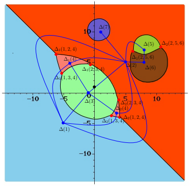
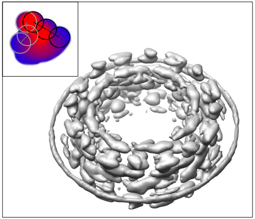
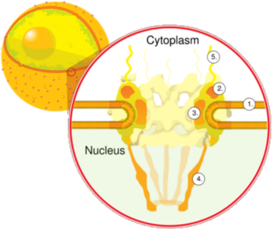
# Balls, sticks, triangles and molecules

Frederic.Cazals@sophia.inria.fr

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

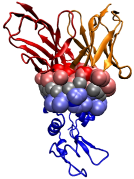


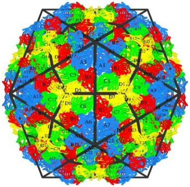




# Structure to Function:
# Challenges in Structural Bioinformatics

▷ Protein complexes are ubiquitous





Stability and specificity
of macro-molecular complexes?

Prediction ?
(with little/no structural information)

▷ Structural information is scarce

$\#$ non redundant sequences $\sim 100$ $\#$ structures

▷ Computer science perspective: improving the prediction of complexes

– How does bio-physics constrain macro-molecular geometry?

– How does one integrate suitable parameters into learning procedures?

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

# Why should we get involved?

- ▷ Computational Structural Biology, key features
  - $O(10^8)$ (unique) genes $\gg O(10^6)$ structures $\gg O(10^3)$ biological complexes
  - Known structures are mainly static...
    but the entropic contribution to the free energy if often key
  - Size of large molecular machines : up to millions of atoms
  - Experimental insights : a zoo of experimental techniques

- ▷ Physics versus geometry
  - Physical model are mainly borrowed from Newtonian mechanics:
    balls, sticks - springs

- ▷ Contributions from a Computer Scientist
  - Go faster – be more accurate
    Joint work with S. Loriot, M. Teillaud, S. Sachdeva
  - Think differently
    Joint work with R. Gruenberg, J. Janin, C. Prevost
  - Change the (modeling) paradigm
    Joint work with T. Dreyfus

# Why should we get involved?

Go faster – be more accurate

Think differently

Change the (modeling) paradigm

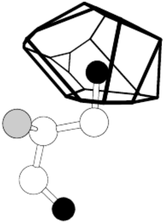
# On the Volume of Union of Balls (Algorithms)

▷ Context: discriminating native vs non-native states

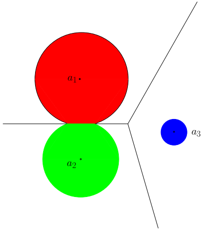
– Describing the packing properties of atoms : surfaces and volumes

– 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 (Algorithms)

▷ Strategy developed: certified volume calculation

– Proved a simple formula for computing the volume of a restriction

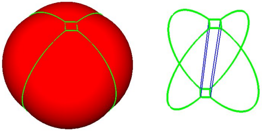
– Analyzed the predicates and constructions involved

– 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; Submitted

# Why should we get involved?

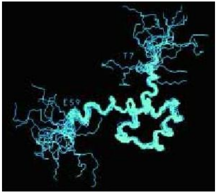
Go faster – be more accurate

Think differently

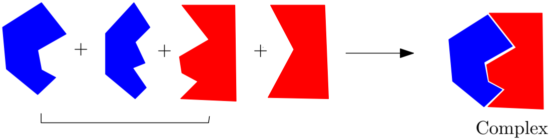
Change the (modeling) paradigm

# Conformer Selection for Docking (Proof-of-concept)

- Context: mean-field theory based docking algorithms
  - Select a diverse subset of *s* conformers out of a pool of *n* conformers

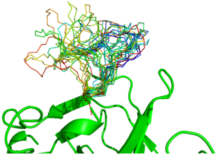


Conformer selection, Monod-Wyman-Changeux, 1965

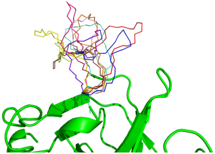


- STAR: RMSD-based or energy based conformer selection strategies

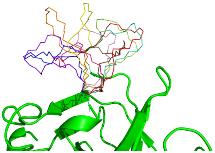
- Conformational diversity: RMSD vs geometric optimization



*n* conformers
pool to choose from



10 conformers:
**diverse** selection



10 conformers:
**redundant** selection

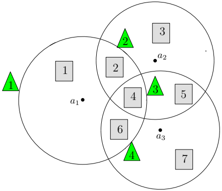
# Conformer Selection for Docking Cont'd (Proof-of-concept)

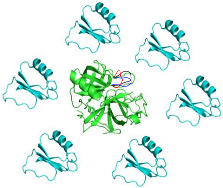
▷ Strategy developed: shape matters

– Choose the selection occupying the biggest possible volume
–                                   exposing the largest possible surface area

▷ Contributions

– Geometric versions of max-k-cover (NP-complete) + greedy strategy
– Computation of cell decompositions to run the optimizations
– Coarse-grain docking validations





▷ Assessment

– Significant improvement for geometric and topological diversity
– Moderate for coarse-grain docking

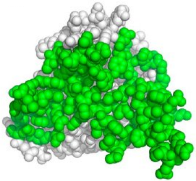
▷Ref: Cazals, Loriot; CGTA 42; 2009

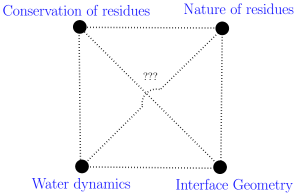
▷Ref: Cazals, Loriot, Machado, Teillaud; CGTA 42; 2009

▷Ref: Loriot, Sachdeva, Bastard, Prevost, Cazals; ACM TCBB; 2011

# Mining Protein - Protein Interfaces (Structural studies)

▷ Context: key interface residues; key properties / correlations?





▷ STAR

| | |
|---|---|
| Energy | Directed mutagenesis / point-wise $\Delta\Delta G$; incomplete<br>Free energy calculations; biological time scale beyond reach |
| Evolution | Conserved residues;<br>may not apply, database dependent,conserved res. not at interface |
| Structure | Shape, size, position of atoms; some general facts |

▷Ref: Bahadur, Chakrabarti, Rodier, Janin; JMB 336; 2004

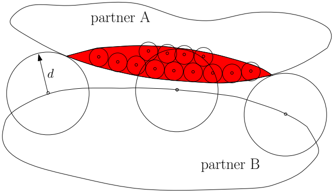
▷Ref: Reichmann et al.; PNAS 102; 2005

▷Ref: Guharoy, Chakrabarti; PNAS 102; 2005

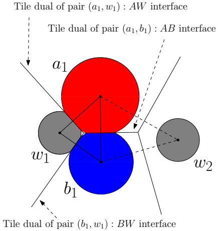
▷Ref: Mihalek, Lichtarge; JMB 369; 2007

# About Interface Models

▷ Distance threshold (geometric footprint)



▷ Contacts between Voronoi restrictions


Tile dual of pair $(a_1, w_1)$ : $AW$ interface
Tile dual of pair $(a_1, b_1)$ : $AB$ interface
$a_1$
$w_1$
$w_2$
$b_1$
Tile dual of pair $(b_1, w_1)$ : $BW$ interface

▷ The Voronoi interface model

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

▷ More applications

- – Shelling and depth orders
- – Discrete level sets, contour tree, partial shape matching

▷Ref: Cazals; Conf. on Pattern Recognition in Bioinformatics; 2010

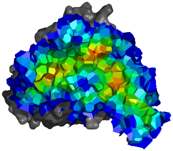
# Mining Protein - Protein Interfaces Cont'd (Structural Studies)

▷ Strategy developed: discrete interface parameterization

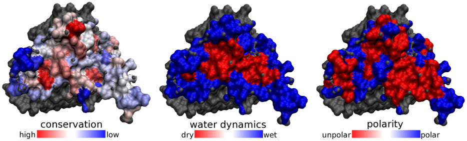
– **V**oronoi **S**helling **O**rder: interface partitioning into concentric shells

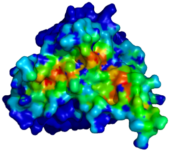
– Integer valued depth of atoms at interface (vs core - rim)

– Statistics (P-values, Fisher meta analysis) for various correlations



▷ Conservation vs dryness vs polarity





▷ Assessment: statements from global → per-complex

– depth and water dynamics: significant **per-complex**

– conservation vs core/rim: **global trend**

– polarity and depth : **global trend**

▷Ref: Cazals, Proust, Bahadur, Janin; Protein Science 15; 2006

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

# Why should we get involved?

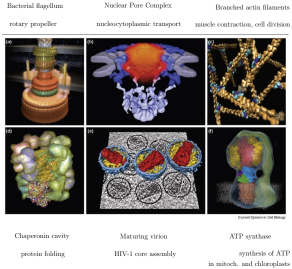
Go faster – be more accurate

Think differently

Change the (modeling) paradigm

# Structural Dynamics of Macromolecular Processes

## Reconstructing Large Macro-molecular Assemblies



Bacterial flagellum

rotary propeller

Nuclear Pore Complex

nucleocytoplasmic transport

Branched actin filaments

muscle contraction, cell division

Current Opinion in Cell Biology

Chaperonin cavity

protein folding

Maturing virion

HIV-1 core assembly

ATP synthase

synthesis of ATP
in mitoch. and chloroplasts

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

▷ Difficulties

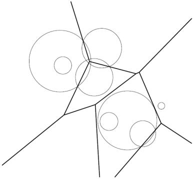
Modularity
Flexibility

▷ Core questions

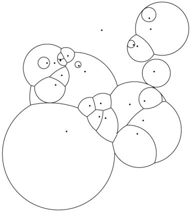
Reconstruction / animation
Integration of (various) experimental data
Coherence model vs experimental data

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

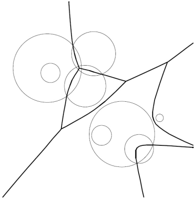
# The Zoo of curved Voronoi diagrams



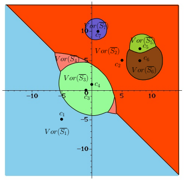
▷ Power diagram:
$d(S(c,r),p) = \|c-p\|^2 - r^2$



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

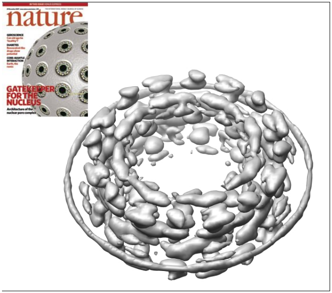


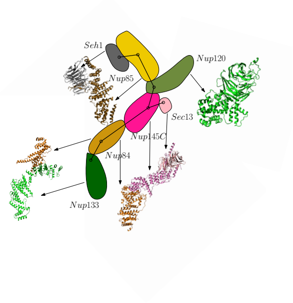
▷ Apollonius diagram:
$d(S(c,r),p) = \|c-p\| - r$



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

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


nature
GATEKEEPER FOR THE NUCLEUS
Architecture of the nuclear pore complex



Alber et al; Nature; 450; 2007

Blobel et al; Nature SMB; 2009

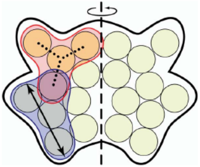
# Reconstructing Large Assemblies: a NMR-like Data Integration Process

▷ 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,. . . )



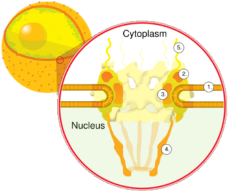
▷ Restraints, experimental data and . . . ambiguities:

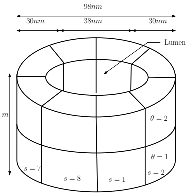
| | | | |
|---|---|---|---|
| Assembly | : shape | cryo-EM | fuzzy envelopes |
| Assembly | : symmetry | cryo-EM | idem |
| Complexes: | : interactions | TAP (Y2H, overlay assays) | stoichiometry |
| Instance: | : shape | Ultra-centrifugation | rough shape (ellipsoids) |
| Instances: | : locations |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

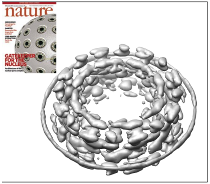
▷`Ref: Alber et al, Ann. Rev. Biochem. 2008 + Structure 2005`

# 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

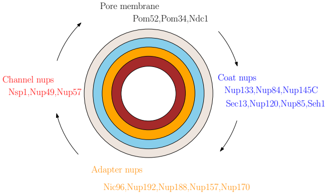
▷ Qualitative results:

*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`

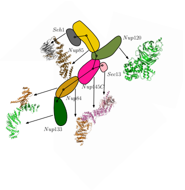
# Putative Models of Sub-complexes: the Y-complex

▷ Symmetric core of the NPC



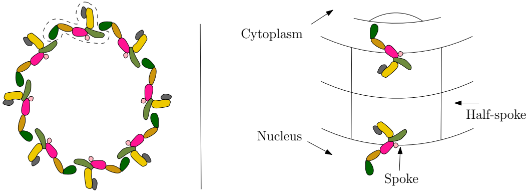
▷Ref: Blobel et al; Cell; 2007

▷ The Y-complex: pairwise contacts



▷Ref: Blobel et al; Nature SMB; 2009

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



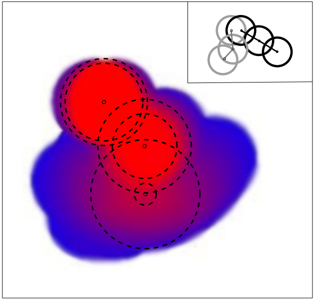
▷Ref: Seo et al; PNAS; 2009

▷Ref: Brohawn, Schwarz; Nature MSB; 2009

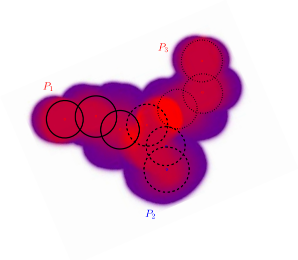
⇒ Bridging the gap between both classes of models?

Prologue; I; II; III-a; III-b; Epilogue

# Building toleranced models (Embracing the geometric noise.)



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



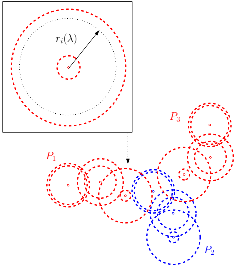
▷ Probability Density Map of a Flexible Complex:

– 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

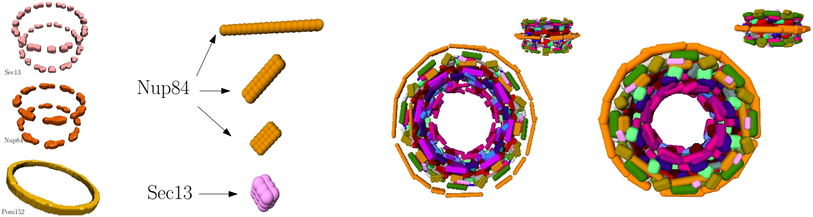
▷Ref: Cazals, Dreyfus; Symp. Geom. Processing; 2010

# Toleranced Models for the NPC

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

  → assign protein instances to pronounced local maxima of the maps
- ▷ Geometry of instances:

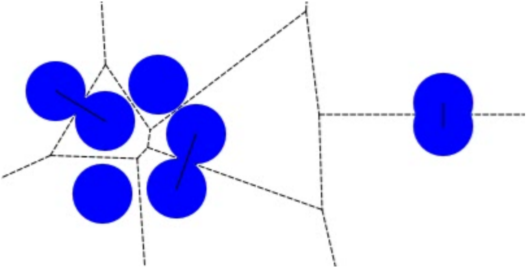
  four canonical shapes. . .



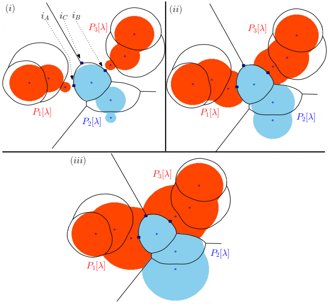
(i) Canonical shapes (ii) NPC at $\lambda = 0$ (iii) NPC at $\lambda = 1$

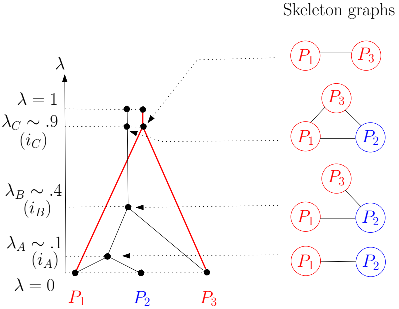
# Growing toleranced models and enumerating their finite set of topologies (Spotting stable structures.)

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



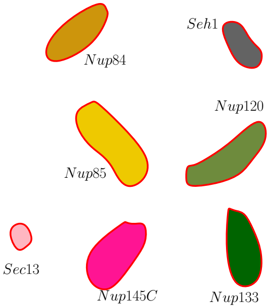
# Multi-scale Analysis of Toleranced Models: Finite Set of Topologies and Hasse Diagram



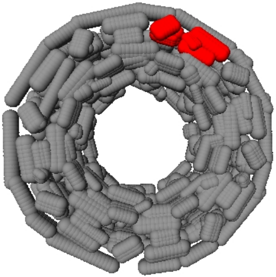


- ▷ Red-blue bicolor setting: red proteins are types singled out (e.g. TAP)
- ▷ Complexes and skeleton graphs: Hasse diagram
- ▷ Finite set of topologies: encoded into a Hasse diagram
  - – Birth and death of a complex
  - – Topological stability of a complex $s(c) = \lambda_d(C) - \lambda_b(C)$
- ▷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

# ASSESSING A TOLERANCED MODEL W.R.T. A SET OF PROTEIN TYPES



$Y$-complex : protein types



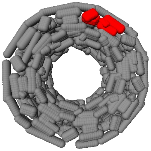
$Y$-complex : instance

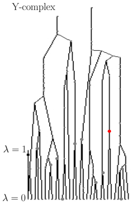
# Assessment w.r.t. a Set of Protein Types: Geometry, Topology, Biochemistry

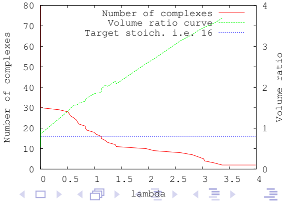
- ▷ Input:
  - – Toleranced model
  - – *T*: set of proteins types, the red proteins (TAP, types involved in sub-complex)
- ▷ Output, overall assembly:
  - – Geometry - biochemistry:

    number of copies – symmetry analysis

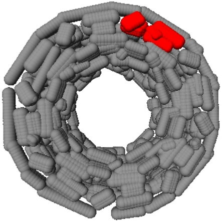
    TAP data: complex or mixture?
  - – Topological stability: death date - birth date (cf $\alpha$-shape demo)
- ▷ Output, per complex:
  - – Biochemistry: stoichiometry of protein instances
  - – Geometry: volume occupied vs. expected volume



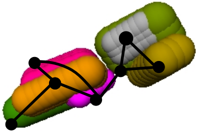




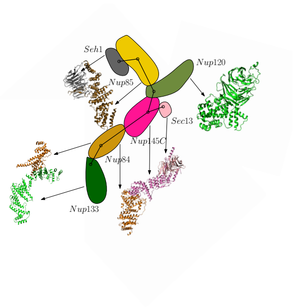
# Assessing a toleranced model w.r.t a high-resolution structural model



Assembly



Complex: skeleton graph



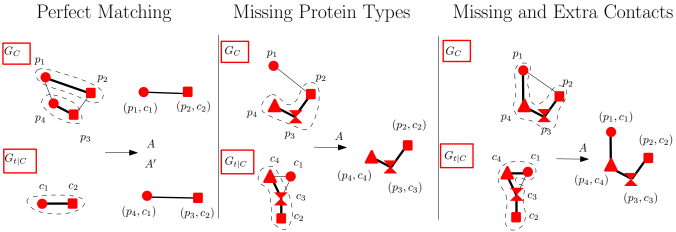
Template: skeleton graph

# 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) × (proteins/contacts)



Perfect Matching

Missing Protein Types

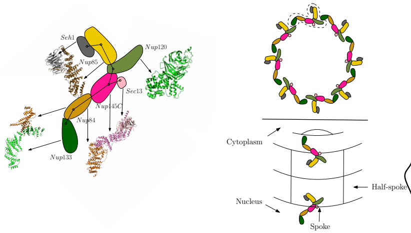
Missing and Extra Contacts

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

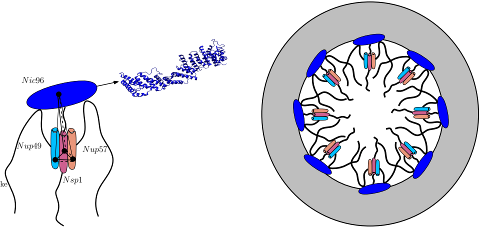
▷Ref: Koch; Theoretical Computer Science; 250 (1-2), 2001

# Prologue; I; II; III-a; III-b; Epilogue

## Insights on the NPC...



$Y$-complex

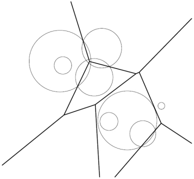


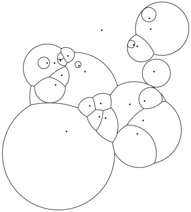
$T$-complex

# CW Voronoi : algorithms

▷Ref: Cazals, Dreyfus; SGP; 2010

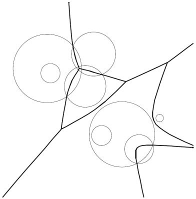
# The Zoo of curved Voronoi diagrams

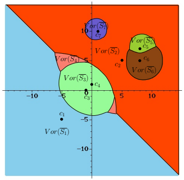




▷ Power diagram:
$d(S(c,r),p) = \|c-p\|^2 - r^2$

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





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

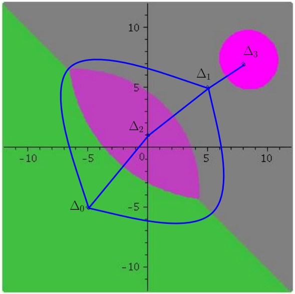
▷ Compoundly Weighted Voronoi diagram:
$d(S(c,\mu,\alpha),p) = \mu\|c-p\| - \alpha$

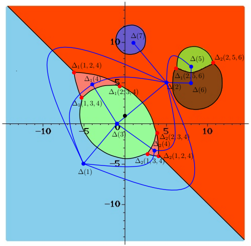
# Voronoi Diagram : Topological Complications

- ▷ Partition of the space:

$$Vor(\overline{S_i}) = \{p \in \mathbb{R}^3 / \lambda(\overline{S_i}, p) \leq \lambda(\overline{S_j}, p)\}$$

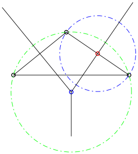
- ▷ Voronoi region in generality:
  - Neither connected : collection of faces
  - Nor simply connected
- ▷ Dual complex:
  - Apollonius complication:
    Lens sand-witched region.
    Exple (**Top**): $\Delta_1(0,1,2)$ and $\Delta_2(0,1,2)$
  - CW Diagram complications:
    Edges without triangles.
    Exple (**Top**): $\Delta(1,3)$
    $\neq$ triangles that share the same edges.
    Exple (**Bottom**): $\Delta_1(1,4,5)$ and $\Delta_2(1,4,5)$





# Toleranced Tangent and Conflict Free Balls

▷ Rationale. Delaunay triangulation:



- Conflict Free ball
- Smallest Circumscribed ball
  empty: Gabriel simplex

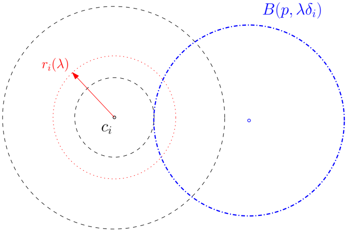
▷ Generalization to the CW case:

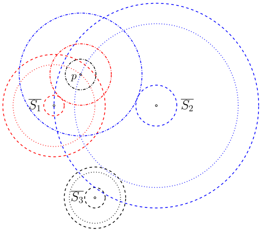
- Toleranced tangent ball $B(p, \lambda)$:

$$|| \, pc_i \, || - r_i^- - \lambda\delta_i = 0. \qquad (1)$$

- Conflict Free ball $B(p, \lambda)$:

$$|| \, pc_i \, || - r_i^- - \lambda\delta_i > 0. \qquad (2)$$





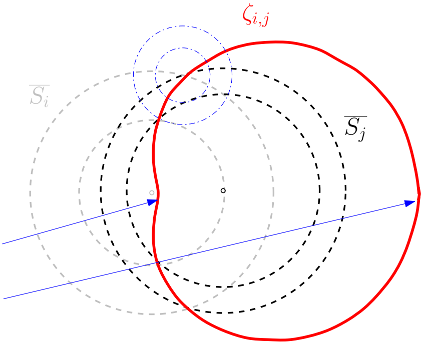
▷ Remark: Conditions (1) and (2) are parametrized by $\delta_i$

# Bisector of Two Toleranced Balls

- Bisector $\zeta_{i,j}$: set of centers of balls toleranced tangent to $\overline{S_i}$ and $\overline{S_j}$.
- Existence of $\zeta_{i,j}$: $\overline{S_i}$ is trivial wrt $\overline{S_j}$ iff

$$\delta_i \leq \delta_j \quad \text{and} \quad \lambda(\overline{S_j}, c_i) < -\frac{r_i^-}{\delta_i} \tag{3}$$

- Geometry of $\zeta_{i,j}$. Four cases:
  - Apollonius
    Hyperboloid
    Hyperplane
    Half straight line
  - CW Voronoi
    Four degree bounded curve
    $\Rightarrow$ Two extremal Toleranced Tangent balls
    minimal: $\overline{S_i}$ and $\overline{S_j}$ are tangent
    maximal: $\delta_i \leq \delta_j \Rightarrow \overline{S_i}$ included in $\overline{S_j}$

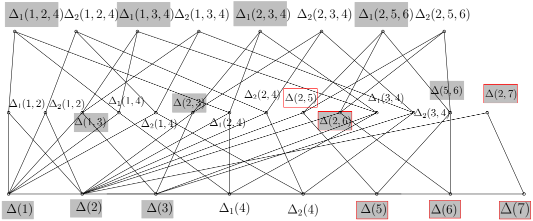


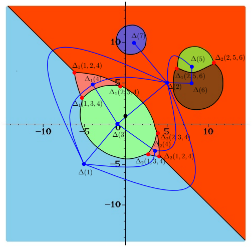
# Representation of the dual as a Hasse diagram

- Focus is on:

  on the intersection between Voronoi regions
  rather than the embedding of the dual

- Several faces for a tuple $T_k(\overline{S_{i_0}}, \ldots, \overline{S_{i_k}})$:
  – $\Delta_1(T_k), \Delta_2(T_k), \ldots$
- Gray box:
  – Smallest Toleranced Tangent ball is Conflict Free
- Red box:
  – Largest Toleranced Tangent ball is Conflict Free





# Classification of simplices in the $\lambda$-complex:

## Two New Cases wrt the Affine Setting

▷ Notations:

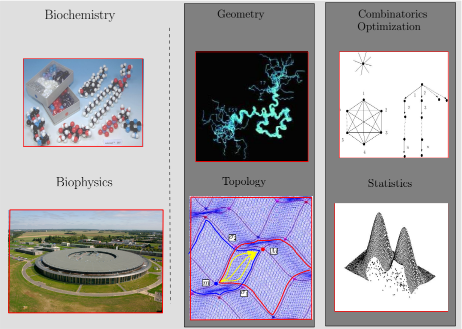
- $\underline{\rho}_{T_k}$: smallest Toleranced Tangent weight
- $\underline{\mu}_{\Delta(T_k)}$: min of $\underline{\rho}_{T_k}$ among co-faces
- $\overline{\mu}_{\Delta(T_k)}$: max of $\underline{\rho}_{T_k}$ among co-faces
- $\overline{\rho}_{T_k}$: largest Toleranced Tangent weight

▷ Classification:

| | Singular | Regular | Interior |
|---|---|---|---|
| $\Delta(T_k) \in \partial(CH(\overline{S}))$,Gabriel, non Dominated | $(\underline{\rho}_{T_k}, \underline{\mu}_{\Delta(T_k)}]$ | $(\underline{\mu}_{\Delta(T_k)}, +\infty]$ | |
| $\Delta(T_k) \in \partial(CH(\overline{S}))$,non Gabriel, non Dominated | | $(\underline{\mu}_{\Delta(T_k)}, +\infty]$ | |
| $\Delta(T_k) \notin \partial(CH(\overline{S}))$, Gabriel, non Dominated | $(\underline{\rho}_{T_k}, \underline{\mu}_{\Delta(T_k)}]$ | $(\underline{\mu}_{\Delta(T_k)}, \overline{\mu}_{\Delta(T_k)}]$ | $(\overline{\mu}_{\Delta(T_k)}, +\infty]$ |
| $\Delta(T_k) \notin \partial(CH(\overline{S}))$,non Gabriel, non Dominated | | $(\underline{\mu}_{\Delta(T_k)}, \overline{\mu}_{\Delta(T_k)}]$ | $(\overline{\mu}_{\Delta(T_k)}, +\infty]$ |
| $\Delta(T_k) \notin \partial(CH(\overline{S}))$ Gabriel, Dominated | $(\underline{\rho}_{T_k}, \underline{\mu}_{\Delta(T_k)}]$ | $(\underline{\mu}_{\Delta(T_k)}, \overline{\rho}_{T_k}]$ | $(\overline{\rho}_{T_k}, +\infty]$ |
| $\Delta(T_k) \notin \partial(CH(\overline{S}))$,non Gabriel, Dominated | | $(\underline{\mu}_{\Delta(T_k)}, \overline{\rho}_{T_k}]$ | $(\overline{\rho}_{T_k}, +\infty]$ |

# Our Vision

## ▷ Experiments and Modeling


Biochemistry

Geometry

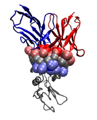
Combinatorics Optimization

Biophysics

Topology

Statistics

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