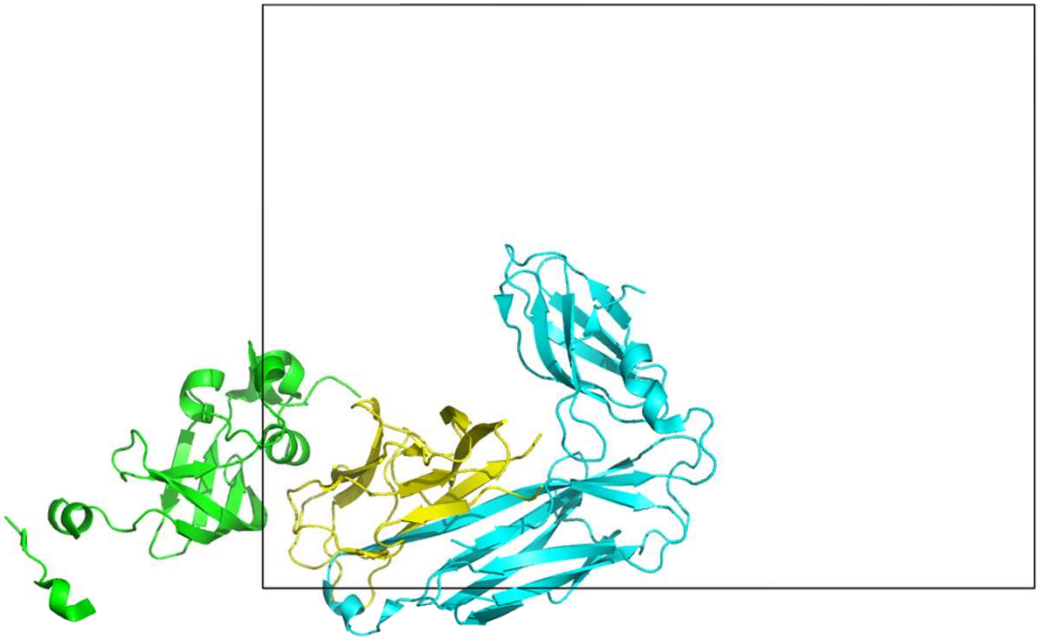
PAUL SCHERRER INSTITUT

Distinguishing biologically relevant interfaces from lattice contacts in protein crystals

Guido Capitani
Paul Scherrer Institut, Villigen, Switzerland

RAMC2013 meeting, Le Bischenberg, 10 September 2013

Crystal packing and biologically relevant interfaces



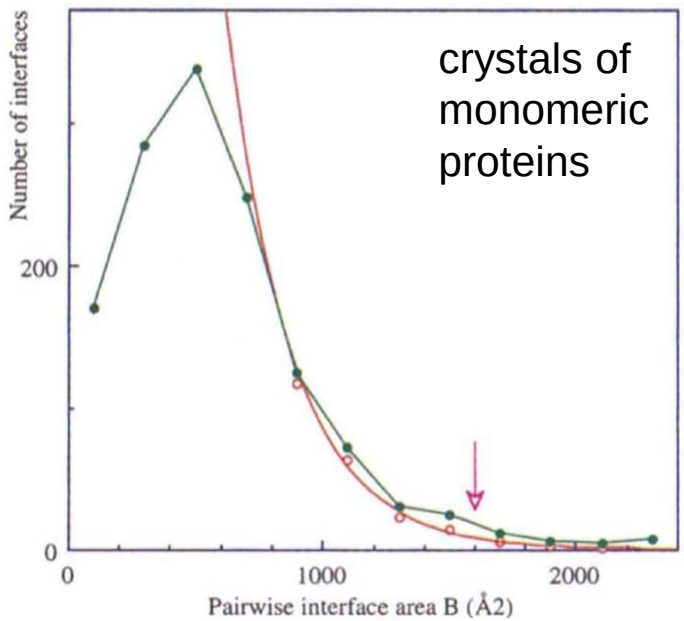
Nishiyama et al., *EMBO J.*, 2005

The first approach: interface area

Interface area (*B*) statistics

For $B > 700$ Å^2, ( )

probability to find a non-specific interface (crystal contact) burying more than B Å^2



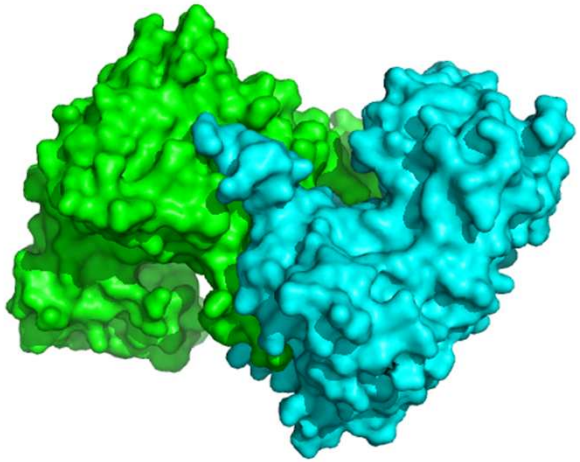
Example: for $B = 1000$ Å^2 $p \approx 9\%$

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

Janin, *Nat Struct Biol*, 1997

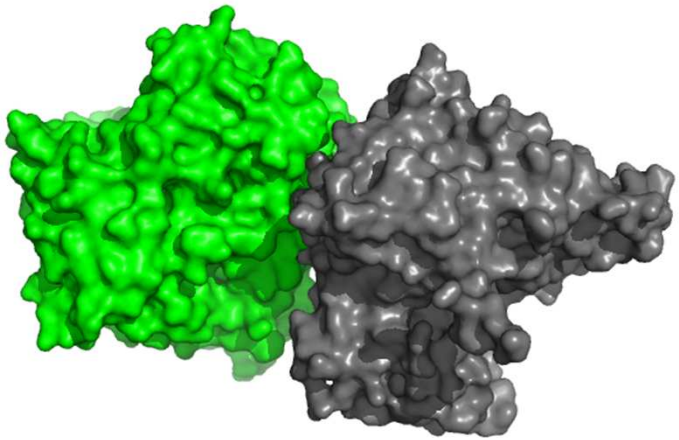
Easy cases

2780 Å^2



large → Biological interface

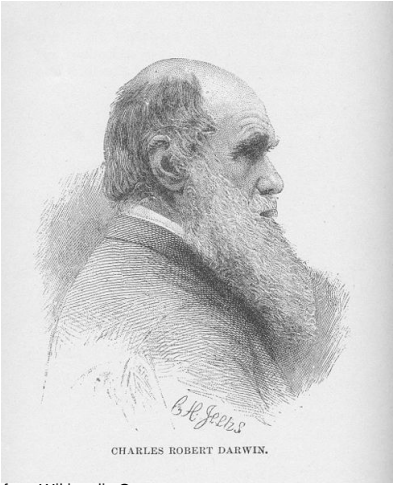
370 Å^2



small → Crystal contact

Evolutionary approach to the problem

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

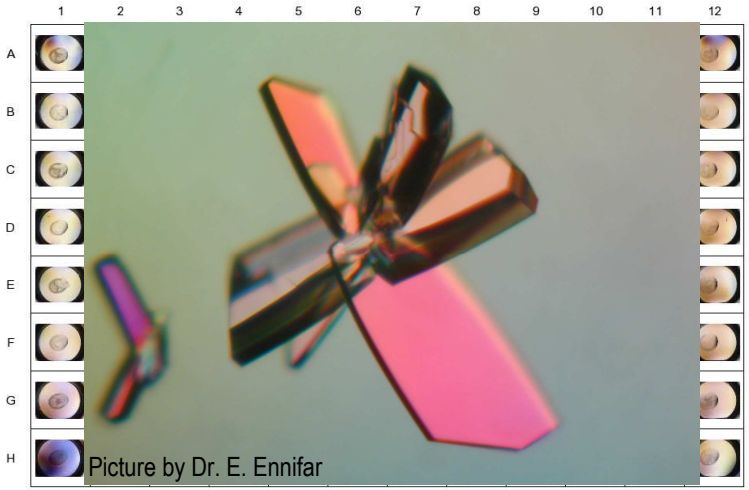


from Wikimedia Commons

Detectable signature of selection pressure

Crystal contacts are not

from Wikimedia Commons

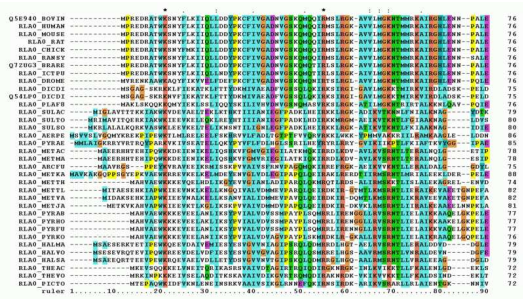


No detectable signature of selection pressure

Entropy ratios for interface classification

Valdar & Thornton[1]: comparison of selection pressure on **interface** versus **surface** residues. Similar approach at the same time by Elcock & McCammon[2]

Sequence entropy of MSA of homologues as metrics of selection pressure[2]:

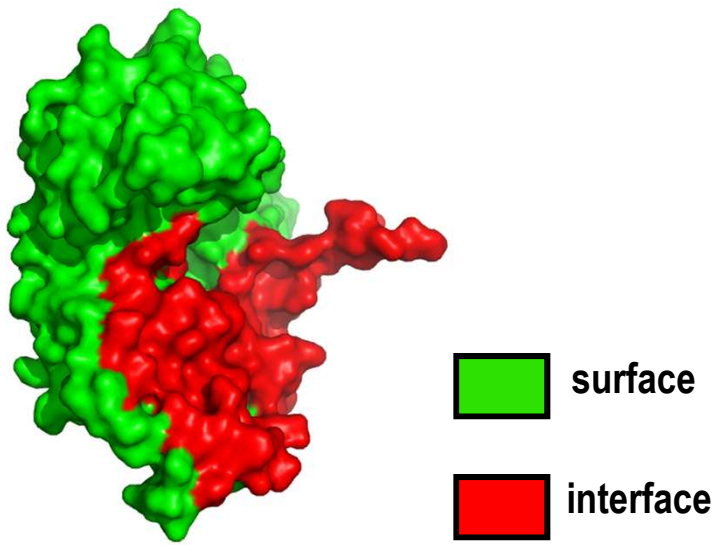


$$s(i) = -\sum_k p_i(k)\log(p_i(k))$$

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

For **biological** interface: $\frac{<s>_{interface}}{<s>_{surface}} < 1$

For **crystal** contact: $\frac{<s>_{interface}}{<s>_{surface}} > 1$

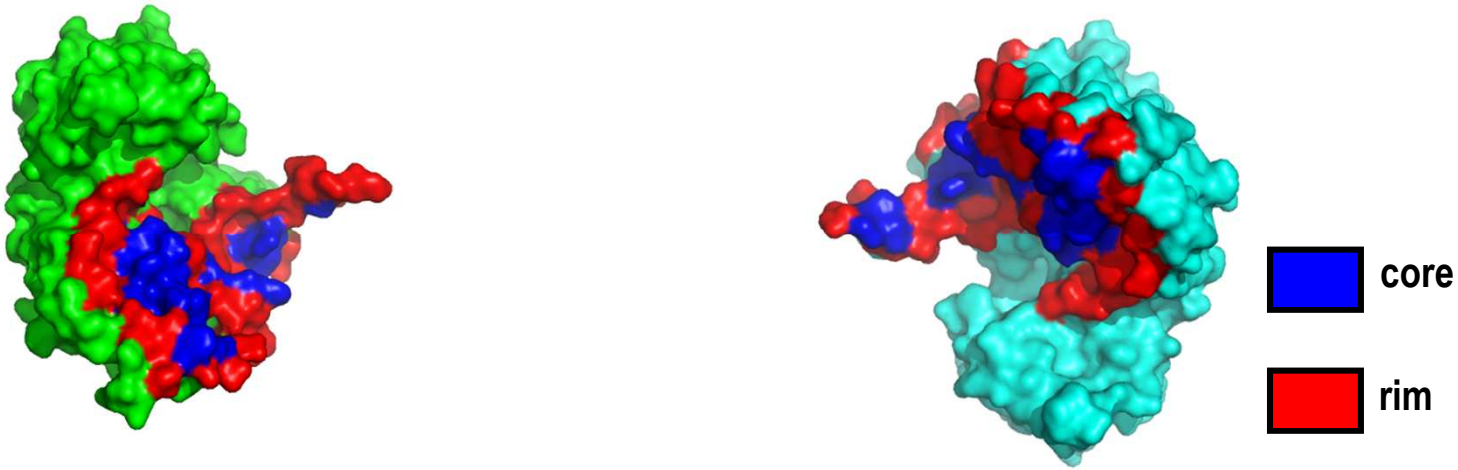


[1]Valdar & Thornton, *J Mol Biol*, 2001

[2]Elcock & 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: $\frac{<s>_{core}}{<s>_{rim}} < 1$

For **crystal** contact: $\frac{<s>_{core}}{<s>_{rim}} > 1$

[1]Guharoy & Chakrabarti, *PNAS*, 2005 [2]Chakrabarti & Janin, *Proteins*, 2002

New datasets: DCxtal and DCbio

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

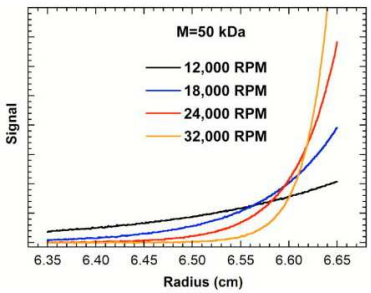
Two new datasets[1]:

- **DC xtal**: large crystal contacts, with BSA 1000 Å^2 and above
- **DC bio**: small biological interfaces

Criteria:

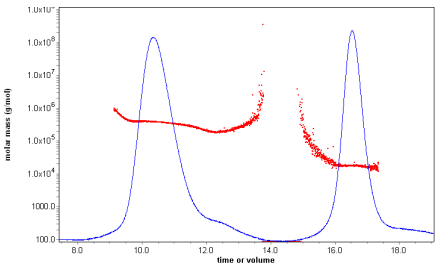
- Good quality structures by strict filtering (resolution, R-free, data deposition)
- Clear **biophysical evidence** for oligomeric state in solution from literature
- No domain swaps or other ambiguous cases

AUC



from Cole, *MethodsCell Biol*, 2008

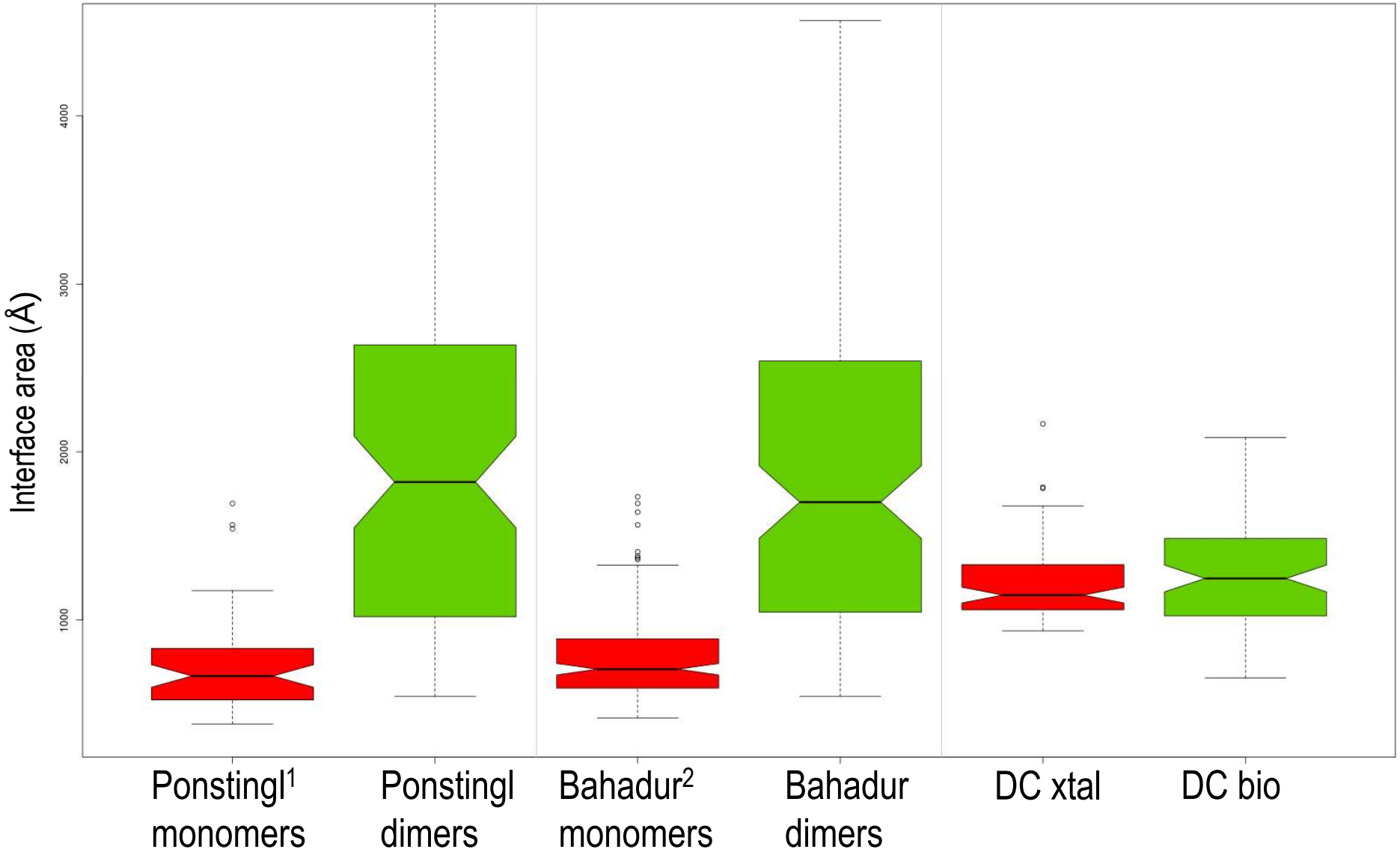
Light scattering



from Gut et al, *JMB*, 2009

[1]Duarte et al., *BMC Bioinformatics*, 2012

Area distributions of different interface datasets



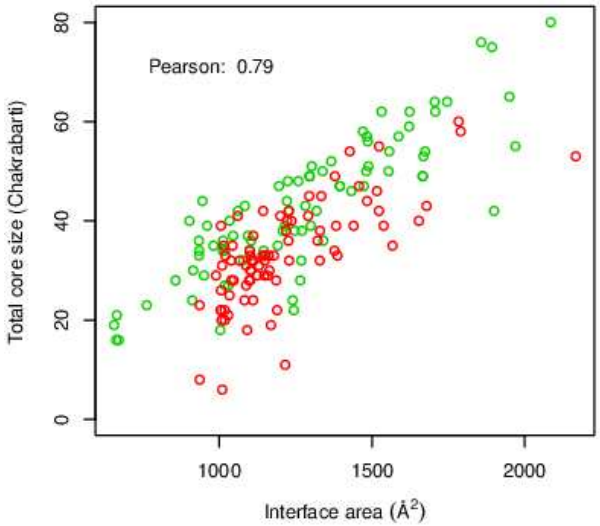
[1]Ponstingl et al., *J Appl Cryst,* 2003 [2]Bahadur et al., *J Mol Biol,* 2004

Core size: a geometric predictor

Core size: # of core residues

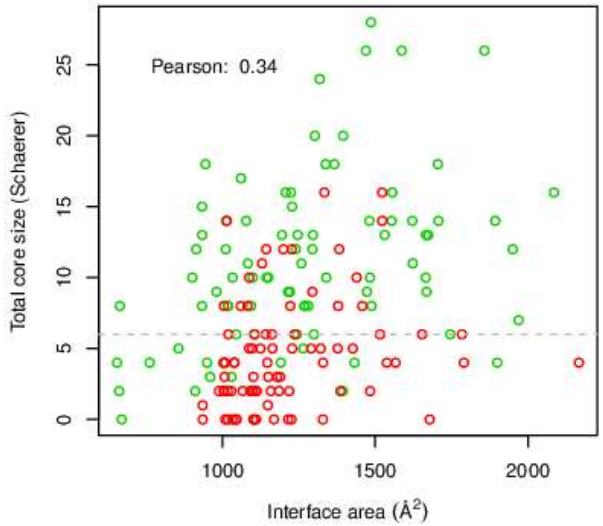
DC xtal set
DC bio set

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**

Our entropy-based interface scoring indicators

Core vs rim:

Introduced by Guharoy and Chakrabarti (2005)

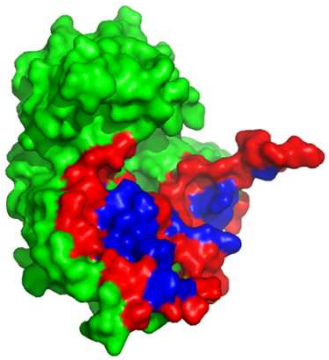
Our core definition (2010)

- Measure: ratio

$$\frac{\langle s_{core} \rangle}{\langle s_{rim} \rangle}$$

< threshold: bio

> threshold: xtal



Interface vs surface:

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

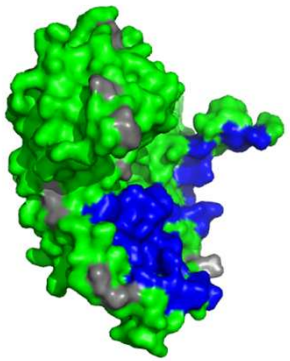
Our modifications:

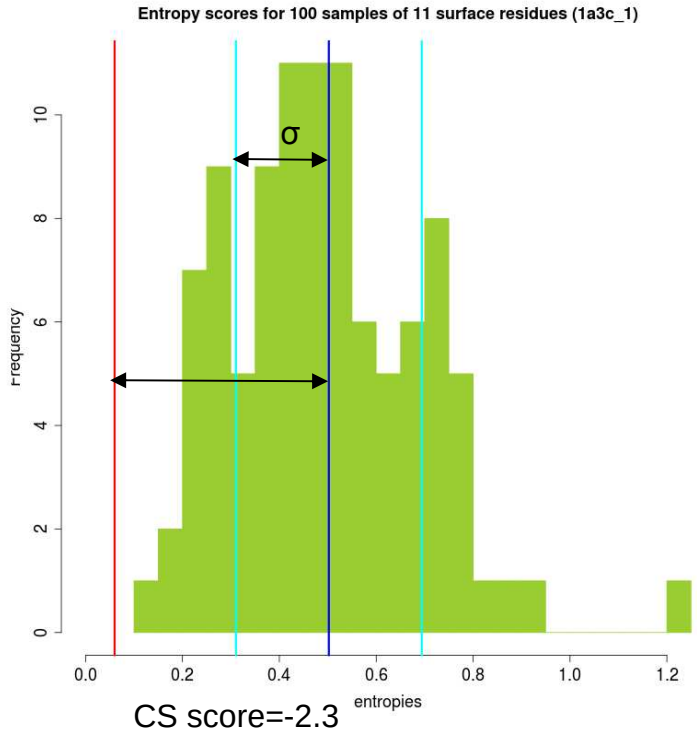
- **core** vs. surface (CS score)
- our core definition
- measure:

$$\frac{\langle s_{core} \rangle - \mu_{surface}}{\sigma_{surface}}$$

< threshold: bio

> threshold: xtal





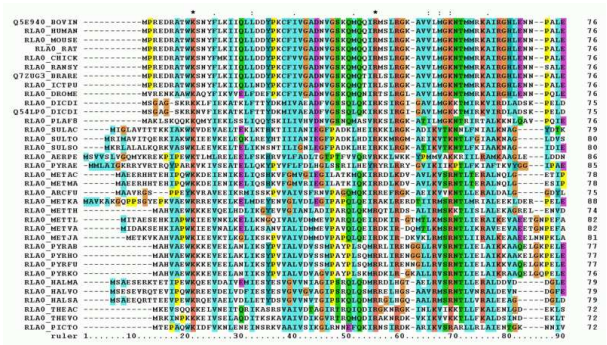
CS score=-2.3

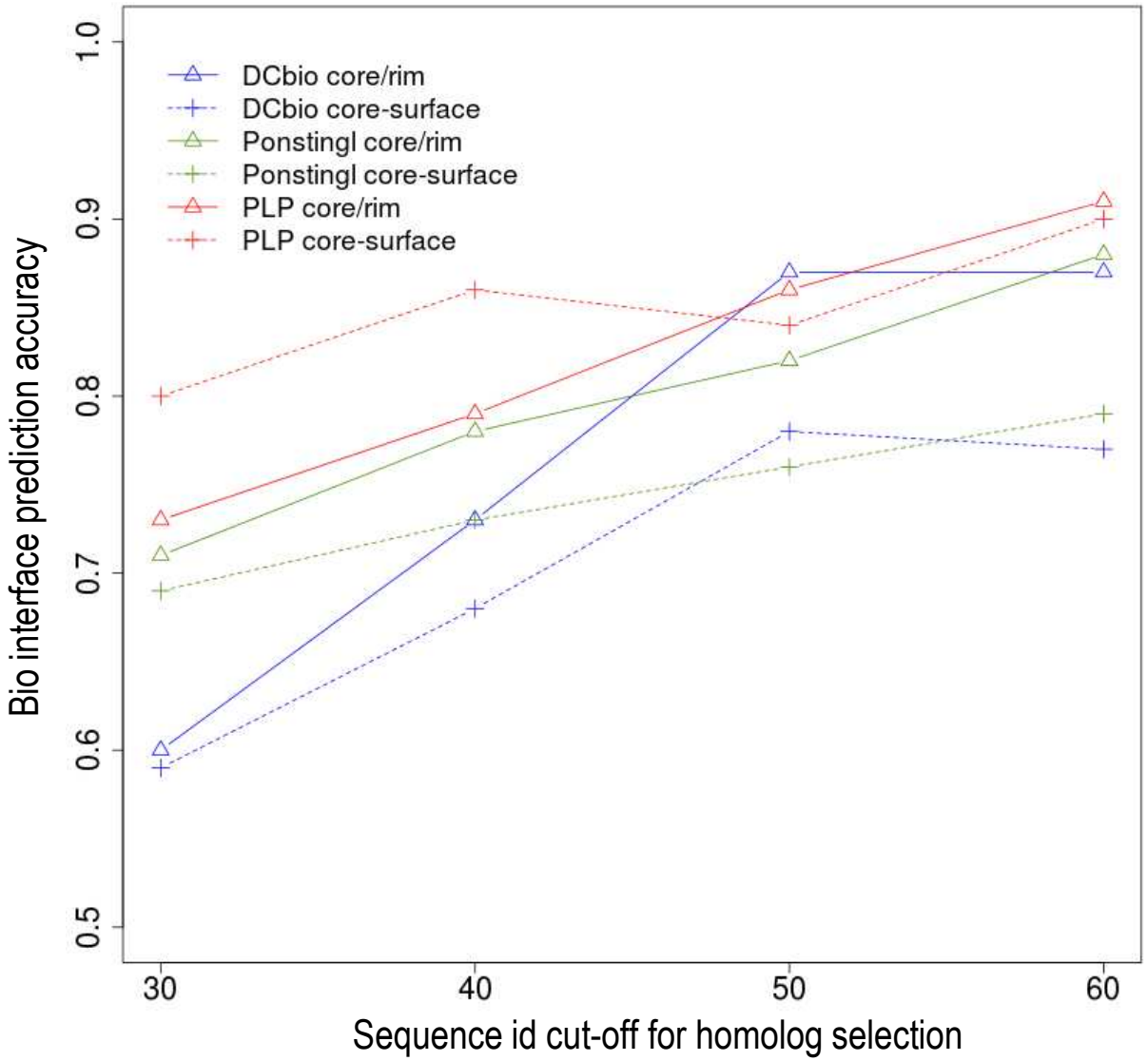
Homolog selection for MSA

Most approaches until now try to pool in as much data as possible: alignments include low sequence identity homologs

We take a conservative approach (now feasible): less data to get better accuracy

Redundancy reduction: UniRef100 and sequence clustering per alignment

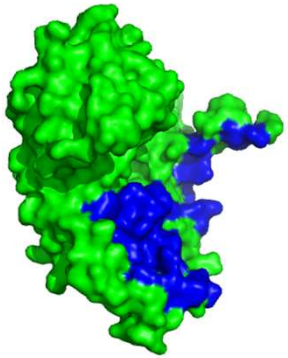




Combining geometry and evolutionary information

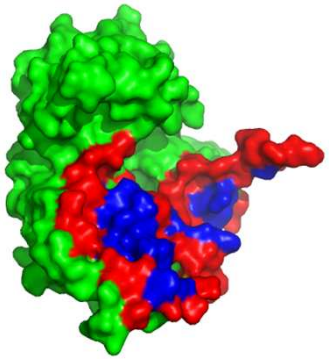
1) Core size:

# of core residues < threshold1: xtal
> threshold1: bio



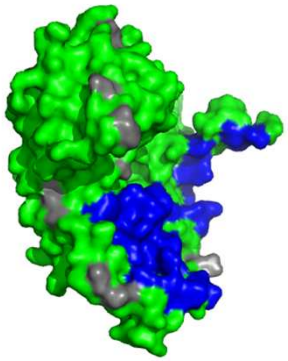
2) Core vs. rim:

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



3) Core vs. surface:

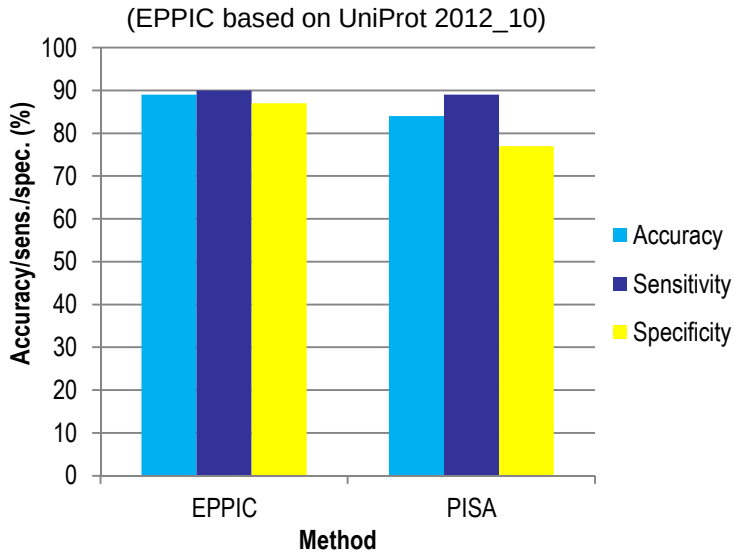
$$\frac{\langle s_{core} \rangle - \mu_{surface}}{\sigma_{surface}}$$ < threshold3: bio
> threshold3: xtal



Final call:
by majority of the three criteria

Thresholds chosen by systematic runs vs DCbio and DCxtal

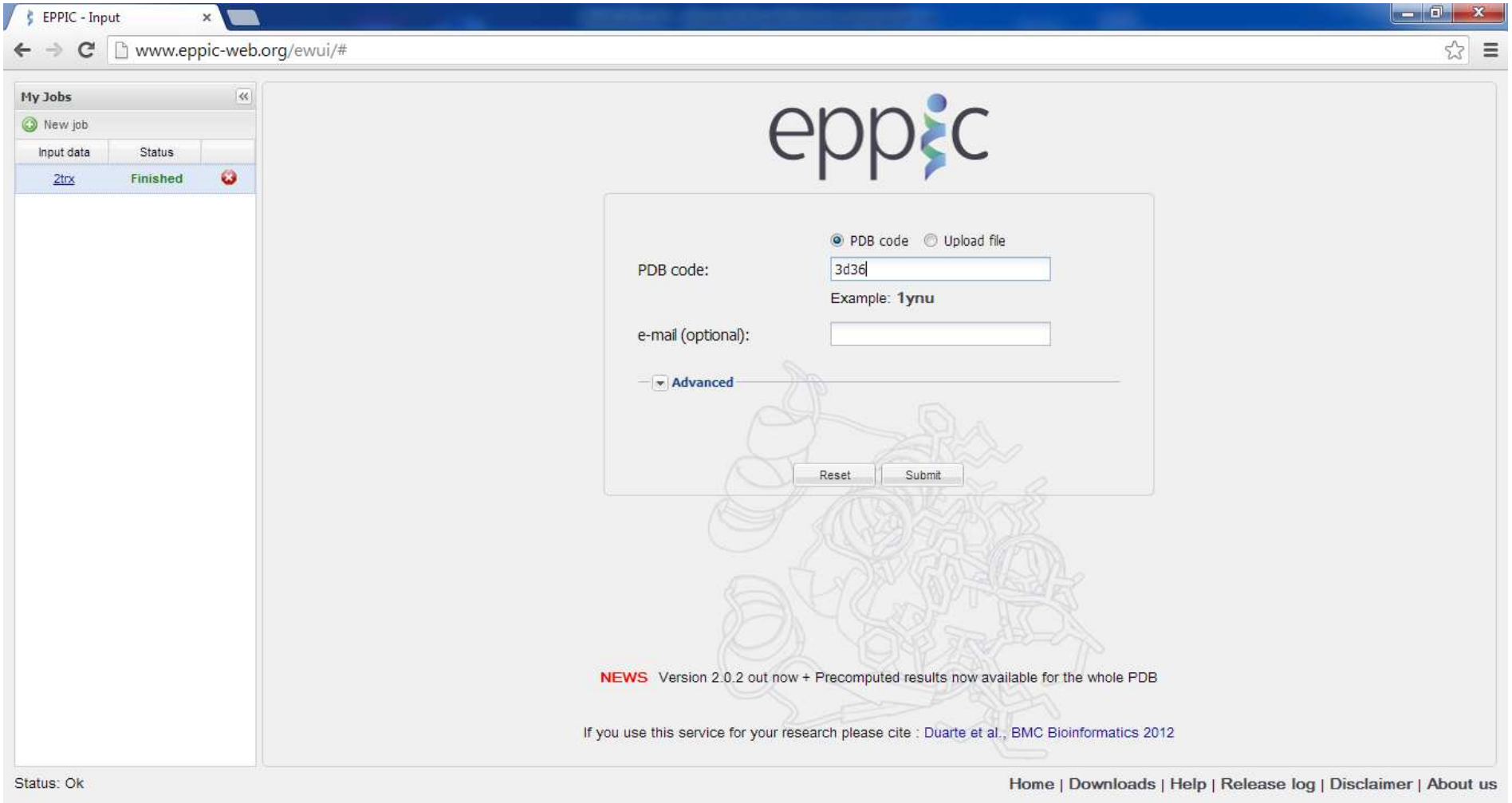
Benchmarking:



Dataset: Ponstingl

Web server: www.eppic-web.org

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



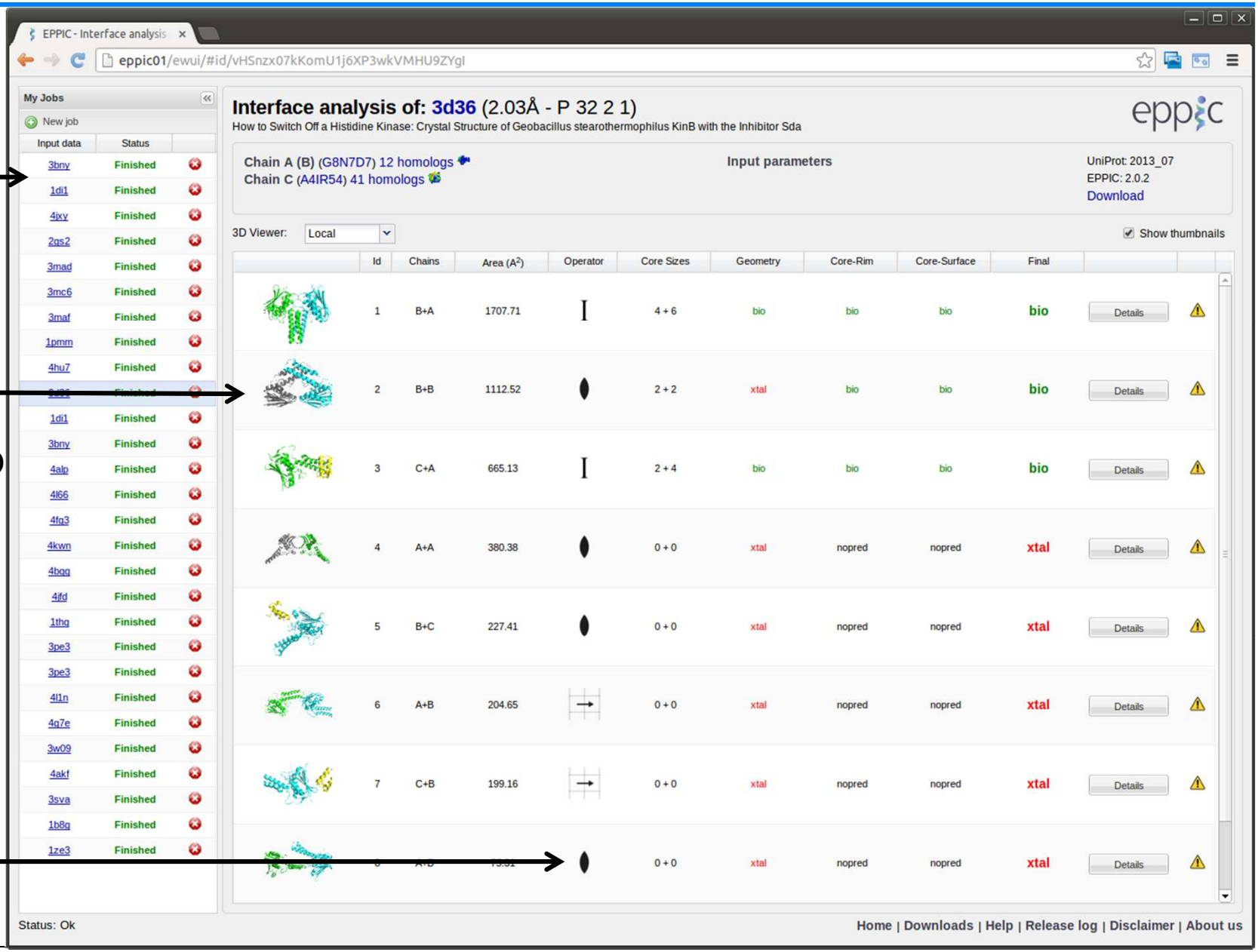
- Output: lists all interfaces with main features and predictions (bio or xtal)

Web server: output page for PDB entry 3d36

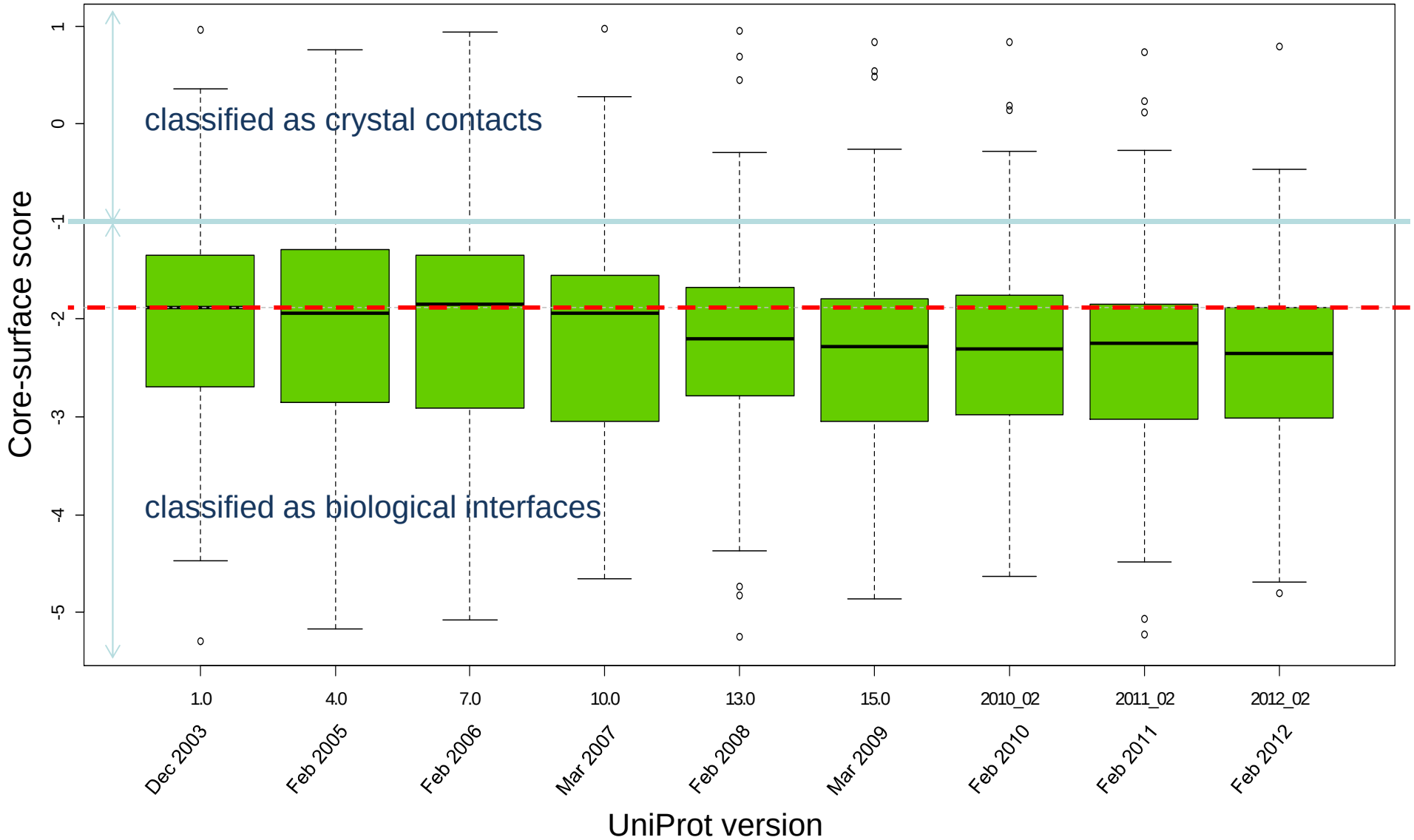
Jobs panel

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

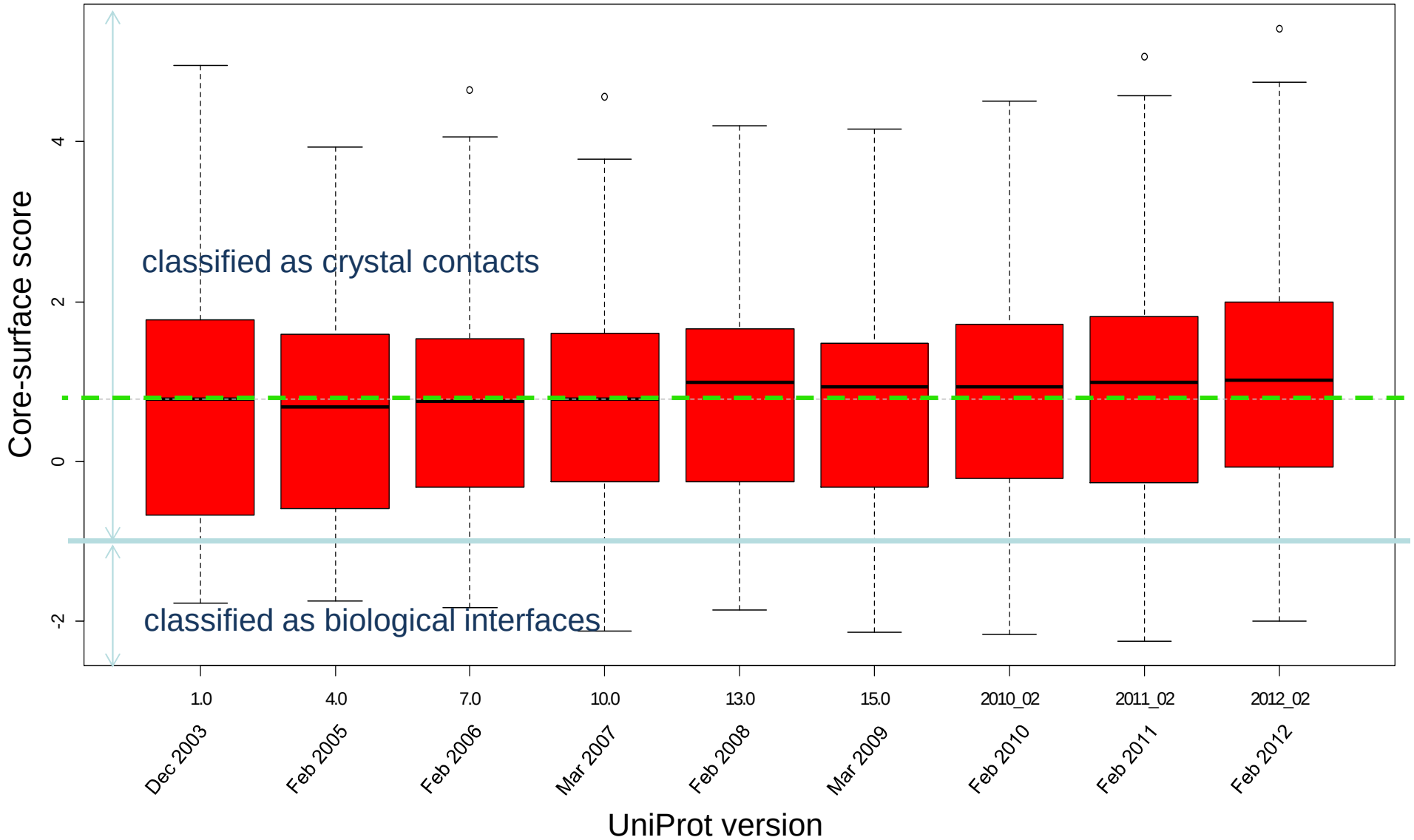
Interface symmetry operator



Bio interfaces: core-surface scores over the years



Xtal interfaces: core-surface scores over the years



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)

University of Zurich

Prof. Amedeo Caflisch
Prof. Andreas Wagner

Funding and support:

Forschungskommission PSI

SyBIT
SystemsX.ch
Biology IT

FNSNF
Swiss National Science Foundation