doi: 10.1038/nature06942

Assembly reflects evolution of protein complexes

Emmanuel D. Levy, Elisabetta Boeri Erba, Carol V. Robinson and Sarah A. Teichmann

Supplementary Discussion

Section S1 It has been suggested that the energetic cost to compensate for the entropic loss when bringing together and holding two proteins is proportional to the log of the masses of the proteins[1,2]. Although simplifications were made in this model, simulations confirmed this proposition[3]. Recent experiments also demonstrated the importance of the entropic loss upon protein association[4]. Thus, for a given protein, the energetic cost of dimer formation is about the same as that of tetramer formation. In other words, in the dihedral tetramer shown Figure 2B, the strength of the green interface, which forms a first dimer, should be comparable to the sum of the strengths of the blue interfaces, which form the dimer of dimers. A single blue interface should have a strength about half that of a green interface. If we assume that interface size reflects interface strength, we can substantiate this by looking at the ratio of interface sizes in tetramers with two contacts per subunit. The distribution of this ratio shows a mean of 1.6 and a median of 2.2, which is compatible with the above hypothesis. However, it does not prove that the larger interface has appeared first during the course of evolution, but we show that this is generally the case in Figure 3b.

Section S2 When looking at membrane bound or transmembrane proteins, the picture changes: we did not find a single transmembrane protein with a dihedral symmetry, while there were 20 with a cyclic symmetry (including C2). A likely explanation for this observation is that dihedral complexes do not have a particular orientation in the cell, while membranes generally separate two different environments[5]. In addition, we expect the energetic barrier of cyclic complex formation to be lower for membrane proteins as they are limited to moving in two dimensions.

Supplementary Figures

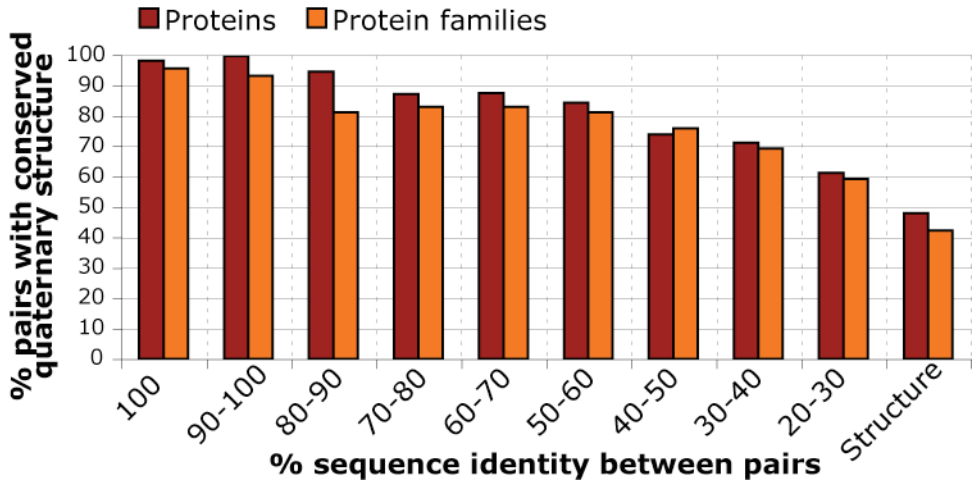


Figure S1 Histogram showing the conservation of QS as a function of protein sequence similarity. A non-redundant set of structure pairs was derived for each range of sequence identity considered (see methods). Red bars indicate the fraction of pairs with the same QS, as defined by the internal symmetry of the complex. Orange bars also represent the fraction of pairs with same QS, except that a single pair is considered per protein family. For sequence identities above 90%, conservation is nearly 100%. The conservation then decreases progressively reaching 70% in the range of 30-40% sequence identity. At sequence identities below 30%, where there is only structural similarity, the conservation drops to ~50%. For these proteins, the conservation may be underestimated due to potential errors in the quaternary structure assignments (we did not systematically curate the matches below the 30% identity threshold, as described in more detail in the Methods section). Thus above 30% sequence identity, QS is well conserved. This result can be integrated with other levels of structural conservation, such as domain-domain geometry[6-8], for structural homology modelling of the cellular machinery[9].

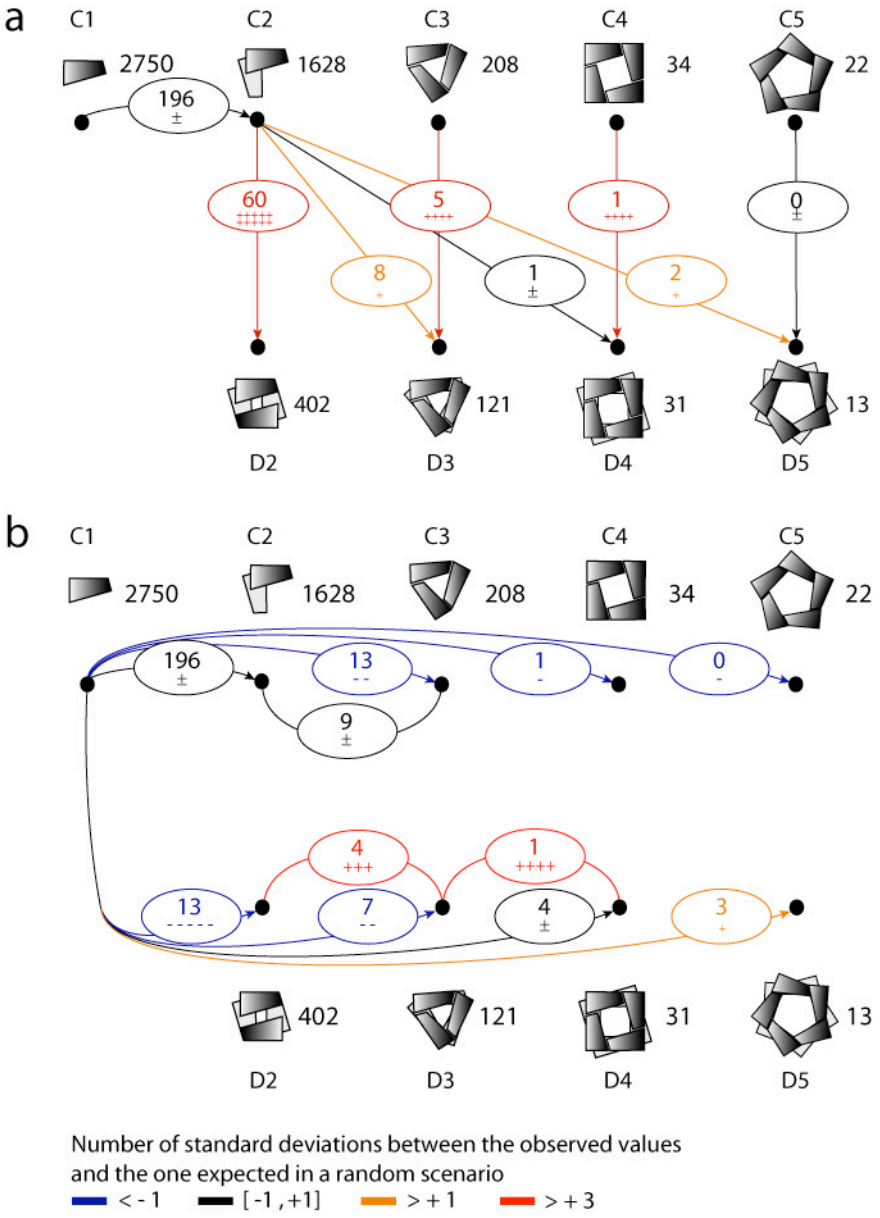
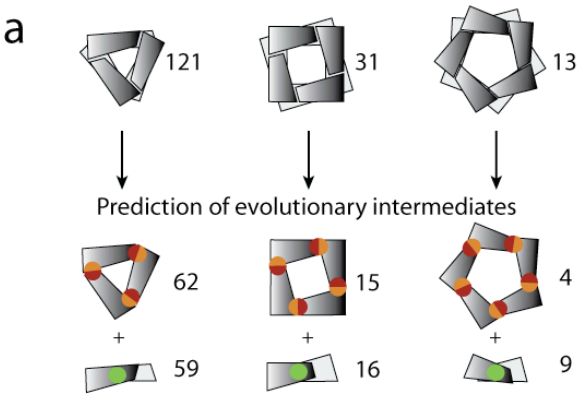


Figure S2 a, The figure shown is the same as that in the main text except that numbers of homologs shared between the different quaternary structure types are given. **b**, The same information as in panel a, for evolutionary relationships that are less favourable.



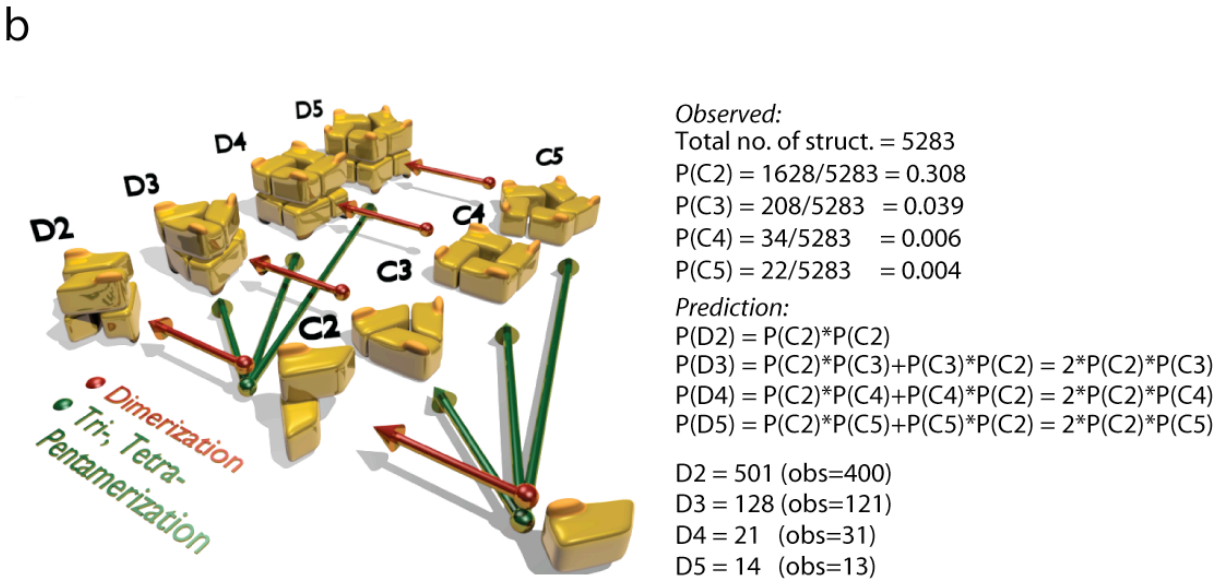


Figure S3. a For these three QS types, we predict that about half evolved from dimers and the other half evolved from their cyclic counterpart, suggesting that these two potential evolutionary routes are not significantly different ($\chi 2$ test; p=0.2). This observation may seem surprising, as one expects dimers to be found more frequently as intermediates than larger cyclic complexes. However there is a straightforward explanation for this result: dimerization is frequent during evolution, while formation of cyclic complexes is more rare. Therefore, once a dimer is formed, it will be difficult for it to form a dihedral complex (larger than D2). By contrast, a cyclic trimer will frequently shift its oligomeric state towards a dihedral hexamer by a dimerization event. In other words, to evolve towards a dihedral complex, there is an "easy" step (dimerization), and a "limiting" step (cyclization), but the order in which the steps are achieved does not matter (commutative property). To test this hypothesis further, we formulated a model of homomer evolution based on this commutative property (part b of the Figure). Consistent with our model, we can successfully predict the observed abundances of dihedral complexes from the abundances of cyclic complexes only. **b** Model for homomer evolution. We have seen in Fig. S3a that the order of events in forming a particular dihedral complex does not matter: dimeric and larger cyclic homomers are equally probable evolutionary intermediates to form a dihedral complex. This led us to propose the model of homomer evolution shown on the left, where the probability of forming a dihedral complex is simply the product of the probabilities of the events involved. That is, the probability that a monomer or a trimer dimerize to form a dimer and a hexamer respectively are the same. We

validate our model by using these "universal" probabilities to predict the abundance of each type of dihedral complex, as shown on the right of the figure. These predictions are accurate: a chi^2 test indicates that predicted and observed numbers are not significantly different, (p=0.2). According to our model, the probability of obtaining a dihedral hexamer is simply p(D3) = p(C2)*p(C3) + p(C3)*p(C2), where p(C2) and p(C3) are the probabilities of forming a dimer and a trimer respectively. These probabilities are estimated by the number of dimers (C2) and trimers (C3), divided by the total number of structures T, which includes monomers: p(C2) = (C2/T), and p(C3) = (C3/T). Then, the number of dihedral hexamers D3 expected with the model can be simply expressed as: D3 = 2*T*P(C2)*P(C3), and the number of dihedral tetramers as: D2 = $T*P(C2)^2$. Therefore, this model explains the abundances observed for the different symmetry types, and also accounts for our previous observation that the two evolutionary routes leading to dihedral complexes are equally favoured.

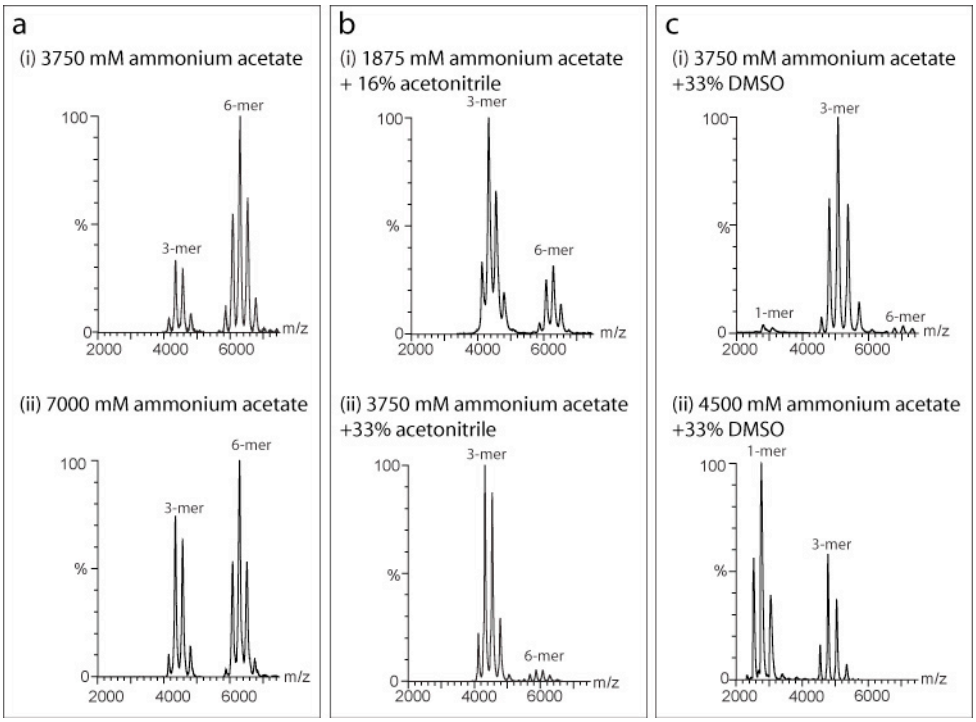


Figure S4 Illustration of the search for the conditions to disrupt a protein complex and generate sub-complexes. Here we show the mass spectra of the hexameric AU binding homolog of enoyl-CoA hydratase (1hzd) and the progressive increase in proportions of sub-complexes (trimer and monomer) obtained after addition of organic solvents and/or change in ionic strength. **a**. Increasing the ionic strength two-fold was found to disrupt the hexamer-trimer interface (i -> ii). This effect is more marked as the concentration of acetonitrile is increased (b (i -> ii)). Addition of DMSO disrupts the complex to a similar extent as acetontrile but in this case a minor population of monomeric subunits is also visible (c(i)). This monomeric dissociation product becomes the predominant species when the ionic strength is increased (c(ii)).

Supplementary Tables

Table S1 Evolutionary relationships between pairs of proteins with different quaternary structures that share an interface. Note that each domain architecture is considered only once. Interface size was measured by the number of amino acids in contact (please refer to [10] for details of contact calculation). The interface ratio is the result of the division of the smaller interface by the larger one, so is always less than 100%. When more than two interfaces are present in a tetramer, we also indicate a second ratio that corresponds to the sum of the two smaller sizes divided by the largest one, which can yield a ratio greater than 100%.

No.	Code1 (D2)	Code2 (C2)	%id	Larger Interface?	Interfaces ratio
1	1non	1a3c	73	Y	55%
2	1t2a	1db3	60	Y	90%-110%
3	3mds	1gv3	58	Y	86%
4	1vjp	1j5p	56	Y	37%
5	1o58	1d6s	52	Y	78%
6	1b9b	2btm	49	Y	24%
7	3pgm	1e58	48	Y	81%
8	1bj4	1kkp	45	Y	18%
9	1ub3	1o0y	45	Y	38%
10	1lk5	1m0s	41	Y	92%-116%
11	1b25	1aor	39	N	89%
12	1sru	1se8	39	Y	28%
13	1m3k	1afw	35	Y	23%-32%
14	1inl	1mjf	35	Y	78%
15	1f8f	1e3i	34	Y	38% - 77%
16	1j1y	1ixl	32	Y	44%-62%
17	1f8w	1lvl	31	Y	5%
18	1a4s	1ad3	30	Y	32%-61%
19	1a16	1pv9	30	Y	30%
20	1eyi	1dk4	30	Y	46%
21	1rli	1x77 (1rtt)	30	Y	58%
22	1fo6	1f89	29	Y	28%-42%
23	1cbl	1umo	28	Y	56%-78%
24	1m41	1tvl	28	Y	48%-68%
25	1qsm	1bo4	28	Y	14%
26	1pfk	1kzh	28	N	72%
27	1bfd	1jsc	27	Y	47%-63%
28	1j2r	1nf9	27	Y	43%
29	1nuq	1kam	26	N	94%
30	1x94	1nri	26	Y	31%
31	1sjw	1ocv	25	Y	30%
32	1dq8	1qax	24	Y	25%-39%
33	1uwt	1cbg	24	Y	75%
No.	**Code1 (D3)**	**Code2 (C2)**	**%id**	**Larger Interf.**	**Ratio**
1	1nqb	1moe	69	Y	18%
2	1cks	1puc	51	Y	68%
3	1tqj	1h1y	44	Y	50%
4	1kr2	1kam	40	Y	34%
5	1k3p	1aj8	36	Y	15%

6	1pjc	1qp8	36	Y	91%
7	1tzf	1vgz	36	Y	82%
8	1a3g	1daa	29	Y	60%
9	1nw4	1jys	28	Y	62%
10	1u1z	1mka	27	Y	73%
11	1uzb	1ad3	25	Y	53%
12	1pmm	1ajr	22	Y	65%
13	1on3	1od2	21	Y	50%
No.	**Code1 (D3)**	**Code2 (C3)**	**%id**	**Larger Interf.**	**Ratio**
1	1pi2	1k9b	67	Y	50%
2	1t3d	1xat	39	Y	14%
3	1p8c	1knc	37	Y	76%
4	1uiy	1jxz	34	Y	39%
5	1mkz	1uuy	32	Y	42%
6	1gq6	1hqf	29	Y	80%

Table S2 Summary of complexes for which (dis)assembly was studied. Percentages are given as v/v.

PDB code	Protein	Disassembly pathway	Partial denaturing conditions	Protein complexes were provided by:
1m3u	Ketopantoate Hydroxymethyltransferase	10 to 2	25% methanol, 25% DMSO, 525 mM ammonium acetate	Prof. Abell (University of Cambridge, United Kingdom)
1m8p	P. chrysogenum ATP Sulfurylase	6 to 2	1500 mM ammonium acetate	Prof. Fisher (University of California, USA)
1vea	HutP, an RNA binding antitermination protein	6 to 1	25% DMSO, 2500 mM ammonium acetate	Prof. Kumar (National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan)
1j2p	Alpha-ring from the proteasome from archaeoglobus fulgidus	7 to 1	33% methanol, 33% DMSO 33 mM ammonium acetate	Prof. Groll (Charite´ Medical School of the Humboldt-University, Berlin, Germany)
1hkx	Calcium Calmodulin-dependent protein kinase	14 to 2	25% methanol, 25% DMSO 50 mM ammonium acetate	Prof. Kuriyan (University of California, USA)
1pvv	Ornithine carbamoyltransferase from Pyrococcus furiosus	12,9,6 to 3	33% ACN, 366 mM ammonium acetate	Dr. Massant (Vrije Universiteit Brussel, Belgium)
1umg	Fructose-1,6-bisphosphatase	8 to 2	25% methanol 25% DMSO, 1900 mM ammonium acetate	Dr. Takayoshi Wakagi and Hiroshi Nishimasu (University of Tokyo, Japan)
1hzd	AU binding homolog of enoyl-CoA hydratase	6 to 3	33% DMSO, 4500 mM ammonium acetate	Prof. Yutaka Muto (University of Tokyo, Japan)
1ekr	Molybdenum cofactor biosynthesis protein C	6 to 2	100 mM ammonium acetate	Prof. Schindelin (University of Würzburg, Germany)
1grl	GroEL	14 to 7	25mM Tris	Sigma Aldrich

Table S3 Summary of complexes for which information on (dis)assembly was found in the literature, and for which there is also a crystal structure, or a close homolog in the case of 1nhk.

PDB Code	Protein name	Reference with information on assembly or disassembly	Agrement with prediction (larger interface forms first or break last)
1pfk	Phosphofructokinase I	[11]	Yes
1nhk	Nucleotide diphosphate kinase	[12]	No (no intermediate observed)
1t3d	Serine Acetyltransferase	[13]	Yes
1aon	GroES	[14]	Yes
1di0	Lumazine synthase	[15]	Yes
1ogf	Ribose transport protein RbsD	[16]	Yes

References

1 J. Janin and C. Chothia, *J Mol Biol* **100** (2), 197 (1976).
2 A. V. Finkelstein and J. Janin, *Protein Eng* **3** (1), 1 (1989).
3 J. M. Swanson, R. H. Henchman, and J. A. McCammon, *Biophys J* **86** (1 Pt 1), 67 (2004).
4 K. K. Frederick, M. S. Marlow, K. G. Valentine et al., *Nature* **448** (7151), 325 (2007).
5 D. S. Goodsell and A. J. Olson, *Annu Rev Biophys Biomol Struct* **29**, 105 (2000).
6 J. H. Han, N. Kerrison, C. Chothia et al., *Structure* **14** (5), 935 (2006).
7 F. P. Davis and A. Sali, *Bioinformatics* **21** (9), 1901 (2005).
8 P. Aloy, H. Ceulemans, A. Stark et al., *J Mol Biol* **332** (5), 989 (2003).
9 P. Aloy, B. Bottcher, H. Ceulemans et al., *Science* **303** (5666), 2026 (2004).
10 E. D. Levy, J. B. Pereira-Leal, C. Chothia et al., *PLoS Comput Biol* **2** (11), e155 (2006).
11 W. Teschner and J. R. Garel, *Biochemistry* **28** (4), 1912 (1989).
12 A. Giartosio, M. Erent, L. Cervoni et al., *J Biol Chem* **271** (30), 17845 (1996).
13 V. J. Hindson, P. C. Moody, A. J. Rowe et al., *J Biol Chem* **275** (1), 461 (2000).
14 T. Higurashi, K. Nosaka, T. Mizobata et al., *J Mol Biol* **291** (3), 703 (1999).
15 V. Zylberman, P. O. Craig, S. Klinke et al., *J Biol Chem* **279** (9), 8093 (2004).
16 Y. Feng, W. Jiao, X. Fu et al., *Protein Sci* **15** (6), 1441 (2006).