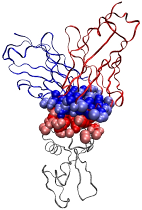
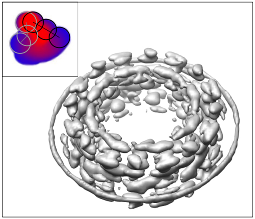
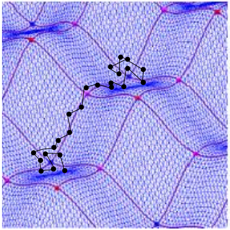
Assessing the Reconstruction of Macro-molecular Assemblies: the Example of the Nuclear Pore Complex

F. Cazals, Algorithms - Biology - Structure, INRIA Sophia-Antipolis
T. Dreyfus, Algorithms - Biology - Structure, INRIA Sophia-Antipolis
V. Doye, Institut Jacques Monod, CNRS, Paris







INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

Centre de recheche
SOPHIA-ANTIPOLIS-MÉDITERRANÉE

- Reconstructing large protein assemblies
- Modeling with uncertainties: Toleranced Models
- Assessing the reconstruction of assemblies
    - Mining contacts: contact probabilities
    - Mining complexes: stoichiometry, volume ratio
    - Mining complexes: graphical models
- Conclusion
- Software of potential interest

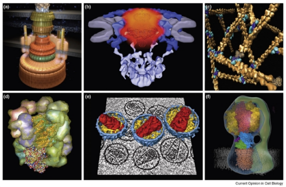
Structural Dynamics of Macromolecular Processes

Reconstructing Large Macro-molecular Assemblies

Bacterial flagellum — rotary propeller

Nuclear Pore Complex — nucleocytoplasmic transport

Branched actin filaments — muscle contraction, cell division



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

Reconstructing Large Assemblies: a NMR-like Data Integration Process

▷ Four ingredients

– Experimental data

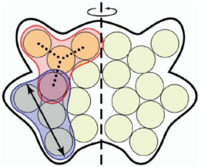
– Model: collection of balls

– Scoring function: sum of restraints

restraint : function measuring the agreement

≪model vs exp. data≫

– Optimization method (simulated annealing,. . .)



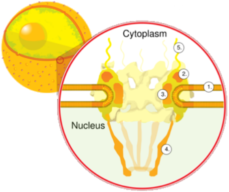
▷ Restraints, experimental data and . . . ambiguities:

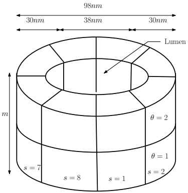
Assembly	:	shape	cryo-EM	fuzzy envelopes
Assembly	:	symmetry	cryo-EM	idem
Complexes:	:	interactions	TAP (Y2H, overlay assays)	stoichiometry
Instance:	:	shape	Ultra-centrifugation	rough shape (ellipsoids)
Instances:	:	locations	Immuno-EM	positional uncertainties

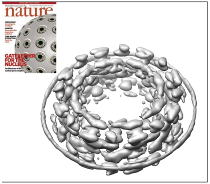
▷Ref: Alber et al, Ann. Rev. Biochem. 2008 + Structure 2005

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:

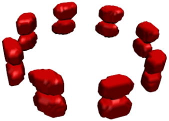
Our map is sufficient to determine the relative positions within NPC
...limited precision; not to be mistaken with the density map from EM
The localization volumes . . . allow a visual interpretation of proximities

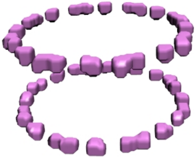
▷`Ref: Alber et al; Nature; 450; 2007`

NPC: Example Density Maps

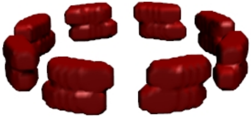
Stoichiometry vs number of connected components

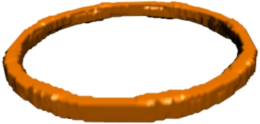
▷ Cases: equal (Nup157); larger (Sec13)





▷ Cases: smaller (Nup170, Pom152)





▷ Two types of problems:

number of connected components vs stoichiometry

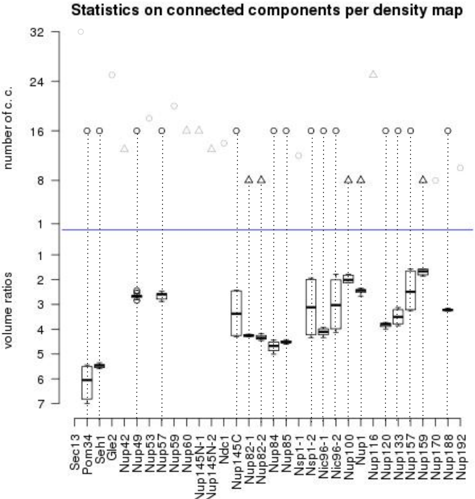
volume of each connected component vs. volume estimated from the sequence

▷`Ref: Alber et al; Nature; 450; 2007`

Uncertainties of the Density Maps

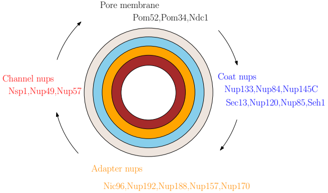
- ▷ Volume of connected components of non empty voxels vs. reference volume (estimated from the sequence)

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



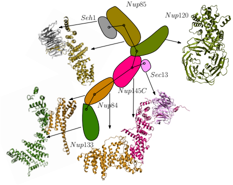
Putative Models of Sub-complexes: the Y-complex

▷ Symmetric core of the NPC



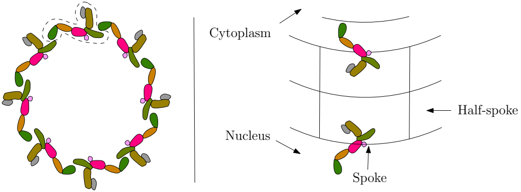
▷Ref: Blobel et al; Cell; 2007

▷ The Y-complex: pairwise contacts



▷Ref: Blobel et al; Nature SMB; 2009

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

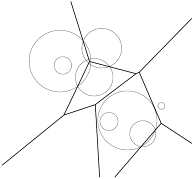


▷Ref: Seo et al; PNAS; 2009

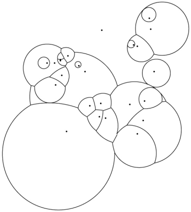
▷Ref: Brohawn, Schwarz; Nature MSB; 2009

⇒ Bridging the gap between both classes of models?

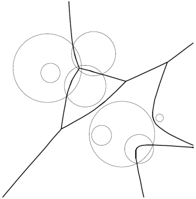
The Zoo of curved Voronoi diagrams



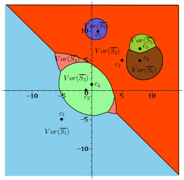
▷ Power diagram:
$d(S(c,r),p) = \|c-p\|^2 - r^2$



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

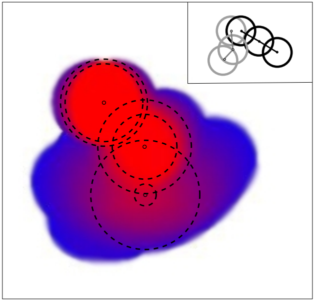


▷ Apollonius diagram:
$d(S(c,r),p) = \|c-p\| - r$



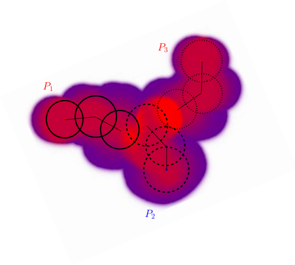
▷ Compoundly Weighted Voronoi diagram:
$d(S(c,\mu,\alpha),p) = \mu\|c-p\| - \alpha$

Building toleranced models (Embracing the geometric noise.)



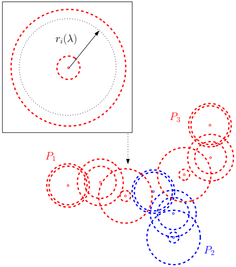
Uncertain Data and Toleranced Models: the Example of Molecular Probability Density Maps

- ▷ 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



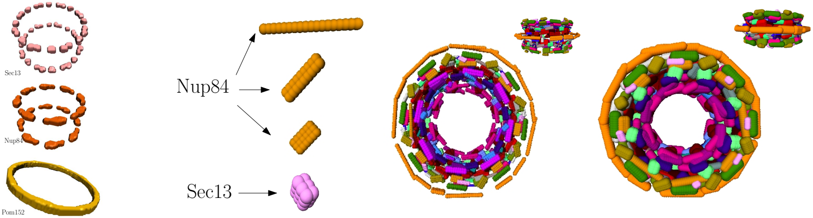
- ▷ Space-filling diagram $\mathcal{F}_\lambda$: a continuum of models
  - – Radius interpolation: $r_i(\lambda) = r_i^- + \lambda(r_i^+ - r_i^-)$

- ▷ 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
  - – controlling $r_i^+ - r_i^-$: w.r.t volume estimated from the sequence



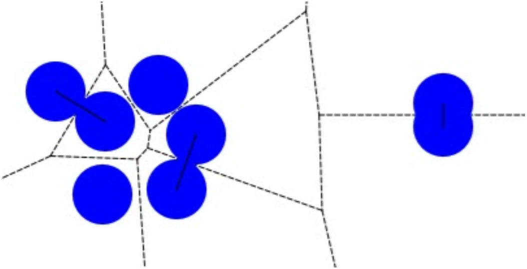
(i) Canonical shapes

(ii) NPC at $\lambda = 0$

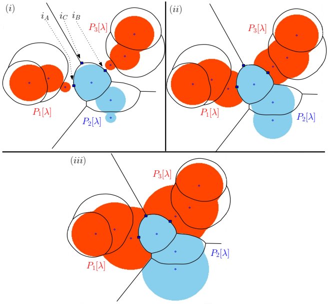
(iii) NPC at $\lambda = 1$

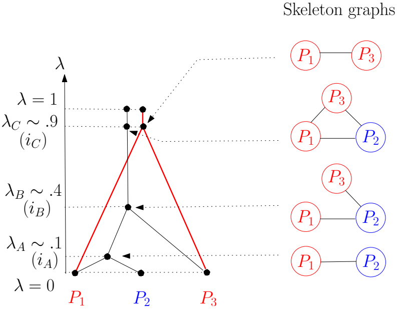
Growing toleranced models and enumerating their finite set of topologies (Spotting stable structures.)

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

The Union-Find Algorithm

▷ How many clusters?

▷ The Union-Find algorithm

Dynamic maintenance of
    the connected components (c.c.)
    of an evolving graph

▷ Three operations

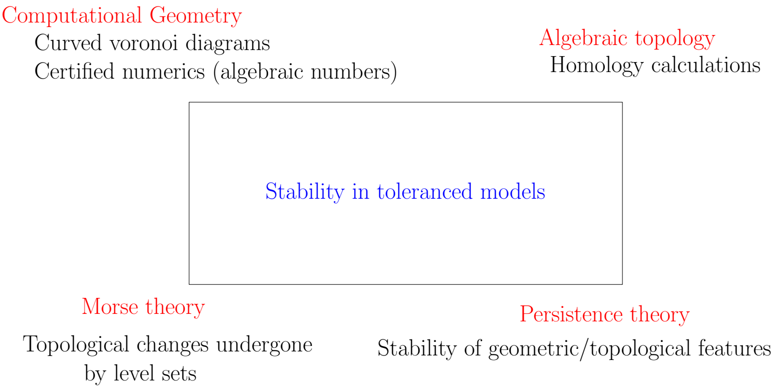
Make_set
Find the leader of a c.c.
Union two components

▷ Complexity: almost linear

$m\alpha(m, n)$

▷`Ref: R.E. Tarjan; Data Structures and Network Algorithms; 1983`

On Intersecting Balls. . .



Computational Geometry

Curved voronoi diagrams

Certified numerics (algebraic numbers)

Algebraic topology

Homology calculations

Stability in toleranced models

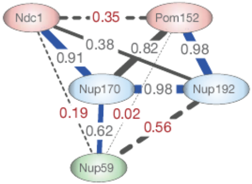
Morse theory

Topological changes undergone by level sets

Persistence theory

Stability of geometric/topological features

PROEMINENT CONTACT FREQUENCIES OUT OF THE $\binom{30}{2} + 30 = 465$ PAIRS OF PROTEIN TYPES



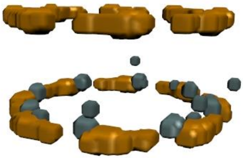
– Contact frequency:
fraction of the 1000 models with $\geq$ one contact between instances of these types

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

– Limitations:
contact can be shallow
stoichiometry missing

Contact Probabilities versus Contact Probabilities

- ▷ Over-represented in Sali et al:
  $Nup84 - Nup60 : f_{ij} = 0.07$



- ▷ Under-represented in Sali et al:
  $Nup192 - Pom152 : f_{ij} = 0.98$

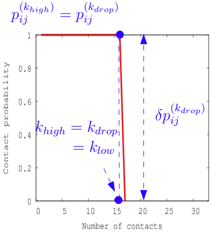
- ▷ Contacts for two types p_i and p_j

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

– Define: $\lambda(p_i, p_j)$: smallest λ
  $\exists$ k contacts between p_i and p_j

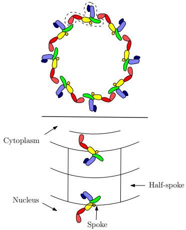
– Contact proba.: $p_{ij}^{(k)} = 1 - \lambda(p_i, p_j)/\lambda_{\max}$

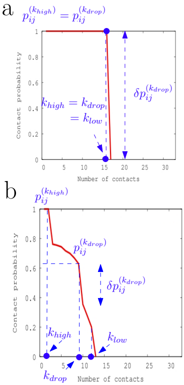
– Contact curve: $p_{ij}^{(k)} = f(k)$



Note: $\lambda_{\max}$ tuned to match the uncertainties on the input

Contact Curves: Insights on (models of) the Y-complex





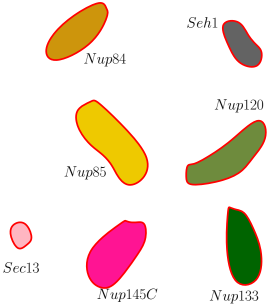
c

Protein types	f_{ij}	k_{high}	k_{drop}	$p_{ij}^{(k_{drop})}$	$s(k_{drop})$	$\min \overline{V}_{\lambda_{k_{drop}}}$
(Nup133, Nup84)	0.571	16	16	1.00	1.00	0.76
(Nup145C, Nup84)	1.000	16	16	1.00	1.00	0.79
(Nup120, Seh1)	0.837	16	16	1.00	1.00	0.82
(Nup133, Nup145C)	0.589	16	16	1.00	1.00	0.83
(Nup120, Nup85)	0.569	16	16	1.00	1.00	0.88
(Nup85, Seh1)	1.000	11	16	0.83	1.21	2.30
(Nup84, Sec13)	0.66	10	14	0.79	1.26	2.63
(Nup145C, Sec13)	0.503	12	12	1.00	1.00	0.81
(Nup133, Sec13)	0.381	10	12	0.96	1.04	1.06
(Nup120, Sec13)	0.284	4	12	0.77	1.31	2.25
(Nup120, Nup84)	0.487	2	10	0.67	1.49	1.79
(Nup133, Nup85)	0.478	1	9	0.82	2.55	2.82
(Nup84, Seh1)	0.376	2	9	0.63	3.63	3.08
(Sec13, Seh1)	0.233	4	4	1.00	1.00	0.56
(Nup85, Sec13)	0.227	4	4	1.00	1.00	0.78
(Nup120, Nup133)	0.465	1	3	0.89	2.91	1.57
(Nup84, Nup85)	0.543	2	2	1.00	2.27	0.83
(Nup120, Nup145C)	0.498	1	2	0.95	1.86	1.16

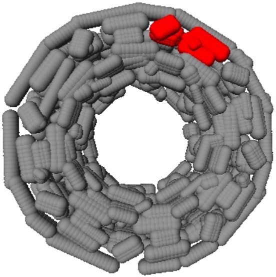
▷ Insights:

contact probabilities sharper than frequencies (Sali et al)
3/6 contacts from Blobel et al confirmed
closure of the rings: Nup120 - Nup133 not prominent

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



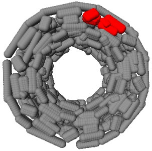
Y-complex : protein types

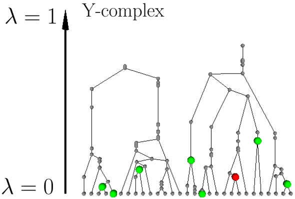


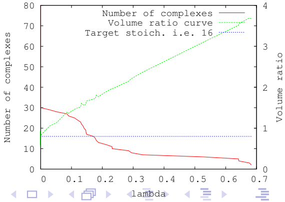
Y-complex : instance

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

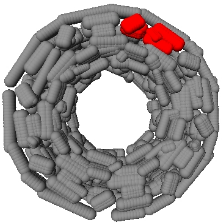
- Input:
  - Toleranced model
  - *T*: set of proteins types, the red proteins (TAP, types involved in sub-complex)
- Output, overall assembly:
  - Geometry - biochemistry:
    number of isolated copies – symmetry analysis
    TAP data: complex or mixture?
  - Topological stability: death date - birth date (cf α-shape demo)
- Output, per complex:
  - Biochemistry: stoichiometry of protein instances per copy
  - Geometry, volume ratio: volume occupied vs. expected volume



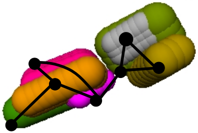




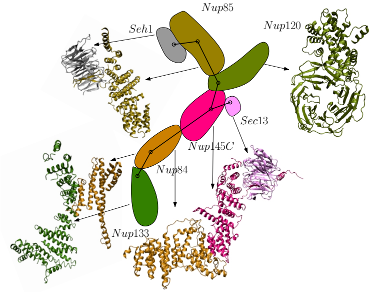
Assessing a toleranced model w.r.t a high-resolution structural model



Assembly



Complex: skeleton graph

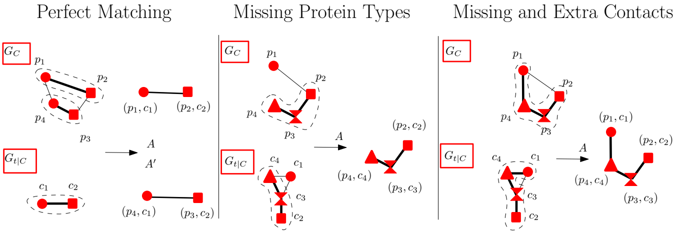


Template: skeleton graph

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

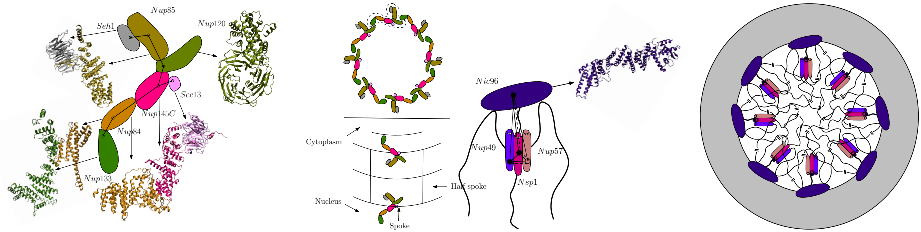
Missing Protein Types

Missing and Extra Contacts

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

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

Insights on the NPC



Y-complex

T-complex

Key Facts on the Y-complex and the T-complex

- ▷ Contacts analysis: 36 over-represented pairs

- ▷ Analysis w.r.t. a set of protein types

  Y-complex:

  Poor positioning of Sec13

  No isolation of copies of the Y-complex: contacts across copies prevail

  T-complex:

  16 isolated copies found: contacts intra-copies prevail

- ▷ Analysis w.r.t. a 3D template

  Y-complex:

  Support for Blobel's model: Y-complexes for two rings

  Contact involved in closure; role of Nup85

  T-complex:

  Asymmetry of the interactions (Nic96,Nup49) [strong] (Nic96,Nic57) [weak]

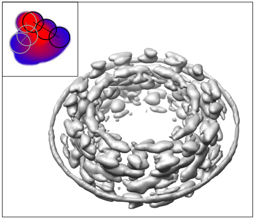
  New 3D template for (Nic96,Nsp1,Nup49,Nup57)

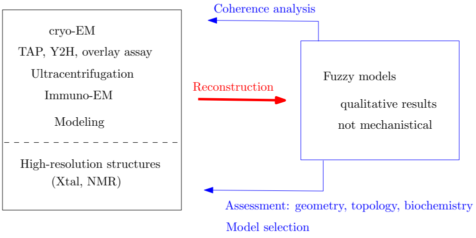
- ▷ The global model of Sali et al does convey precise information...

  when coupled to appropriate tools to probe it; in particular

THE TOLERANCED MODEL ALLOWS TESTING ANY SUB-COMPLEX/PULLOUT

Toleranced Models for Large Assemblies: Positioning





cryo-EM

TAP, Y2H, overlay assay

Ultracentrifugation

Immuno-EM

Modeling

High-resolution structures (Xtal, NMR)

Coherence analysis

Reconstruction

Fuzzy models

qualitative results

not mechanistical

Assessment: geometry, topology, biochemistry

Model selection

- ▷ Methodology: modeling with uncertainties
  - – Toleranced models: continuum of shapes vs fixed shapes
  - – Topological and geometric stability assessment
    Curved α-shapes
- ▷ Applications to toleranced complexes
  - – A-I. Contact probabilities (stoichiometry)
  - – A-II. Analysis of sub-complexes (symmetries, volume ratio)
  - – A-III. Contacts within sub-complexes (graphical models of sub-complexes)

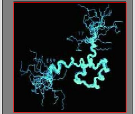
Our Vision

▷ Experiments and Modeling

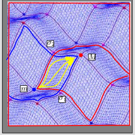
Biochemistry

Biophysics

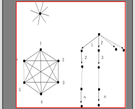
Geometry



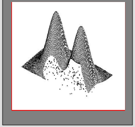
Topology



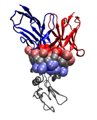
Combinatorics
Optimization



Statistics



Structure-to-Function



Docking (and Folding)

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

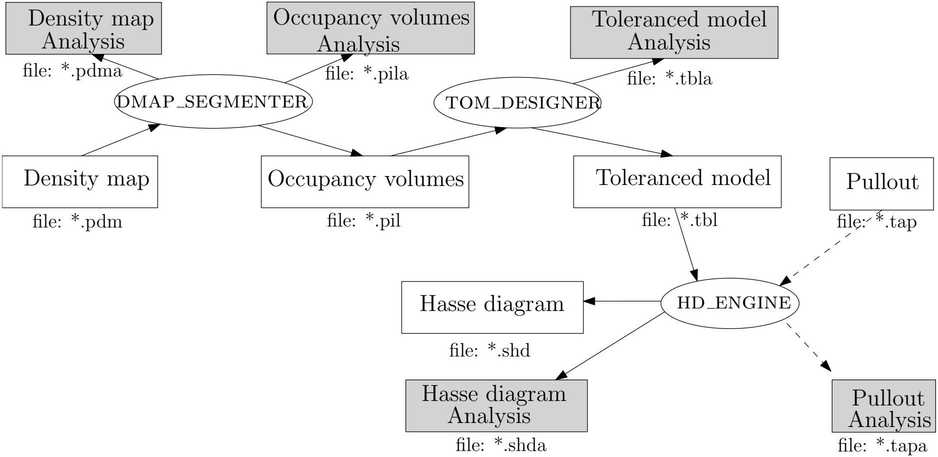
▷ Questions

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

▷ Partial answers from

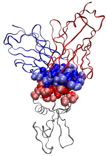
- Geometric - topological modeling stability analysis
- Graph theory matching algorithms
- Statistical testing
- Dimensionality reduction investigating correlations

Sotware: Modeling Large Assemblies



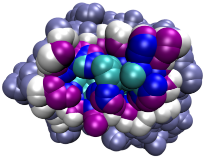
Sotware: Modeling Protein Interfaces

▷ intervor: modeling protein - protein interfaces

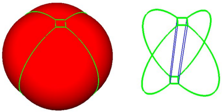


http://cgal.inria.fr/abs/Intervor; Bioinformatics; 26 2010

▷ vorpatch: topological encoding of binding patches

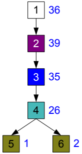


▷ vorlume: certified molecular surfaces and volumes



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

▷ compatch: comparing binding patches

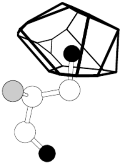


Sotware: Misc

▷ Geomsel:
selection of diverse conformers

ACM Trans. CBB; 2011

▷ ESBTL: C++ template library
data model / geometry



http://esbtl.sf.net;
Bioinformatics 26; 2010

▷ Computational Geometry Algorithms Library: 3D spherical kernel

http://www.cgal.org