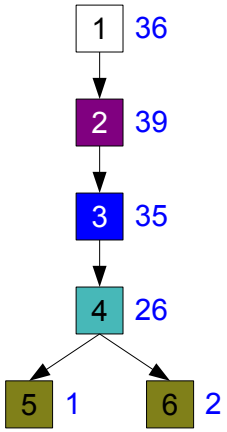


Shelling a Binding Patch yields a Topological Encoding

- 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
- Semantics of the 3 operations: problem dependent
- Complexity, using dynamic programming: time: $O(n^3)$; space: $O(n^2)$

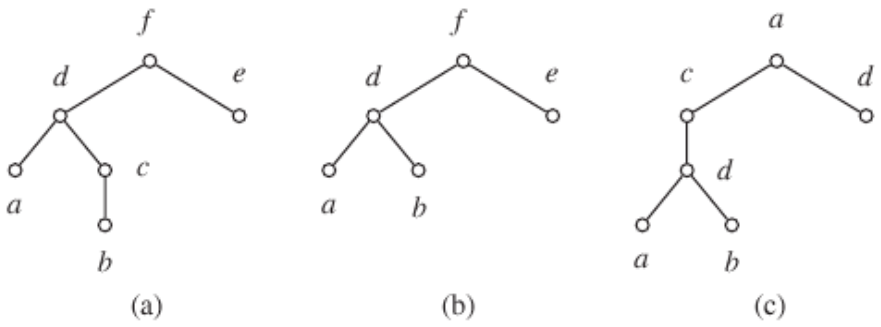
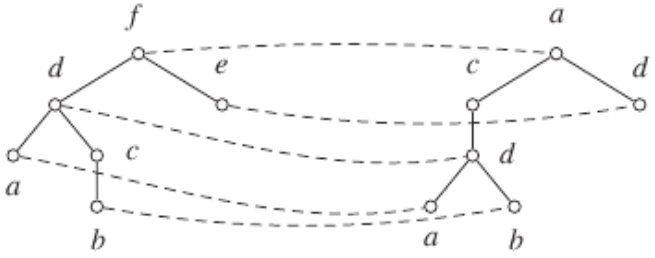


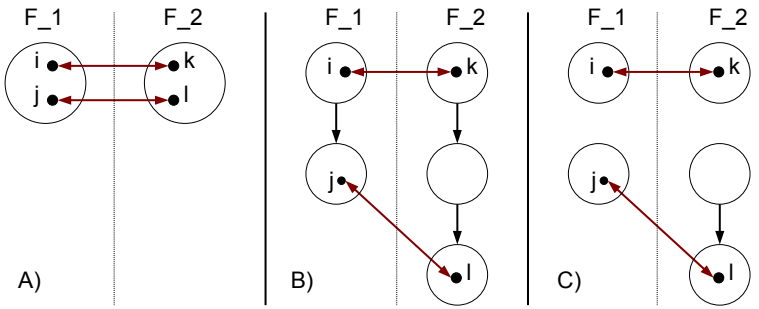
Fig. 2. Transforming (a) into (c) via editing operations. (a) A tree. (b) The tree after deleting the node labeled c. (c) The tree after inserting the node labeled c and relabeling f to a and e to d.



- 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:
  $\mathrm{SIM}_t(T_1, T_2)$: number of atoms matched
  $\mathrm{TED}_t(T_1, T_2) = |\ T_1\ | + |\ T_2\ | - 2\ \mathrm{SIM}_t(T_1, T_2)$
- Corresponding dissimilarity $\in 0..1$
  $\mathrm{DIS}_t(T_1, T_2) = \mathrm{TED}_t(T_1, T_2)/(|\ T_1\ | + |\ T_2\ |)$



Application 2: Geometric Comparison of Patches

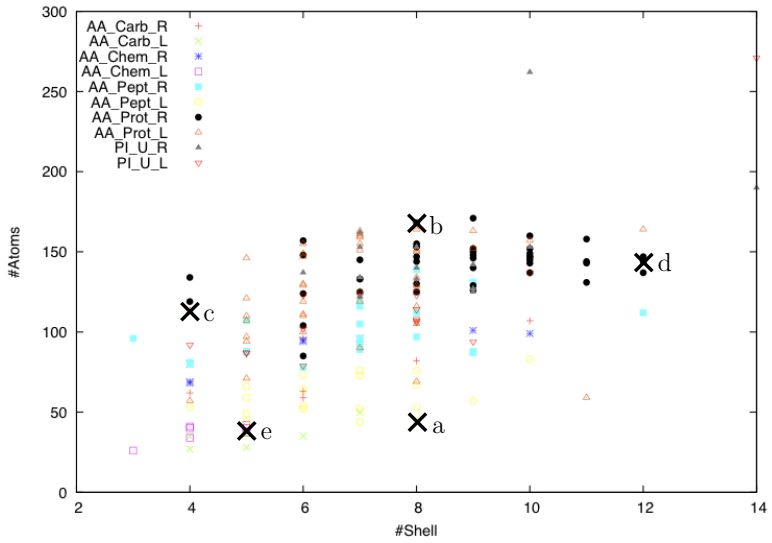
- Restrict the Max-Clique calculation to shells

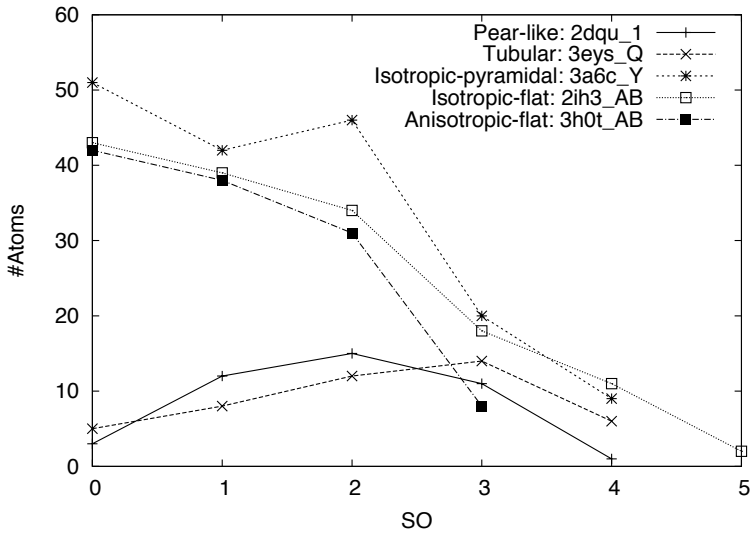
- Atoms matched are called isotopologic:
  $\mathrm{SIM}_g(T_1, T_2)$: number of atoms matched
  $\mathrm{TED}_g(T_1, T_2) = \mid T_1 \mid + \mid T_2 \mid - 2\ \mathrm{SIM}_g(T_1, T_2)$
- Corresponding dissimilarity $\in$ 0..1:
  $\mathrm{DIS}_g(T_1, T_2) = \mathrm{TED}_g(T_1, T_2)/(\mid T_1 \mid + \mid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

- 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: $\mathrm{DIS}_g(T_1, T_2) \geq \mathrm{DIS}_t(T_1, T_2)$

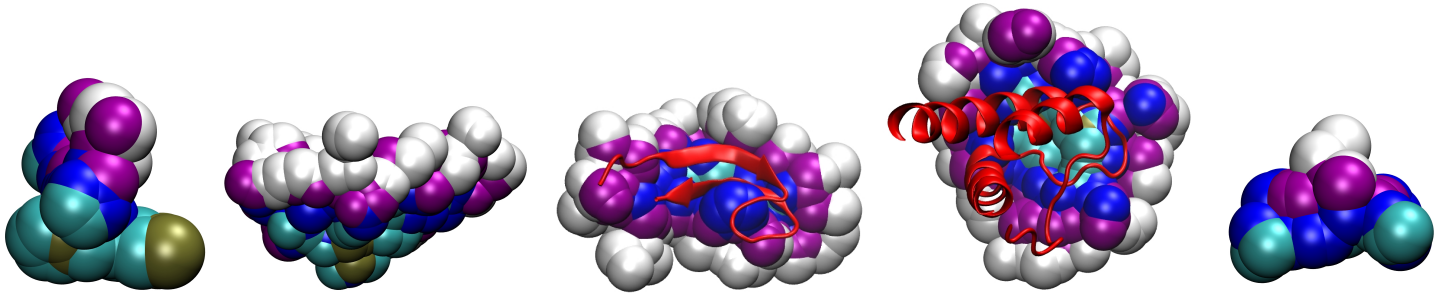
Binging Patches: Typical Morphologies

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



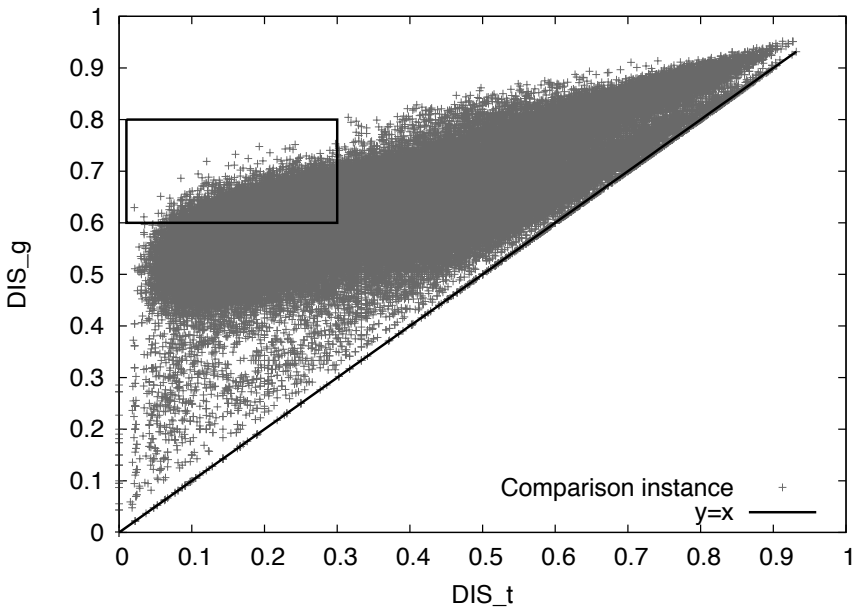


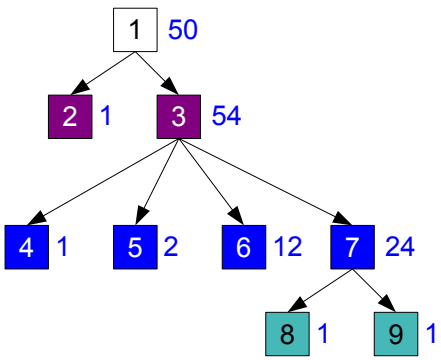
- Typical morphologies:

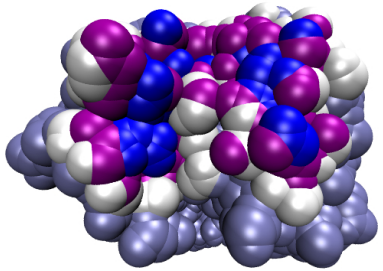


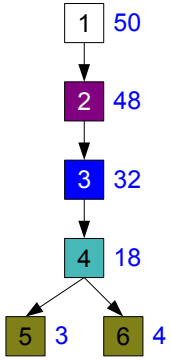
(a) tubular (b) isotropic-pyramidal (c) anisotropic-flat (d) isotropic-flat (e) pear-like

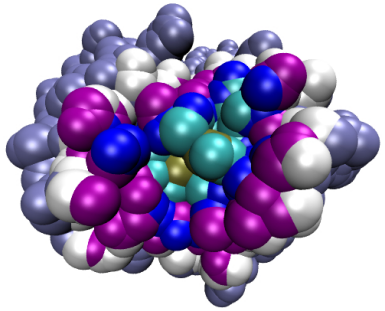
Similar Topology, Dissimilar Geometry





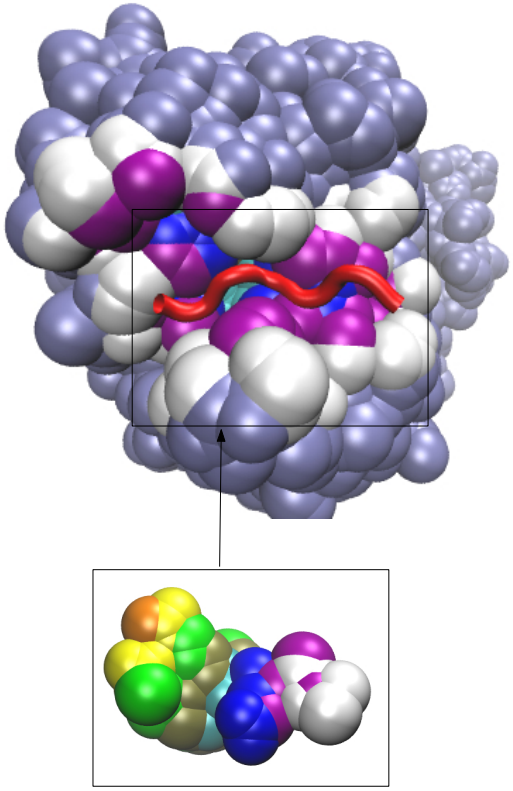






Symmetry of Patches and Homogeneity of Families

▷ Anisotropic vs tubular



▷ Identification favors the family rather than the complement

DB decomposition:

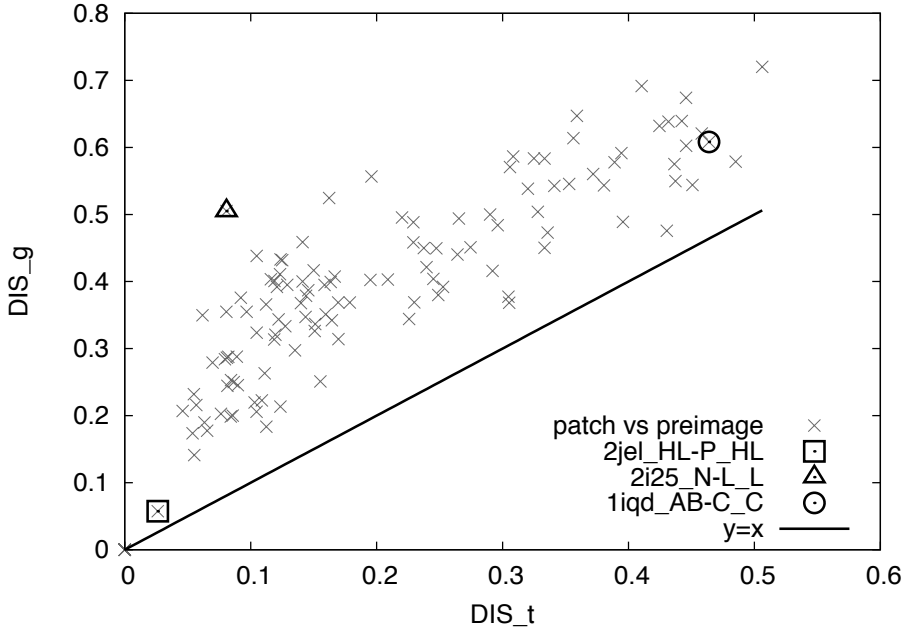
$$\mathcal{P} = p \cup P_{\setminus p} \cup \overline{P} \cup P^c$$

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

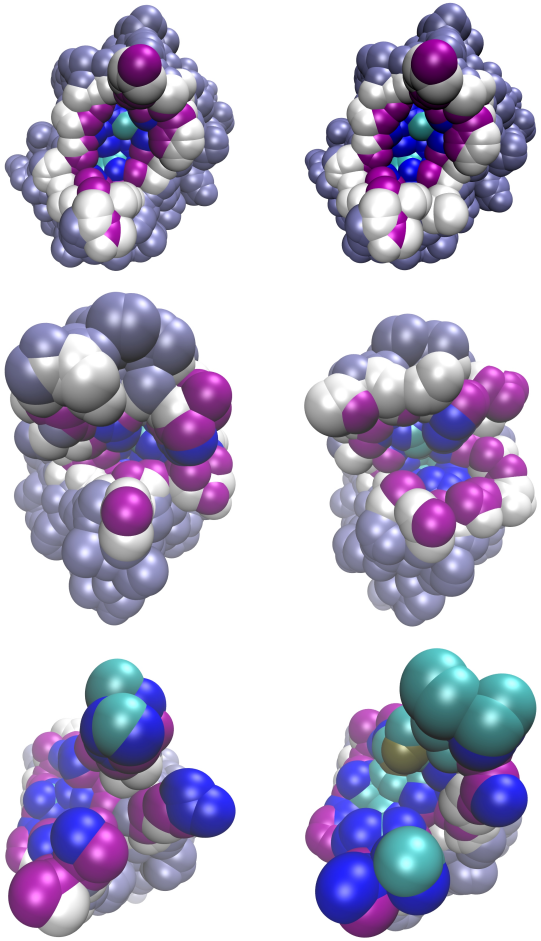
- Patch vs. prepatch on unbound partner



- Topologically rigid patches:

a third tier

- Rigid, topo-rigid, flexible



Affinity Benchmark: Predicting Binding Affinities

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 ($-\ln K_d$).

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, Journal of proteome Research, 9 (5); 2010

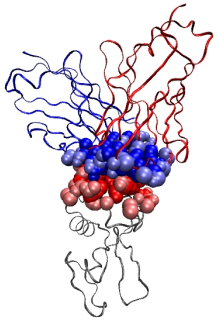
▷Ref: Kastritis et al; Protein Science (20), 2011

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

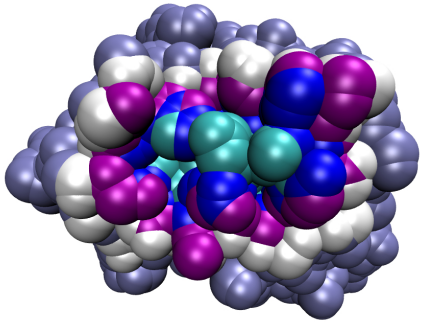
Sotware: Modeling Protein Interfaces

▷ intervor: modeling protein - protein interfaces

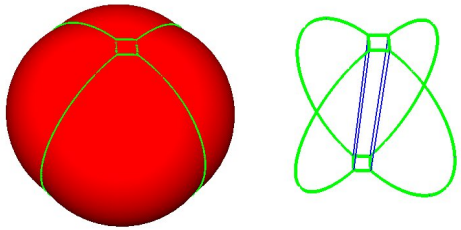


http://cgal.inria.fr/abs/Intervor; Bioinformatics; 26 2010

▷ vorpatch: topological encoding of binding patches

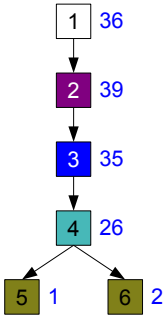


▷ vorlume: certified molecular surfaces and volumes

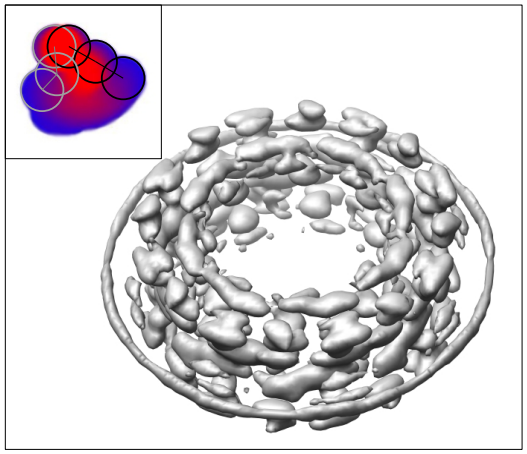


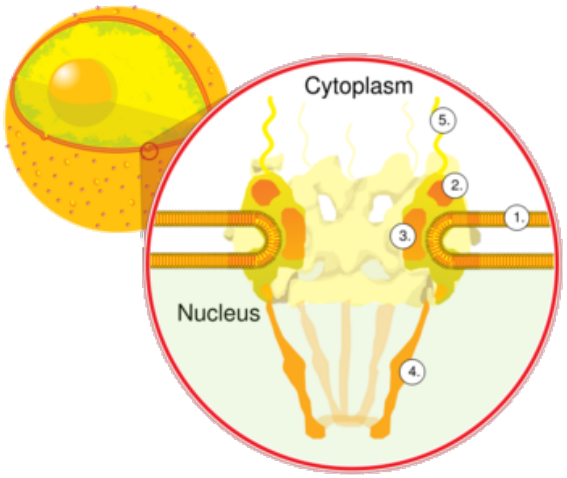
http://cgal.inria.fr/abs/Vorlume; ACM Trans. Math Softw.; 2011

▷ compatch: comparing binding patches



Part II: Modeling Large Protein Assemblies





Voronoi Diagrams Again

Reconstruction by Data Integration

Toleranced Models

Assessing the Reconstruction of Fuzzy Models

- Contact probabilities
- Isolated copies
- Pairwise contacts

Modeling Large Protein Assemblies

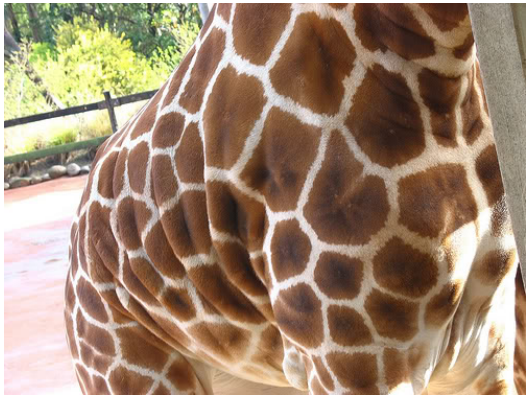
Voronoi Diagrams Again

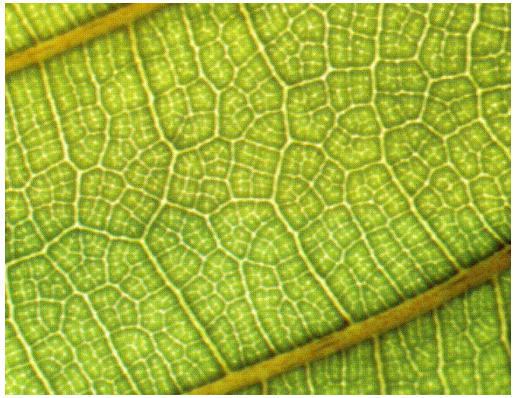
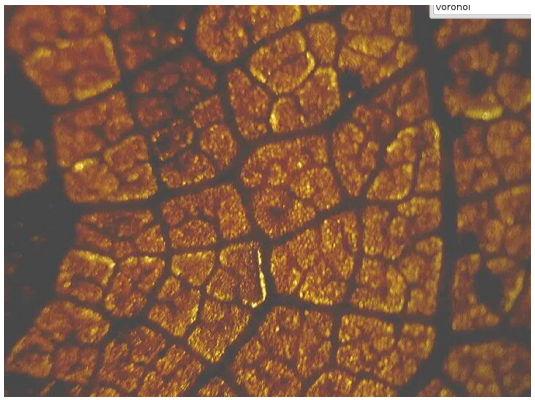
Reconstruction by Data Integration

Toleranced Models

Assessing the Reconstruction of Fuzzy Models
- Contact probabilities
- Isolated copies
- Pairwise contacts

Voronoi diagrams in Science and Growth Processes: Gallery

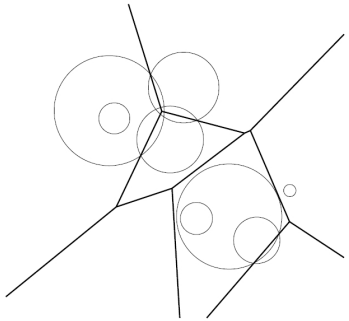


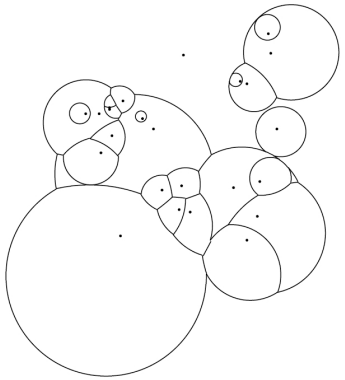



http://forum.woodenboat.com/showthread.php?112363-Voronoi-Diagrams-in-Nature

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

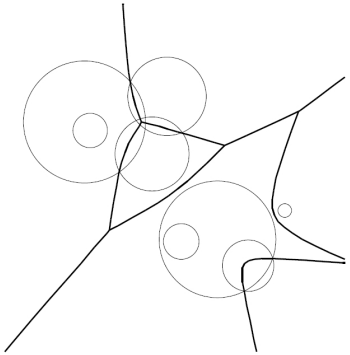
The Zoo of curved Voronoi diagrams

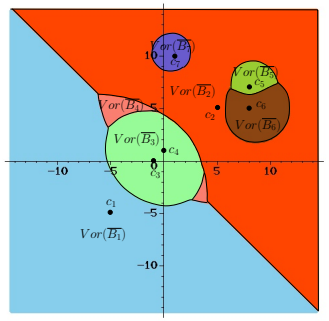




▷ Power diagram:
$d(S(c,r),p) = \|c-p\|^2 - r^2$

▷ 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; Effective Comp. Geom.; 2006

▷Ref: Cazals, Dreyfus; Symposium on Geometry Processing; 2010

Modeling Large Protein Assemblies

Voronoi Diagrams Again

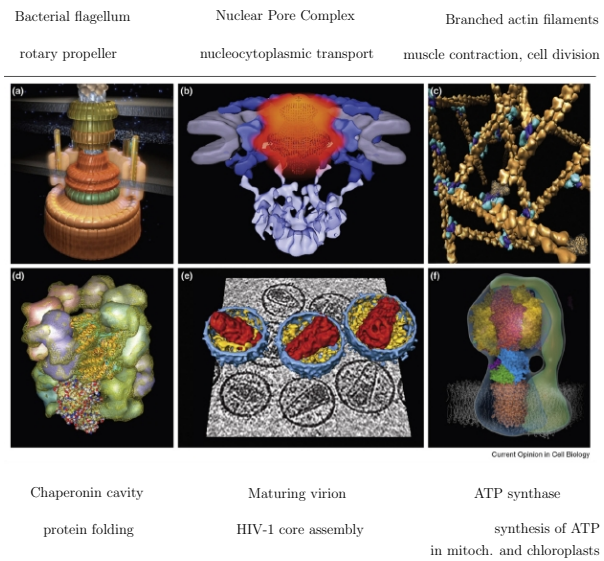
Reconstruction by Data Integration

Toleranced Models

Assessing the Reconstruction of Fuzzy Models
- Contact probabilities
- Isolated copies
- Pairwise contacts

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

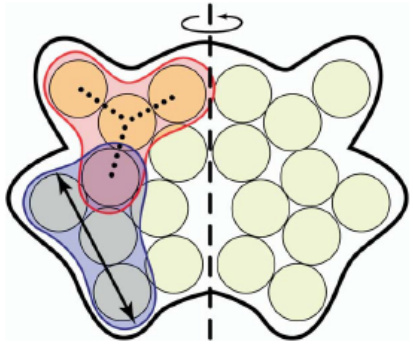
▷ Difficulties

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,. . .)



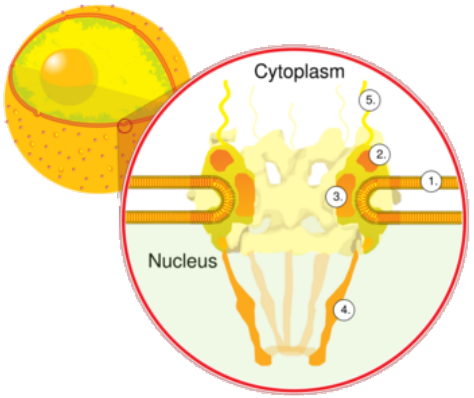
▷ Restraints, experimental data and . . . ambiguities:

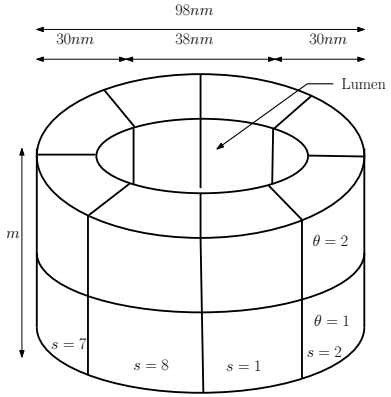
Assembly	: shape	cryo-EM	fuzzy envelopes
Assembly	: symmetry	cryo-EM	idem
Complexes:	: interactions	TAP (Y2H, overlay assays)	stoichiometry
Instance:	: shape	Ultra-centrifugation	rough shape (ellipsoids)
Instances:	: locations	Immuno-EM	positional uncertainties

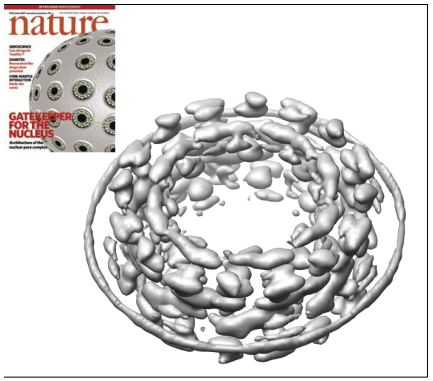
▷Ref: Alber et al, Ann. Rev. Biochem. 2008 + Structure 2005

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

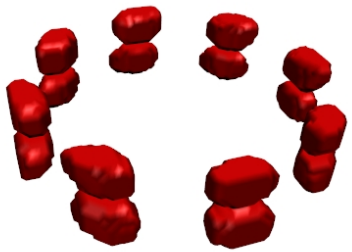
▷Ref: Alber et al; Nature; 450; 2007

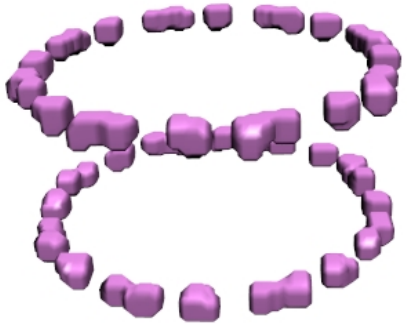
NPC: Example Density Maps

Stoichiometry vs number of connected components

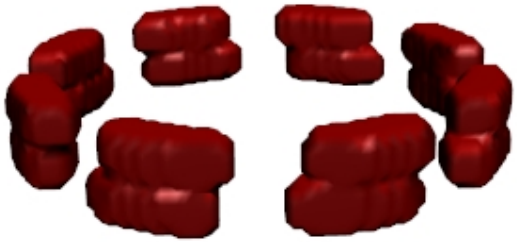
- Two types of problems:
  number of connected components vs stoichiometry
  volume of each connected component vs. volume estimated from the sequence

- Cases: equal (Nup157); larger (Sec13)





- Cases: smaller (Nup170, Pom152)

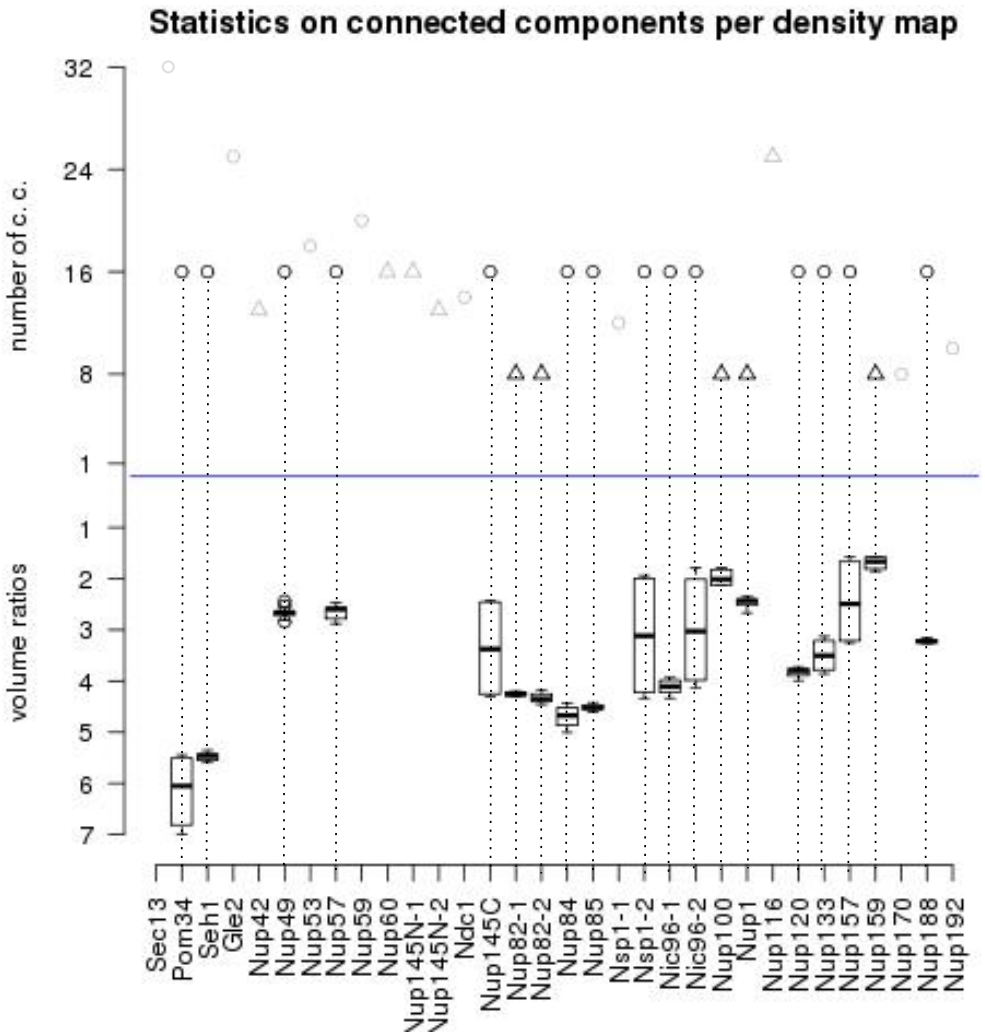


- Ref: Alber et al; Nature; 450; 2007

Uncertainties of the Density Maps

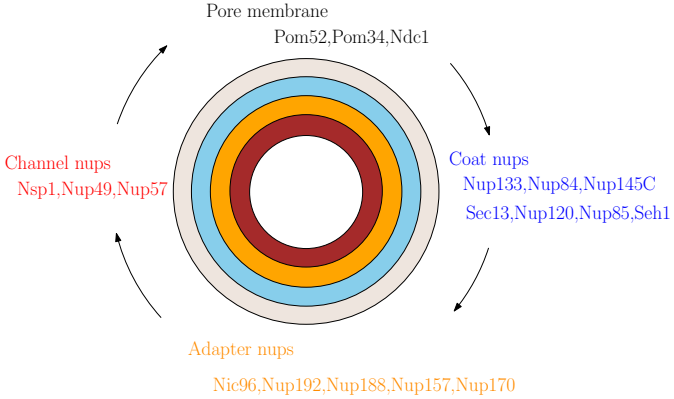
- Volume of connected components of non empty voxels vs. reference volume (estimated from the sequence)

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



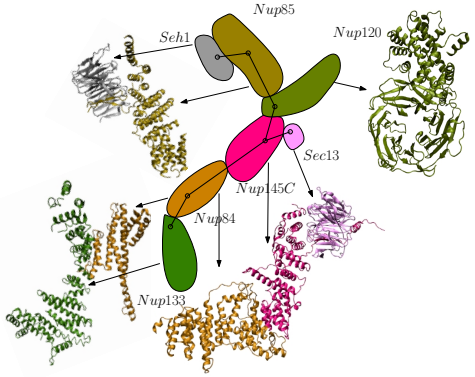
Putative Models of Sub-complexes: the Y-complex

▷ Symmetric core of the NPC



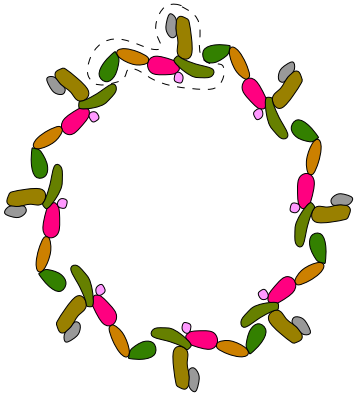
▷Ref: Blobel et al; Cell; 2007

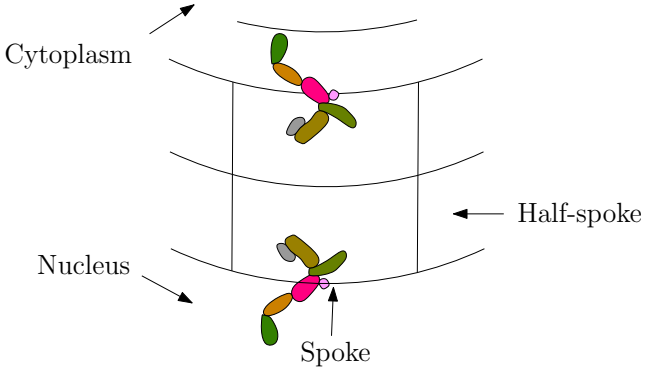
▷ The Y-complex: pairwise contacts



▷Ref: Blobel et al; Nature SMB; 2009

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



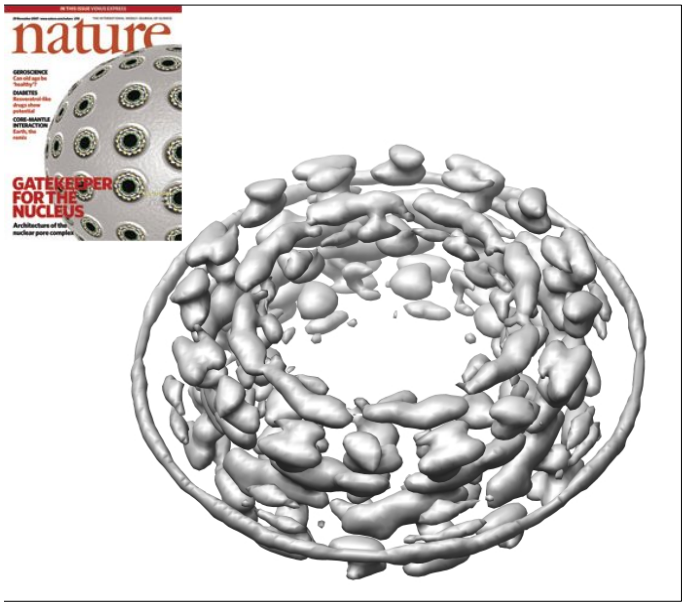


▷Ref: Seo et al; PNAS; 2009

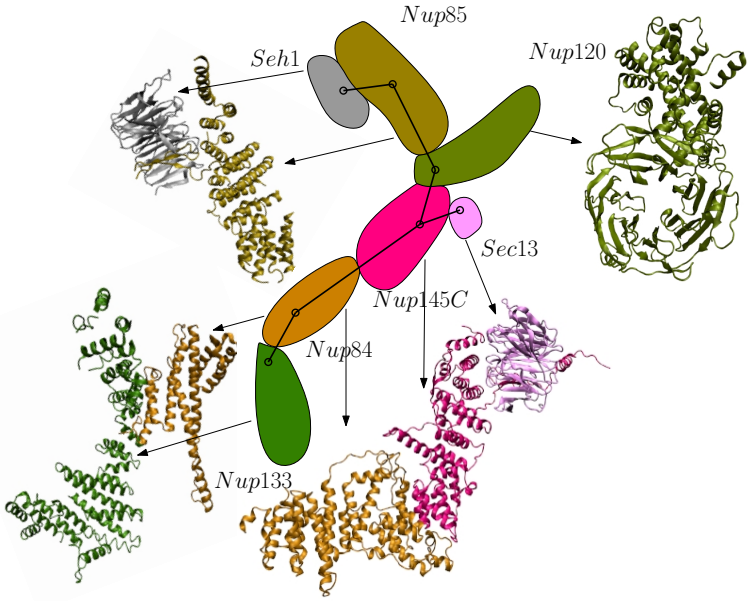
▷Ref: Brohawn, Schwarz; Nature MSB; 2009

⇒ Bridging the gap between both classes of models?

Reconstruction of large assemblies: global - qualitative models versus local - atomic-resolution models



Alber et al; Nature; 450; 2007



Blobel et al; Nature SMB; 2009

Modeling Large Protein Assemblies

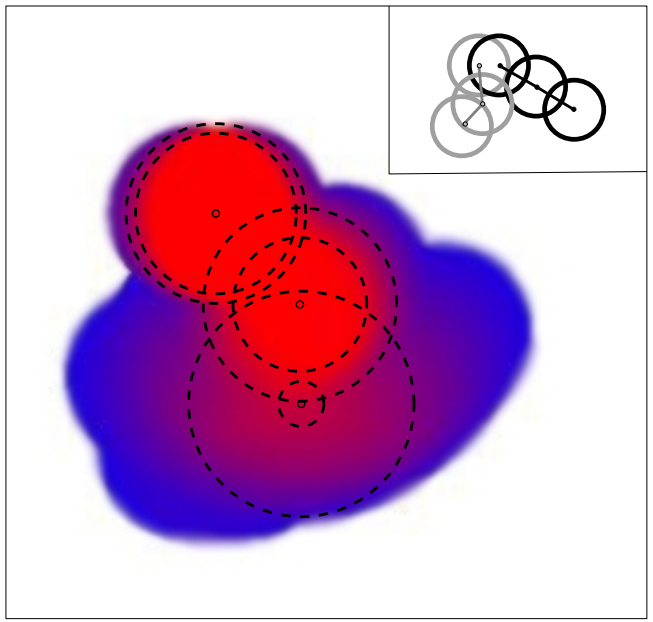
Voronoi Diagrams Again

Reconstruction by Data Integration

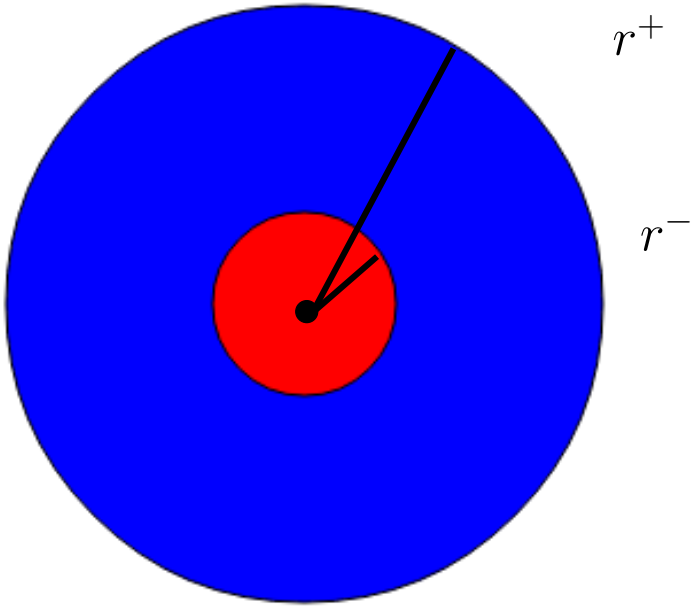
Toleranced Models

Assessing the Reconstruction of Fuzzy Models
- Contact probabilities
- Isolated copies
- Pairwise contacts

Building toleranced models (Embracing the geometric noise.)



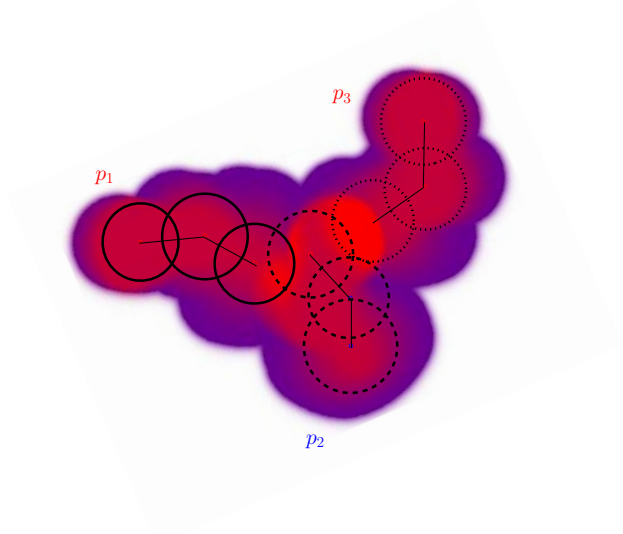
A Toleranced Ball



VIDEO/tol-ball-animation.html

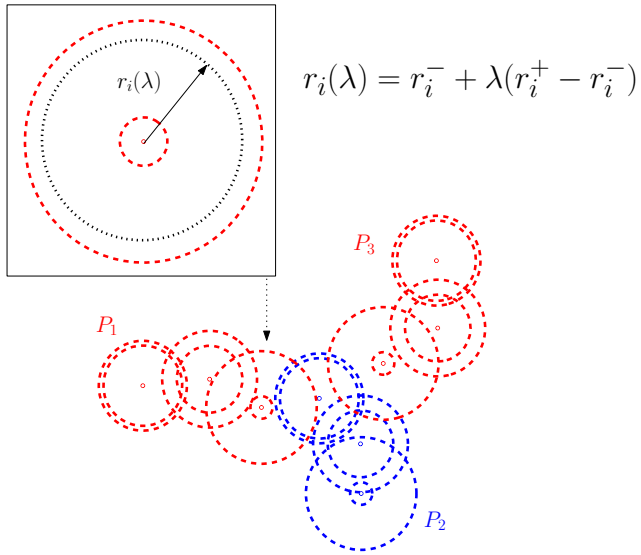
Uncertain Data and Toleranced Models: the Example of Molecular Probability Density Maps

- ▷ 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



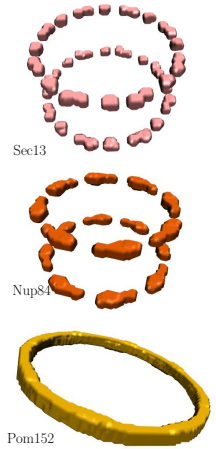
- ▷ Space-filling diagram $\mathcal{F}_\lambda$: a continuum of models
  - Radius interpolation: $r_i(\lambda) = r_i^- + \lambda(r_i^+ - r_i^-)$

- ▷ Multiplicative weights required

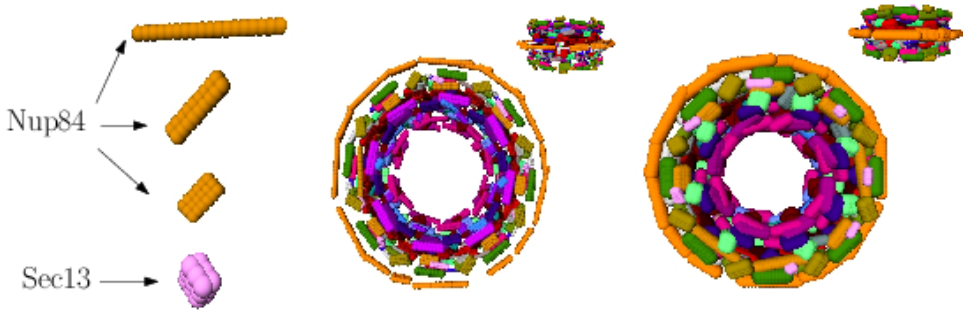
▷Ref: Cazals, Dreyfus; Symp. Geom. Processing; 2010

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



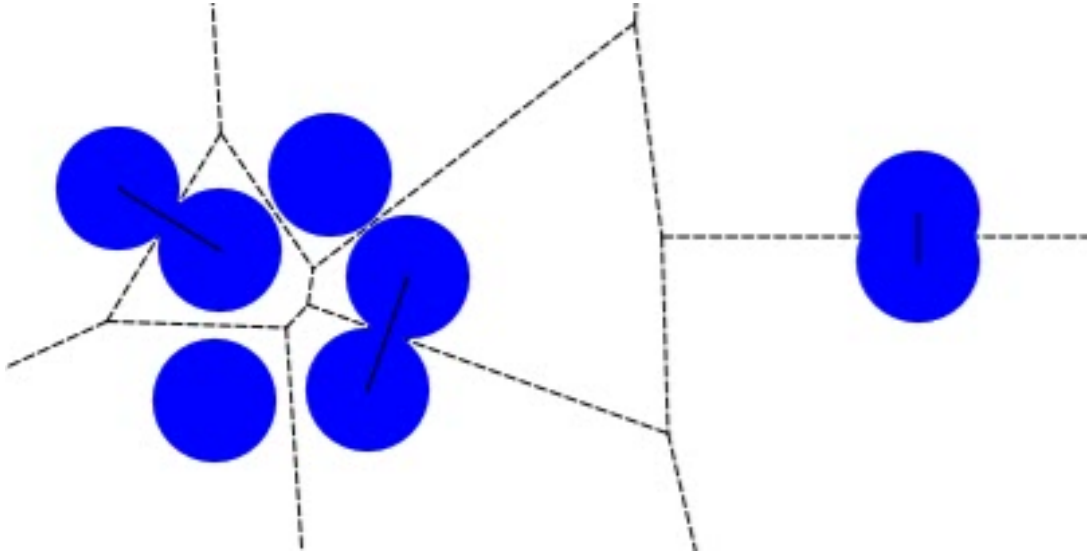
(i) Canonical shapes



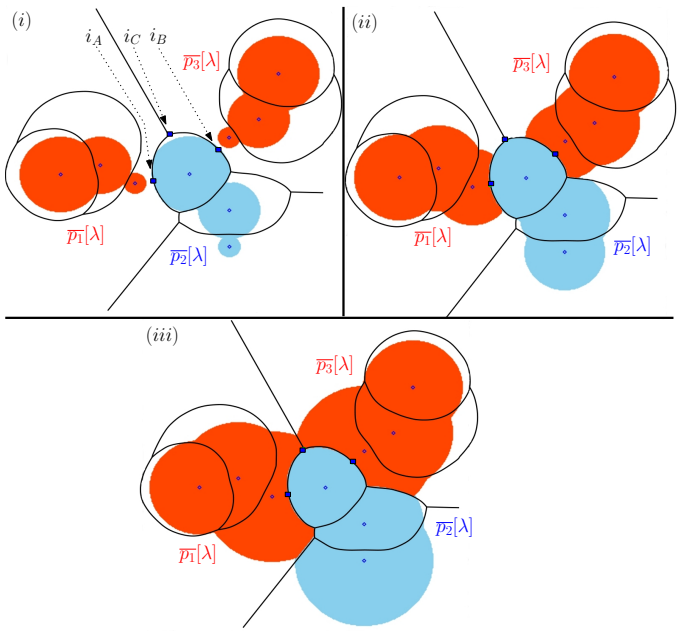
(ii) NPC at $\lambda = 0$ (iii) NPC at $\lambda = 1$

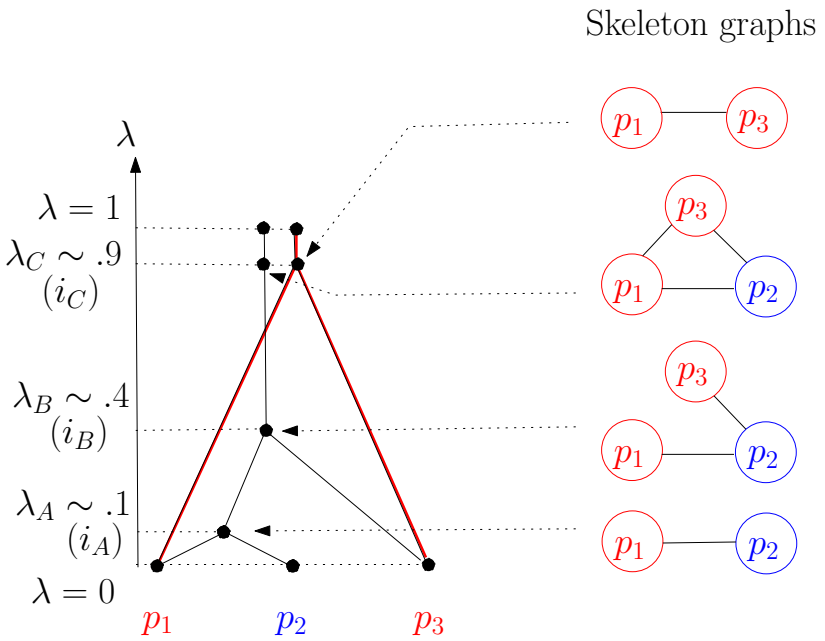
Growing toleranced models and enumerating their finite set of topologies (Spotting stable structures.)

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



Multi-scale Analysis of Toleranced Models: Finite Set of Topologies and Hasse Diagram

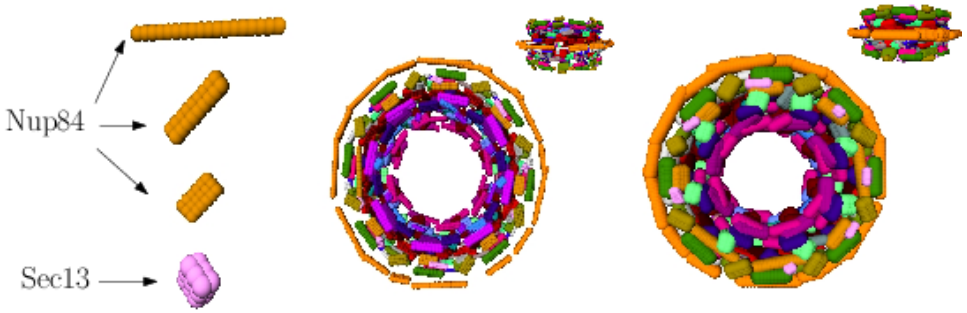




- Red-blue bicolor setting: red proteins are types singled out (e.g. TAP)
- Complexes and skeleton graphs: Hasse diagram
- Finite set of topologies: encoded into a Hasse diagram
  - Birth and death of a complex
  - Topological stability of a complex $s(c) = \lambda_d(C) - \lambda_b(C)$
- Computation: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

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 Large Protein Assemblies

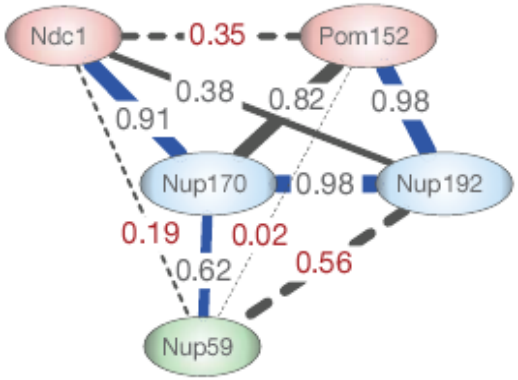
Voronoi Diagrams Again

Reconstruction by Data Integration

Toleranced Models

Assessing the Reconstruction of Fuzzy Models
- Contact probabilities
- Isolated copies
- Pairwise contacts

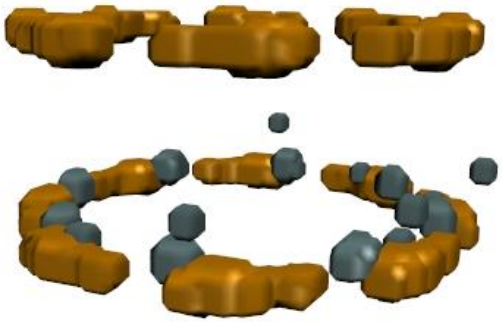
Proeminent contact frequencies out of the $\binom{30}{2} + 30 = 465$ pairs of protein types



- 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

Over- and Under-represented pairs for $a = 0.1$ and $b = 0.9$

▷ Over-represented pair:
$Nup84 - Nup60$:
$f_{ij} = 0.07, p_{ij}^{(4)} = p_{ij}^{(1)} = 1$



▷ Under-represented pair:
$Nup192 - Pom152 : f_{ij} = 0.98, p_{ij}^{(1)} = 0$

Contact	f_{ij}	$p_{ij}^{(1)}$	λ_{max}
Nup59 Nup59	0	1	0
Pom34 Pom34	0.02	1	0
Nsp1 Nsp1	0.02	1	0
Nup60 Nup145N	0.03	1	0
Nup60 Pom34	0.03	1	0
Nup145N Nup49	0.04	1	0
Nup1 Nup145N	0.05	1	0
Nup60 Ndc1	0.06	1	0
Nup84 Nup60	0.07	1	0
Nsp1 Nup145N	0.07	1	0
Nup145C Nup60	0.08	1	0
Sec13 Nup159	0.08	1	0
Nsp1 Nup60	0.08	1	0
Nup49 Nup116	0.08	1	0
Nup57 Nup145N	0.08	1	0
Nsp1 Nup42	0.09	1	0
Nup60 Nup59	0.09	1	0
Nup42 Nup116	0.09	1	0
Nup57 Nup116	0.09	1	0
Sec13 Nup145N	0.1	1	0
Nup59 Pom34	0.03	0.9	0.15
Seh1 Nup60	0.06	0.9	0.18
Gle2 Nup57	0.08	0.9	0.21

Contacts	f_{ij}	$p_{ij}^{(1)}$	λ_{max}
Nup192 Pom152	0.98	0	1
Nup170 Ndc1	0.91	0.1	0.35
Nup188 Nic96	1	0.1	0.32
Pom152 Pom34	1	0.1	0.28

Contact Probabilities: Sharpening the Contact Frequencies

- Toleranced model (TM) is a continuum:
  $\longrightarrow$ contact probability analogous to frequency

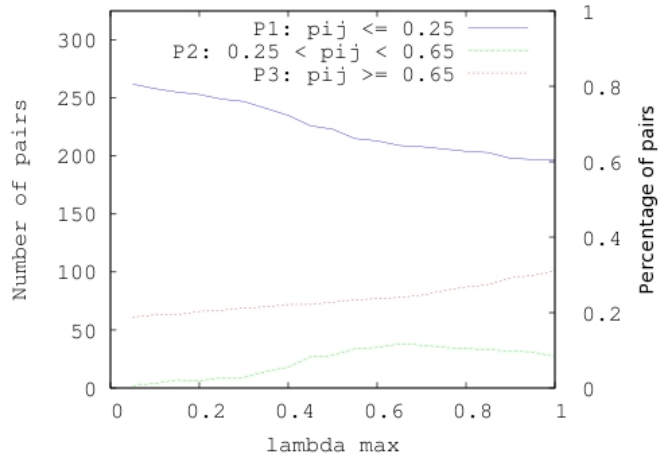
- Contact probability for types p_i, p_j and stoichio. k:
  - If k contacts at $\lambda(p_i, p_j)$
    contact probability: $p_{ij}^{(k)} = 1 - \lambda(p_i, p_j)/\lambda_{\max}$
  - Else i.e. strictly less than k contacts: $p_{ij}^{(k)} = 0$
  - Note: $p_{ij}^{(k)}$ strictly increasing with $\lambda_{\max}$

- Partitioning of all pairs into 3 classes:
  $P_1^{(k)} : p_{ij}^{(k)} \leq a; P_2^{(k)} : a < p_{ij}^{(k)} < b; P_3^{(k)} : b \leq p_{ij}^{(k)}$

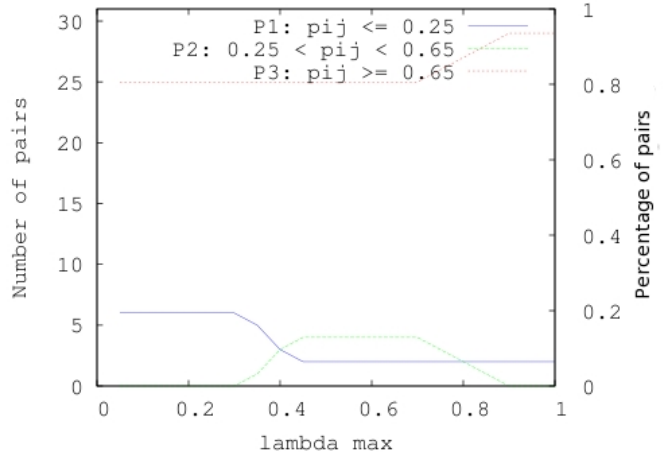
- Over-represented pairs in the TM:
  $(p_i, p_j) \in F_1$ but $\in P_3^{(1)}$

- Under-represented pairs in the TM:
  $(p_i, p_j) \in F_3$ but $\in P_1^{(1)}$

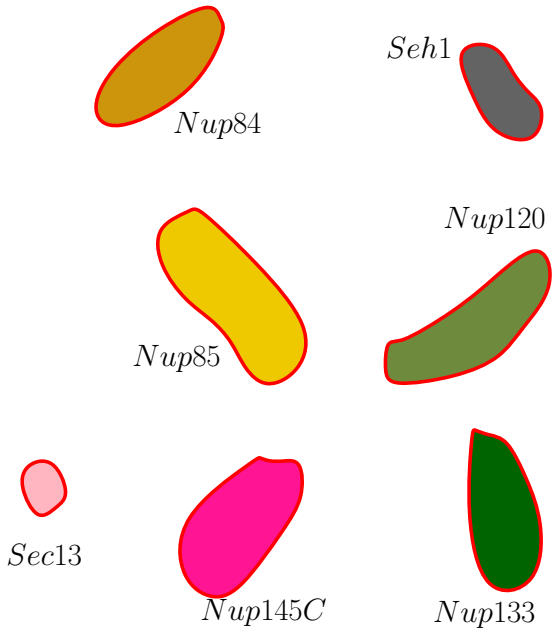
- Pairs in F_1 vs $P_i^{(1)}$:



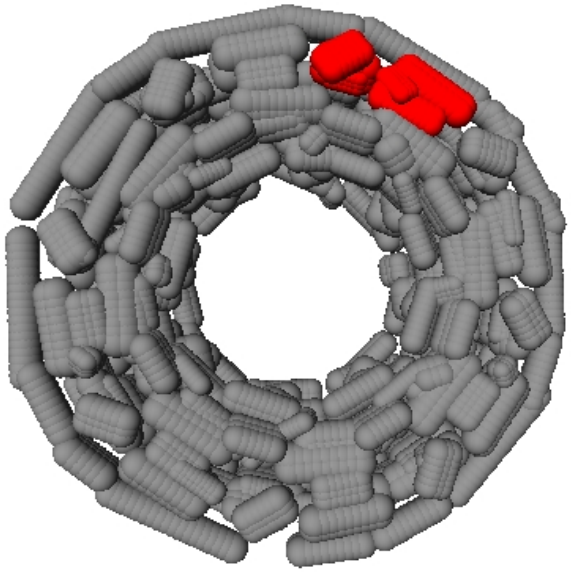
- Pairs in F_3 vs $P_i^{(1)}$:



Assessing a toleranced model w.r.t. a set of protein types



Y-complex : protein types

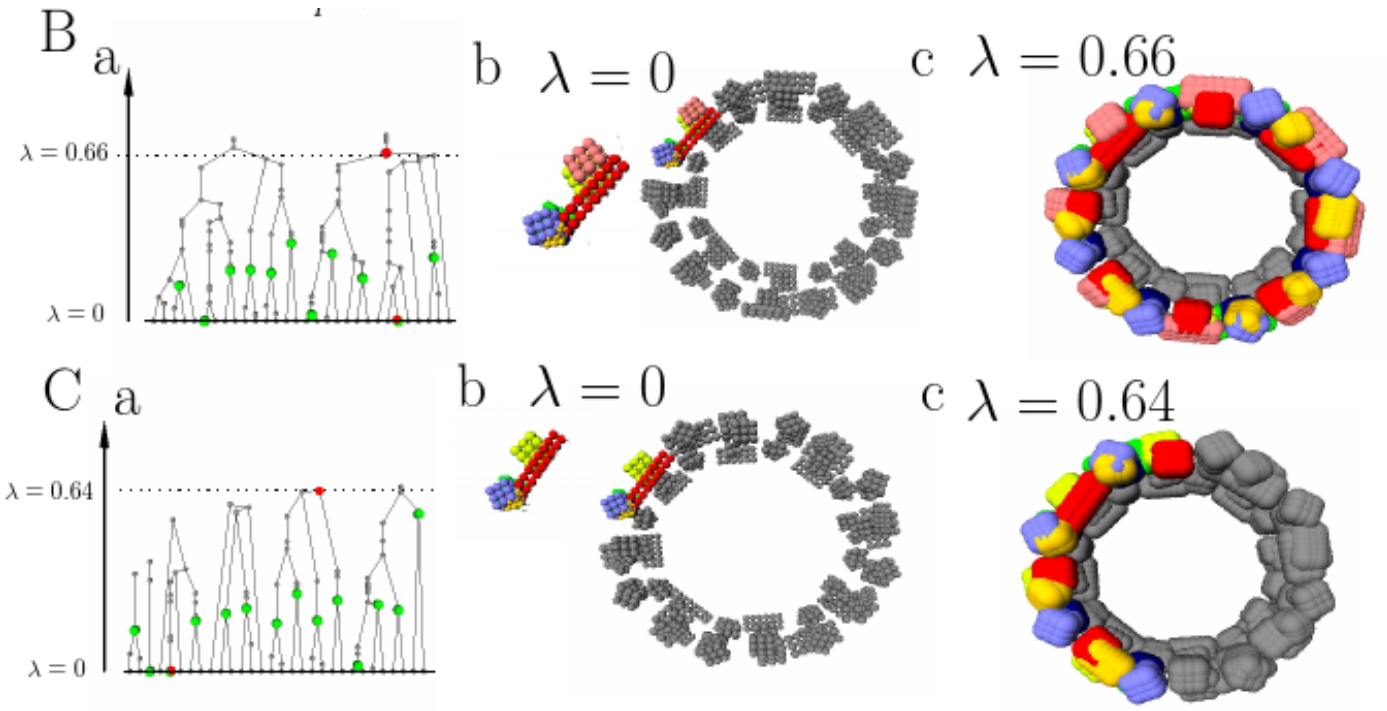


Y-complex : instance

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

Geometry, Topology, Biochemistry

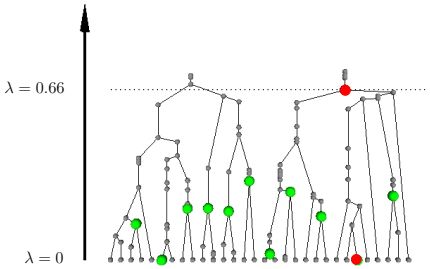
- Input:
  - Toleranced model
  - T: set of proteins types, the red proteins (TAP, types involved in sub-complex)
- Output, overall assembly:
  - number of isolated copies: symmetry analysis
  - their topological stability: death date - birth date (cf α-shape demo)



- 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

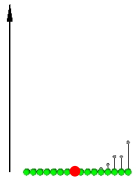
9 (Nup133, Nup85) $\lambda \in [0.09, 0.70]$
5 (Nup84, Nup85) $\lambda \in [0.52, 0.69]$
1 (Nup133, Nup120) $\lambda = 0$
1 (Nup84, Nup120) $\lambda = 0.06$

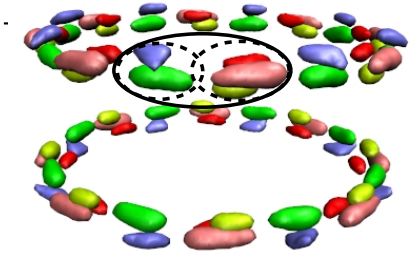
Nup85 involved in 14 / 16 contacts

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

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

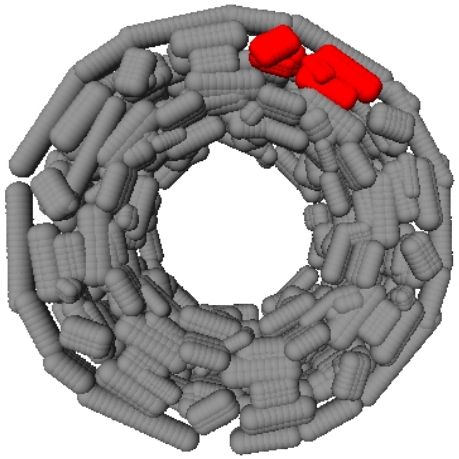
(Nup120, Nup85, Seh1)



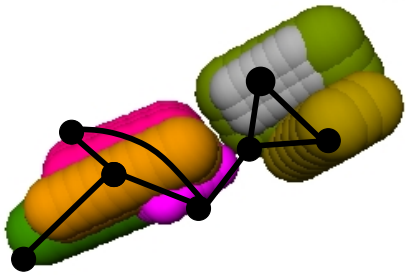


(Nup84, Nup145C, Nup133)

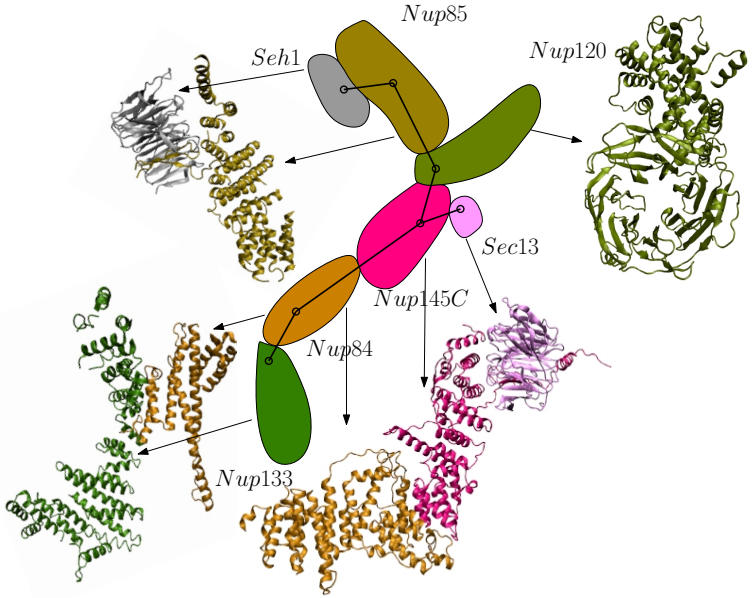
Assessing a toleranced model w.r.t a high-resolution structural model



Assembly



Complex: skeleton graph



Template: skeleton graph

Comparing a Skeleton Graph against a Template: Matchings

Depleted matchings:

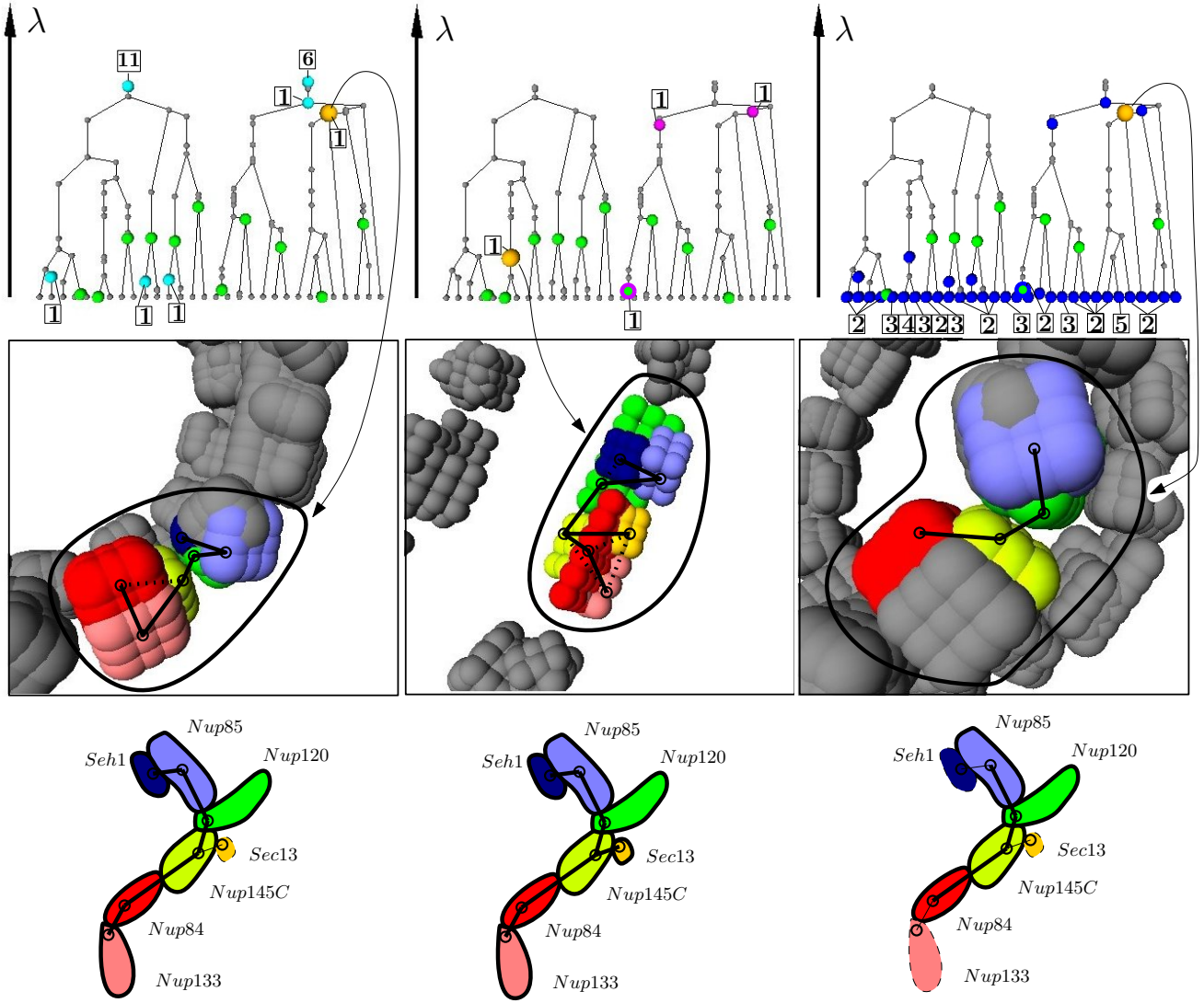
- missing nodes...
- problems on edges

Complete matchings:

- all nodes...
- problems on edges

Exact matchings:

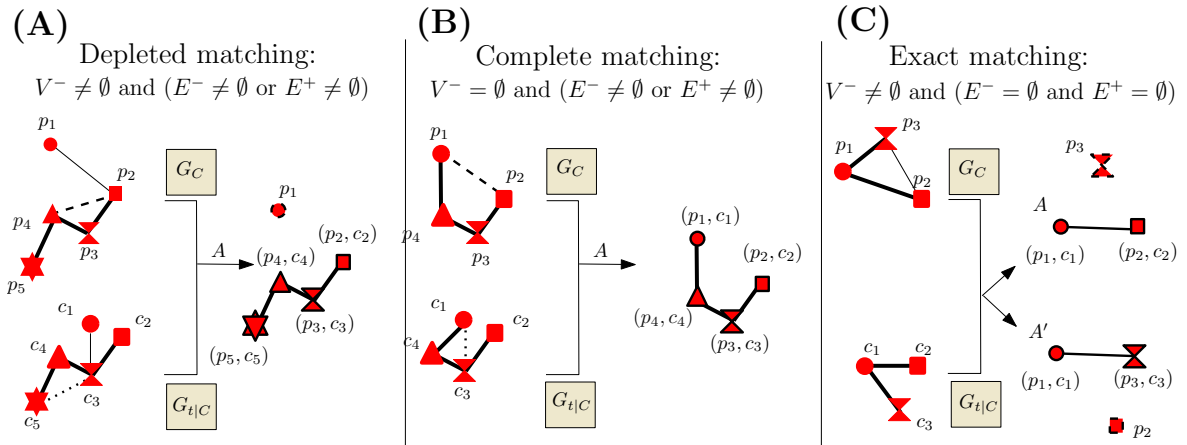
- missing nodes...
- but edges perfect for these nodes



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

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

- Input: two skeleton graphs
  - template G_t, the red proteins : contacts within an atomic resolution model
  - complex G_C: skeleton graph of a complex of a node of the Hasse diagram
- Output: graph comparison, complex G_C versus template G_t: (common/missing/extra) $\times$ (proteins/contacts)



Signature \ Matching		A
$V^\sim$	Matching Protein Types	(p_2, c_2) (p_3, c_3) (p_4, c_4) (p_5, c_5)
V^-	Missing Protein Types	c_1
$E^\sim$	Matching Contacts	$(c_2, c_3) \leftrightarrow (p_2, p_3)$ $(c_3, c_4) \leftrightarrow (p_3, p_4)$ $(c_4, c_5) \leftrightarrow (p_4, p_5)$
E^-	Missing Contacts	(p_2, p_4)
E^+	Extra Contacts	(c_3, c_5)

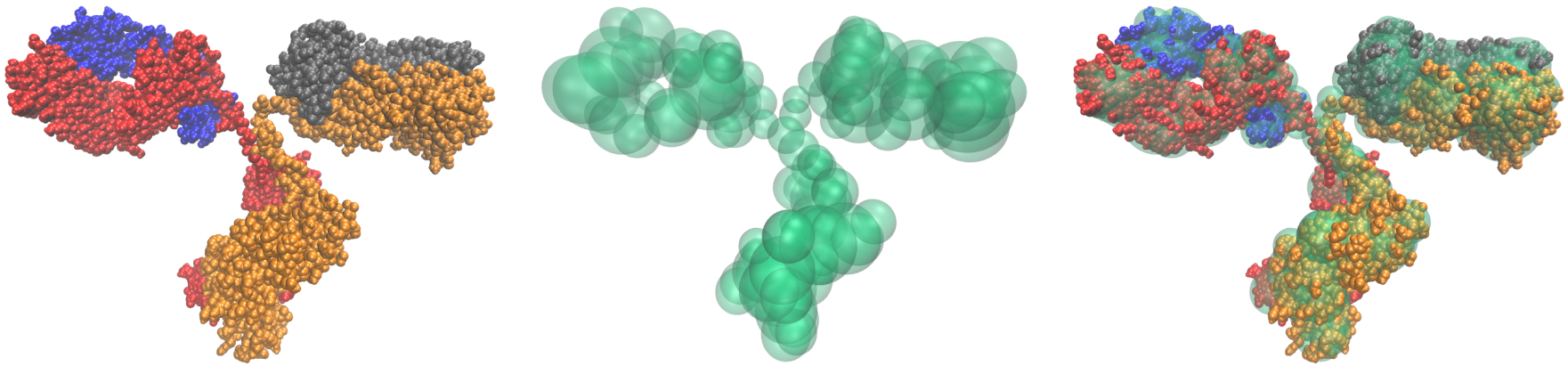
Signature \ Matching		A
$V^\sim$	Matching Protein Types	(p_1, c_1) (p_2, c_2) (p_3, c_3) (p_4, c_4)
V^-	Missing Protein Types	$\emptyset$
$E^\sim$	Matching Contacts	$(c_2, c_3) \leftrightarrow (p_2, p_3)$ $(c_3, c_4) \leftrightarrow (p_3, p_4)$ $(c_1, c_4) \leftrightarrow (p_1, p_4)$
E^-	Missing Contacts	(c_1, c_3)
E^+	Extra Contacts	(p_1, p_2)

Signature \ Matchings		A	A'
$V^\sim$	Matching Protein Types	(p_1, c_1) (p_2, c_2)	(p_1, c_1) (p_3, c_3)
V^-	Missing Protein Types	c_3	c_2
$E^\sim$	Matching Contacts	(c_1, c_2) $(p_1 \leftrightarrow p_2)$	(c_1, c_3) $(p_1 \leftrightarrow p_3)$
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

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
  - Morse theoretical analysis of density maps
  - Geometric max-k-cover

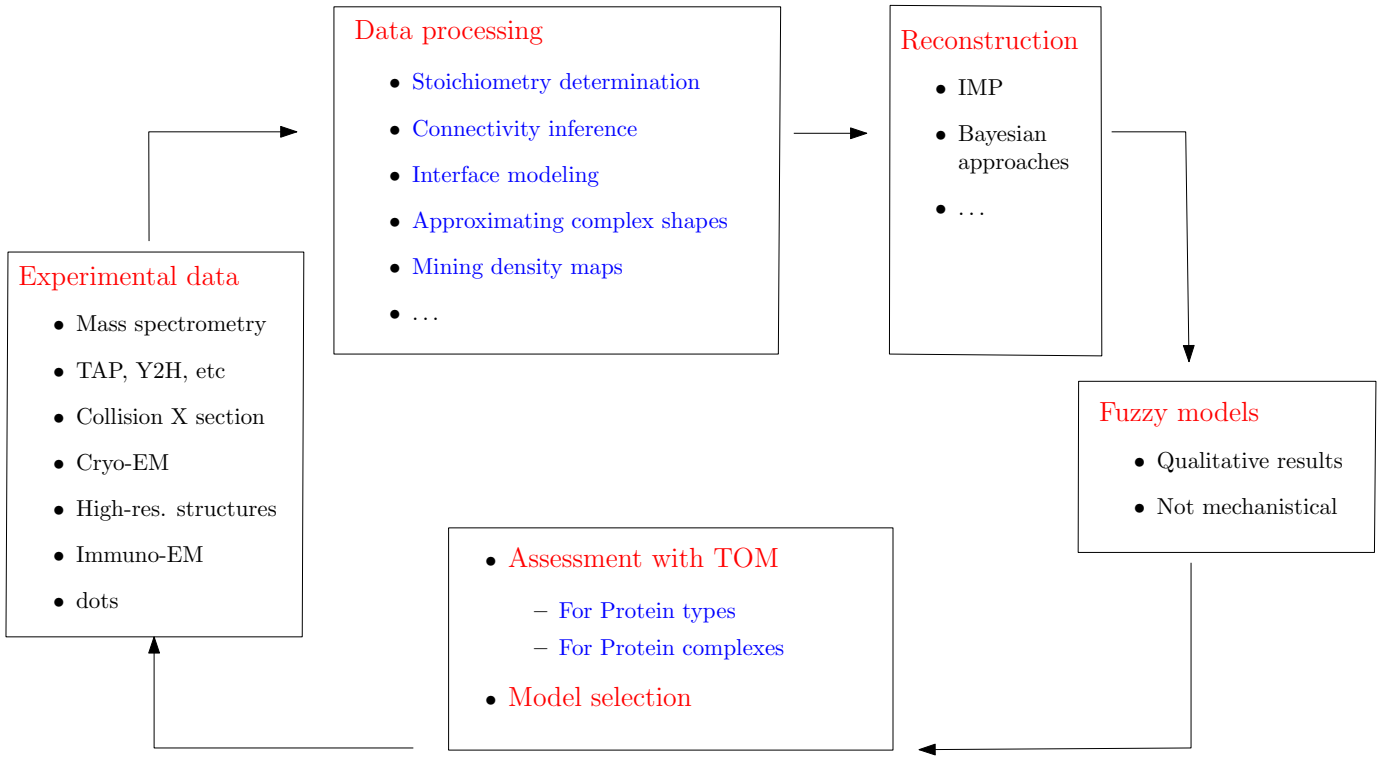


▷Ref: F. Cazals and T. Dreyfus and S. Sachdeva and N. Shah; About to be submitted

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, applications
  : Proteins 2012, Submission 2012

  Toleranced Models, theory
  : Symp. on Geometry Processing 2010

  Collections of balls
  : ACM Trans. on Math. Soft. 2010, ACM IEEE Trans. CBB 2011

  Graphs
  : Theoretical Computer Science 2005 + 2008

  Mass spectrometry
  : Submissions 2012

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