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 \(B > 700 \text{ Å}^2\),

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

Example: for \(B = 1000 \text{ Å}^2\) \(p \approx 9\%\)

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

*Janin, Nat Struct Biol, 1997*
Easy cases

2780 Å² → large Biological interface

370 Å² → small Crystal contact
Biologically relevant interfaces are the result of evolution

Crystal contacts are not...
Valdar & Thornton\textsuperscript{1}: comparison of selection pressure on interface versus surface residues. Similar approach at the same time by Elcock & McCammon\textsuperscript{2}
Sequence entropy of MSA of homologues as metrics of selection pressure\textsuperscript{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 \textbf{biological} interface:
\[
\frac{<s>_{\text{interface}}}{<s>_{\text{surface}}} < 1
\]

For \textbf{crystal} contact:
\[
\frac{<s>_{\text{interface}}}{<s>_{\text{surface}}} > 1
\]
Entropy ratio of interface core and rim

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

For \textbf{biological} interface:\[ \frac{<s>_{\text{core}}}{<s>_{\text{rim}}} < 1 \]

For \textbf{crystal} contact:\[ \frac{<s>_{\text{core}}}{<s>_{\text{rim}}} > 1 \]

\textsuperscript{1}Guharoy & Chakrabarti, PNAS, 2005 \hspace{1cm} \textsuperscript{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 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

¹Duarte et al., *BMC Bioinformatics*, 2012
Area distributions of different interface datasets

- Ponstingl monomers
- Ponstingl dimers
- Bahadur monomers
- Bahadur dimers
- DC xtal
- DC bio

1Ponstingl et al., *J Appl Cryst*, 2003  
2Bahadur et al., *J Mol Biol*, 2004
Core size: a geometric predictor

Core size: # of core residues


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

\[
\frac{\langle s_{core} \rangle}{\langle s_{rim} \rangle} < \text{threshold: bio} \\
\frac{\langle s_{rim} \rangle}{\langle s_{core} \rangle} > \text{threshold: xtal}
\]

Interface vs surface:
Introduced by Valdar/Thornton, Elcock/McCammon (2001)
Our modifications:

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

\[
\frac{\langle s_{core} \rangle - \mu_{surface}}{\sigma_{surface}} < \text{threshold: bio} \\
\frac{\langle s_{rim} \rangle}{\langle s_{core} \rangle} > \text{threshold: xtal}
\]
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 $< \text{threshold}_1$: xtal
- $> \text{threshold}_1$: bio

2) Core vs. rim:
- $\langle s_{\text{core}} \rangle < \text{threshold}_2$: bio
- $\langle s_{\text{rim}} \rangle > \text{threshold}_2$: xtal

3) Core vs. surface:
- $\langle s_{\text{core}} \rangle - \mu_{\text{surface}} \sigma_{\text{surface}} < \text{threshold}_3$: bio
- $\langle s_{\text{core}} \rangle - \mu_{\text{surface}} \sigma_{\text{surface}} > \text{threshold}_3$: 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
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)
Bio interfaces: core-surface scores over the years

The graph shows the distribution of core-surface scores over different UniProt versions. Scores are classified as crystal contacts (upper region) and biological interfaces (lower region). The scores are plotted over the years, with each bar representing a different version of UniProt.

- Classified as crystal contacts
- Classified as biological interfaces

UniProt version:
- Dec 2003
- Feb 2005
- Feb 2006
- Mar 2007
- Feb 2008
- Mar 2009
- Feb 2010
- Feb 2011
- Feb 2012

Core-surface score:
- Range from -5 to 5
Xtal interfaces: core-surface scores over the years

- Classified as crystal contacts
- Classified as biological interfaces

Core-surface scores over the years, classified as crystal contacts and biological interfaces.
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](http://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

University of Zurich

Prof. Amedeo Caflisch
Prof. Andreas Wagner

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

Derek Feichtinger & Valeri Markushin
(SciComp group, PSI)

Funding and support:
Forschungskommission PSI