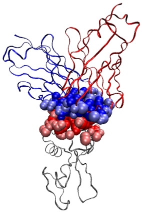
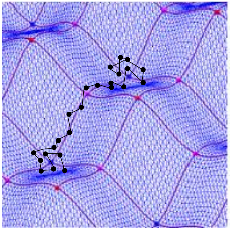
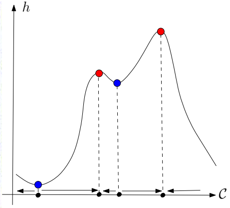
Geometric Models for the Description of 3D Molecular Systems (part I) and High-dimensional Point Cloud Data (part II)

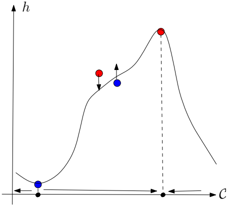
F. Cazals
Algorithms - Biology - Structure
INRIA Sophia-Antipolis







Original



Simplified

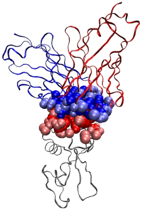
INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

Centre de recheche
SOPHIA-ANTIPOLIS-MÉDITERRANÉE

Describing 3D Molecular Systems

(Bias towards protein complexes)

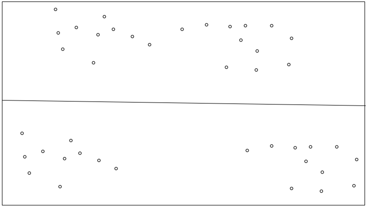


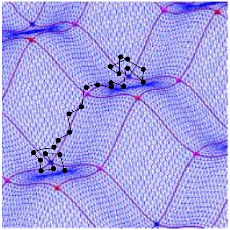
- Shape - topology:
  # connected components, holes, voids / cavities [Homology]
  fat, skinny, dumbbell-like

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

- Correlations with bio-physical quantities

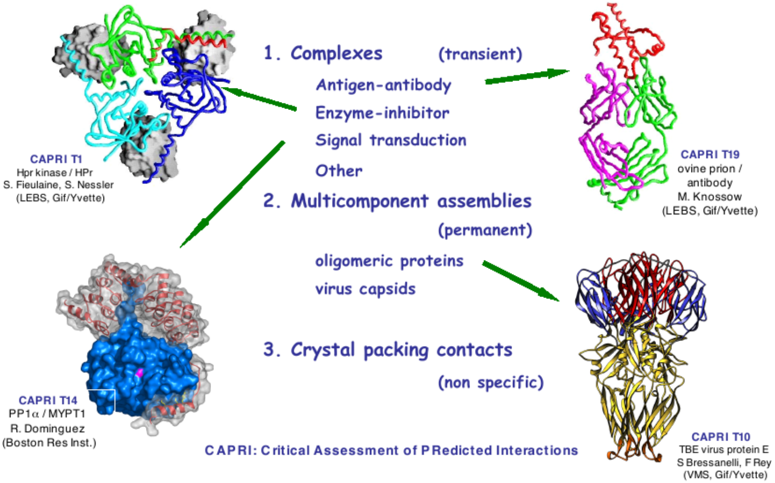
Describing (High-dimensional) Point Cloud Data





- (Related) goals
  - Reconstructing a sampled shape: connect the dots
  - Stratifying sampled landscapes (cf David Wales' remarks)
  - Performing a Multi-scale analysis
- Sampling density versus guarantees
  - Topology : homotopy / homeomorphy / isotopy
  - Geometry : Hausdorff distance
- STAR
  - 3D, surface: under control (challenging cases: cf P. Salamon's talk)
  - 3D, stratified complex: this talk
  - nD, manifold or stratified: in progress

Structure to Function: Diversity of Protein Assemblies

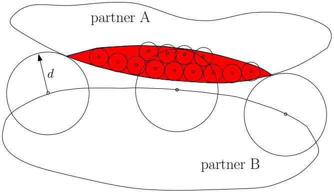


[J. Janin]

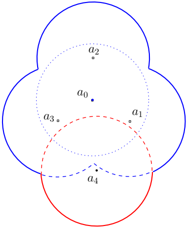
- ▷ Up to × 500,000 atoms (NPC, etc)

About Interface Models

▷ Distance threshold
(geometric footprint)



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



▷ The Voronoi interface model

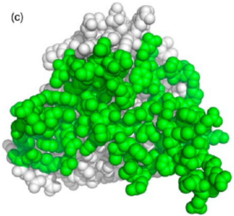
A parameter free interface model
Singles out a single layer of atoms
Is amenable to geometric and topological calculations

▷ Applications

In general : many!
In the context of landscapes:
beyond the fraction of native contacts and the-like (?)

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

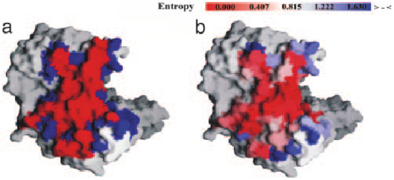
Structure Shape, size, position of atoms; hot residues tend to be located in the interface *core*

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

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

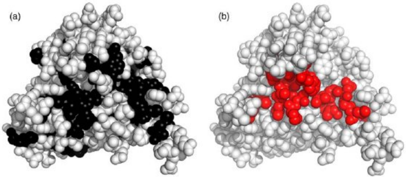
Core residues more conserved

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

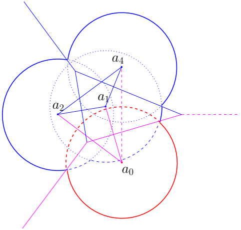
Conclusion

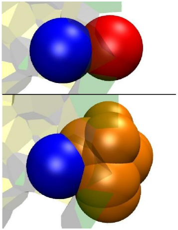
Conservation detects dry $\gg$ Conservation geom. footprint

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

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

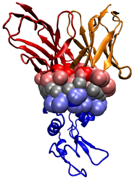
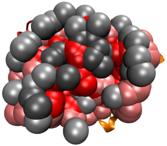
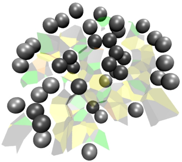
▷Ref: Cazals, Proust, Bahadur, Janin; Protein Science; 2006

Demo!

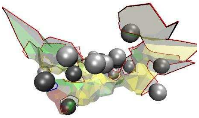
Voronoi Interfaces : Illustrations

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

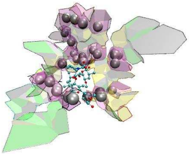
- ▷ Role of strutural water –antobody-antigen

- ▷ Curvature –protease-inhbitor



- ▷ Multi-patch structure –signal transduction



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

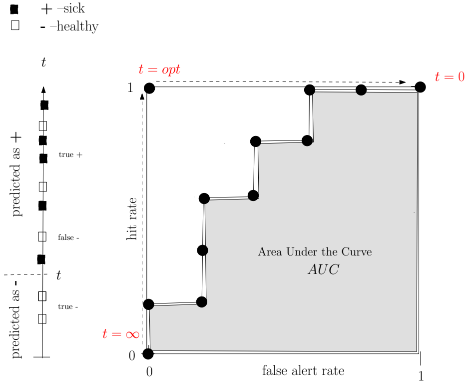
Receiver Operating Characteristic (ROC) curves

- 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:

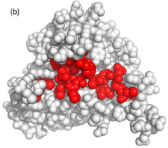


- p-value calculation for a particular value AUC_0:
  AUC_0 vs. distribution of areas over all permutations of + and -

Water Traffic and Conservation of Residues at Protein - Protein Interfaces

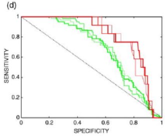
- Dry A.A. tend to be more *important*
- Protocol: MD simulation; A.A. s.t. $\Delta ASA > 0$
- Traffic intensity for A.A. i: $I_i = \frac{1}{T} \sum_w \frac{1}{\tau_w}$
- Dry residue w.r.t.traffic intensity:
  - $I_i \leq 0.005 ps^{-2}$ for homodimers
  - $I_i \leq 0.01 ps^{-2}$ for heterodimers
- Assessment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

- 2DOR: dry residues



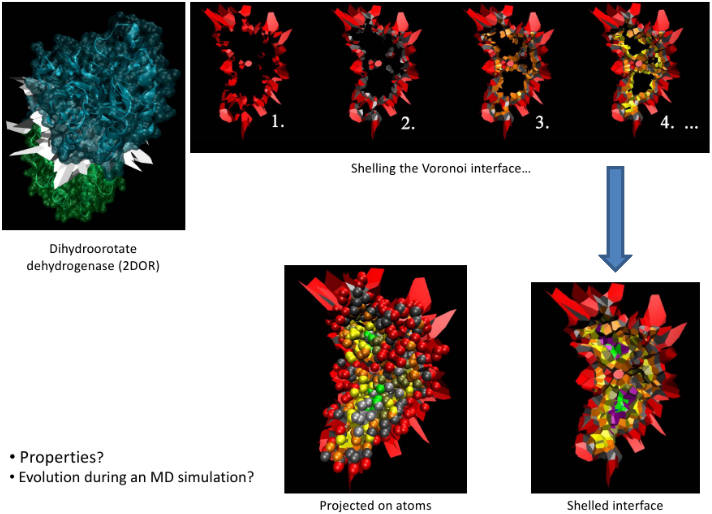
conservation predicts dryness versus conservation predicts geom. footprint

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



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

Shelling the Voronoi Interface: Illustration

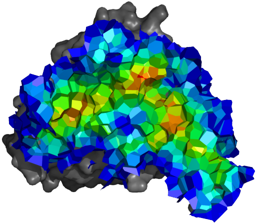


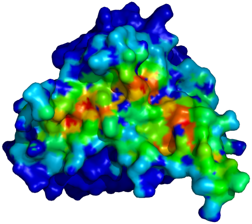
- Properties?
- Evolution during an MD simulation?

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

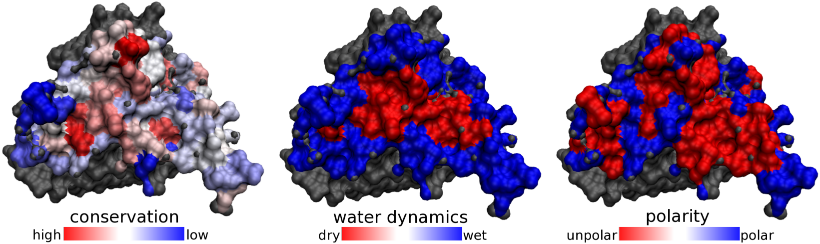
VSO versus Dryness – 2DOR

▷ 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.]→ dryness [threshold] conserv. [cont.] → dryness [threshold]

conserv. [cont.]→ VSO [threshold] VSO [cont.] → 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

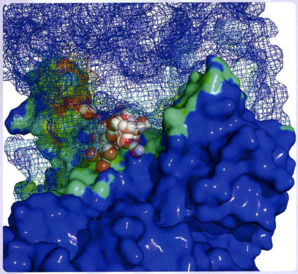
conserv.→dryness (cf Lichtarge et al, JMB 369, 2007) [no p-values]
conserv.→VSO (cf Chakrabarti et al, PNAS 102, 2005) [combined p-values only]
VSO→unpolar
global trend . . . but prediction often fails on an individual basis
binary core/rim interface models do not account for the subtlety of distributions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

Shelling Voronoi interfaces: Conclusion and Outlook

▷ Interface models

– Binary core/rim model :
  does not account for the geometry of conservation / polarity;
  mining correlations : global signals only
– Voronoi based model
  discrete interface parameterization;
  statistics on a per-complex basis

▷ Water dynamics mainly shaped by geometry

...as opposed to force fields and properties of residues

▷ Future work

–Simple percolation models on Voronoi lattices?
Percolare (≪ filtrer, passer≫) de per- (≪au travers ≫) et colare (≪ couler ≫).
–Connexion to interface properties: dynamic interfaces (MD), $\Delta\Delta G$
–Connexion to free energy calculations (structure of the solvent)

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

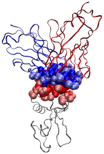
Geometry-based Quantitative Models for Water Traffic

- Punchline: cost would incommensurable wrt molecular dynamics
- Static setting : consider an irregular Voronoi lattice embedded in 3D
  - Overall representation of the lattice
    Tree encoding changes in the topology of the level sets of the VSO
  - Attributes of a tile
    depth
    uniform(?) probability : ability to accommodate a W molecule
    cf packing properties of dual atoms
  - Interface surrounded by a water bulk
- Questions
  - Water-centric : behaviour of a water molecule entering the lattice
  - A.A.-centric : water traffic as a function of VSO
  - Interface-centric: formation of *channels* from the bulk to the core
    overall hydration of the interface
- Dynamic setting
  - The probability field varies over time, maybe as a function of depth
- Remarks
  - Asymptotic regime . . . VSO bounded by 10

Software

▷ Computational Geometry Algorithms Library: 3D spherical kernel

▷ Intervor: modeling protein - protein interfaces

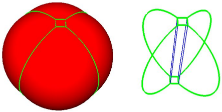


cgal.inria.fr/abs/Intervor/;
Bioinformatics 26 2010

▷ Geomsel:
selection of diverse conformers

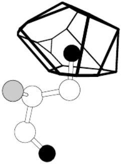
Not released yet;
ACM Trans CBB 2010

▷ Vorlume: certified molecular surfaces and volumes



cgal.inria.fr/abs/Vorlume/;
ACM Trans. Math Softw. 2010

ESBTL: C++ template library
data model / geometry



esbtl.sf.net;
Bioinformatics 26, 2010

ABS : Synergy Between Algorithms and Structural Biology

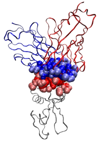
▷ Work-package 1: modeling protein complexes

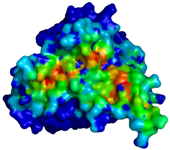
–Macro-molecular interfaces, from fine descriptions to scoring:

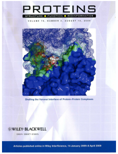
Description: interface geometry vs a.a. conservation vs solvent dynamics

Scoring: discriminating native vs non native complexes

– Modeling large assemblies (Nuclear Pore Complex)



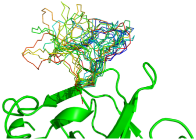


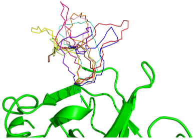


▷ Work-package 2: modeling the flexibility of proteins

– Manipulating conformer ensembles: boosting conformational diversity

– Collective coordinates: beyond normal modes





▷ Work-package 3: algorithmic foundations