

PAUL SCHERRER INSTITUT

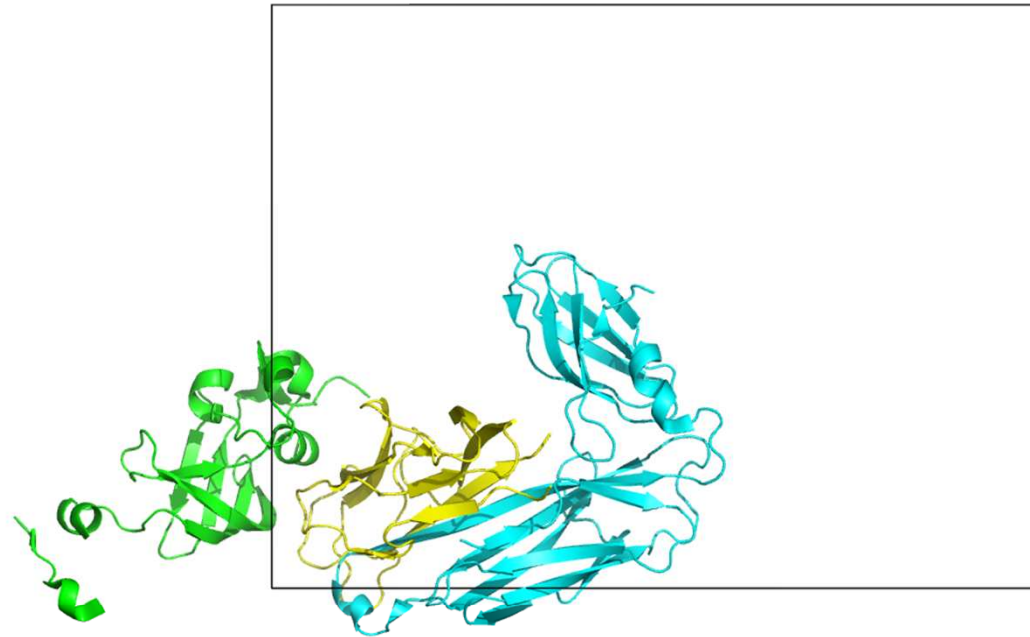


Distinguishing biologically relevant interfaces from lattice contacts in protein crystals

Guido Capitani

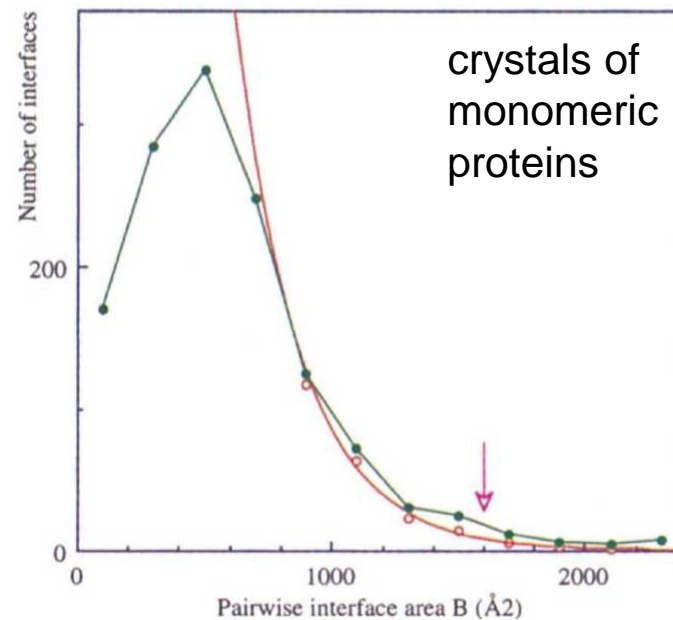
Paul Scherrer Institut, Villigen, Switzerland

RAMC2013 meeting, Le Bischenberg, 10 September 2013



Interface area (B) statistics

For $B > 700 \text{ \AA}^2$, ()

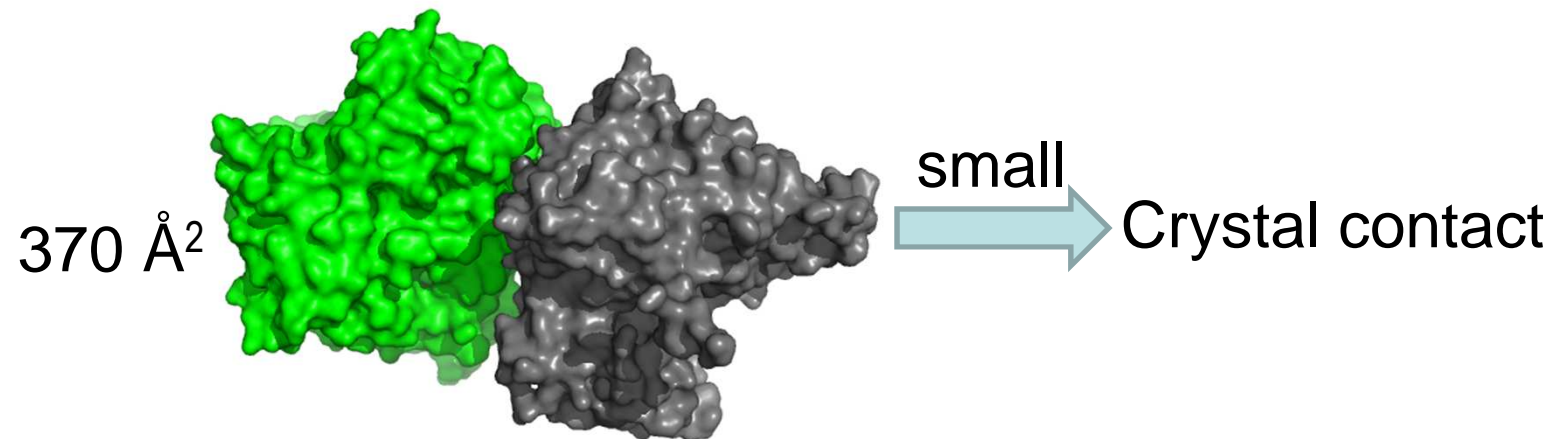
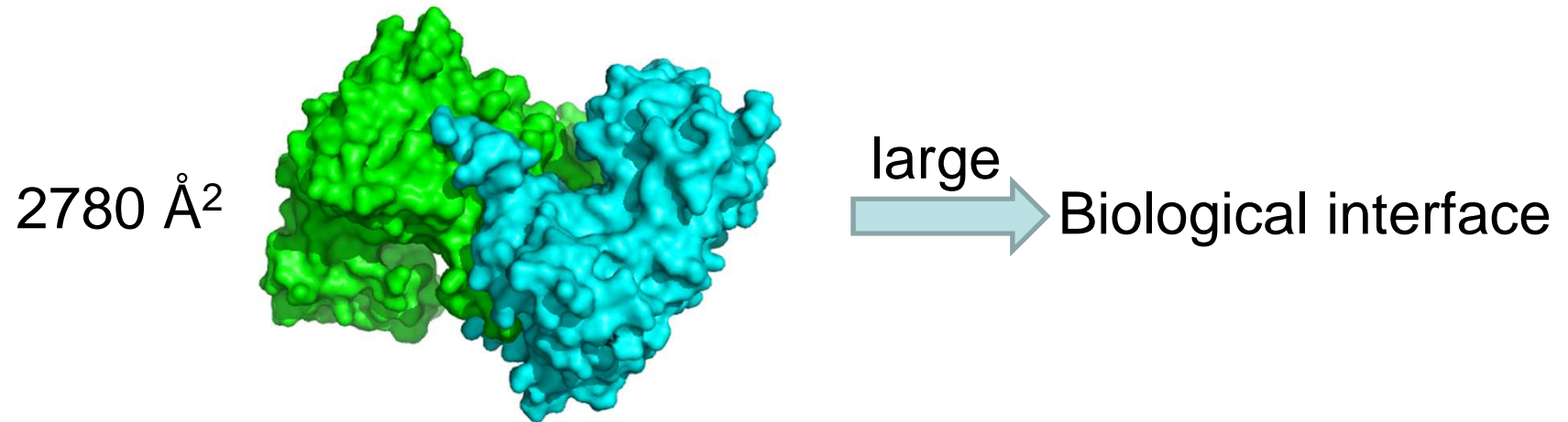


probability to find a non-specific interface (crystal contact) burying more than $B \text{ \AA}^2$

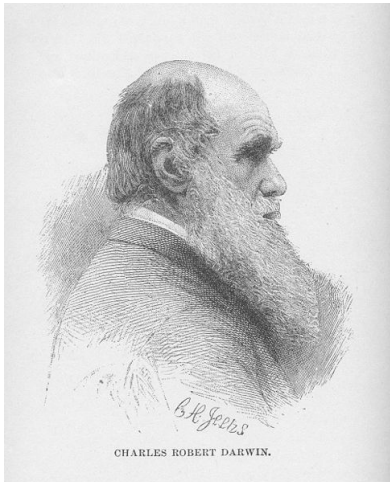
Example: for $B = 1000 \text{ \AA}^2$ $p \approx 9\%$

Bottom line: the **larger** the interface the more likely to be **bio**

Easy cases



Biologically relevant interfaces
are the result of **evolution**



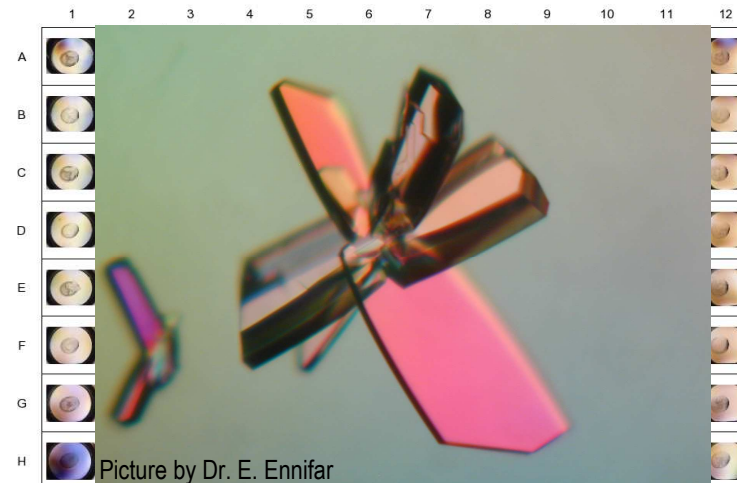
from Wikimedia Commons

Detectable signature of
selection pressure

Crystal contacts are not



from Wikimedia Commons



No detectable signature of
selection pressure

Entropy ratios for interface classification

Valdar & Thornton¹: comparison of selection pressure on **interface** versus **surface** residues. Similar approach at the same time by Elcock & McCammon²
Sequence entropy of MSA of homologues as metrics of selection pressure²:

```

05940_BOVIN -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_HUMAN -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_MOUSE -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_RAT -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_CHICK -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_BABY -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
Q7063_BABAE -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_CTRF -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_DROME -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_DICD -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
05180_DICDI -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_PEARF -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_SULAC -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_SULBO -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_SULSO -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_PYRAE -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_METAC -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_METMA -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_METRA -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_METTH -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_METTI -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_METYA -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_METJA -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_PYRAB -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_PYRBO -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_PYRFO -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_PYRKO -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_PALAN -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_PALYO -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_PALAS -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_THIAC -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_THIYO -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
HLAQ_PICTO -----HAEHREHVRVRELELLELDQCFIVAVGAGVQVQVHHELEKAVVLEGGVNDKAEKQVDR--FAE 76
  
```

$$s(i) = - \sum_k p_i(k) \log(p_i(k))$$

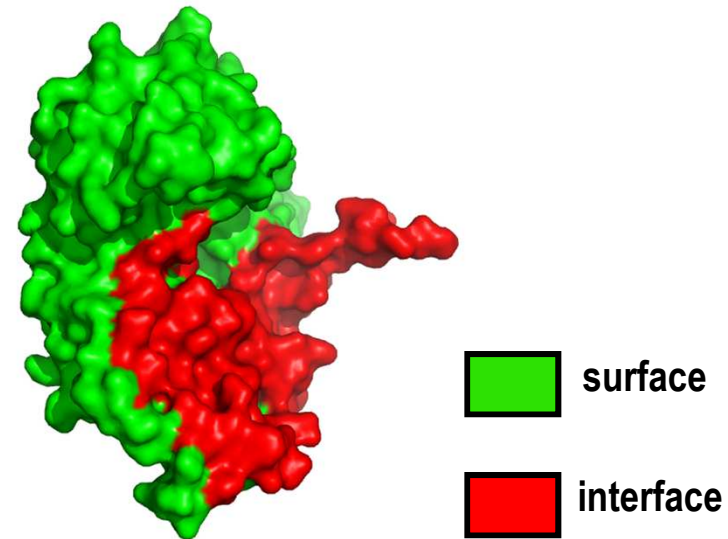
$p_i(k) \rightarrow$ probability of residue of type k at pos. i

For **biological** interface:

$$\frac{\langle S \rangle_{interface}}{\langle S \rangle_{surface}} < 1$$

For **crystal** contact:

$$\frac{\langle S \rangle_{interface}}{\langle S \rangle_{surface}} > 1$$

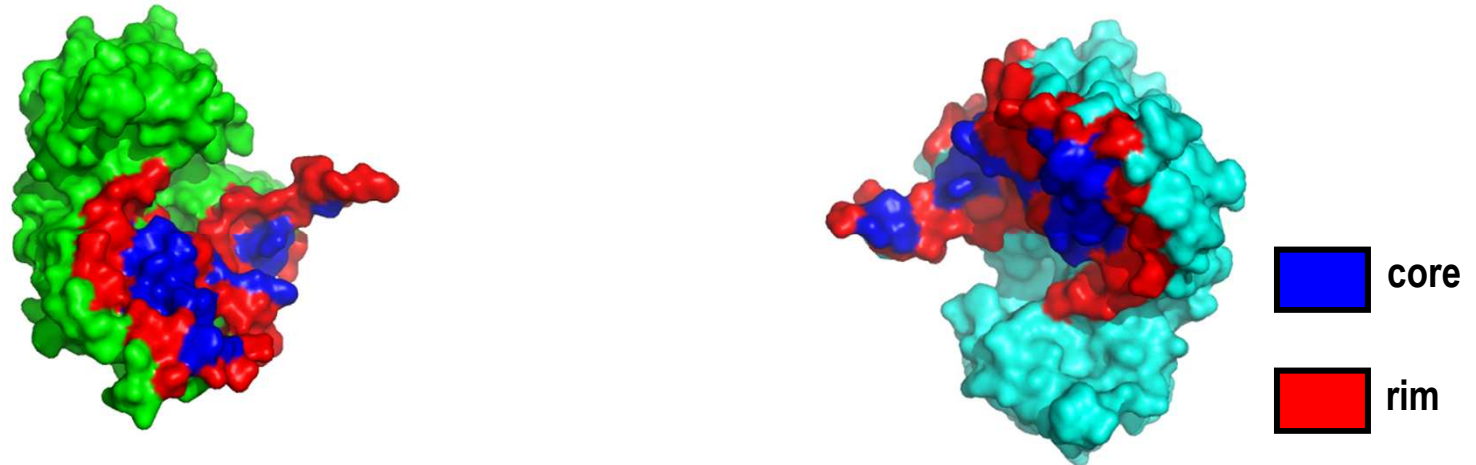


¹Valdar & Thornton, *J Mol Biol*, 2001

²Elcock & McCammon, *PNAS*, 2001

Entropy ratio of interface core and rim

Guharoy & Chakrabarti^{1,2}: Core: interface residues with at least one fully buried atom
Rim: the other interface residues



For **biological** interface: $\frac{\langle S \rangle_{core}}{\langle S \rangle_{rim}} < 1$

For **crystal** contact: $\frac{\langle S \rangle_{core}}{\langle S \rangle_{rim}} > 1$

Need for new datasets for method development

“Classical” datasets contain

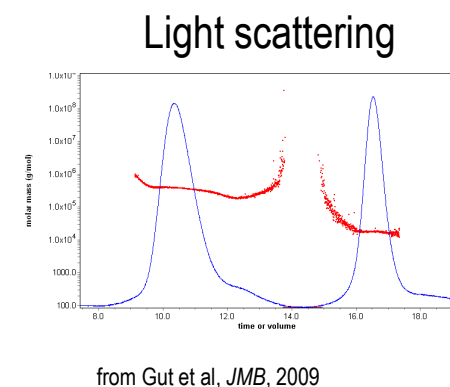
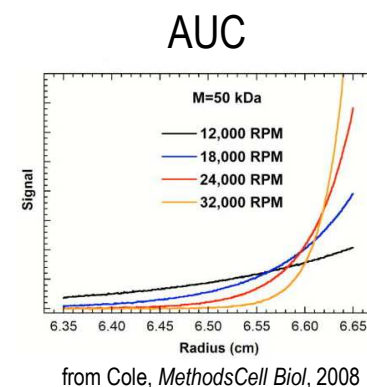
- Too many “easy-to-predict” (obvious by area) interfaces
- Many old structures with no R-free statistics and no deposited structure factors

Two new datasets¹:

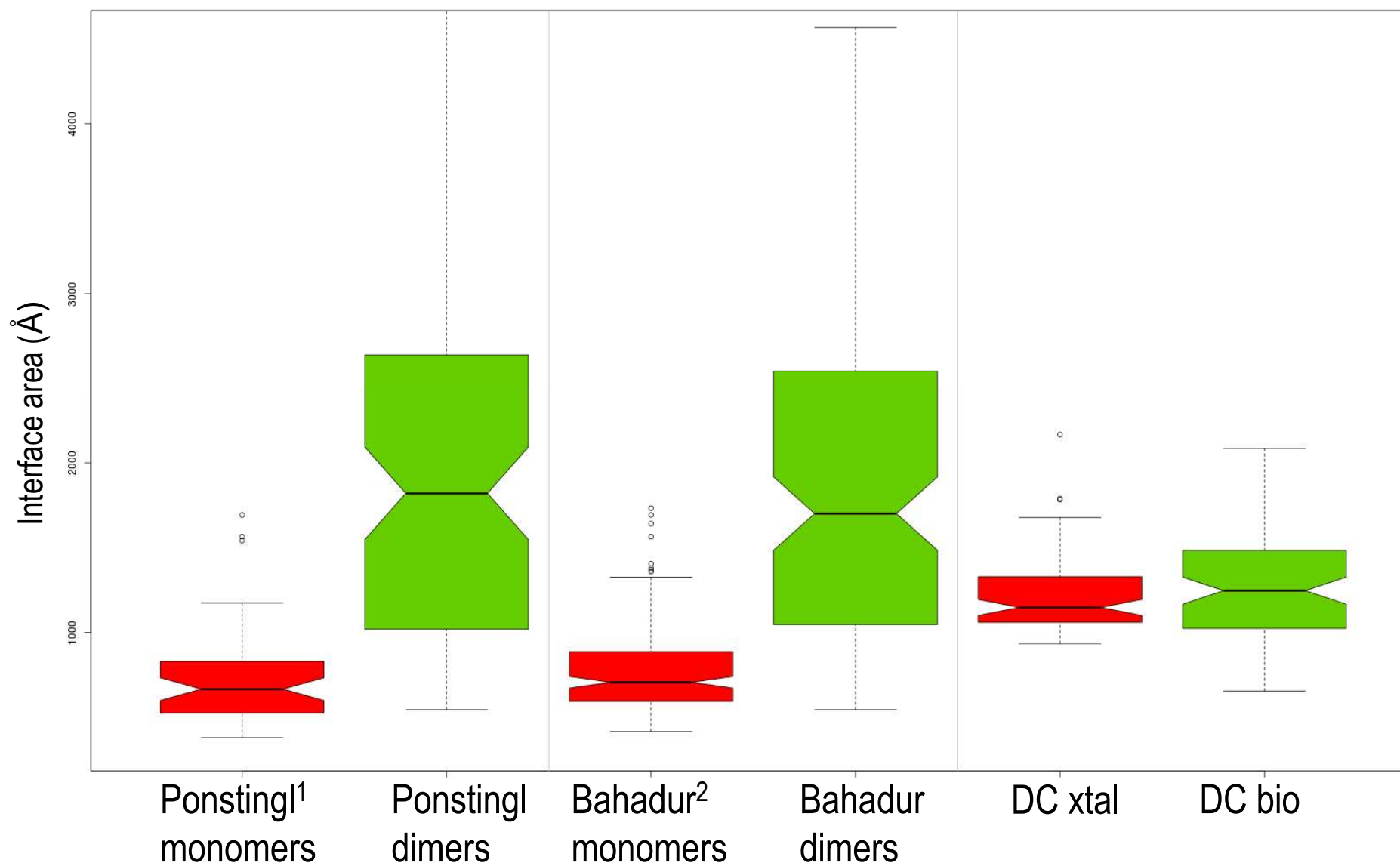
- **DC xtal**: large crystal contacts, with BSA 1000 Å² and above
- **DC bio**: small biological interfaces

Criteria:

- Good quality structures by strict filtering (resolution, R-free, data deposition)
- Clear **biophysical evidence** for oligomeric state in solution from literature
- No domain swaps or other ambiguous cases



Area distributions of different interface datasets

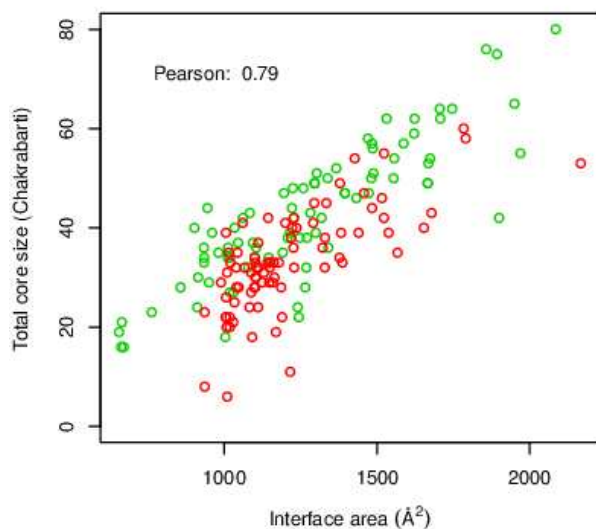


Core size: a geometric predictor

Core size: # of core residues

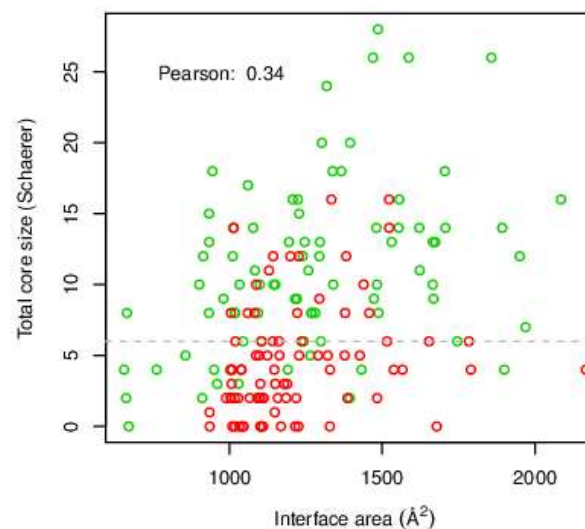
- DC xtal set
- DC bio set

Our definition



Chakrabarti & Janin, *Proteins*, 2002

Core residues: those with >1 fully buried atom



Schärer et al, *Proteins*, 2010

Core residues: the fully buried ones

Core vs rim:

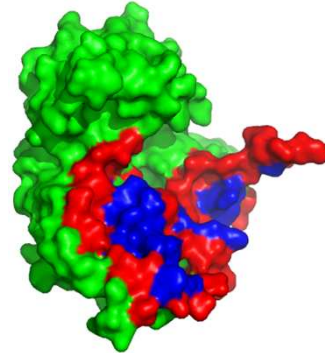
Introduced by Guharoy and Chakrabarti (2005)

Our core definition (2010)

- Measure: ratio

$$\frac{\langle S_{core} \rangle}{\langle S_{rim} \rangle} < \text{threshold: bio}$$

$$\frac{\langle S_{core} \rangle}{\langle S_{rim} \rangle} > \text{threshold: xtal}$$



Interface vs surface:

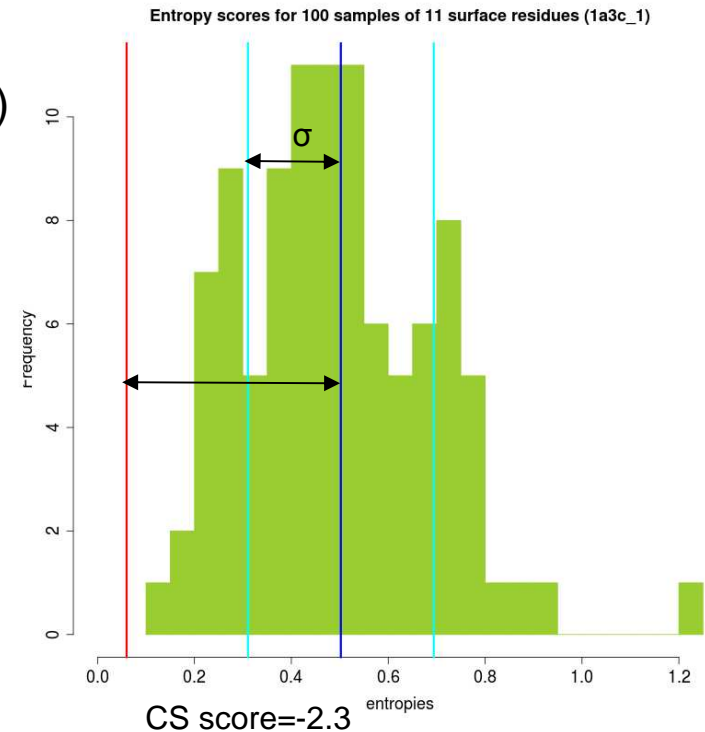
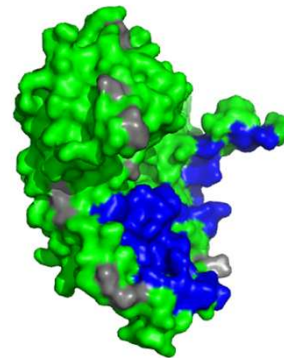
Introduced by Valdar/Thornton, Elcock/McCammon (2001)

Our modifications:

- **core** vs. surface (CS score)
- our core definition
- measure:

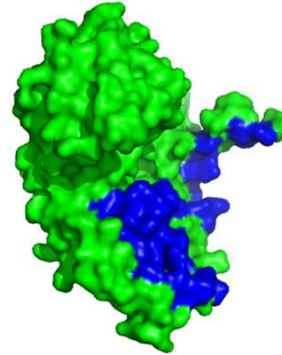
$$\frac{\langle S_{core} \rangle - \mu_{surface}}{\sigma_{surface}} < \text{threshold: bio}$$

$$\frac{\langle S_{core} \rangle - \mu_{surface}}{\sigma_{surface}} > \text{threshold: xtal}$$



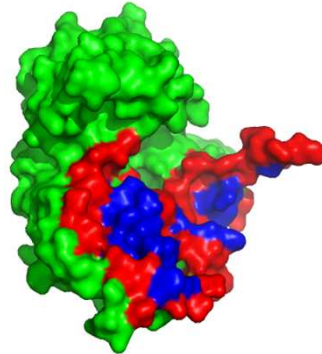
1) Core size:

of core residues < threshold1: xtal
> threshold1: bio



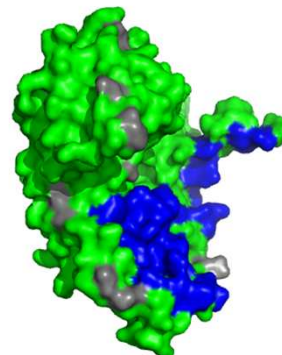
2) Core vs. rim:

$\frac{\langle S_{core} \rangle}{\langle S_{rim} \rangle} < \text{threshold2: bio}$
 $> \text{threshold2: xtal}$



3) Core vs. surface:

$\frac{\langle S_{core} \rangle - \mu_{surface}}{\sigma_{surface}} < \text{threshold3: bio}$
 $> \text{threshold3: xtal}$



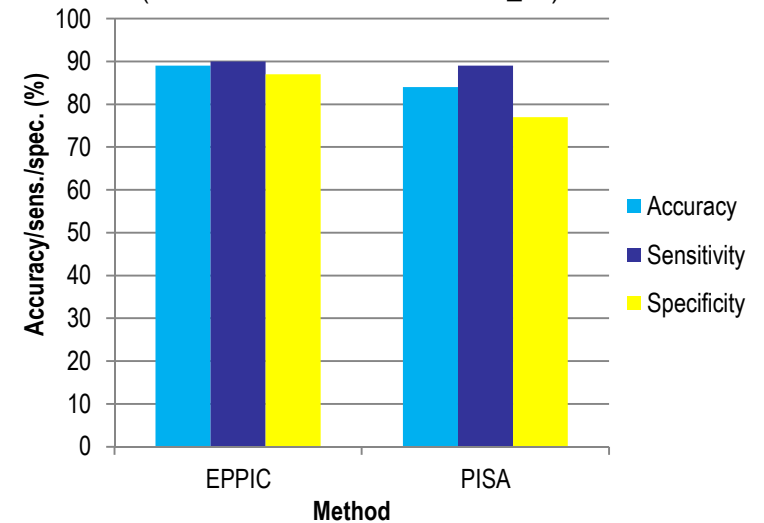
Final call:

by majority of the three criteria

Thresholds chosen by systematic runs vs DCbio and DCxtal

Benchmarking:

(EPPIC based on UniProt 2012_10)



Dataset: Ponstingl

- Input: PDB entry code or PDB file uploaded by the user
- Rich web application (GWT/ext-js)

The screenshot shows the EPPIC web application interface in a browser window. The browser title is "EPPIC - Input" and the address bar shows "www.eppic-web.org/ewui/#". The interface features a "My Jobs" sidebar on the left with a "New job" button and a table listing jobs. The main content area displays the "eppic" logo and a form for inputting a PDB code or uploading a file. The form includes a radio button for "PDB code" (selected), a text input field containing "3d36", and an "Example: 1ynu" label. Below the input field is an "e-mail (optional):" field. A collapsed "Advanced" section is visible below the email field. At the bottom of the form are "Reset" and "Submit" buttons. A faint protein structure is visible in the background. At the bottom of the page, there is a "NEWS" section stating "Version 2.0.2 out now + Precomputed results now available for the whole PDB" and a citation: "If you use this service for your research please cite : Duarte et al., BMC Bioinformatics 2012". The footer contains navigation links: "Home | Downloads | Help | Release log | Disclaimer | About us". The status bar at the bottom left shows "Status: Ok".

My Jobs

New job

Input data	Status
2trx	Finished

eppic

PDB code Upload file

PDB code:

Example: 1ynu

e-mail (optional):

Advanced

Reset Submit

NEWS Version 2.0.2 out now + Precomputed results now available for the whole PDB

If you use this service for your research please cite : Duarte et al., BMC Bioinformatics 2012

Status: Ok

Home | Downloads | Help | Release log | Disclaimer | About us

- Output: lists all interfaces with main features and predictions (bio or xtal)

Web server: output page for PDB entry 3d36

Jobs panel

Interface thumbnail:
click to show in 3D
(Jmol or PyMOL)

Interface symmetry operator

Interface analysis of: 3d36 (2.03Å - P 32 2 1)
How to Switch Off a Histidine Kinase: Crystal Structure of Geobacillus stearothermophilus KinB with the Inhibitor Sda

Chain A (B) (G8N7D7) 12 homologs
Chain C (A4IR54) 41 homologs

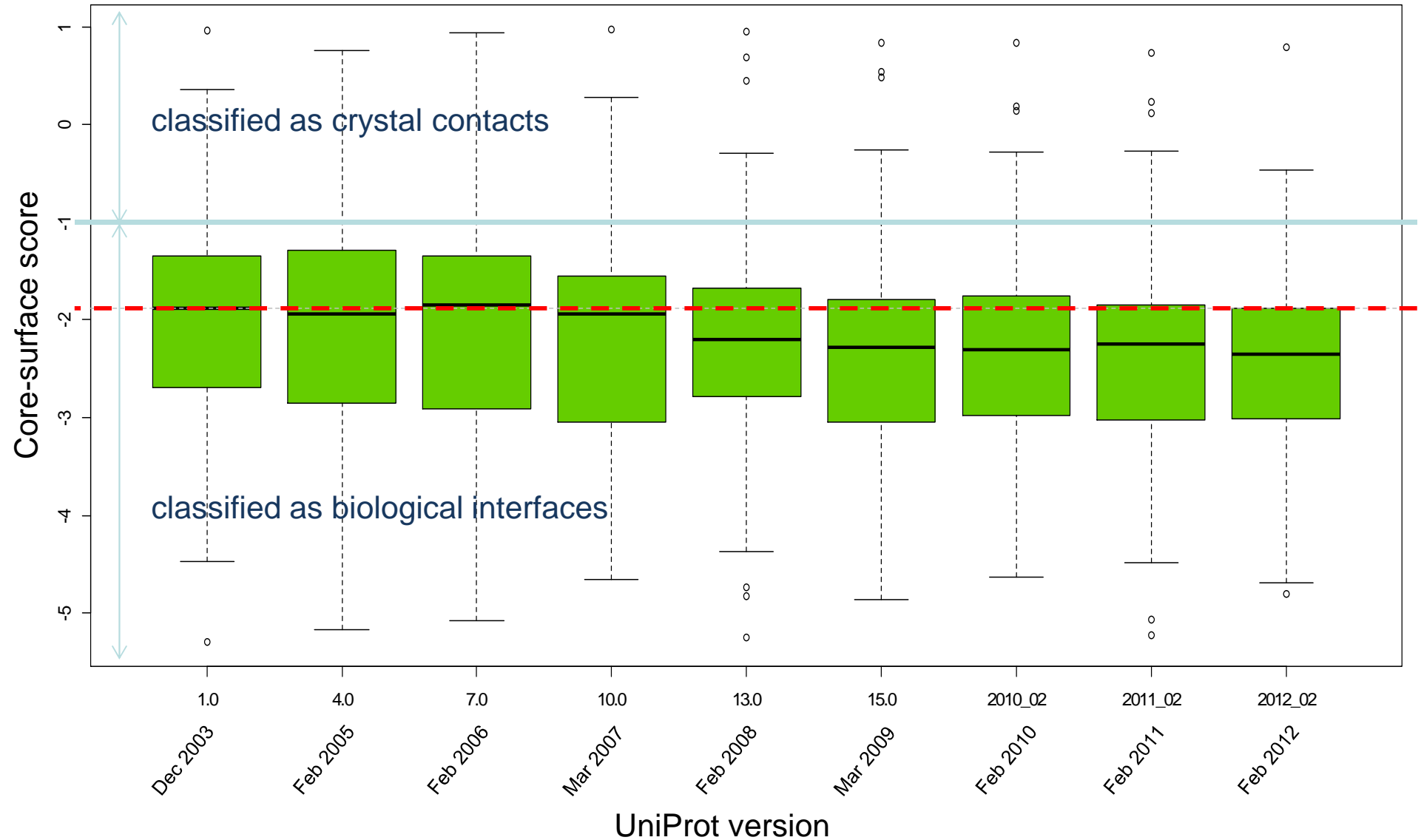
Input parameters: UniProt: 2013_07, EPPIC: 2.0.2, Download

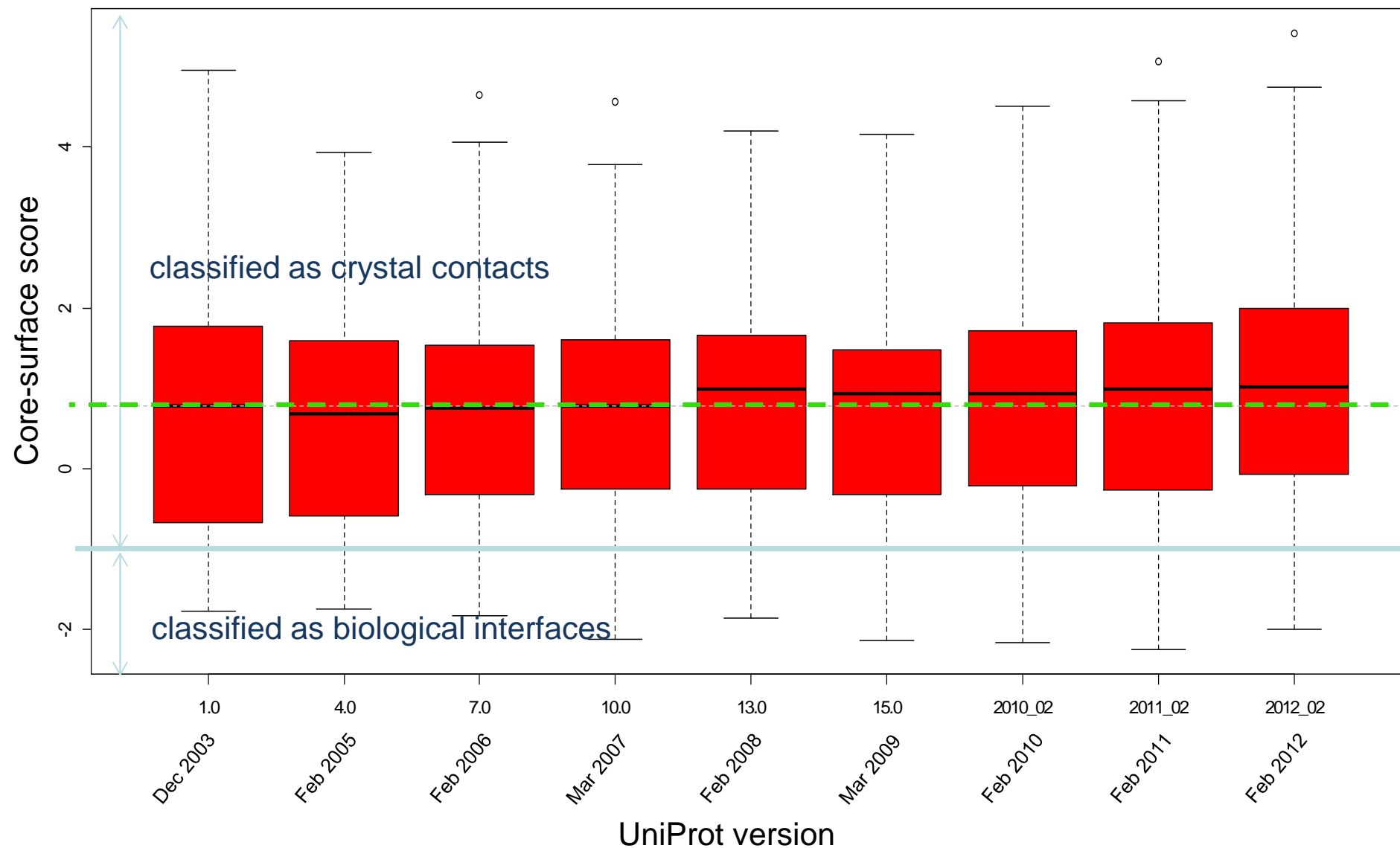
3D Viewer: Local Show thumbnails

	Id	Chains	Area (Å ²)	Operator	Core Sizes	Geometry	Core-Rim	Core-Surface	Final	
	1	B+A	1707.71	I	4 + 6	bio	bio	bio	bio	Details ⚠
	2	B+B	1112.52	⬤	2 + 2	xtal	bio	bio	bio	Details ⚠
	3	C+A	665.13	I	2 + 4	bio	bio	bio	bio	Details ⚠
	4	A+A	380.38	⬤	0 + 0	xtal	nopred	nopred	xtal	Details ⚠
	5	B+C	227.41	⬤	0 + 0	xtal	nopred	nopred	xtal	Details ⚠
	6	A+B	204.65	→	0 + 0	xtal	nopred	nopred	xtal	Details ⚠
	7	C+B	199.16	→	0 + 0	xtal	nopred	nopred	xtal	Details ⚠
	8	A+B	159.92	⬤	0 + 0	xtal	nopred	nopred	xtal	Details ⚠

Status: Ok

Home | Downloads | Help | Release log | Disclaimer | About us





Conclusions and outlook

- **Core size** (# of core residues by our definition) is an important geometric determinant of bio interfaces: good interface packing is essential for a bio interface
 - **Sequence entropy** can be used with satisfactory performance by combining **core/rim** ratio and **core/surface** score and using close homologs only
 - **Evolution-based method**: results can only improve with sequence database growth (numerically shown for 2002-2012)
 - Implemented in a robust, open-source **software** package and web server:
www.eppic-web.org
 - The method works satisfactorily on **membrane proteins** as well (Duarte *et al.*, 2013, under review)
 - In the works/outlook:
 - Prediction of **biological assemblies** based on interface calls and symmetry
 - PDB-wide statistical analysis of bio and xtal interfaces
-

Acknowledgements

**Paul Scherrer Institut
(Biology and Chemistry, LBR)**

Jose M. Duarte
Kumaran Baskaran
Nikhil Biyani
Martin A. Schärer

Bela Hullar (SyBIT/ETH Zurich)
Adam Srebniak (SyBIT/ETH Zurich)

Derek Feichtinger & Valeri Markushin
(SciComp group, PSI)

Funding and support:

Forschungskommission PSI



University of Zurich

Prof. Amedeo Caflisch
Prof. Andreas Wagner

