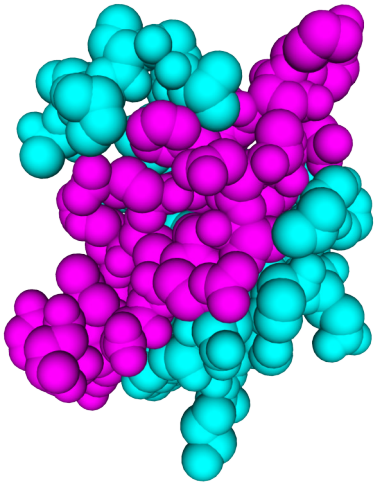
Voronoi tessellation-based analysis of 3D conformations of non-globular proteins

Dr. Kliment Olechnovič

CNRS Laboratoire Jean Kuntzmann, Grenoble, France

Vilnius University Life Sciences Center, Vilnius, Lithuania

2024-05-17



Common problems:

- analyzing how different parts in a molecule interact
- selecting the best prediction of a multimeric complex

Our solutions involve:

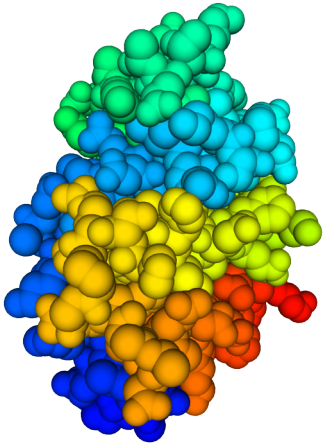
- computational geometry
- machine learning
- developing free software

Today's questions:

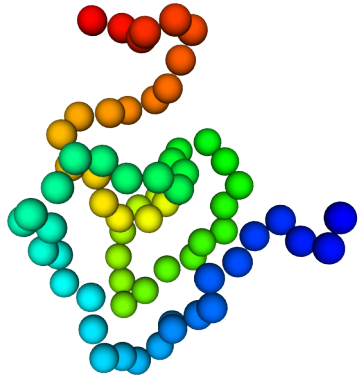
- What can we learn from disordered proteins?
- And would it be useful for analyzing protein-protein interactions?

Data of molecular conformations

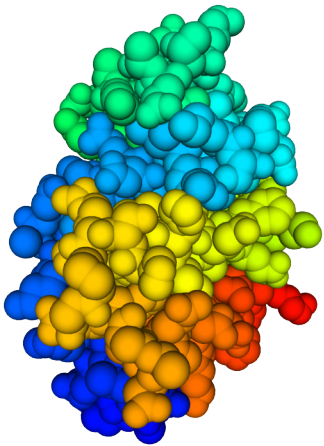
Protein Data Bank (PDB) data



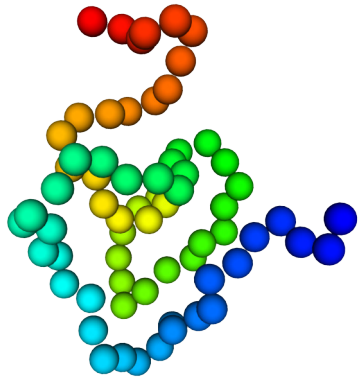
Simulated intrinsically disordered protein (IDP) data



PDB

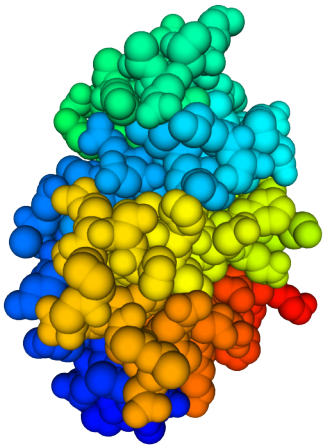


IDP

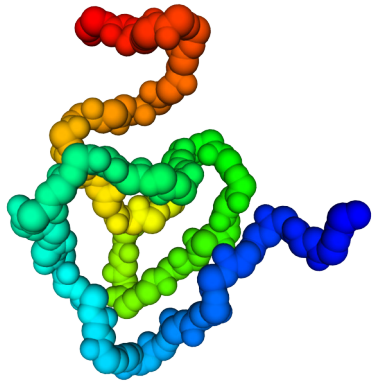


C-alpha atoms

PDB

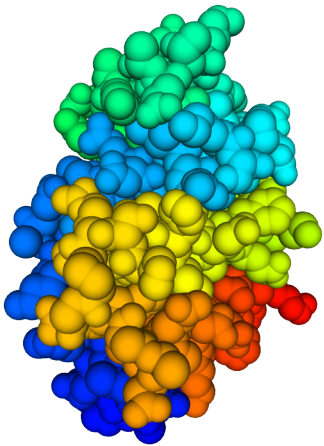


IDP

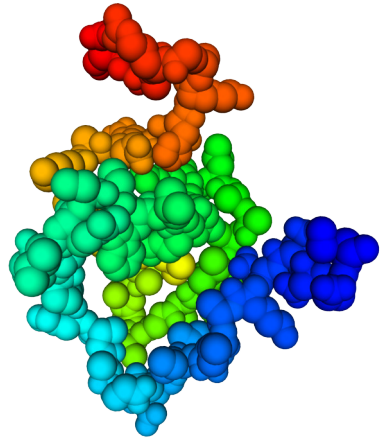


backbone atoms built by Pulchra

PDB

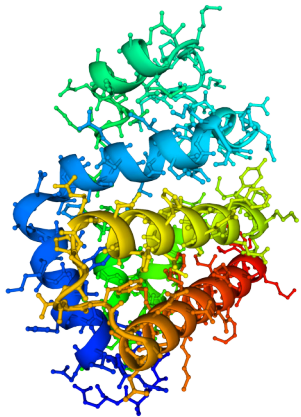


IDP

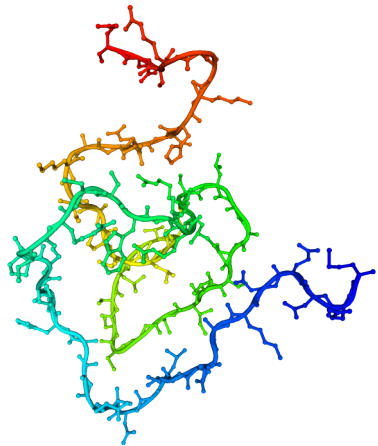


side-chain atoms built by FASPR

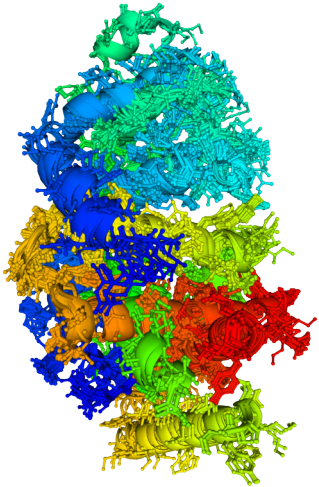
PDB



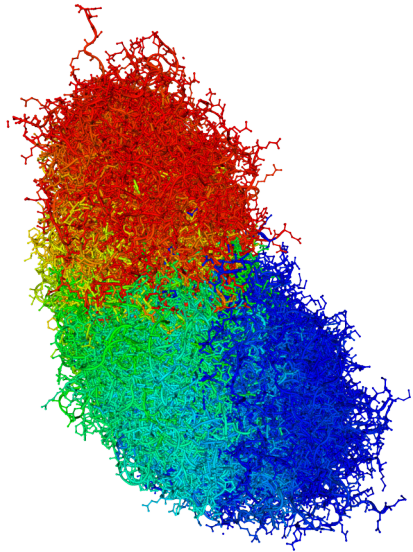
IDP



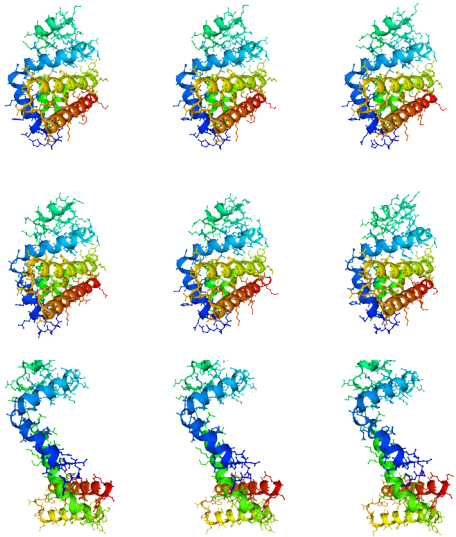
PDB ensemble



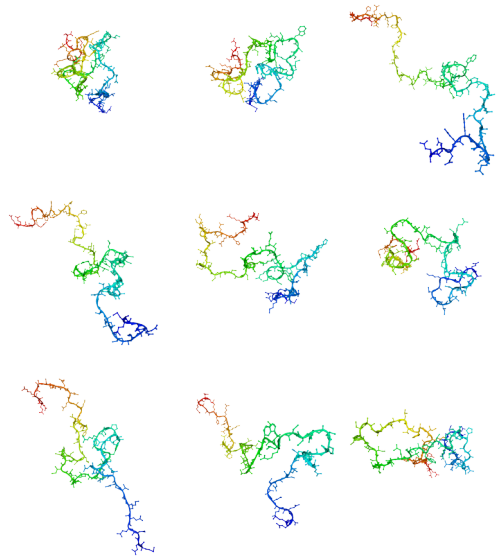
IDP ensemble



PDB ensemble



IDP ensemble



PDB dataset

- From the Protein Data Bank, wwwpdb.org
- Ensembles formed by clustering chain sequences using 90% identity
- We used all **38'807** ensembles
- Ensembles have very different numbers of chains, largest ensemble contains 1413 chains, 9989 ensembles contain only two chains, there are **429'945** protein chains in total.

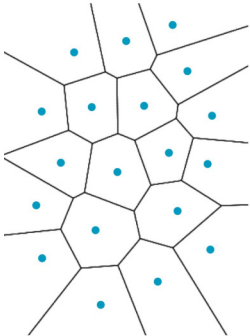
IDP dataset **"IDRome"**

- From Tesei, G., Trolle, A.I., Jonsson, N. et al. *Conformational ensembles of the human intrinsically disordered proteome*. Nature (2024)
- Ensembles generated by running coarse-grained simulations using CALVADOS for 28'058 IDP-like protein sequences
- We used all **16'774** ensembles for chains of 60 to 600 residues in length
- Every ensemble has 1010 conformations, so there **16'941'740** conformations in total

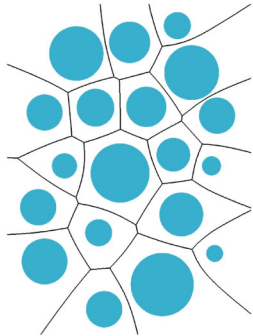
Describing interactions in molecular conformations using the Voronoi tessellation

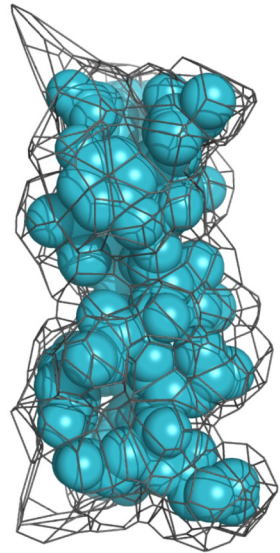
Voronoi diagram of points and balls

"Classic" Voronoi diagram of points



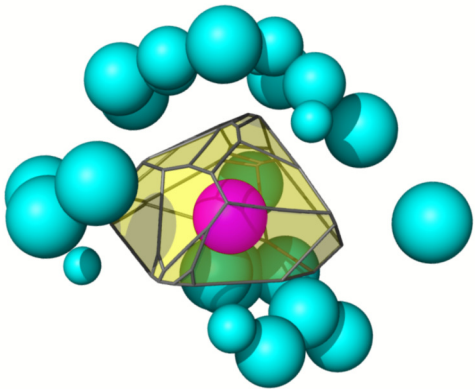
Voronoi diagram of balls



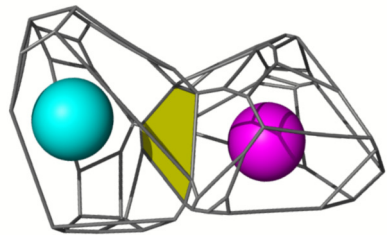


Voronoi tessellation-based analysis of structures

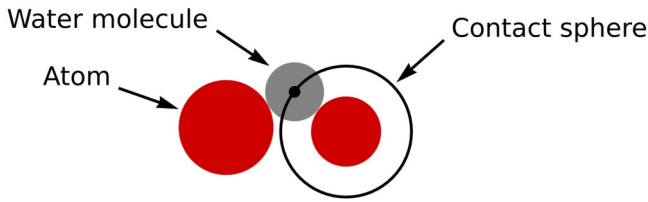
Voronoi cell of an atom surrounded by its neighbors

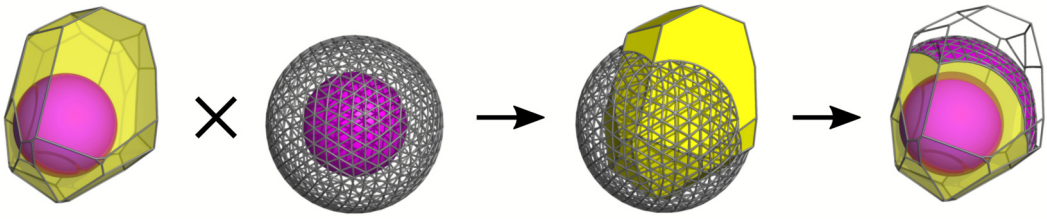


Atom-atom contact surface defined as the face shared by two adjacent Voronoi cells.



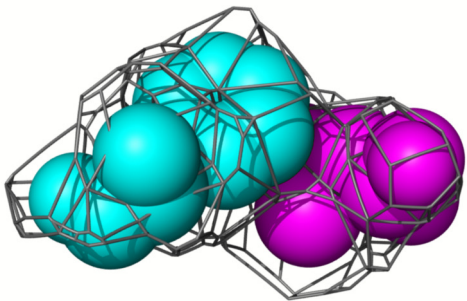
Constrained contacts



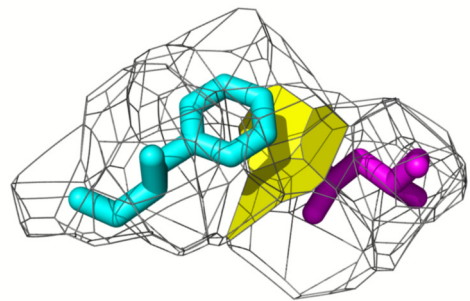


Deriving residue-residue contacts

Voronoi cells of two neighboring residues

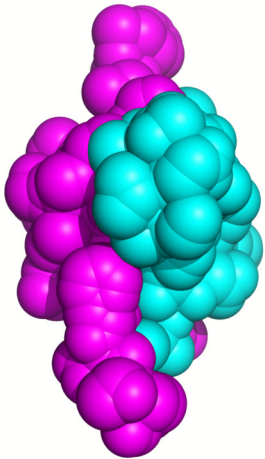


Residue-residue contact surface defined as a union of atom-atom contact surfaces

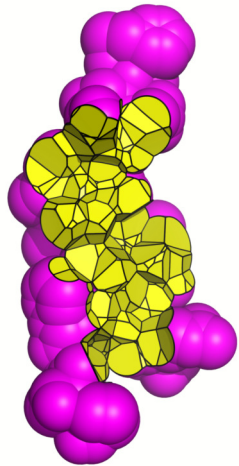


Inter-chain contacts

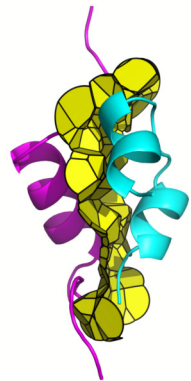
Solvent-accessible surface of an insulin heterodimer PDB:4UNG colored by subunit



The intersubunit interface shown together with the SAS of one subunit

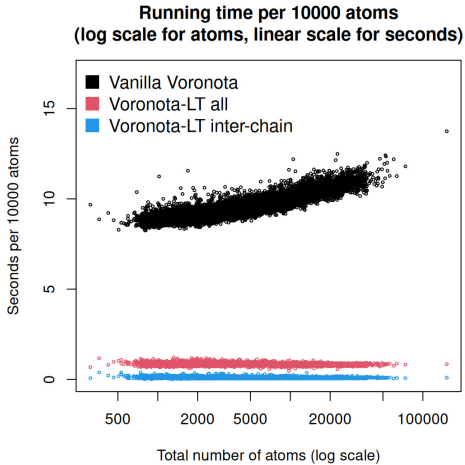


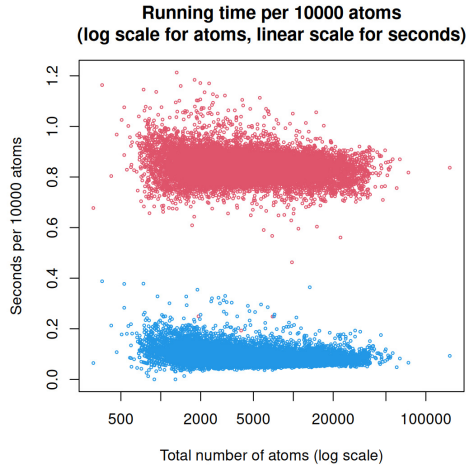
The intersubunit interface shown together with both subunits represented as cartoons



Voronota-LT

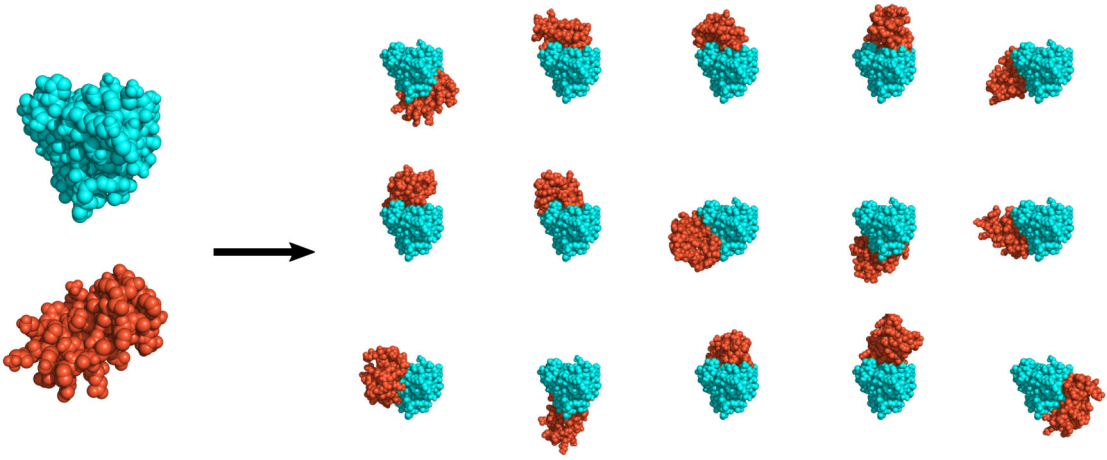
Voronota-LT is a new fast software for constructing tessellation-derived atomic contact areas and volumes. It is significantly faster than its predecessor, Voronota:



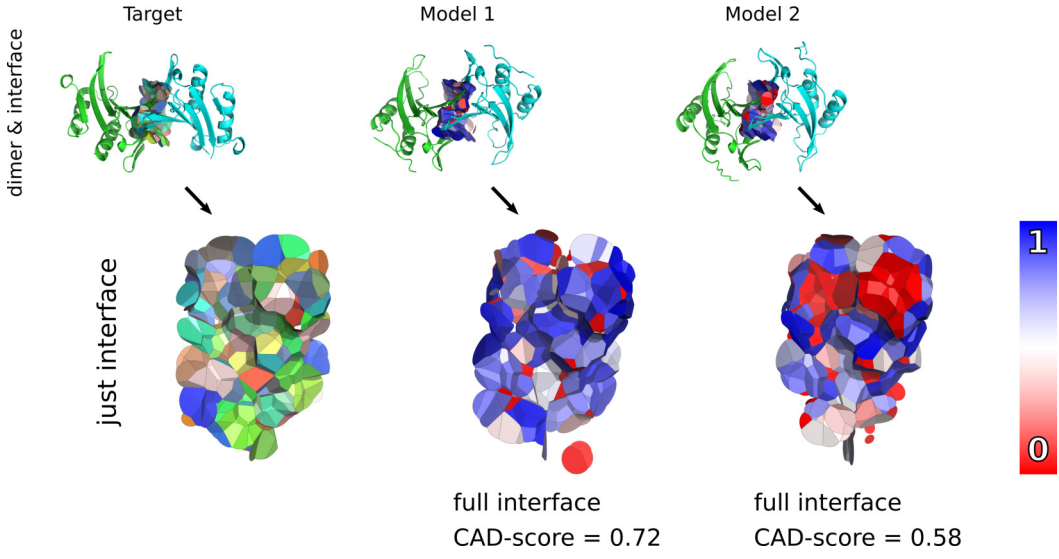


An application of tessellation-based description of interactions

Same chains can have differently modelled interfaces



Comparing interfaces using CAD-score (Contact Area Difference score)

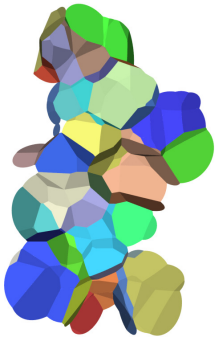


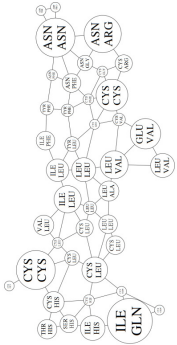
A dataset of correct and incorrect interfaces

- A non-redundant set of 1567 native heterodimers, selected using PPI3D and downloaded from PDB.
- Each native structure (target) was redocked and a set of models of varying quality was selected (about 15-20 models for a target), for example:

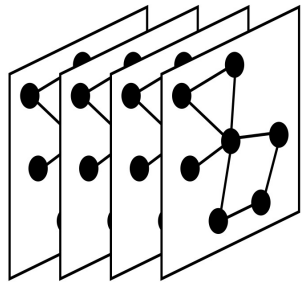
ID	x	y	z	a1	a2	a3	cadscore	site_cadscore
1E50_nat	0	0	0	0	0	0	1	1
1E50_2250	-7	27	4	45	153	90	0.74375	0.87635
1E50_32	-13	25	2	18	153	90	0.63728	0.75543
1E50_2735	-7	28	1	72	162	120	0.53173	0.68644
1E50_15946	-16	26	-2	45	162	120	0.38075	0.55364
1E50_10393	-16	28	5	0	153	90	0.24134	0.47034
1E50_3759	7	29	7	351	117	40	0.13939	0.51889
1E50_17192	24	22	8	315	63	0	0.0386	0.42122
1E50_15006	-13	27	13	342	18	0	0	0.40432
1E50_5533	28	-13	20	0	45	204	0	0.30295
1E50_14280	27	-22	-22	180	126	60	0	0.20266
1E50_532	34	4	-18	207	54	100	0	0.10126
1E50_20368	1	-39	10	324	117	80	0	0.00119
1E50_9297	37	5	-22	261	54	80	0	0

Evaluating interfaces with a graph neural network (e.g. VoroIF-GNN)

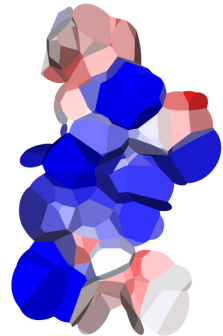




Interface
contacts graph



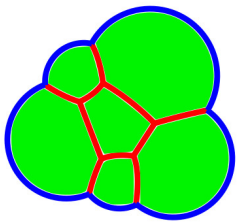
Graph attention
neural network



Predicted
CAD-score

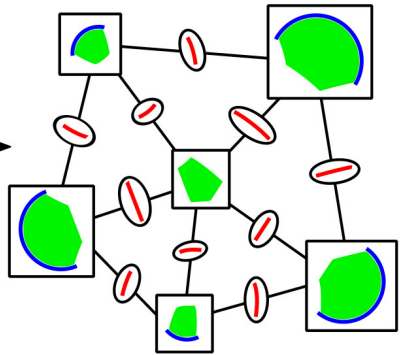
Input interface graph annotation in VoroIF-GNN

Tessellation-derived interface contacts



Contact surface
Contact-solvent border
Inter-contact border

Interface graph



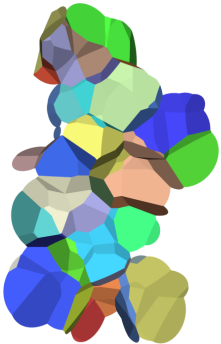
Graph **node** attributes (15 values)

- **Contact surface area**
- **Contact-solvent border length**
- **Sum of inter-contact border lengths**
- **Contact type-dependent descriptors (12 values)**

Graph **edge** attribute (1 value)

Inter-contact border length

Evaluating interfaces with an area-based potential (e.g. VoroMQA)



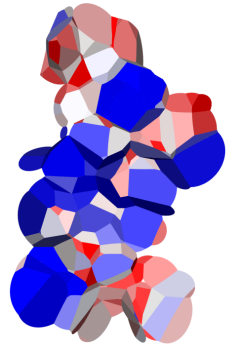
Interface
contact areas

$$\mathrm{E}(a_i, a_j, c_k) = \log \frac{\mathrm{P}_{\mathrm{exp}}(a_i, a_j, c_k)}{\mathrm{P}_{\mathrm{obs}}(a_i, a_j, c_k)} =$$

$$= \log \frac{\mathrm{F}_{\mathrm{exp}}(\mathrm{area}(a_i), \mathrm{area}(a_j), \mathrm{area}(c_k))}{\mathrm{F}_{\mathrm{obs}}(\mathrm{area}(a_i, a_j, c_k))}$$

$$\mathrm{E_n}(\Omega_\phi) = \frac{\sum_{\omega \in \Omega_\phi} \mathrm{E}(\mathrm{type}_\omega) \cdot \mathrm{area}_\omega}{\sum_{\omega \in \Omega_\phi} \mathrm{area}_\omega}$$

Statistical potential
for contact areas



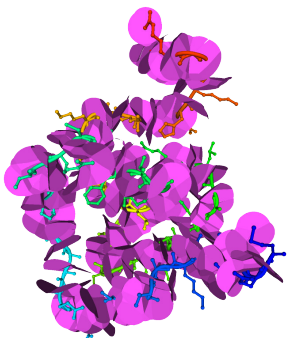
Interface
pseudo-energy

Deriving and using statistics of contact areas from ensembles of conformations

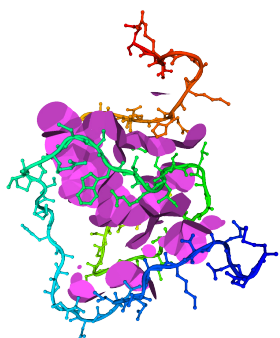
Contacts from a single conformation

A contact type is a tuple *(first atom type, second atom type, contact category)* = (a_1, a_2, c).

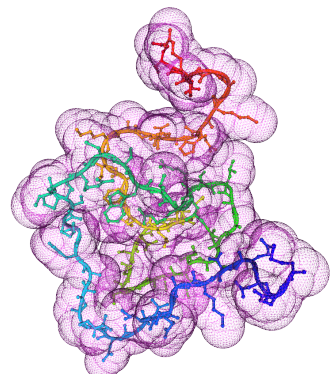
sequence separation <=5



sequence separation >5

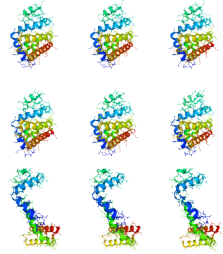


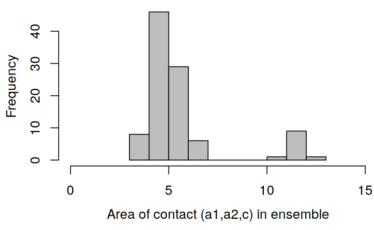
solvent-accessible surface



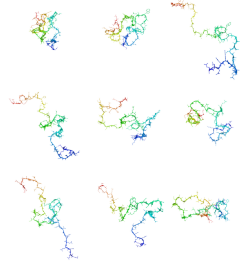
Contact areas from a single ensemble of conformations

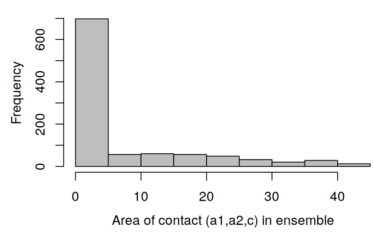
PDB ensemble





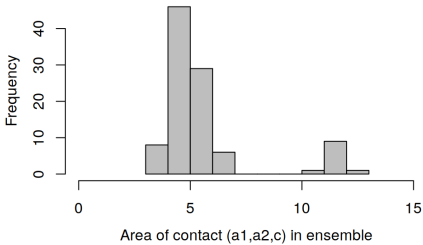
IDRome ensemble





Contact areas from a single ensemble of conformations

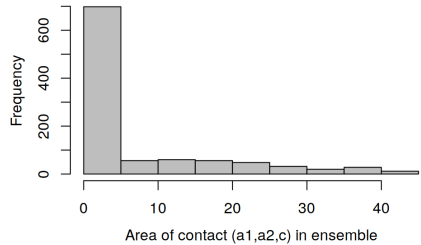
PDB ensemble



We summarize a contact type (a_1, a_2, c) area distribution in a PDB ensemble with:

- v = min(observed contact areas)
- u = max(observed contact areas)

IDRome ensemble



We summarize a contact type (a_1, a_2, c) area distribution in an IDP ensemble with:

- v = mean(observed contact areas)
- u = max(observed contact areas)

Areas of contact types from a multiple ensembles of conformations

v and u values are areas, therefore we can sum them.

For every contact type $t = (a_1, a_2, c)$ from the set of all possible types contact T, we sum the relevant v^t and u^t values from all the ensembles to get V^t and U^t sums.

Observed probabilities of areas of contact types

Observed probability estimate of contact area unit of type $t = (a_1, a_2, c)$ to occur:

$$P^t_{\text{obs}}(\text{occur}) = \frac{V^t + U^t}{\sum_{s \in T}(V^s + U^s)} \tag{1}$$

Observed conditional probability estimate of contact area unit to persist:

$$P^t_{\text{obs}}(\text{persist}|\text{occur}) = \frac{2V^t}{V^s + U^s} \tag{2}$$

Observed probability estimate of contact area unit to occur and persist:

$$P^t_{\text{obs}}(\text{occur and persist}) = P^t_{\text{obs}}(\text{occur}) \cdot P^t_{\text{obs}}(\text{persist}|\text{occur}) \tag{3}$$

Expected probabilities of areas of contact types

Expected probability estimate of contact area unit of type $t = (a_1, a_2, c)$ to occur (modeling the situation where there are no atom type-dependent or contact category-dependent effects):

$$P_{\text{exp}}^{t=(a_1,a_2,c)}(\text{occur}) \sim P_{\text{obs}}^{(a_1,*,*)}(\text{occur}) \cdot P_{\text{obs}}^{(*,a_2,*)}(\text{occur}) \cdot P_{\text{obs}}^{(*,*,c)}(\text{occur}) \cdot \tag{4}$$

Expected conditional probability estimate of contact area unit to persist:

$$P_{\text{exp}}^{t}(\text{persist}|\text{occur}) = \frac{2 \cdot \sum_{s \in T} V^s}{\sum_{s \in T}(V^s + U^s)} \tag{5}$$

Expected probability estimate of contact area unit to occur and persist:

$$P_{\text{exp}}^{t}(\text{occur and persist}) = P_{\text{exp}}^{t}(\text{occur}) \cdot P_{\text{exp}}^{t}(\text{persist}|\text{occur}) \tag{6}$$

Deriving pseudo-energy coefficient from probality estimates

Pseudo-energy coefficient for a contact area unit of type $t = (a_1, a_2, c)$:

$$E^t \sim \log\left(\frac{P^t_{\text{exp}}(\text{occur and persist})}{P^t_{\text{obs}}(\text{occur and persist})}\right) \tag{7}$$

E^t can be written as a weighted sum (weights tobe optimized later):

$$E^t = \alpha_1 \cdot E^t_{\text{obs}}(\text{occur}) + \alpha_2 \cdot E^t_{\text{exp}}(\text{occur}) + \alpha_3 \cdot E^t_{\text{obs}}(\text{persist}|\text{occur}) + \alpha_4 \cdot E^t_{\text{exp}}(\text{persist}|\text{occur}) + \beta \tag{8}$$

where:

$E^t_{\text{obs}}(\text{occur}) = \log P^t_{\text{obs}}(\text{occur})$
$E^t_{\text{exp}}(\text{occur}) = \log P^t_{\text{exp}}(\text{occur})$
$E^t_{\text{obs}}(\text{persist}|\text{occur}) = \log P^t_{\text{obs}}(\text{persist}|\text{occur})$
$E^t_{\text{exp}}(\text{persist}|\text{occur}) = \log P^t_{\text{exp}}(\text{persist}|\text{occur})$

Using pseudo-energy to score inter-chain interfaces

A total pseudo-energy score for a set of contacts G is:

$$S_{\alpha_1,\alpha_2,\alpha_3,\alpha_4,\beta}(G) = \sum_{g \in G} \text{area}(g) \cdot E^{\text{type}}_{\alpha_1,\alpha_2,\alpha_3,\alpha_4,\beta}(g) \tag{9}$$

We used 70% of the docking model sets from our interface decoys dataset to grid-search (primitively, but exhaustively, using a step of 0.1) for the best combination of weighting coefficients $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta)$ for the task of selecting well-modelled interfaces.

We used the remaining 30% of the docking model sets for testing.

Results of inter-chain interfaces scoring

Method	Data	Granularity	Components	Mean CAD-score
Ideal selector				1
Random				0.25
Pseudo energy	PDB	atom-atom	P(occur)	0.78
			P(persist|occur)	0.78
			P(occur)*P(persist|occur)	**0.89**
Pseudo energy	PDB	residue-residue	P(occur)	0.59
			P(persist|occur)	0.59
			P(occur)*P(persist|occur)	0.62
Pseudo energy	IDRome	residue-residue	P(occur)	0.50
			P(persist|occur)	0.53
			P(occur)*P(persist|occur)	0.55

Results of inter-chain interfaces scoring

Method	Data	Granularity	Components	Mean CAD-score
Ideal selector				1.00
Random				0.25
Pseudo energy	PDB	atom-atom	P(occur)	0.78
			P(persist|occur)	0.78
			P(occur)*P(persist|occur)	**0.89**
Pseudo energy	PDB	residue-residue	P(occur)	0.59
			P(persist|occur)	0.59
			P(occur)*P(persist|occur)	0.62
Pseudo energy	IDRome	residue-residue	P(occur)	0.50
			P(persist|occur)	0.53
			P(occur)*P(persist|occur)	0.55
VoroIF-GNN	all	hybrid	all	**0.98**

Results of inter-chain interfaces scoring

When there are no ideal models:

Method	Data	Granularity	Components	Mean CAD-score
Ideal selector				0.78
Random				0.23
Pseudo energy	PDB	atom-atom	P(occur)	0.60
			P(persist|occur)	0.59
			P(occur)*P(persist|occur)	**0.64**
Pseudo energy	PDB	residue-residue	P(occur)	0.54
			P(persist|occur)	0.52
			P(occur)*P(persist|occur)	0.55
Pseudo energy	IDRome	residue-residue	P(occur)	0.49
			P(persist|occur)	0.51
			P(occur)*P(persist|occur)	0.50
VoroIF-GNN	all	hybrid	all	**0.77**

Conclusion

- Ensembles of conformations from PDB provide useful information about contact stability, it can improve scoring protein-protein interfaces.
- Ensembles of simulated conformations of IDPs can also be useful as a source of statistics about tesselation-derived amino acid interactions.
- Different statistical descriptors can be efficiently employed using tessellation-based graph neural network.

Thank you!

CNRS Laboratoire Jean Kuntzmann:

- Sergei Grudinin

Useful links:

- https://www.voronota.com
- https://www.kliment.lt



Funded by
the European Union