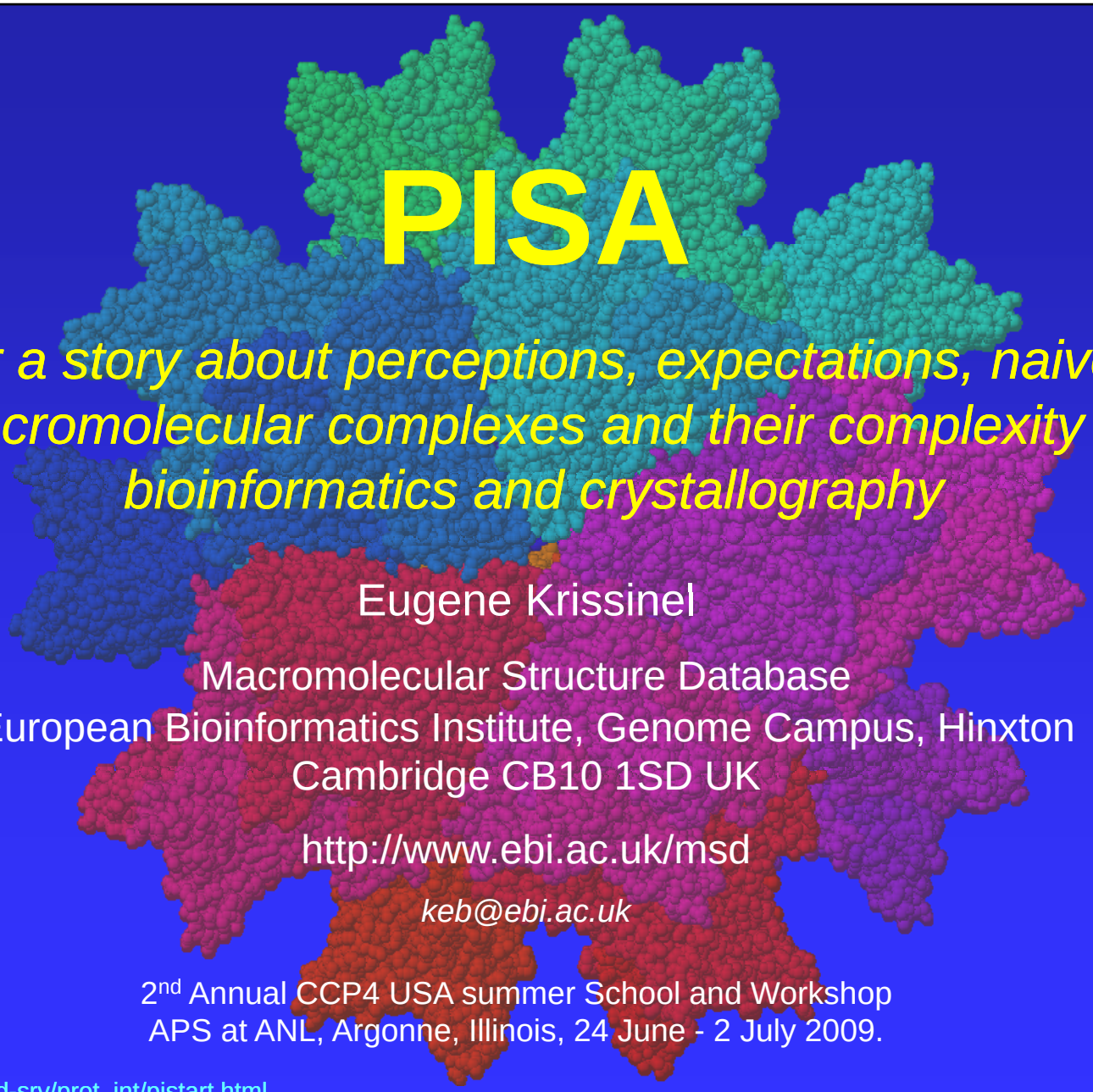


PISA

... or a story about perceptions, expectations, naivety, macromolecular complexes and their complexity in bioinformatics and crystallography

Eugene Krissinel

Macromolecular Structure Database
European Bioinformatics Institute, Genome Campus, Hinxton
Cambridge CB10 1SD UK

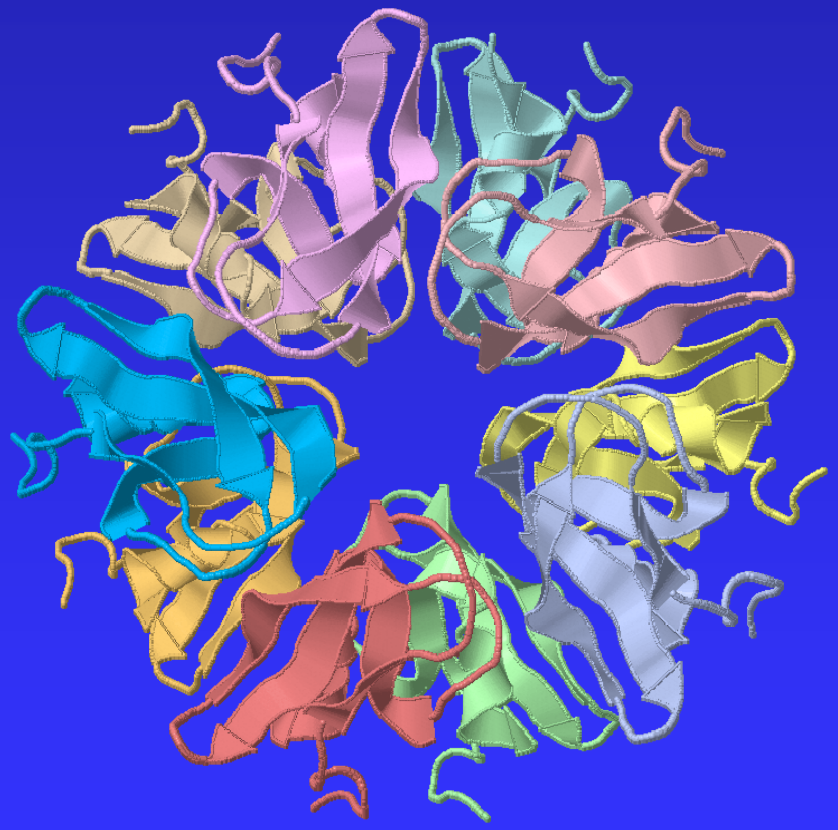
http://www.ebi.ac.uk/msd

keb@ebi.ac.uk

2nd Annual CCP4 USA summer School and Workshop
APS at ANL, Argonne, Illinois, 24 June - 2 July 2009.

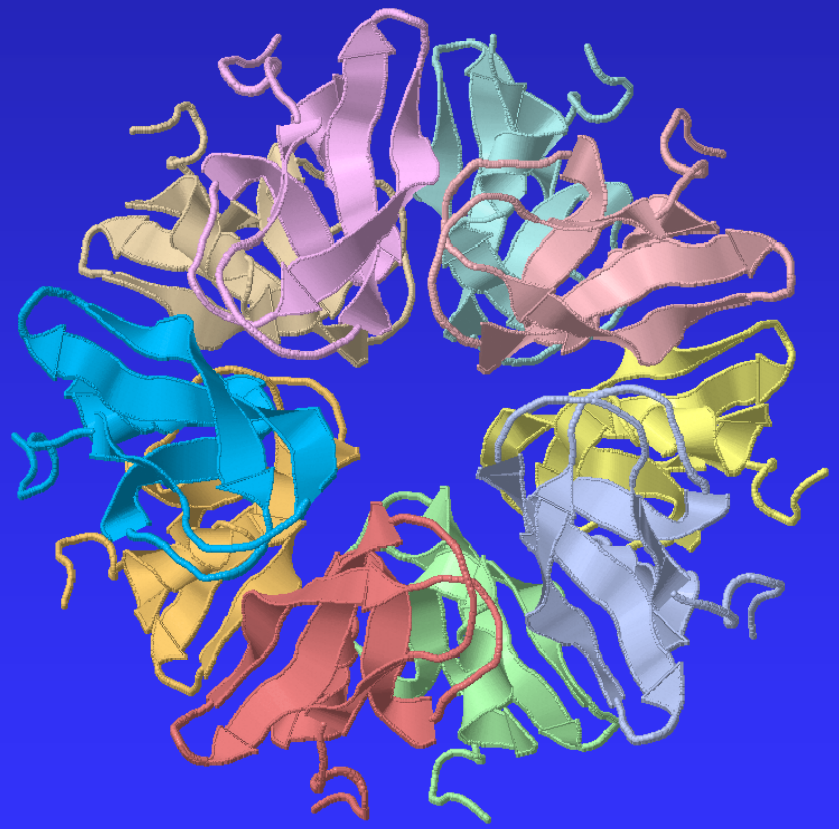
Macromolecular Assemblies

- Complexes of protein, DNA/RNA chains and ligands, stable in native environment
- The way the chains assemble represents the [Protein] Quaternary Structure (PQS)
- Macromolecular assemblies are often the Biological Units, performing certain biochemical functions
- Biological significance of macromolecular assemblies is truly immense



PQS is a difficult object for experimental studies

- *Light / Neutron / X-ray / Small angle scatterings*: mainly composition and multimeric state may be found. 3D shape may be guessed from mobility measurements.
- *Electron microscopy*: not a fantastic resolution and not applicable to all objects
- *NMR* is not good for big chains, even less so for protein assemblies.



Very few quaternary structures have been identified experimentally.

PQS are difficult to calculate

If we know the sequence ...

1 VNKERTFLAVKPDGVARGLVGEIIARYEKKGFVLVGLKQLVPTKDLAESHYAEHKERPFF

then we can calculate ...

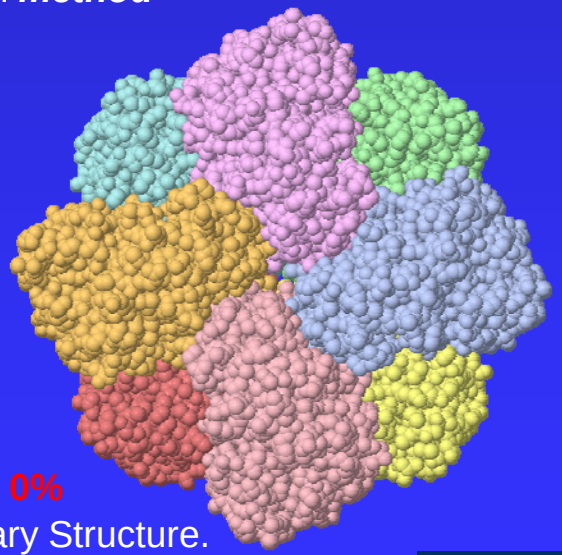
2

50 - 90% Secondary Structure (CASP 5), depending on ***method***

3

10 - 90% Tertiary Structure (CASP 5), depending on method and ***target***

4



Probably 0% Quaternary Structure. Docking of given number of given structures: **5 - 20%** success (CAPRI 5)

http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

European Bioinformatics Institute

But PQS are assigned to many entries in PDB!

RCSB PDB PROTEIN DATA BANK
www.pdb.org

PDBj Protein Data Bank Japan
www.pdbj.org

www.ebi.ac.uk/msd

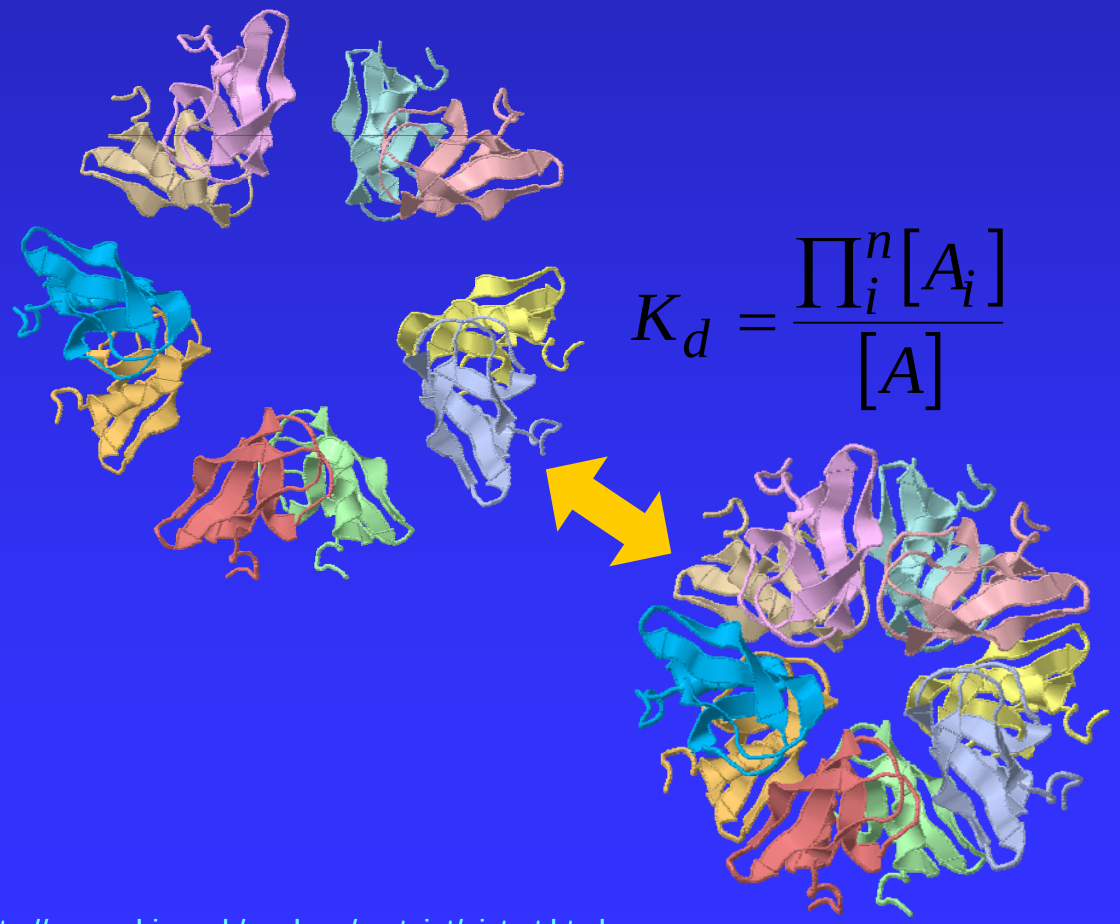
Most of those are ***PROBABLE Q***uaternary ***S***tructures.

The PDB "rules" are:

1. Depositor's say prevails.
2. Accept everything which passes formal validation checks.
3. No experimental evidence for PQS is required.
4. If a depositor does not know or does not care (60-80% of instances for PQS), the curator is to decide.
5. The curator may (EBI-MSD) or may not (Rutgers-RCSB, PDBj) use computing/modeling tools to assist the PQS annotation.

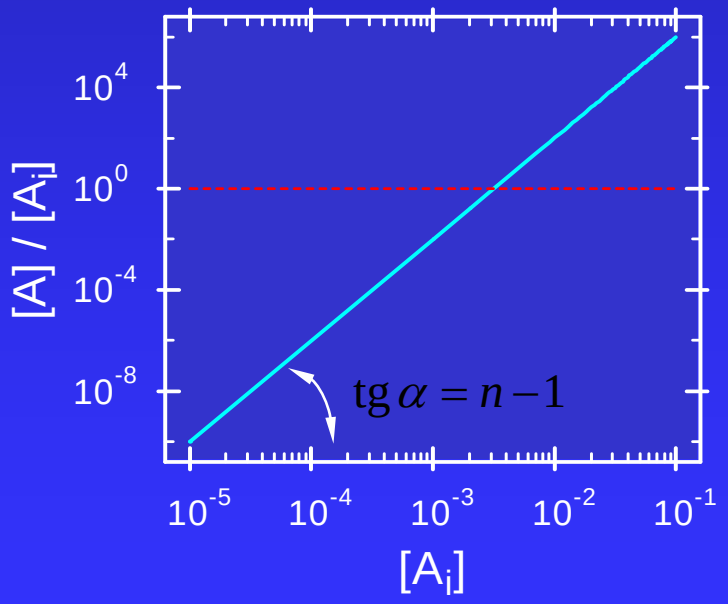
Crystallography is special

Because: A) crystal is made of assemblies

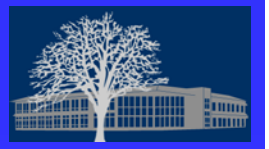


$$K_d = \frac{\prod_i^n [A_i]}{[A]}$$

$$[A]/[A_i] = [A_i]^{n-1} / K_d$$

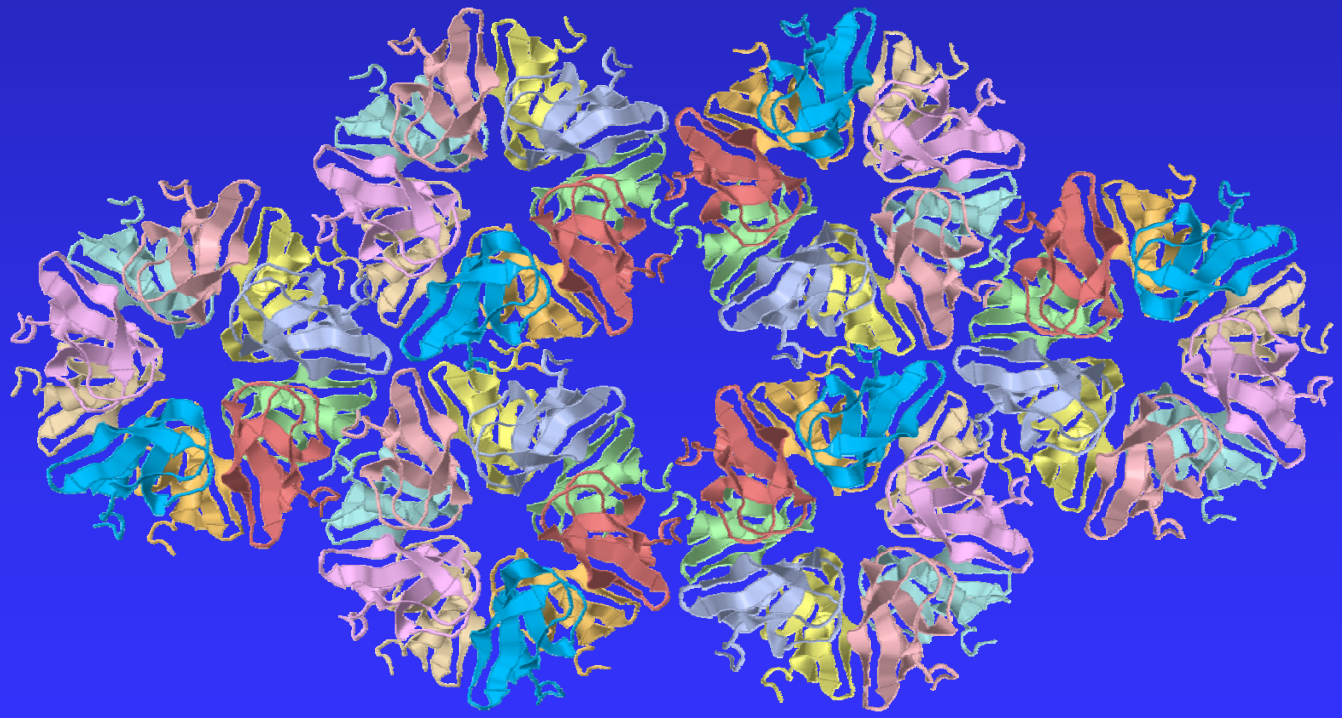


http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html



Crystallography is special

Because: B) there is no need to dock subunits
– the docking is given by crystal structure



Macromolecular interfaces should be viewed as an additional important product of protein crystallography

PDB contains a wealth of experimental data on PQS

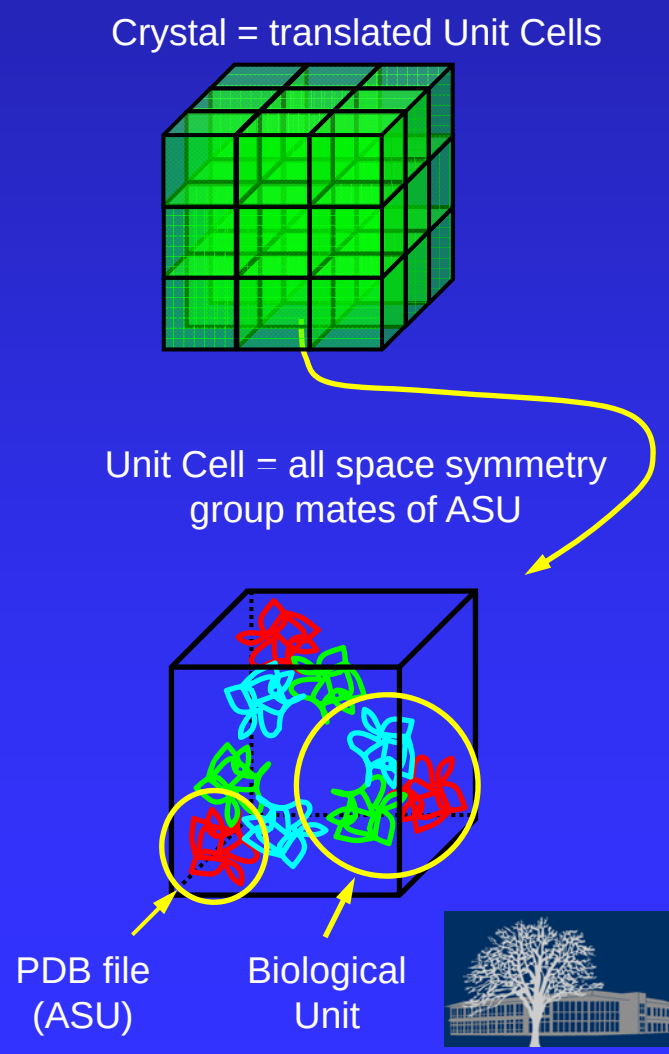
More than 80% of macromolecular structures are solved by means of X-ray diffraction on crystals.

It is reasonable to expect that PQS make construction blocks for the crystal.

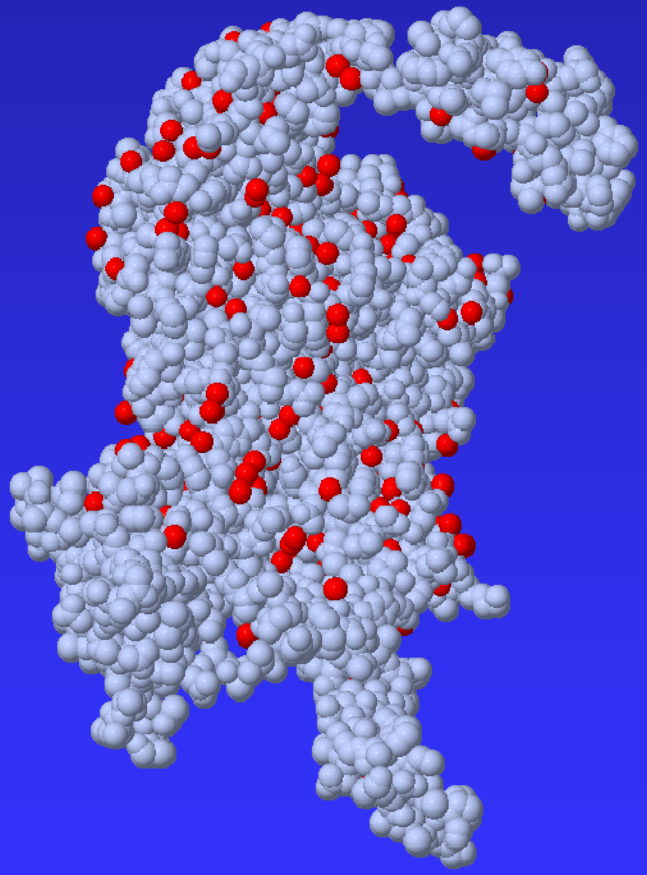
An X-ray diffraction experiment produces atomic coordinates of the Asymmetric Unit (ASU), which is stored as a PDB file.

In general, neither ASU nor Unit Cell has any direct relation to PQS. The PQS may be made of

- a single ASU
- part of ASU
- several ASU
- several ASU parts

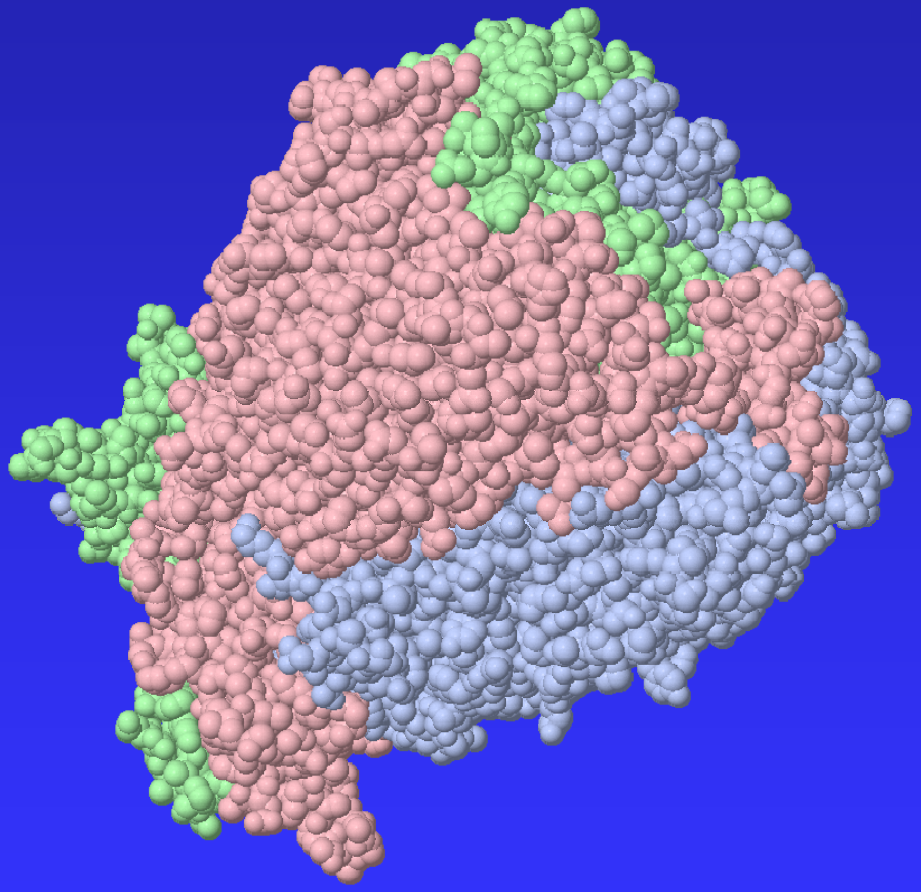


PDB entries and Biological Units



PDB entry 1P30

A monomer?

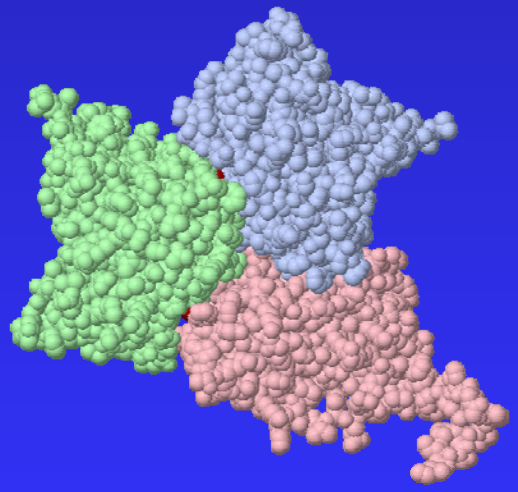


Biological unit 1P30

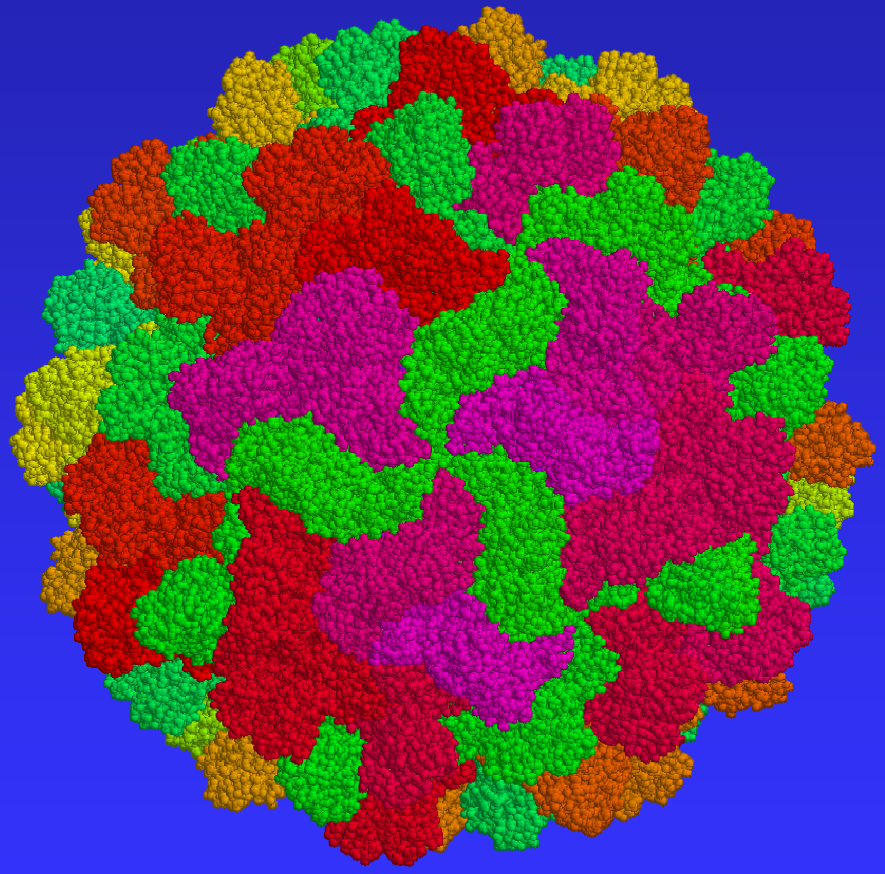
Homotrimer!

http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

PDB entries and Biological Units



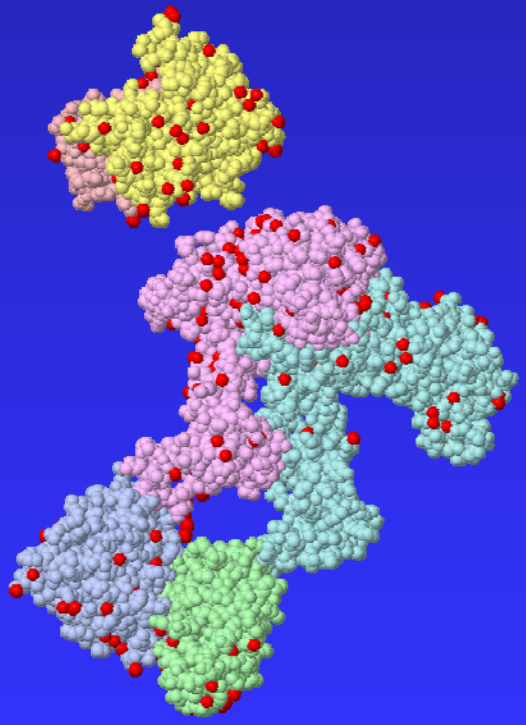
PDB entry 2TBV
A trimer?



Biological Unit 2TBV
180-mer!

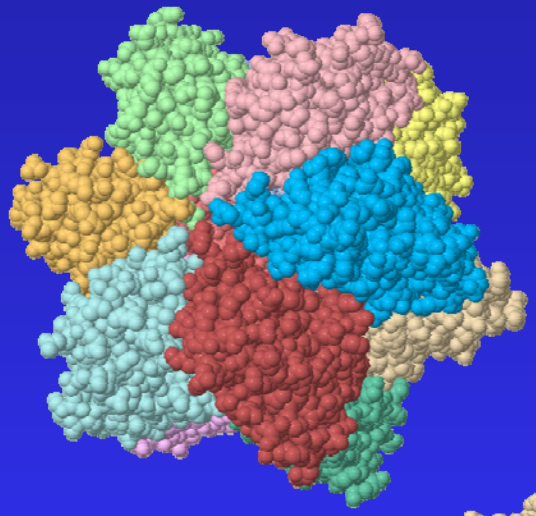
http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

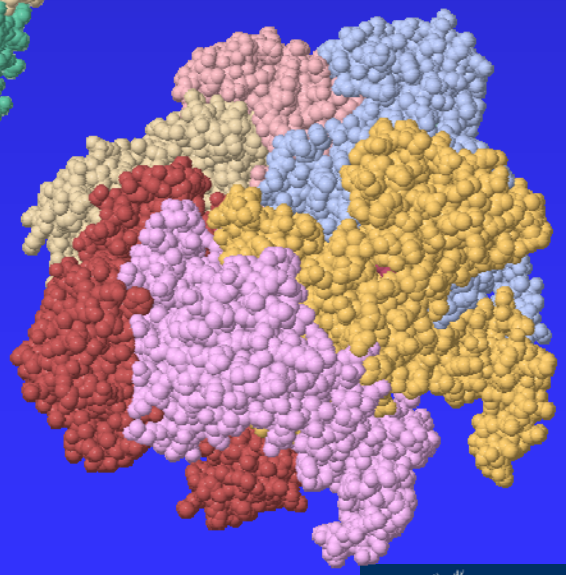
PDB entries and Biological Units



PDB entry 1E94

????????



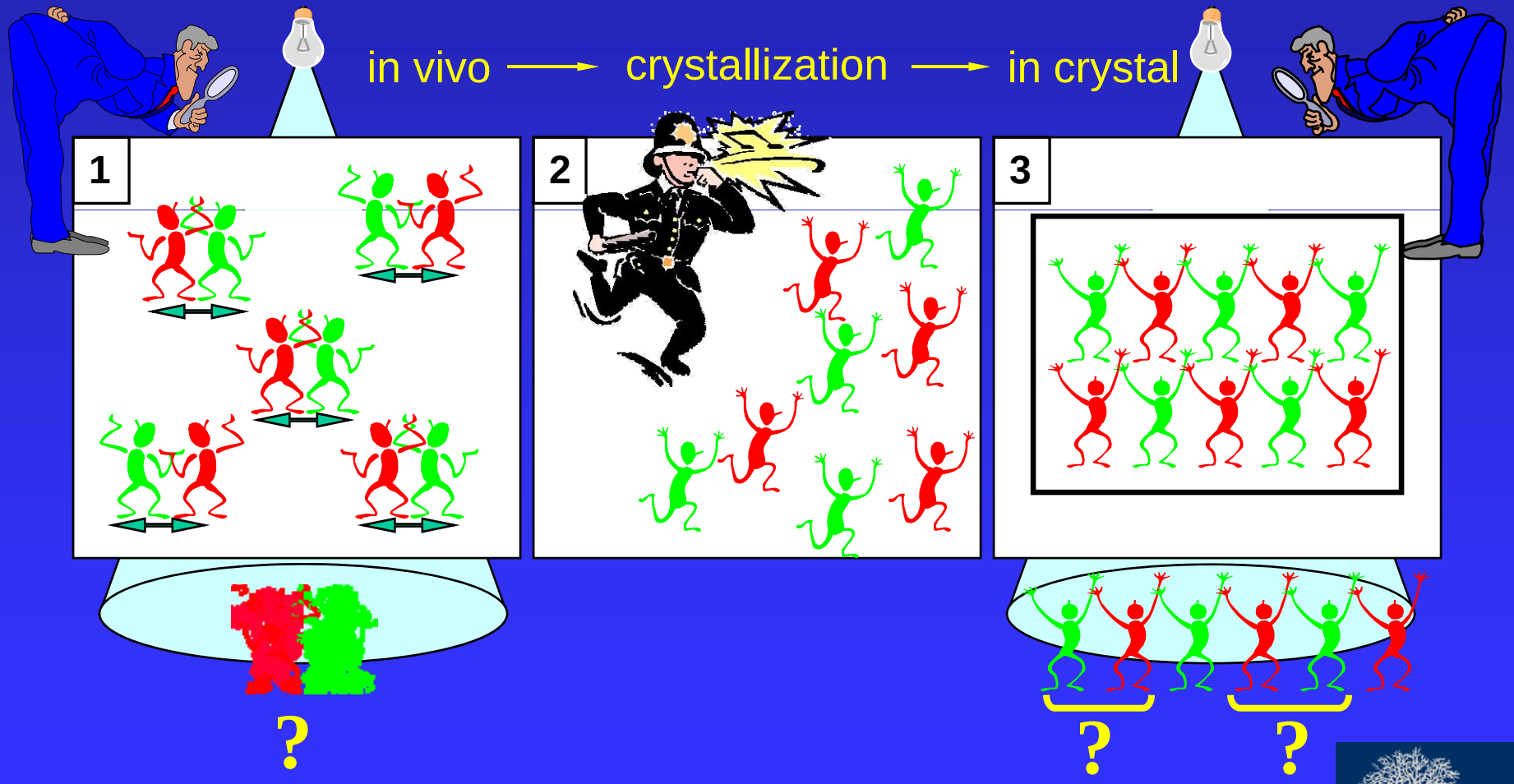


2 Biological Units in 1E94:

A dodecamer and a hexamer!

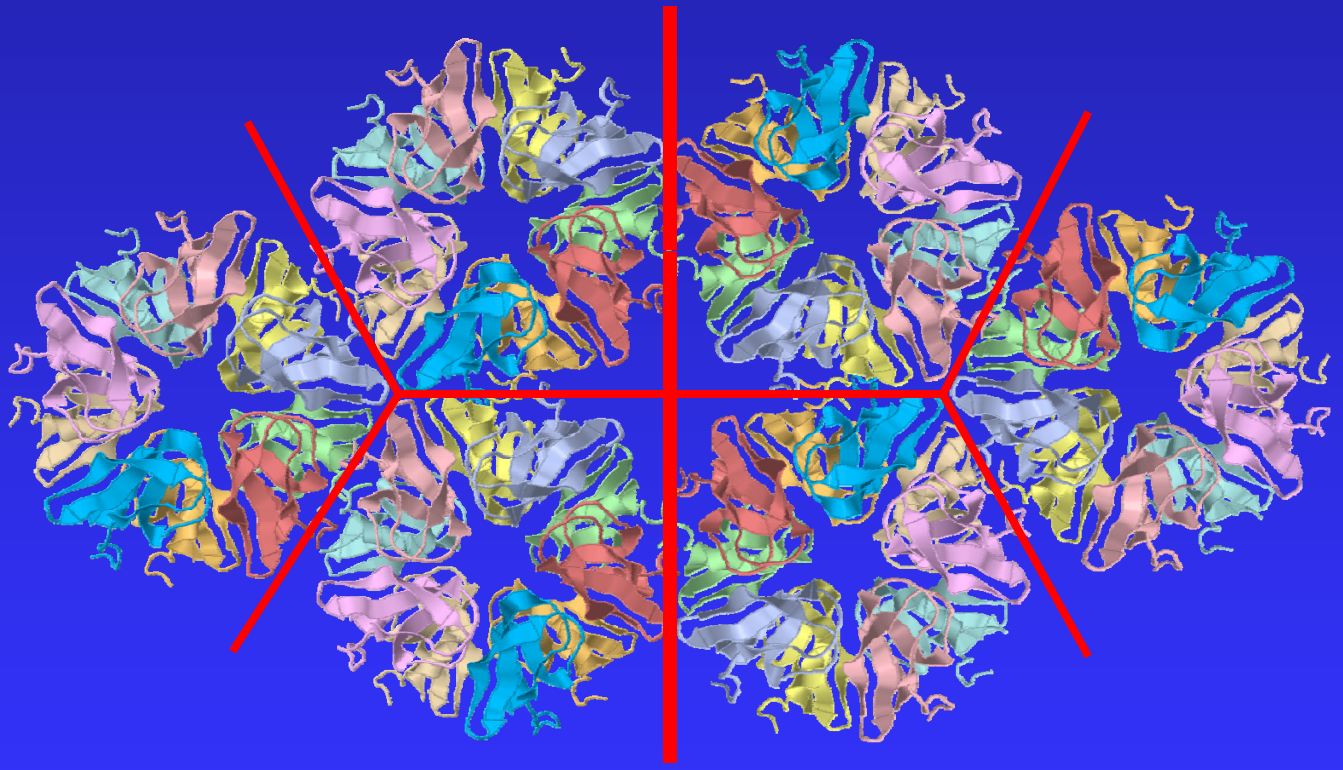
http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

In (very) simple words …



http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

A simple thing to do



PQS server @ EBI (Kim Henrick) *Trends in Biochem. Sci.* (1998) **23**, 358

PITA server @ EBI (Hannes Ponstingl) *J. Appl. Cryst.* (2003) **36**, 1116

http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

What is a significant interface?

Depends on the problem.

Protein functionality: the interface should be engaged in *any* sort of interaction, including transient short-living protein-ligand and protein-protein etc. associations. Obviously important properties:

- Affinity (comes from area, hydrophobicity, electrostatics, H-bond density etc.)

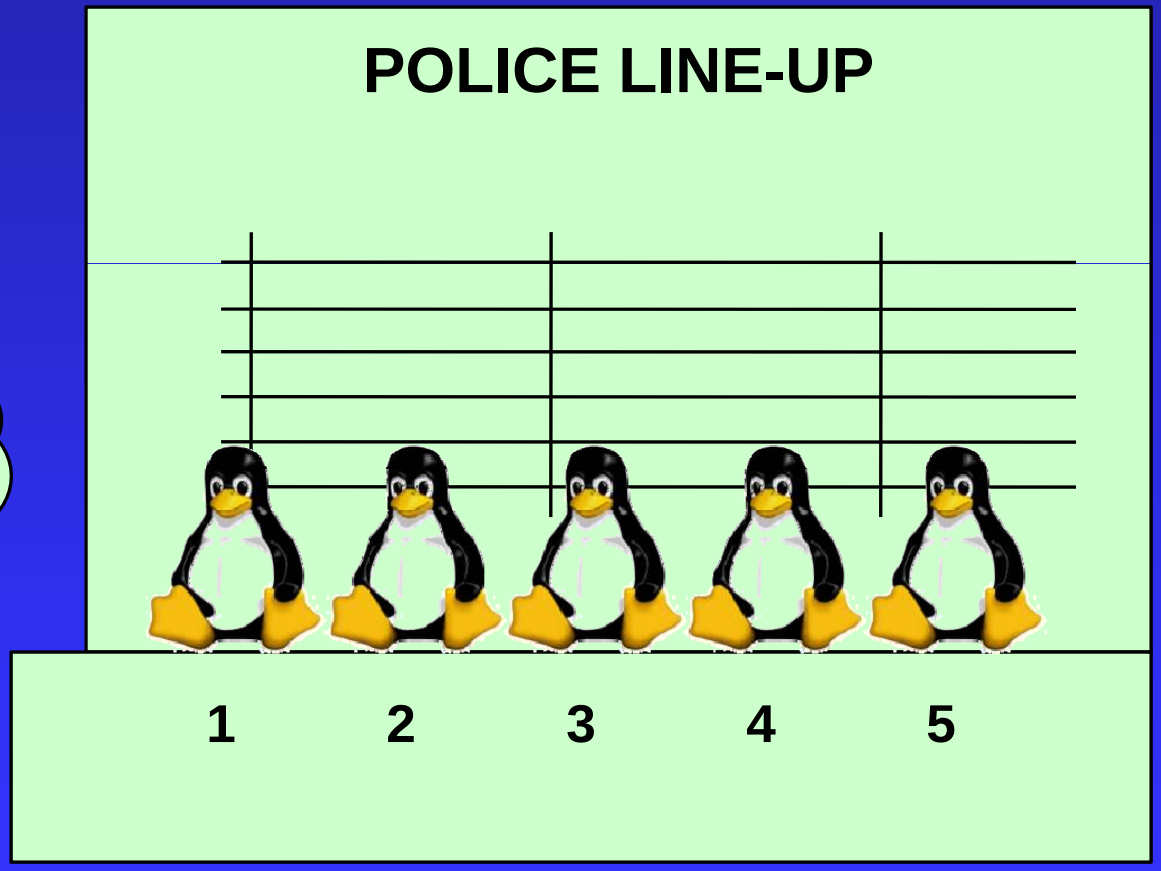
and properties that may be important for *reaction pathway and dynamics*:

- Aminoacid composition
- Geometrical complementarity
- Overall shape, compactness
- Charge distribution
- etc.

Stable macromolecular complexes, PQS: the interface should make a sound binding. Important properties:

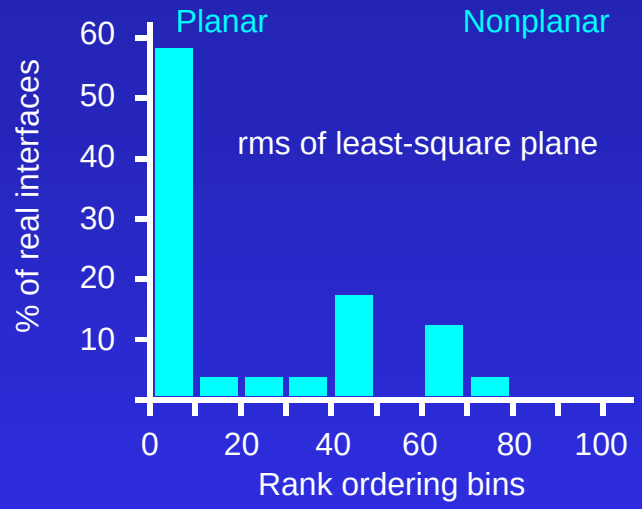
- Sufficient free energy of binding
- something else?

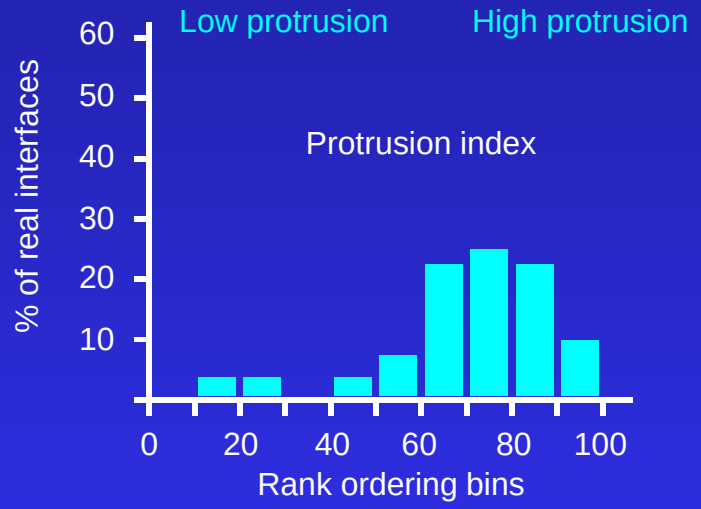
A Common Identification Problem

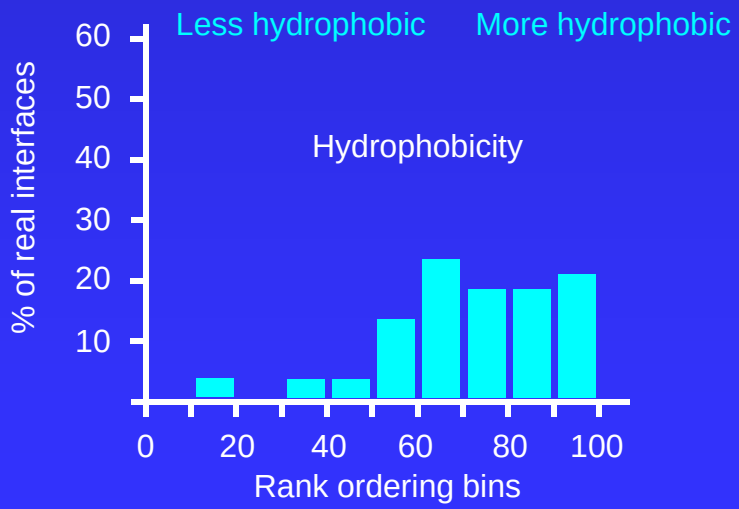


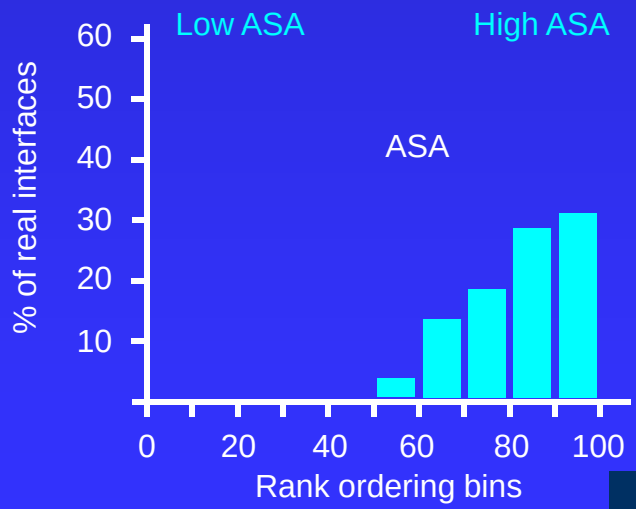
http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

Real and superficial protein interfaces









Jones, S. & Thornton, J.M. (1996) Principles of protein-protein interactions, *Proc. Natl. Acad. Sci. USA*, **93**, 13-20.

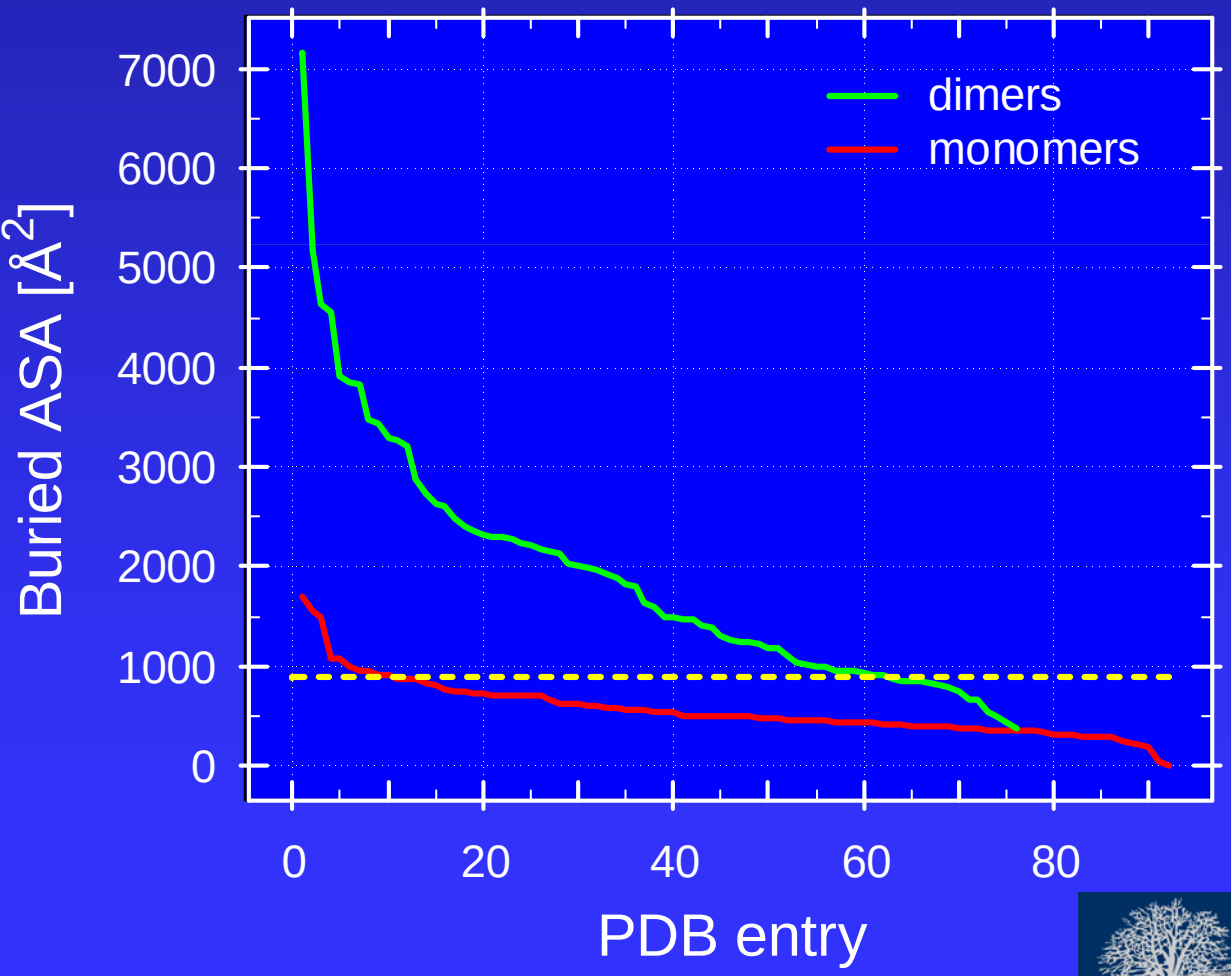
http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

Real and superficial protein interfaces

Most often used discrimination criteria - interface area.

A cut-off at 900 Å^2 gives about 80% success rate of discrimination between monomers and dimers.

Big proteins will be always sticky if this criteria is true …

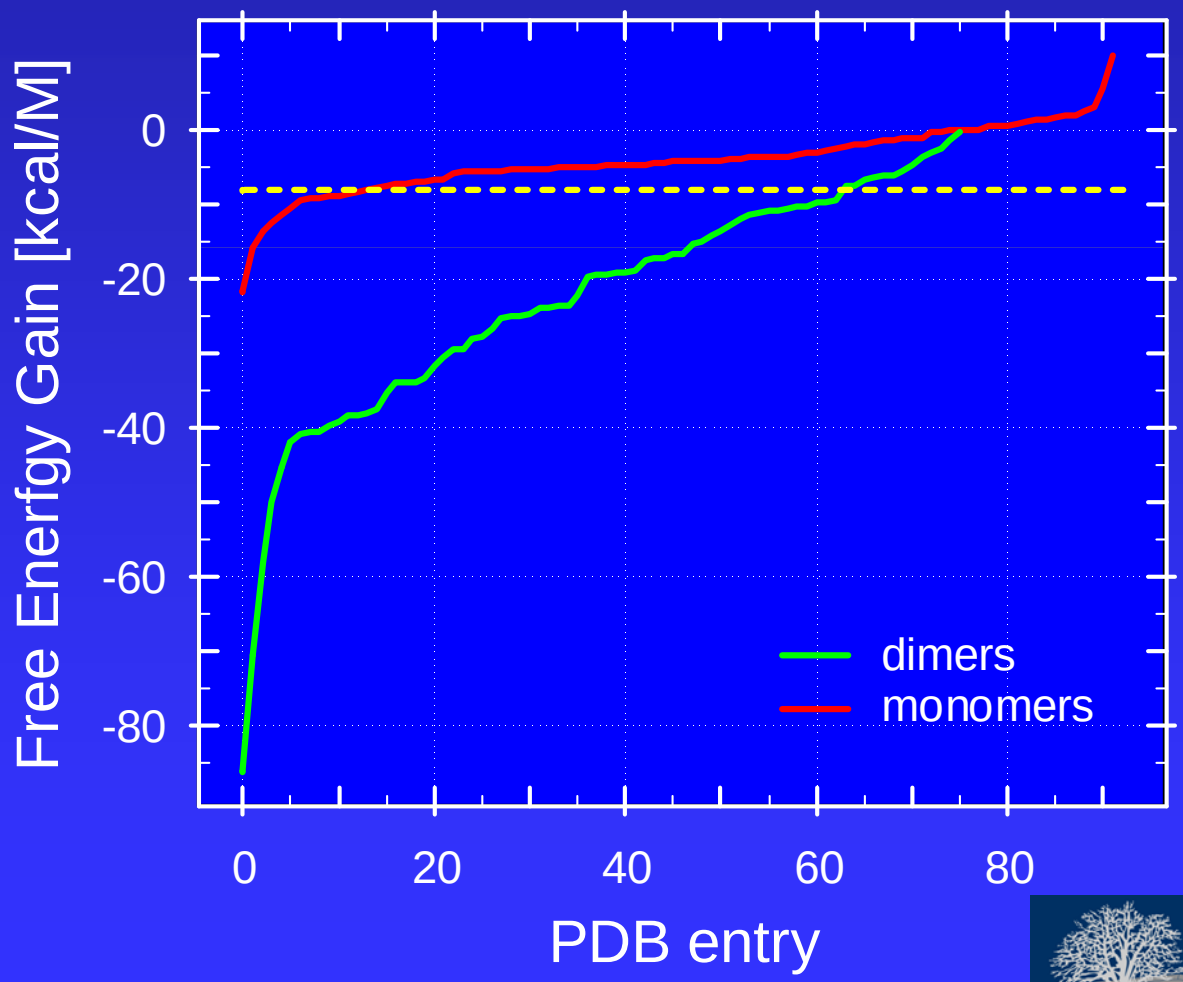


Real and superficial protein interfaces

Free energy gain of interface formation.

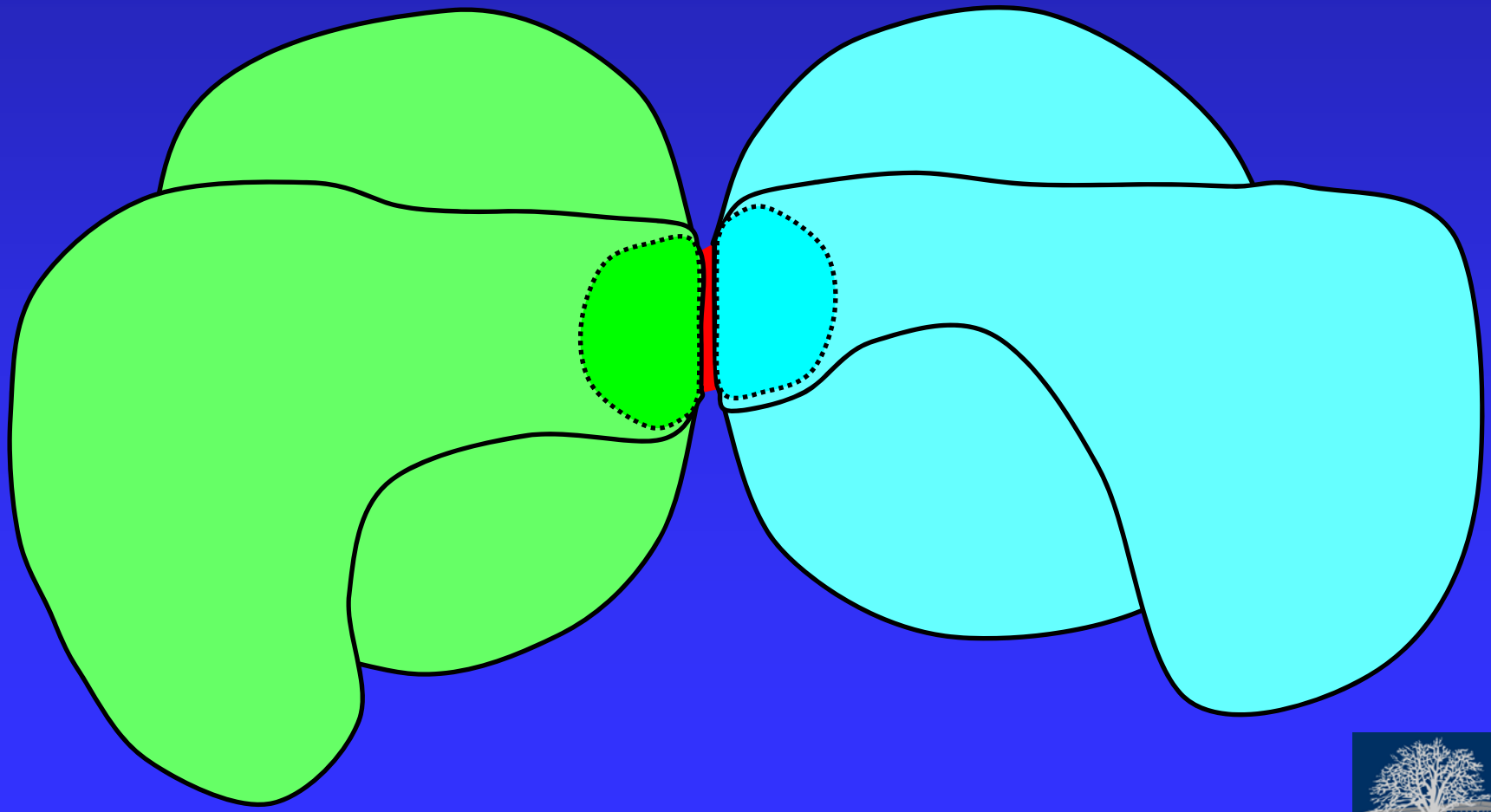
A cut-off at -8 kcal/M gives about 82% success rate of discrimination between monomers and dimers.

Can energy measure be uniform for all weights and shapes?



http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

Is there a (good) measure of interface significance at all?



Real and superficial protein interfaces

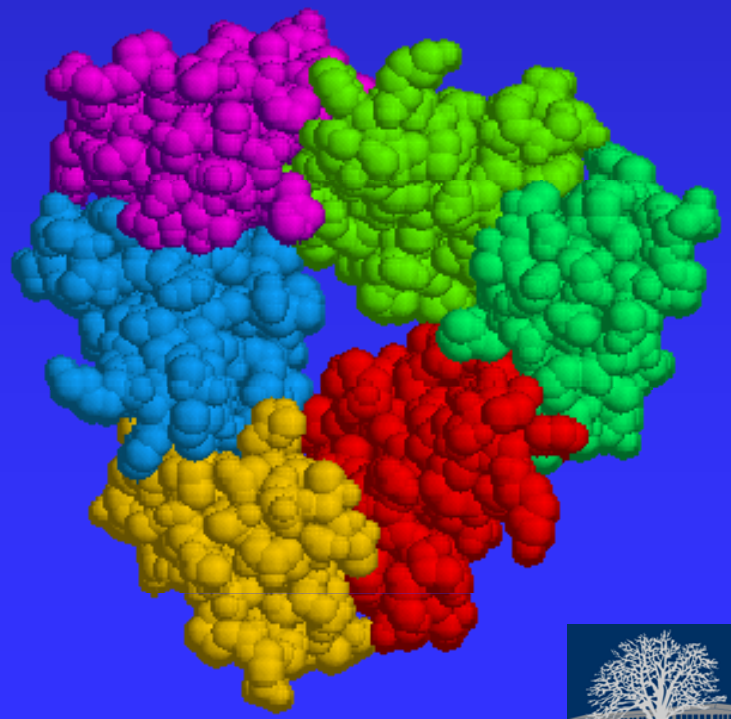
- "*No single parameter absolutely differentiates the interfaces from all other surface patches*"

  Jones, S. & Thornton, J.M. (1996) Principles of protein-protein interactions, *Proc. Natl. Acad. Sci. USA*, **93**, 13-20.

- Formation of N>2 -meric complexes is most probably a corporate process involving a set of interfaces. Therefore significance of an interface should not be detached from the context of macromolecular complex

- "…*the type of complexes need to be taken into account when characterizing interfaces between them.*"

  Jones, S. & Thornton, J.M., ibid.



Chemical stability of macromolecular complexes

- It is not properties of individual interfaces but rather chemical stability of complexes in general that really matters
- *Macromolecular units will most likely associate into largest complexes that are still stable*
- A complex is stable if its free energy of dissociation is positive:

$$\Delta G^{0}_{diss} = -\Delta G_{\text{int}} - T\Delta S > 0$$

Chemical stability of macromolecular complexes

$$\Delta G_{diss} = -\Delta G_{\text{int}} - T\Delta S > 0$$

Binding energy

Solvation energy of protein complex

Solvation energies of dissociated subunits

Free energy of H-bond formation

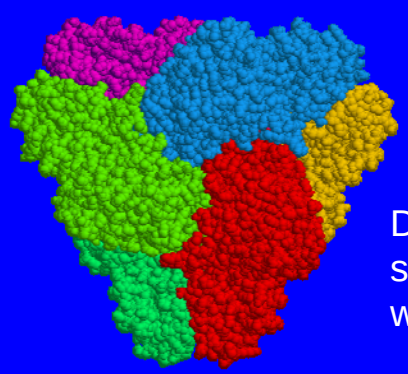
Free energy of salt bridge formation

$$\Delta G_{\text{int}} = \Delta G_{sol}(A_1, A_2 \ldots A_n) - \sum_{i=1}^{n} \Delta G_{sol}(A_i) - E_{hb} N_{hb} - E_{sb} N_{sb}$$

Number of H-bonds between dissociated subunits

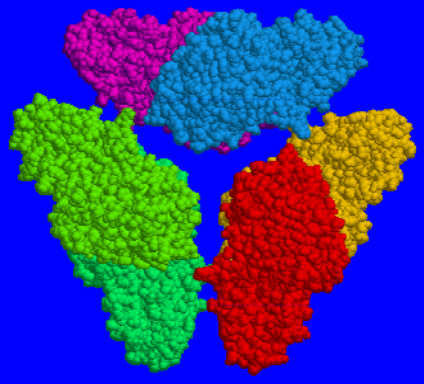
Number of salt bridges between dissociated subunits

Choice of dissociation subunits:



$(A_1 A_2 A_3)$

Dissociation into stable subunits with minimum ΔG_{diss}



$A_1 + A_2 + A_3$

Solvation free energy

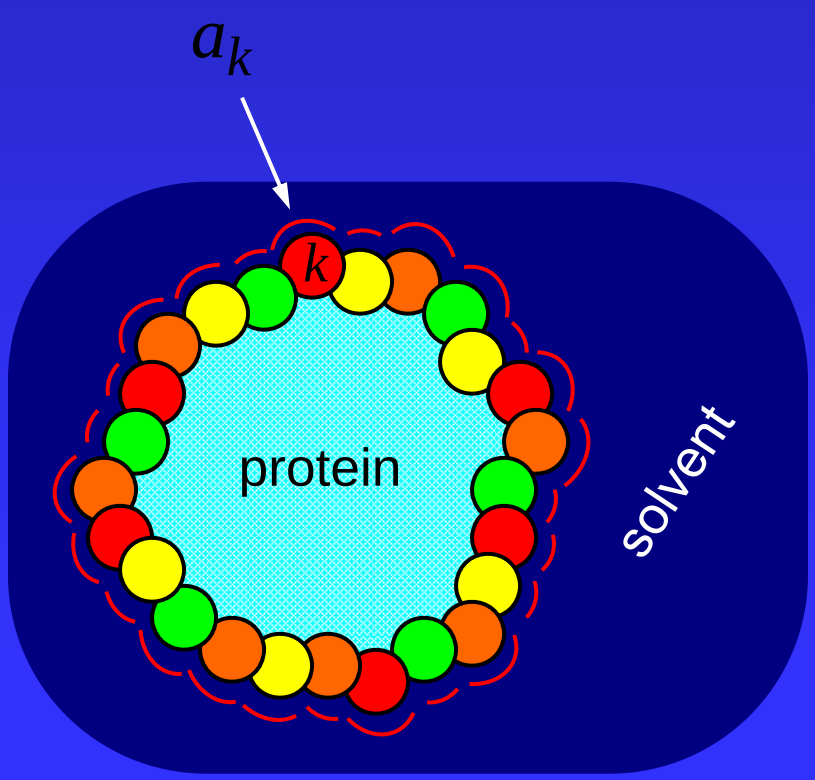
Eisenberg, D. & McLachlan, A.D. (1986)
Nature 319, 199-203.

Atomic solvation parameters

Atom's accessible surface area

$$\Delta G_{sol}(A) = \sum_k \Delta\sigma_k \left(a_k - a_k^r\right)$$

Atom's accessible surface area in the reference (unfolded) state



In this model, binding energy is function of individual interfaces.

Entropy of macromolecules in solutions

Translational entropy *Rotational entropy* *Sidechain entropy*

$$S = S_{trans}(m) + S_{rot}(\hat{I}, \sigma_S) + S_{surf}(a)$$

Mass *Tensor of inertia* *Symmetry number* *Solvent-accessible surface area*

$$S_{trans}(m) \approx c_t + \frac{3R}{2}\log(m)$$

$$S_{rot}(\hat{I}, \sigma_S) \approx c_r + \frac{R}{2}\log\left(I_1 I_2 I_3 / \sigma_S^2\right)$$

Murray C.W. and Verdonk M.L. (2002)
J. Comput.-Aided Mol. Design 16, 741-753.

$$S_{surf}(a) \approx Fa$$

c_t, c_r and F are semiempirical parameters

Entropy of dissociation

$$\Delta S = \sum_{i=1}^{n} S(A_i) - S(A_1, A_2 \ldots A_n)$$

$$= (n-1)C + \frac{3R}{2}\log\left(\frac{\prod_i m_i}{\sum_i m_i}\right) +$$

$$\frac{R}{2}\log\left(\frac{\prod_i \prod_k I_k(A_i) / \sigma_S^2(A_i)}{\prod_k I_k(A_1 \ldots A_n) / \sigma_S^2(A_1 \ldots A_n)}\right) + Fa_{buried}$$

Mass of i-th subunit

k-th principal moment of inertia of i-th subunit

Fitted parameter

Fitted parameter

By its very nature, entropy of dissociation is function of protein complex rather than that of individual interfaces.

Entropy of dissociation

- Drives thermodynamic systems towards most disordered (dissolved) state.
- Makes bias towards less symmetric states. However, in practice, this is overweighed by binding energy, which is normally maximal in most symmetric states.
- Responsible for complex “instability”.

We know how it looks, can we see it?

We now know (or we think that we know) how to evaluate chemical stability of protein complexes:

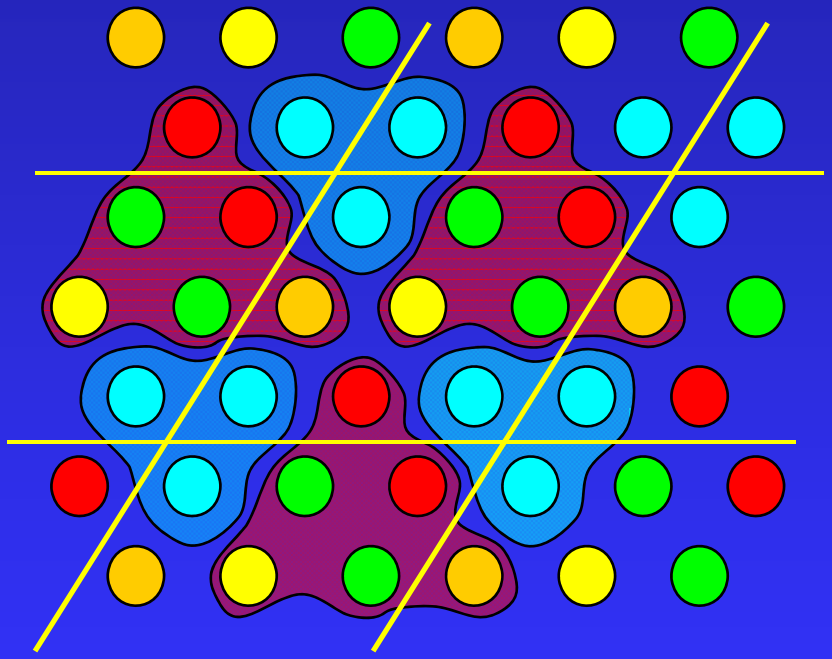
$$\Delta G^{0}_{diss} = -\Delta G_{\text{int}} - T\Delta S > 0$$

which depends on interface properties as well as on the geometry of protein complex and dissociated subunits.

How to find stable complexes in a crystal?

Enumerating assemblies in crystal

- crystal is represented as a periodic graph with monomeric chains as vertices and interfaces as edges
- each set of assemblies is identified by engaged interface types
- all assemblies may be enumerated by a backtracking scheme engaging all possible combinations of different interface types



Example: crystal with 3 interface types

Assembly set	Engaged interface types	
1	000	- only monomers
2	001	- dimer N1
3	010	- dimer N2
4	011	
5	100	- dimer N3
6	101	
7	110	
8	111	- all crystal

http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

Detection of Biological Units in Crystals: Method Summary

1. Build periodic graph of the crystal
2. Enumerate all possibly stable assemblies
3. Evaluate assemblies for chemical stability
4. Leave only sets of stable assemblies in the list and range them by chances to be a biological unit :
    - *Larger assemblies take preference*
    - *Single-assembly solutions take preference*
    - *Otherwise, assemblies with higher* ΔG_{diss} *take preference*

Classification of protein assemblies

Assembly classification on the benchmark set of 218 protein structures published in

Ponstingl, H., Kabir, T. and Thornton, J. (2003) Automatic inference of protein quaternary structures from crystals. J. Appl. Cryst. 36, 1116-1122.

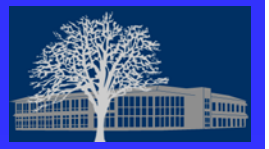
	1mer	2mer	3mer	4mer	6mer	Other	Sum	Correct
1mer	49	3	0	1	1	1	55	89%
2mer	3	71+11	0	2+1	0	0	76+12	93%
3mer	1	0	22	0	1	0	24	92%
4mer	2	2+1	0	26+6	0	1	31+7	84%
6mer	0	0	0	0+1	10+2	0	10+3	92%
196+22	<=>	196 homomers and 22 heteromers				Total:	196+22	**90%**

Fitted parameters:

1. Free energy of a H-bond : E_{hb} = 0.51 kcal/mol
2. Free energy of a salt bridge : E_{sb} = 0.21 kcal/mol
3. Constant entropy term : $T \cdot C$ = 11.7 kcal/mol
4. Surface entropy factor : $T \cdot F$ = $0.57 \cdot 10^{-3}$ kcal/(mol*Å^2)

Classification error in ΔG^0_{diss} : *± 5 kcal/mol*

http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html



Classification of protein-DNA complexes

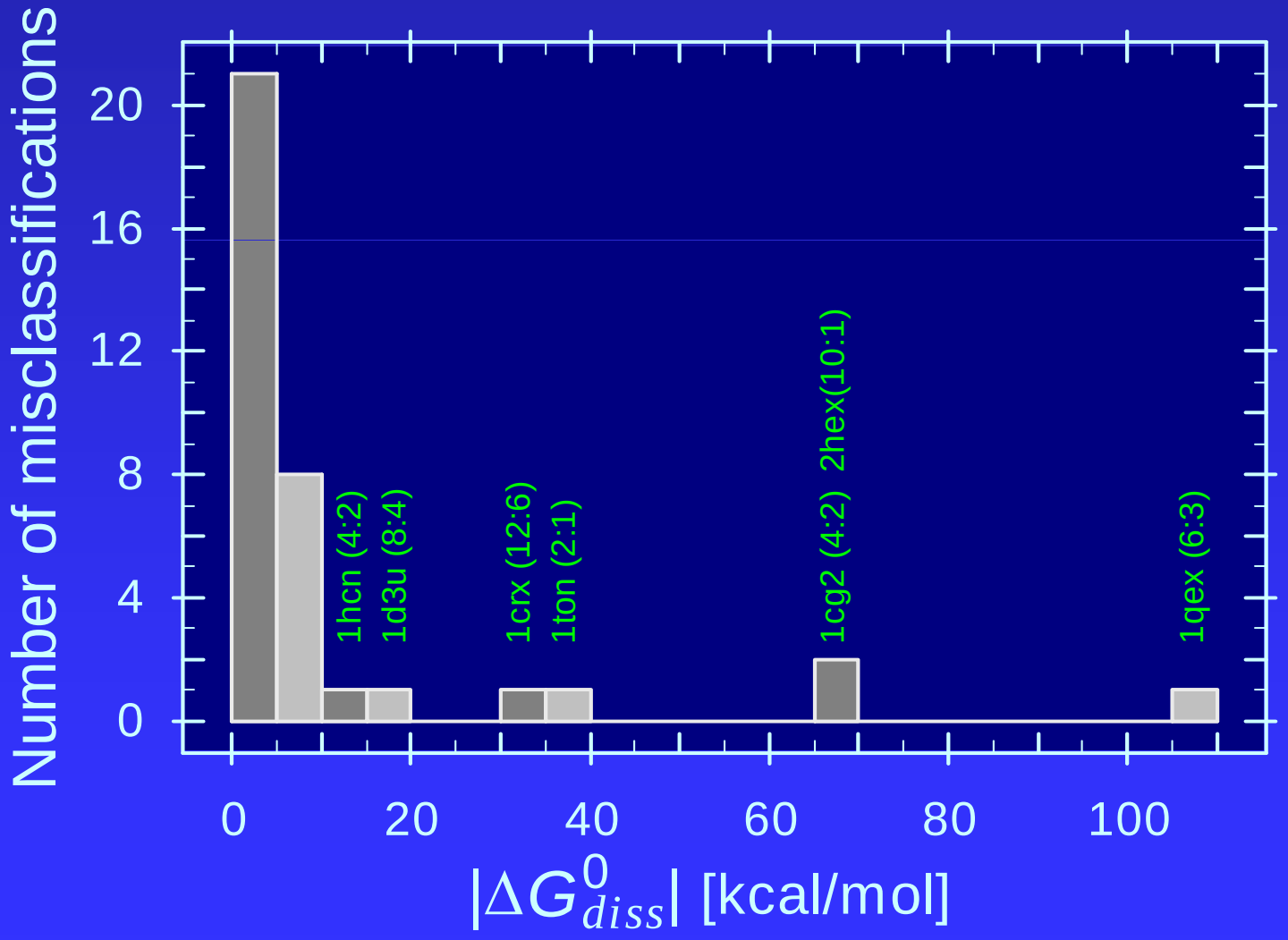
Assembly classification on the benchmark set of 212 protein – DNA complexes published in

Luscombe, N.M., Austin, S.E., Berman H.M. and Thornton, J.M. (2000) An overview of the structures of protein-DNA complexes. Genome Biol. 1, 1-37.

	2mer	3mer	4mer	5mer	6mer	10mer	Other	Sum	Correct
2mer	1	0	0	0	0	0	0	1	100%
3mer	6	96	0	0	1	0	2	105	91%
4mer	0	2	83	0	0	0	0	85	98%
5mer	0	0	2	3	0	0	0	5	60%
6mer	1	0	0	0	13	0	1	15	87%
10mer	0	0	0	0	0	1	0	1	100%
							Total:	212	**93%**

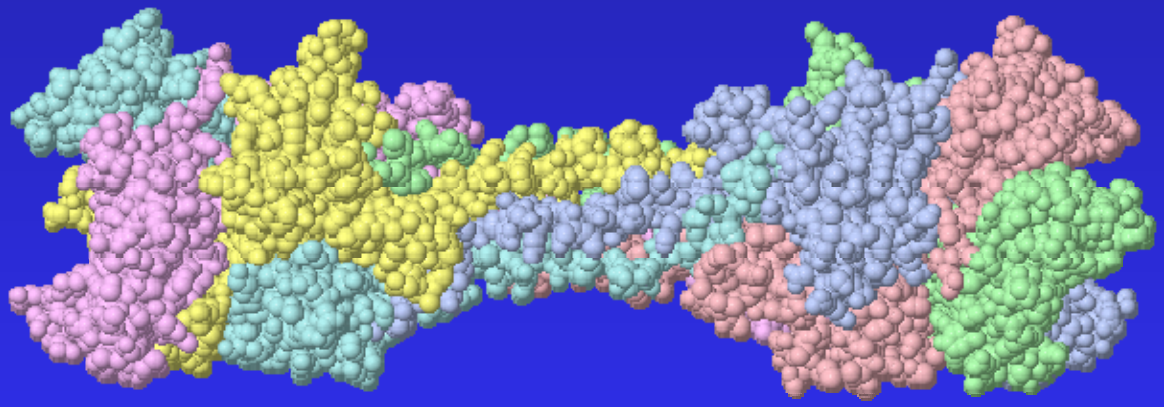
Classification error in ΔG^0_{diss} : *± 5 kcal/mol*

Free energy distribution of misclassifications



Example of misclassification: 1QEX

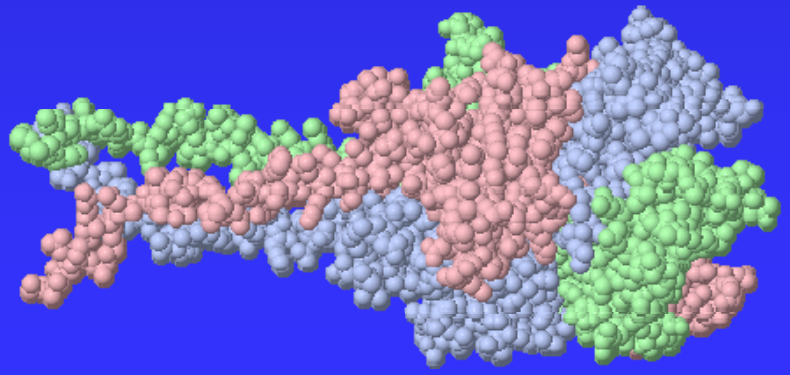
BACTERIOPHAGE T4 GENE PRODUCT 9 (GP9), THE TRIGGER OF TAIL CONTRACTION AND THE LONG TAIL FIBERS CONNECTOR



Predicted: homohexamer

Dissociates into 2 trimers

$\Delta G^0_{diss} \approx$ *106 kcal/mol*



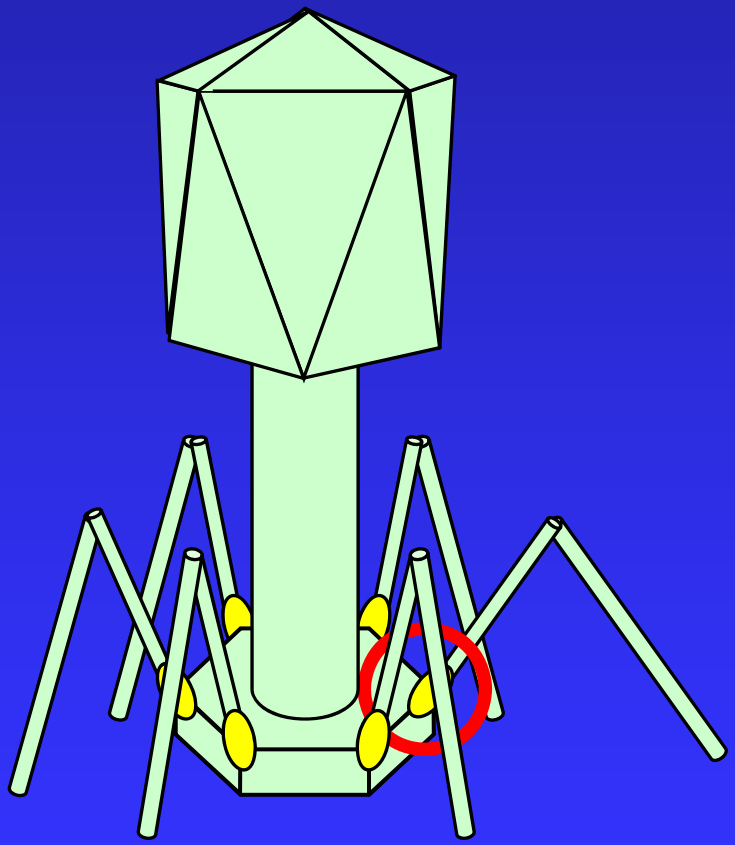
Biolgical unit: homotrimer

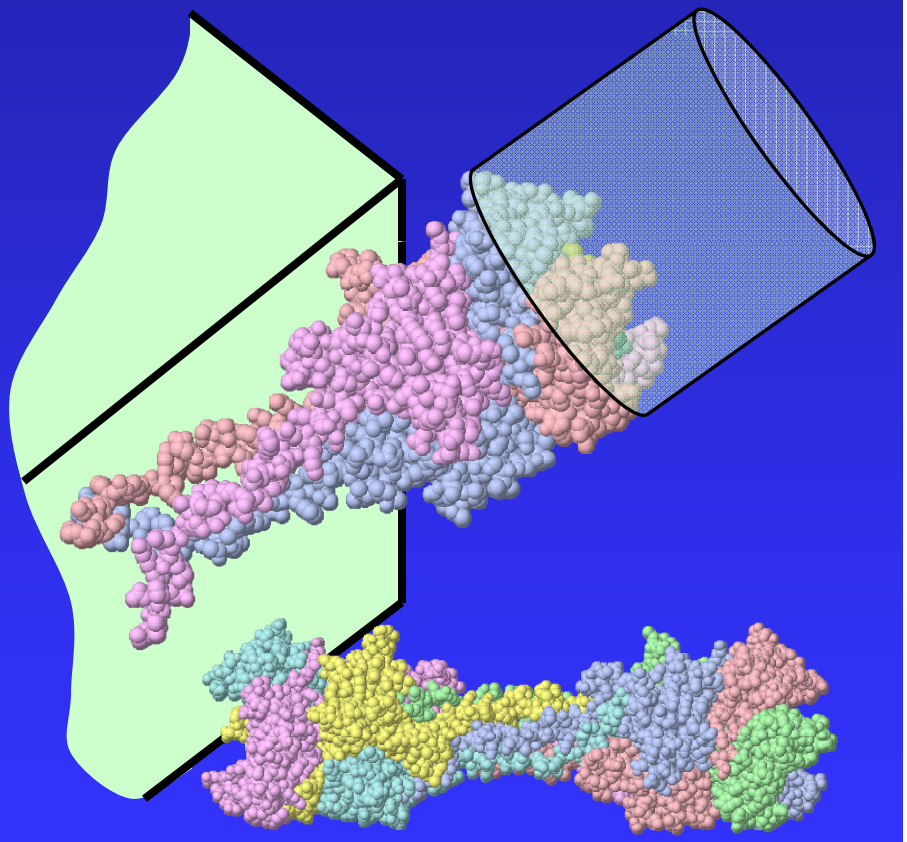
Dissociates into 3 monomers

$\Delta G^0_{diss} \approx$ *90 kcal/mol*

Example of misclassification: 1QEX

BACTERIOPHAGE T4 GENE PRODUCT 9 (GP9), THE TRIGGER OF TAIL CONTRACTION AND THE LONG TAIL FIBERS CONNECTOR

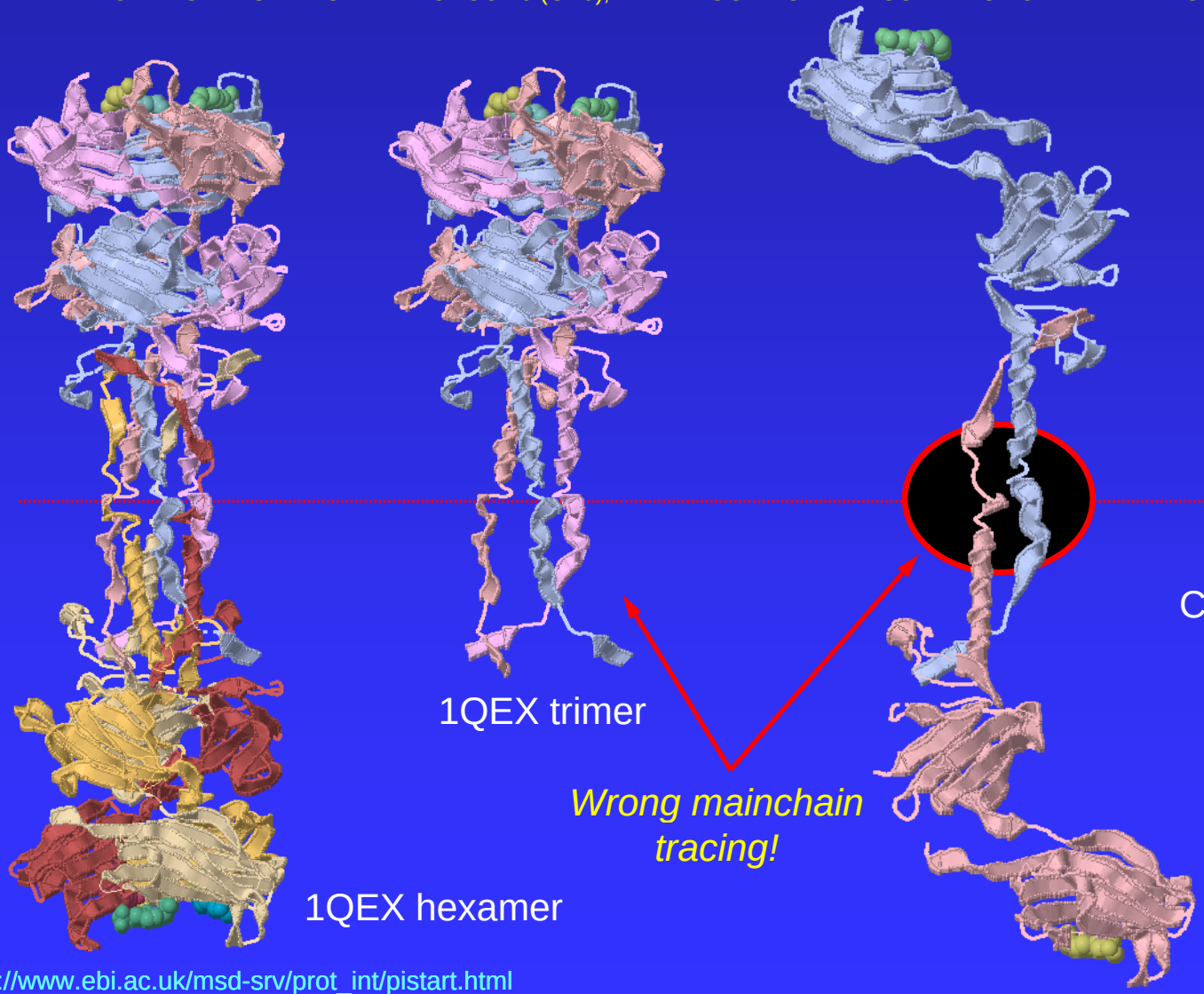




Rossmann M.G., Mesyanzhinov V.V., Arisaka F and Leiman P.G. (2004) *The bacteriophage T4 DNA injection machine*. Curr. Opinion Struct. Biol. **14**:171-180.

Example of misclassification: 1QEX

BACTERIOPHAGE T4 GENE PRODUCT 9 (GP9), THE TRIGGER OF TAIL CONTRACTION AND THE LONG TAIL FIBERS CONNECTOR



1QEX hexamer

1QEX trimer

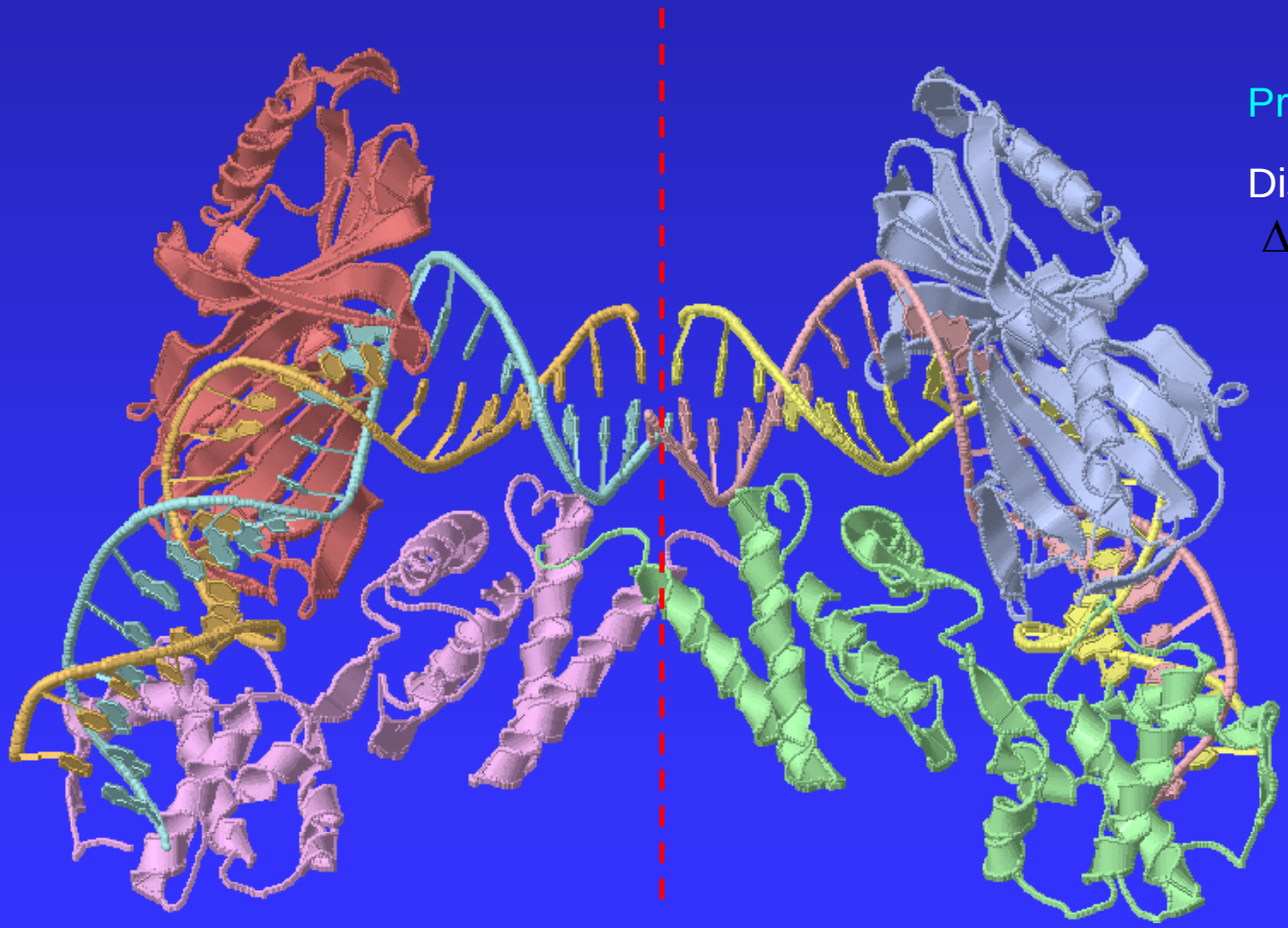
Wrong mainchain tracing!

1S2E trimer
Correct mainchain tracing
Classed correctly

http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

Example of misclassification: 1D3U

TATA-BINDING PROTEIN / TRANSCRIPTION FACTOR



Predicted: octamer

Dissociates into 2 tetramers

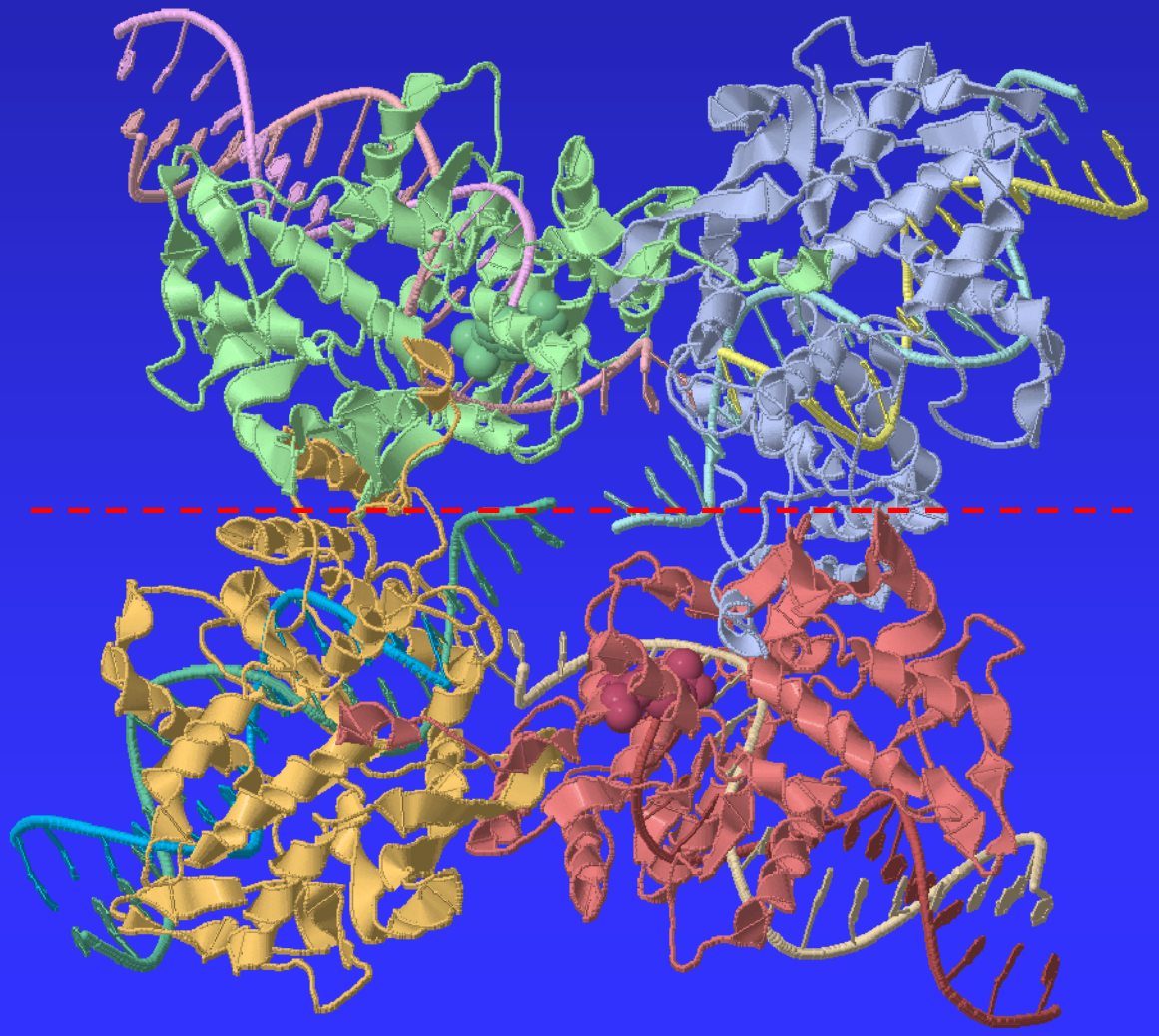
$\Delta G^{0}_{diss} \approx$ *20 kcal/mol*

Functional unit: tetramer

http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

Example of misclassification: 1CRX

CRE RECOMBINASE / DNA COMPLEX REACTION INTERMEDIATE



Predicted: dodecamer

Dissociates into 2 hexamers

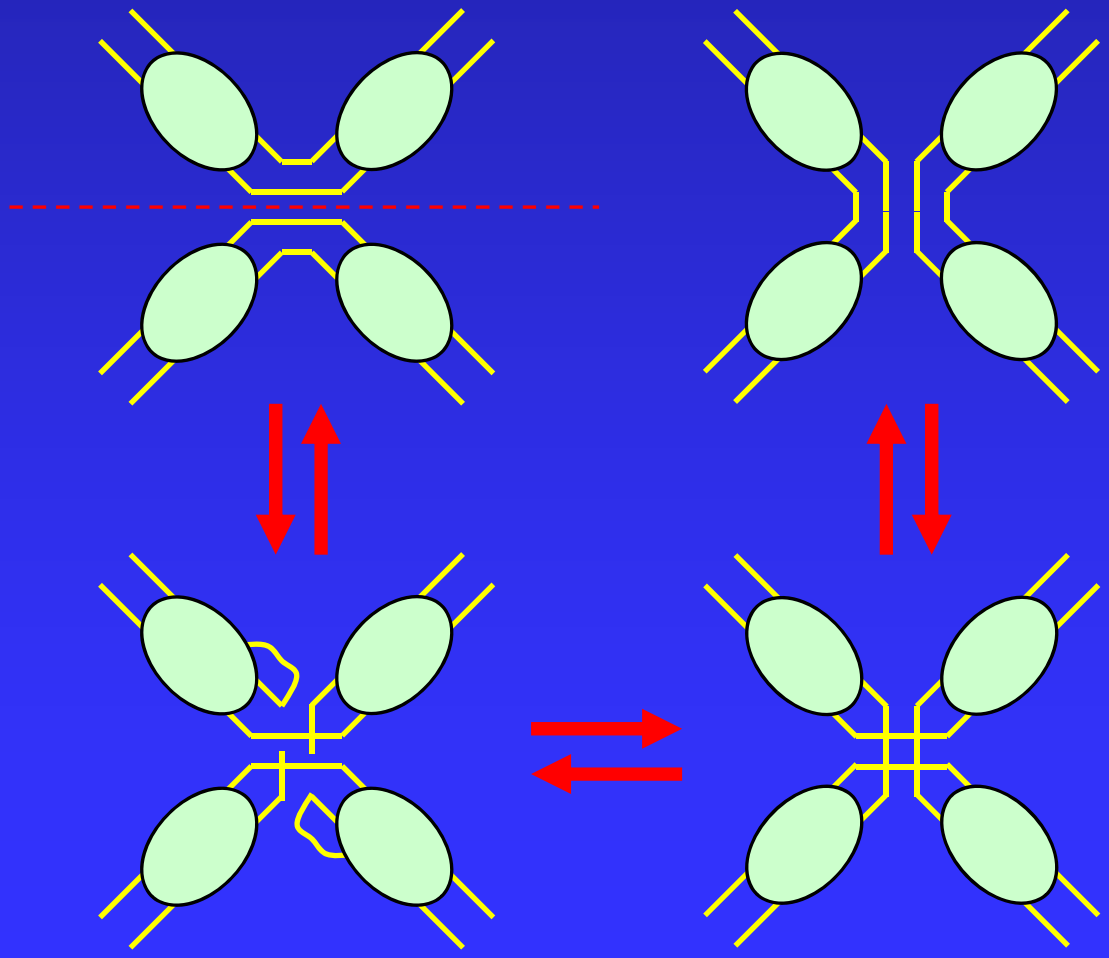
$\Delta G^{0}_{diss} \approx$ *28 kcal/mol*

Functional unit: trimer

http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

Example of misclassification: 1CRX

CRE RECOMBINASE / DNA COMPLEX REACTION INTERMEDIATE



Guo F., Gopaul D.N. and van Duyne G.D. (1997)

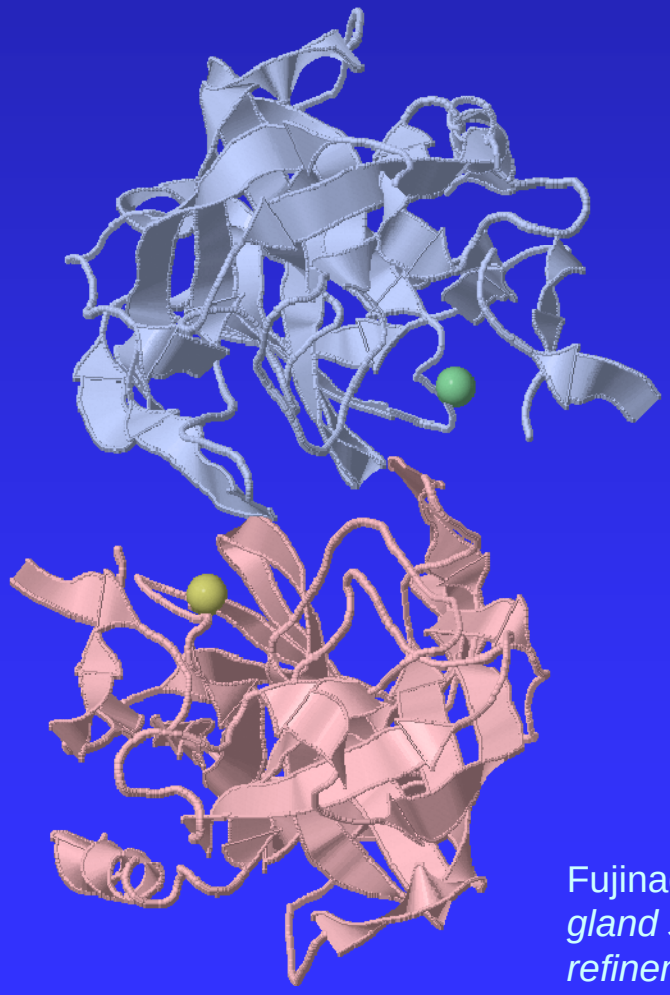
Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse.

Nature **389**:40-46.

http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

Example of misclassification: 1TON

TONIN



Predicted: dimer

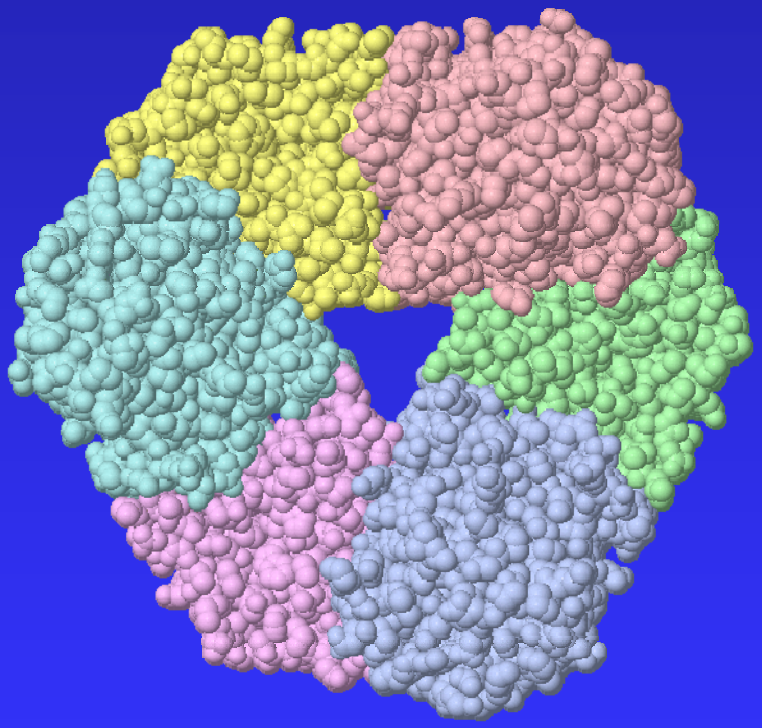
Dissociates at
$\Delta G^{0}_{diss} \approx$ *37 kcal/mol*

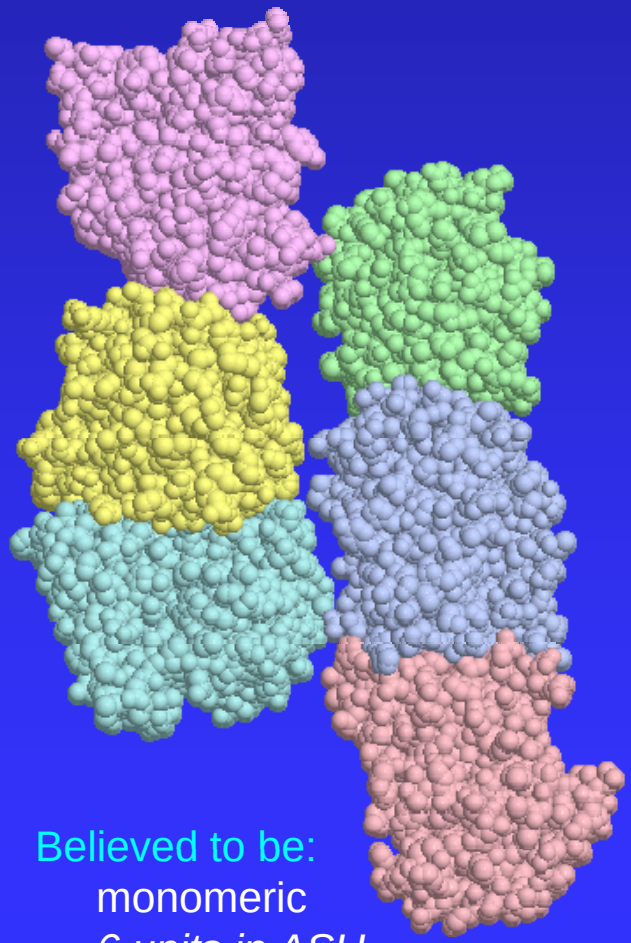
Biological unit: monomer

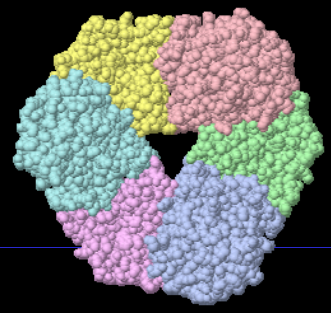
Apparent dimerization is an artefact due to the presence of Zn^{+2} ions added to the buffer to aid crystallization. Removal Zn from the file results in $\Delta G^{0}_{diss} \approx$ *3 kcal/mol*

Fujinaga M., James M.N.G. (1997) *Rat submaxillary gland serine protease, tonin structure solution and refinement at 1.8 Å resolution.* J.Mol.Biol. **195**:373-396.

Example of misclassification: 1YWK







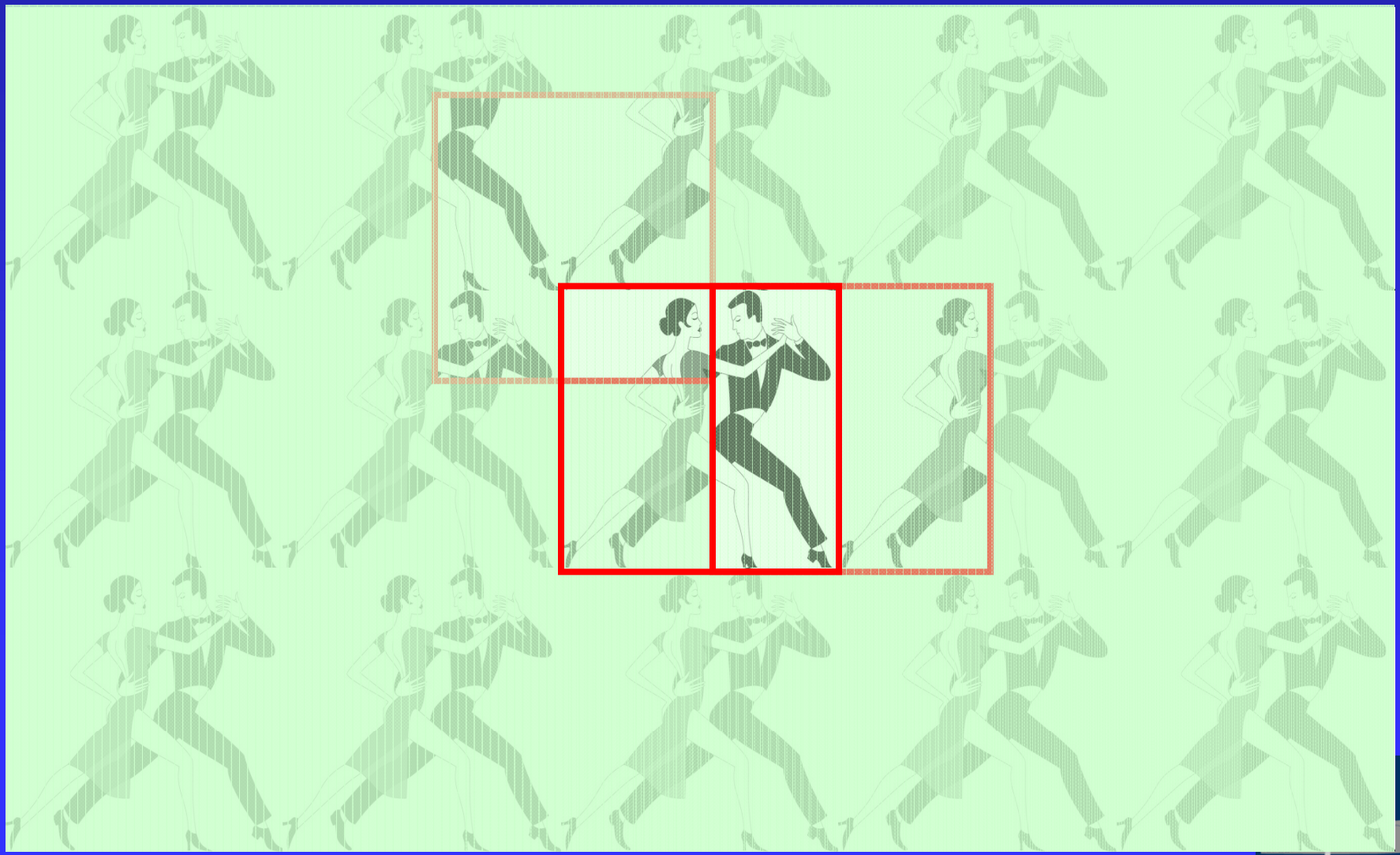
Structural homologue 1XRU:
RMSD ≈ 0.9 Å
Seq.Id ≈ 50%
Homohexameric with
$\Delta G_{diss} \approx 9.3$ *kcal/mol*

Predicted: homohexameric
$\Delta G_{diss} \approx 4.4$ *kcal/mol*
dissociating into 3 dimers

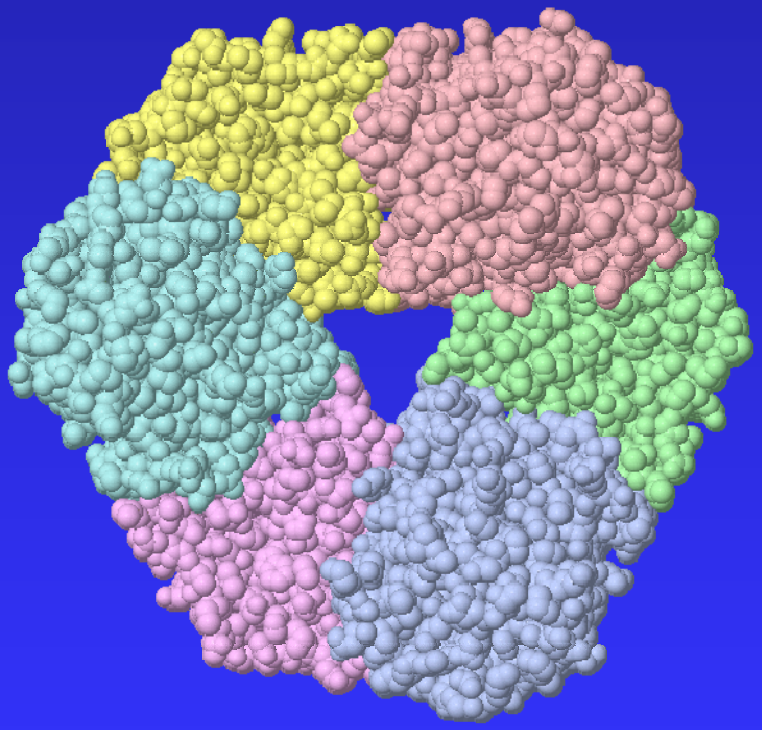
Believed to be:
monomeric
6 units in ASU

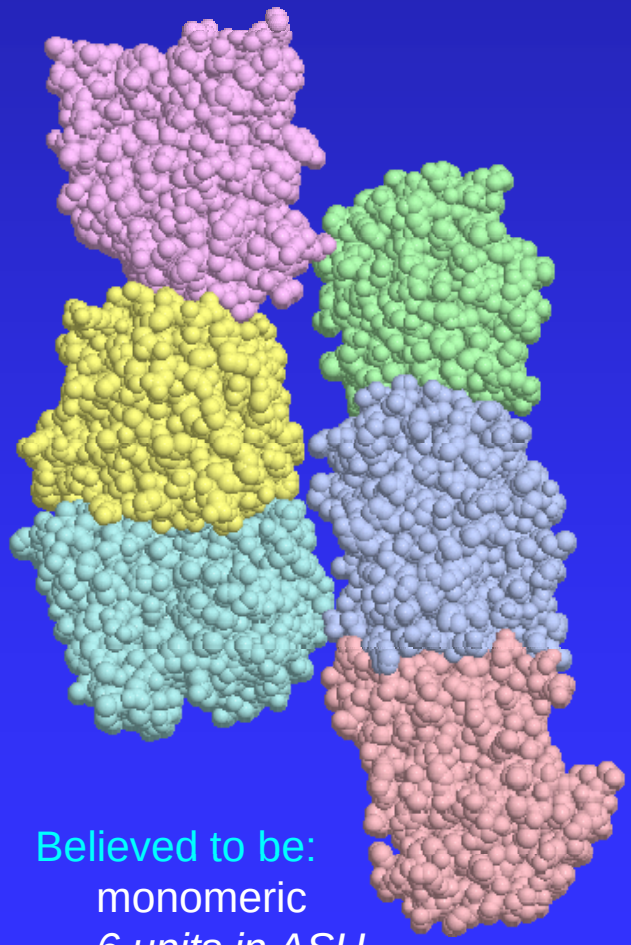
http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

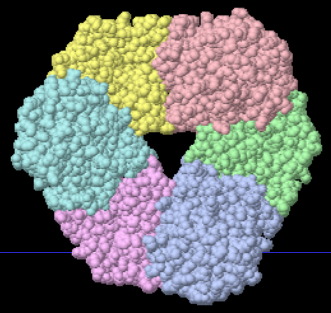
Choice of ASU



Example of misclassification: 1YWK







Structural homologue 1XRU:
RMSD ≈ 0.9 Å
Seq.Id ≈ 50%
Homohexameric with
$\Delta G_{diss} \approx 9.3$ *kcal/mol*

Predicted: homohexameric
$\Delta G_{diss} \approx 4.4$ *kcal/mol*
dissociating into 3 dimers

Believed to be:
monomeric
6 units in ASU

http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

Reasons for the misclassification of Biological Units

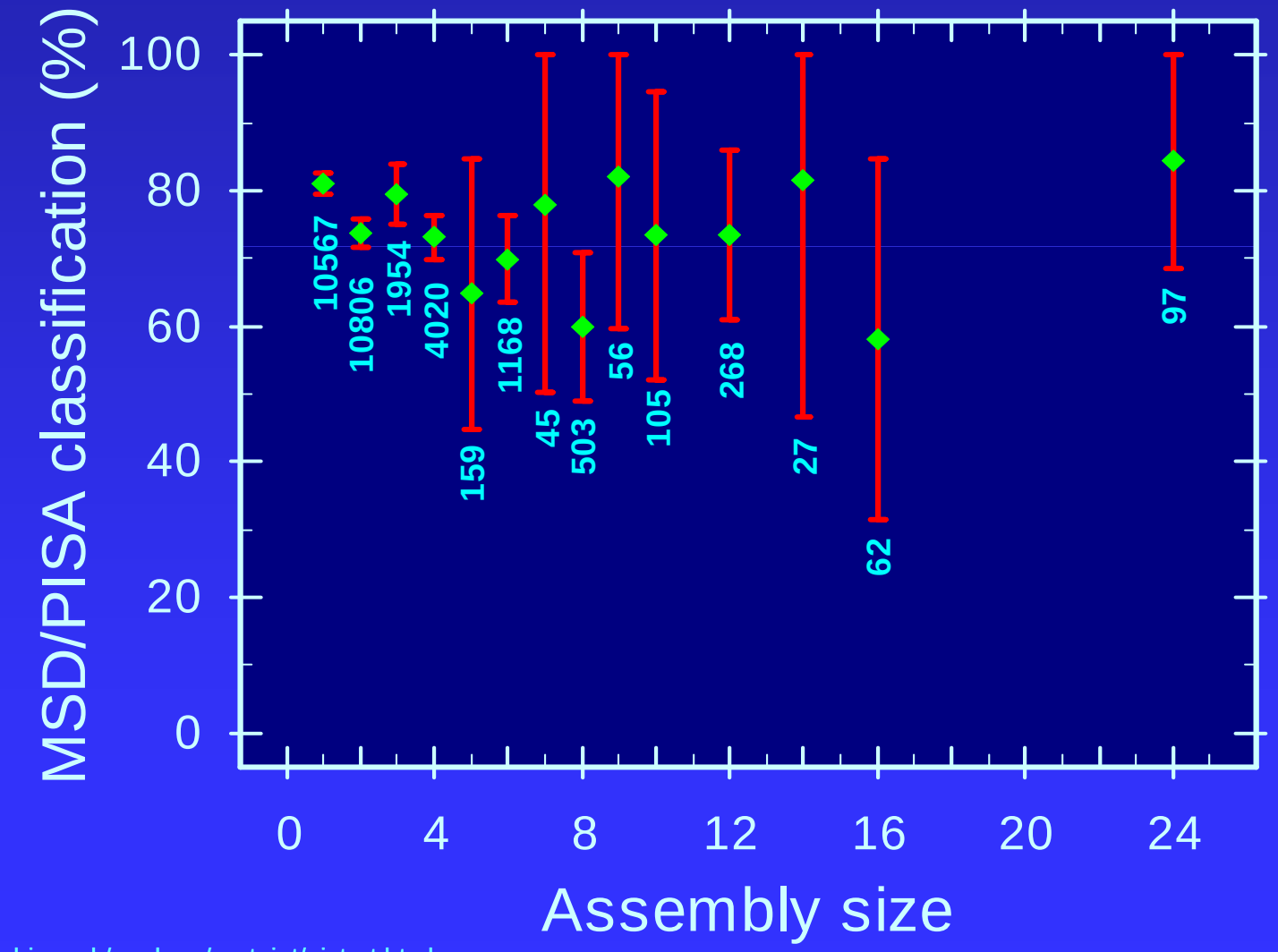
- Crystal packing may introduce interactions that are not engaged in native environment
- Definition of Biological Unit is based on physiological function and may be to a certain degree subjective
- Interaction artifacts are often due to the addition of binding agents in order to aid crystallization

50%

- Theoretical models of macromolecular complexation are simplified
- Experimental data are of a limited accuracy
- No explicit account for concentrations, temperature, pH and ionic strength

50%

Comparison with current PDB annotation

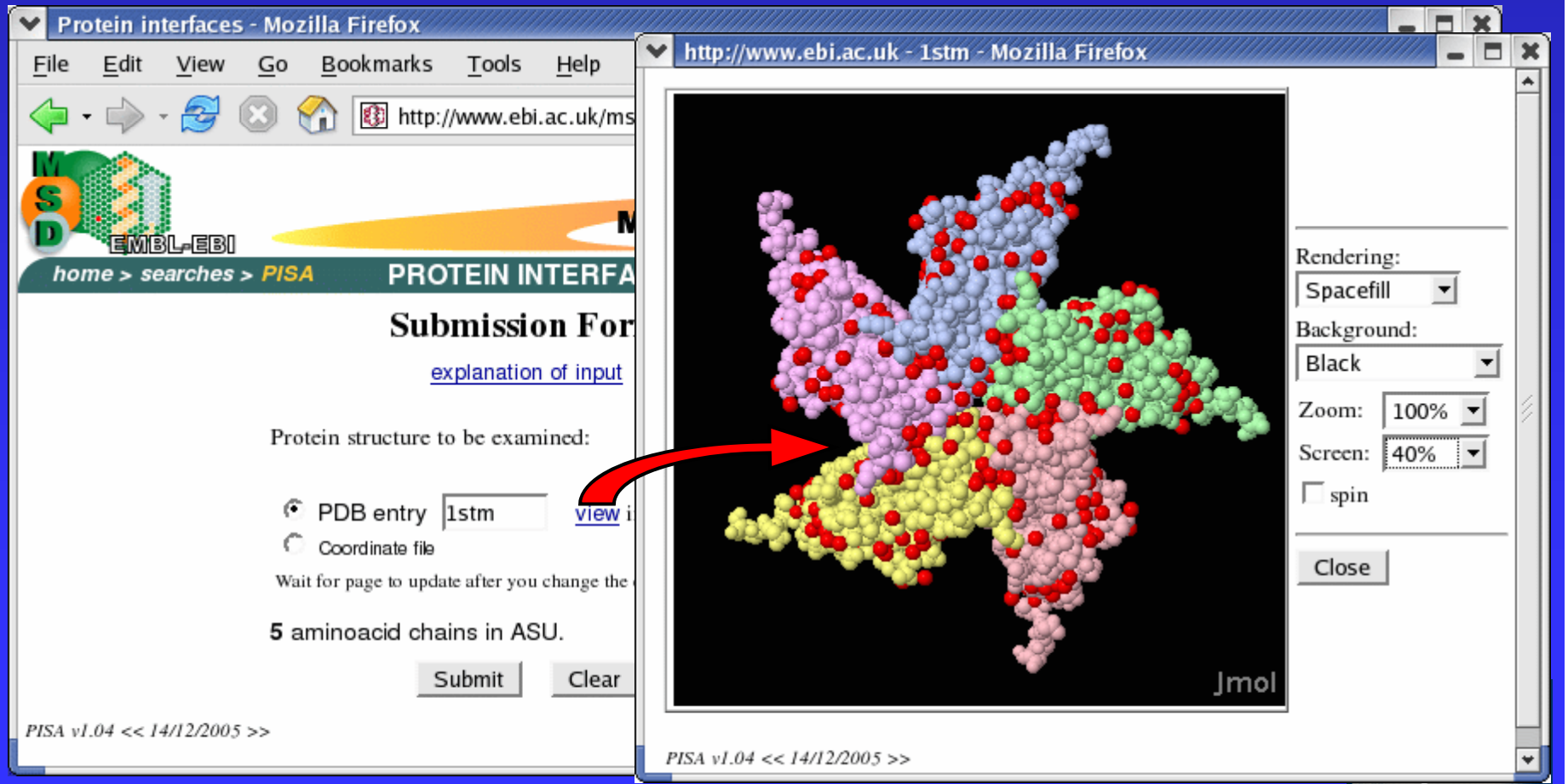


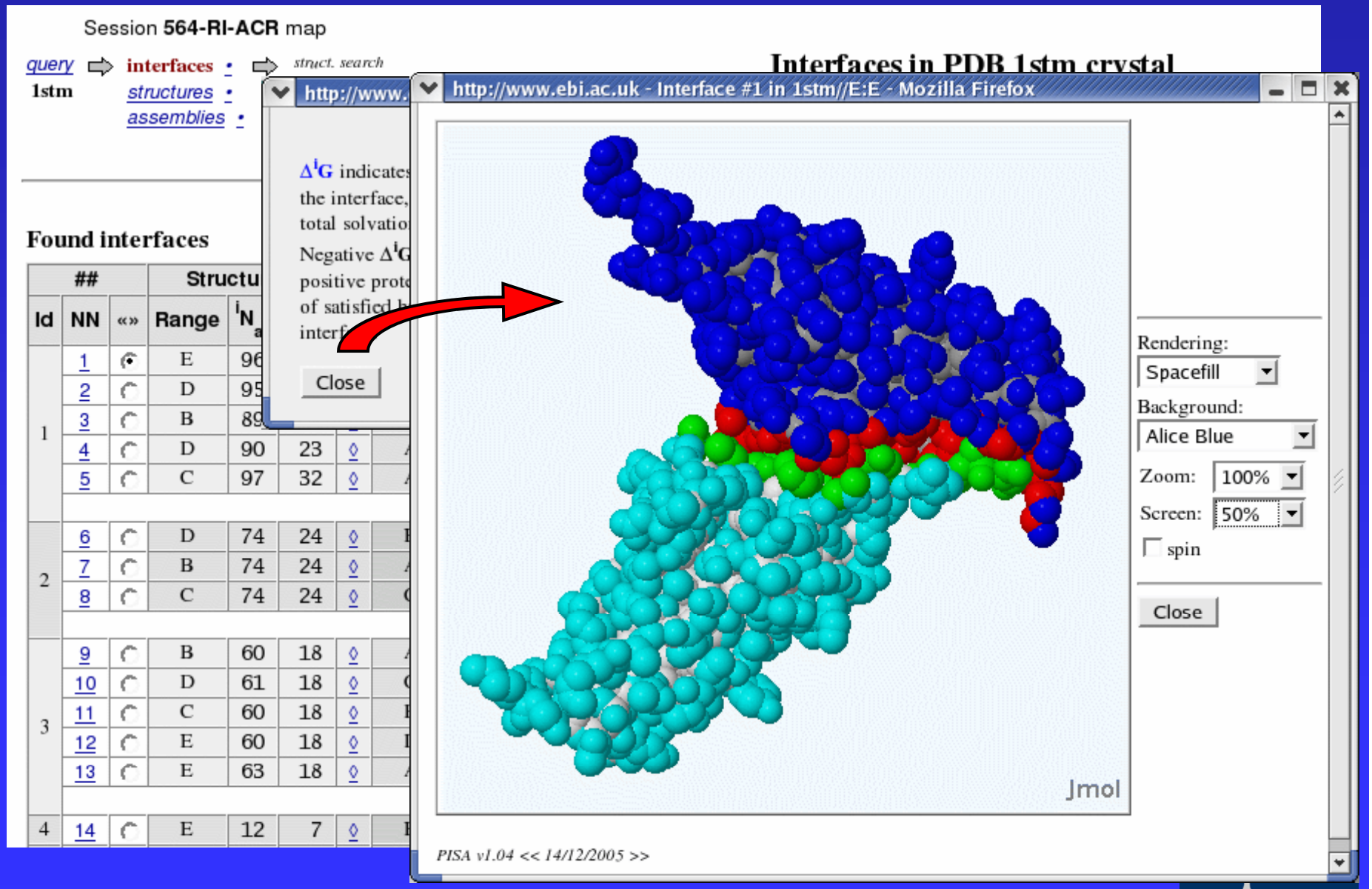
http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

Web-server PISA

http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

A new MSD-EBI tool for working with ***P***rotein ***I***nterfaces, ***S***urfaces and ***A***ssemblies




Session 564-RI-ACR map
query 1stm
interfaces
structures
assemblies
struct. search
Interfaces in PDB 1stm crystal
Found interfaces
Id
NN
Range
Close
http://www.ebi.ac.uk - Interface #1 in 1stm//E:E - Mozilla Firefox
Rendering:
Spacefill
Background:
Alice Blue
Zoom:
100%
Screen:
50%
spin
Close
Jmol
PISA v1.04 << 14/12/2005 >>

Session **564-RI-ACR** map

query **1stm** ⇨ interfaces • structures • assemblies • ⇨ *struct. search* *interfaces* *structures* *assemblies*

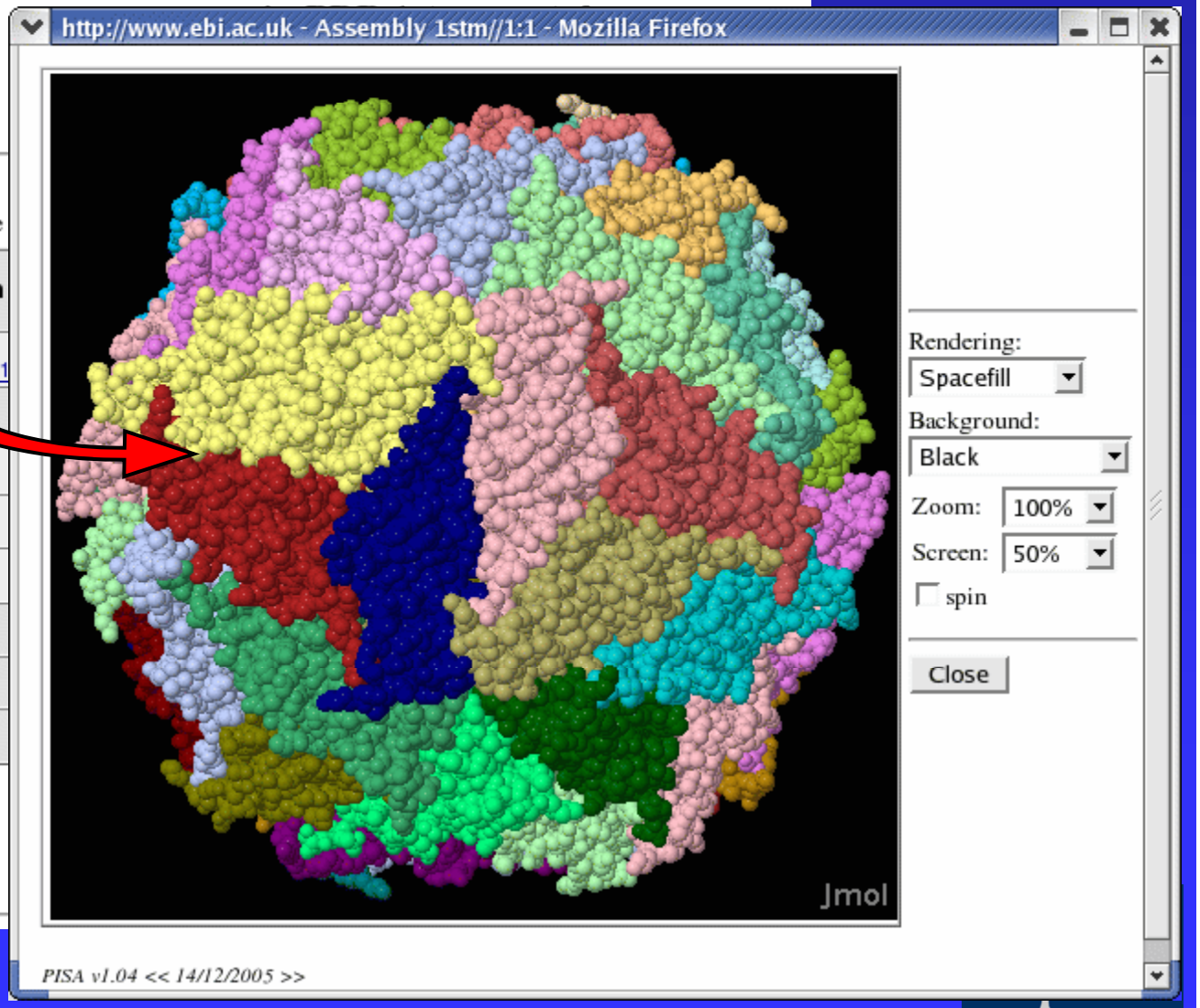
PQS sets 1 to 4 of total 4

Analysis of protein interfaces suggests that the

PQS set NN	«»	Size	Formula	Composition
1	◉	2×60	A_{60}	$A_{12}B_{12}C_{12}D_{12}E_{1}$
2	○	24×5	A_5	ABC
3	○	8×3	A_3	E_3
	○	24×3	A_3	ACD
	○	8×3	A_3	B_3
4	○	24×2	A_2	AB
	○	24×2	A_2	DE
	○	12×2	A_2	C_2

>> view selected assembly
>> details of selected assembly
>> download selected assembly

Probable Quaternary Structures



Conclusions

- Chemical-thermodynamical models for protein complex stability allow one to recover biological units from protein crystallography data at 80-90% success rate
- Considerable part of misclassifications is due to the difference of experimental and native environments and artificial interactions induced by crystal packing
- Functional significance of protein interfaces cannot be reliably inferred only from their properties. Due to entropy contribution and entangled interactions, interface function is also subject to protein complex composition and geometry.
- Protein interface and assembly analysis software (PISA) is available, please use it

Acknowledgements

Kim Henrick *European Bioinformatics Institute*	General introduction, job assignment and gentle push
Mark Shenderovich *Structural Bioinformatics Inc.*	Helpful discussion
Hannes Ponstingl *Sanger Centre*	Sharing the expertise and benchmark data
Sergei Strelkov *University of Leuven*	“Mystery” of bacteriophage T4
MSD & PDB teams *EBI & Rutgers*	Everyday use of PISA, examples, verification and feedback
CCP4 *Daresbury-York-Oxford-Cambridge*	Encouragement and publicity
~4000 PISA users *Worldwide*	Using PISA and feedback
Biotechnology and Biological Sciences Research Council *(BBSRC) UK*	Research grant No. 721/B19544