Science Advances

AAAS

advances.sciencemag.org/cgi/content/full/3/1/e1600552/DC1

Supplementary Materials for

An ambiguity principle for assigning protein structural domains

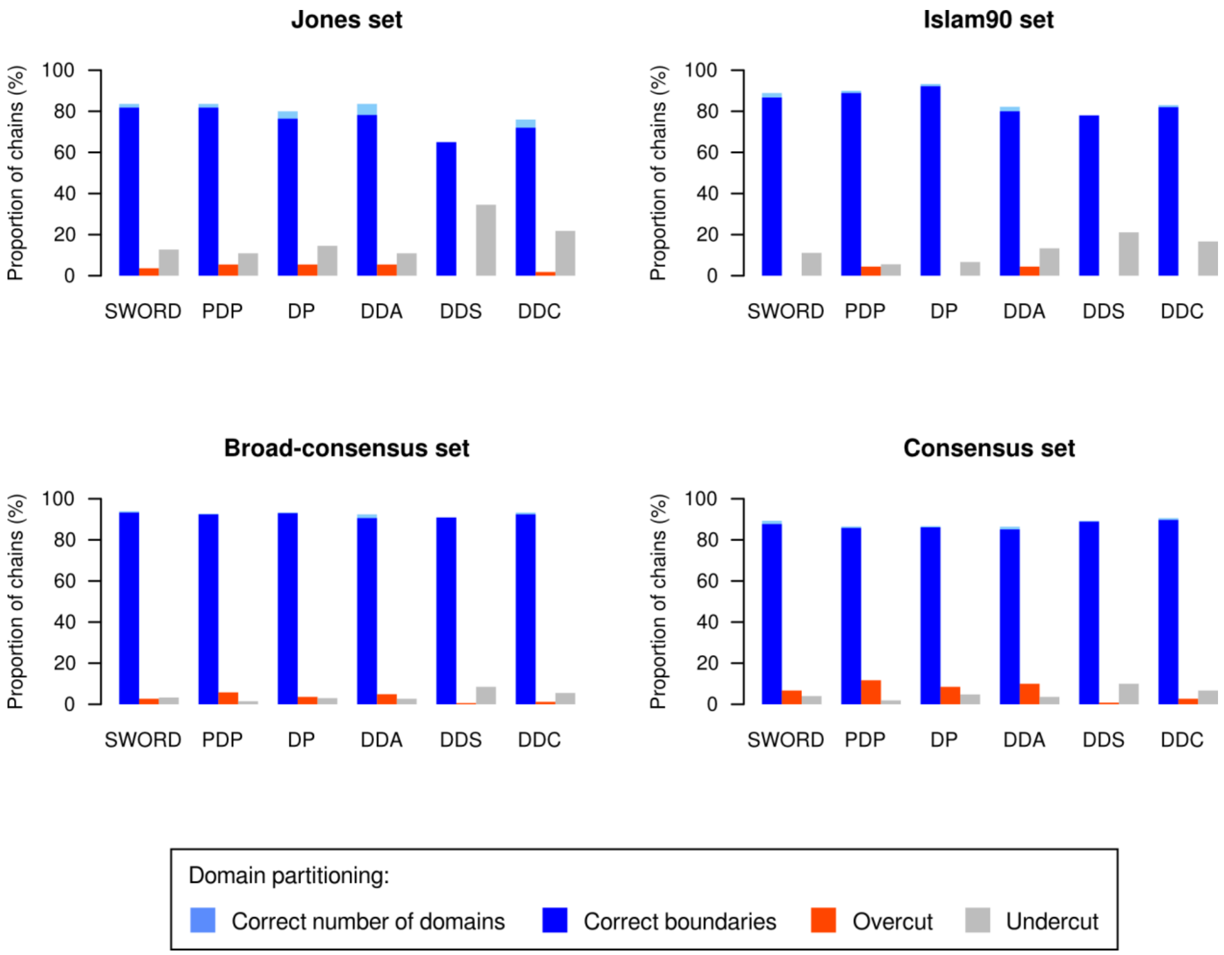
Guillaume Postic, Yassine Ghouzam, Romain Chebrek, Jean-Christophe Gelly

Published 13 January 2017, *Sci. Adv.* **3**, e1600552 (2017)
DOI: 10.1126/sciadv.1600552

This PDF file includes:

- fig. S1. Monopartitioning accuracies of SWORD, PDP, DomainParser, and DDomain.
- fig. S2. Rate of agreement between SWORD and CATH, SCOP, or ECOD annotations, depending on the number of assignments provided, for the structures of the Strong-dissensus data set.
- fig. S3. Representation of the domain assignment model.
- fig. S4. Domain assignments of the 1A8YA protein structure, as displayed by SWORD.
- table S1. Rate of agreement between SWORD and annotations from the five data sets of structural domains, depending on the number of assignments provided.
- table S2. The 34 most ambiguous protein structures of the Consensus set.
- table S3. The *P* values of the Mann-Whitney-Wilcoxon tests comparing the A-index means of the Consensus, Dissensus, and Strong-dissensus sets.
- table S4. The *P* values of the Mann-Whitney-Wilcoxon and Pearson's χ^2 tests comparing the A-index distributions of the Consensus and Dissensus sets.
- equation S1. The contact probability between two PUs.

SUPPLEMENTARY MATERIALS



	SWORD	PDP	DP	DDA	DDS	DDC
Jones	83.6; 81.8; 3.6; 12.7	83.6; 81.8; 5.5; 10.9	80.0; 76.4; 5.5; 14.5	83.6; 78.2; 5.5; 10.9	65.5; 65.5; 0.0; 34.5	76.4; 72.7; 1.8; 21.8
Islam90	88.9; 86.7; 0.0; 11.1	90.0; 88.9; 4.4; 5.6	93.3; 92.2; 0.0; 6.7	82.2; 80.0; 4.4; 13.3	78.9; 78.9; 0.0; 21.1	83.3; 82.2; 0.0; 16.7
Consensus	89.3; 87.7; 6.7; 4.0	86.5; 85.8; 11.7; 1.9	86.7; 86.1; 8.5; 4.8	86.4; 85.1; 10.0; 3.6		
Broad-consensus	93.9; 93.3; 2.7; 3.3	92.7; 92.4; 5.8; 1.5	93.3; 93.0; 3.6; 3.0	92.4; 90.6; 4.9; 2.7		

fig. S1. Monopartitioning accuracies of SWORD, PDP, DomainParser, and DDomain. In each cell of the table, the four values (%) separated by semicolons represent, from left to right, i) the proportion of correct assignments without the boundaries overlap criterion, ii) the proportion of correct assignments with the boundaries overlap criterion, iii) the proportion of overcut assignments and iv) the proportion of undercut assignments.

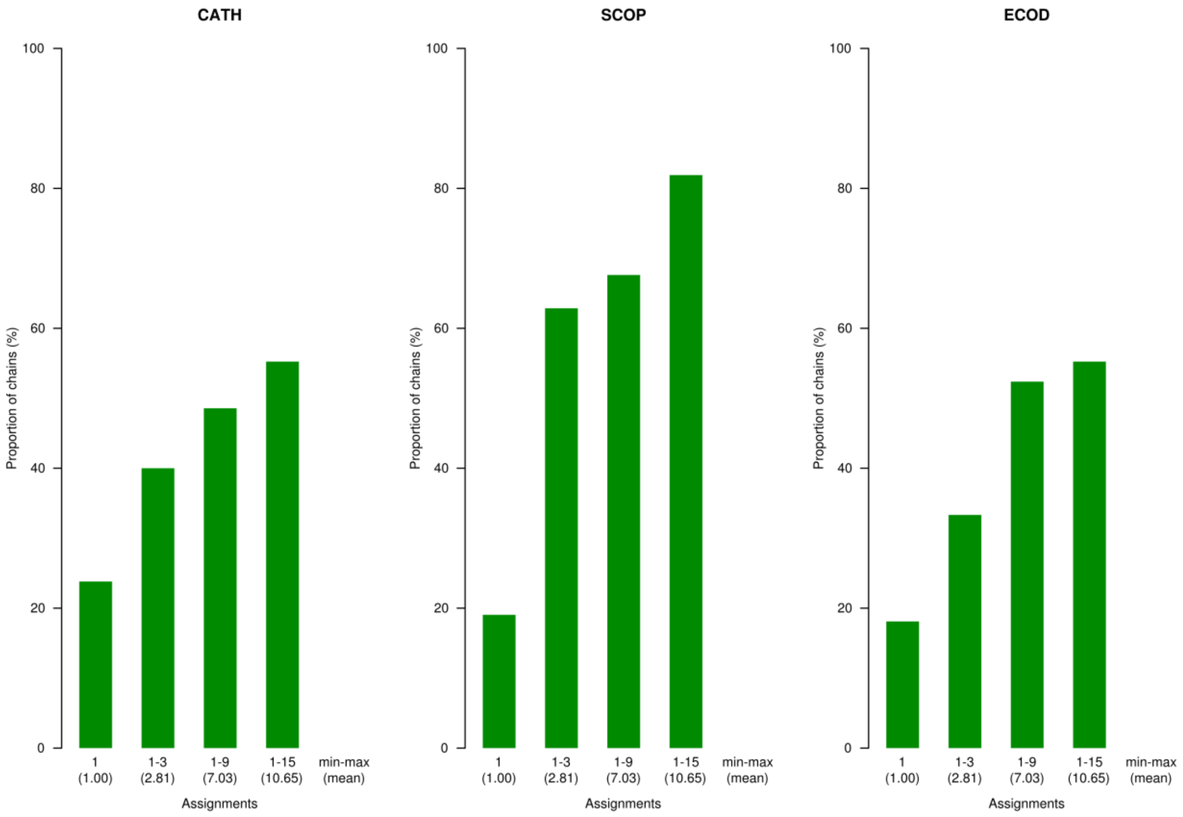


fig. S2. Rate of agreement between SWORD and CATH, SCOP, or ECOD annotations, depending on the number of assignments provided, for the structures of the Strong-dissensus data set. (values are given in table S1).

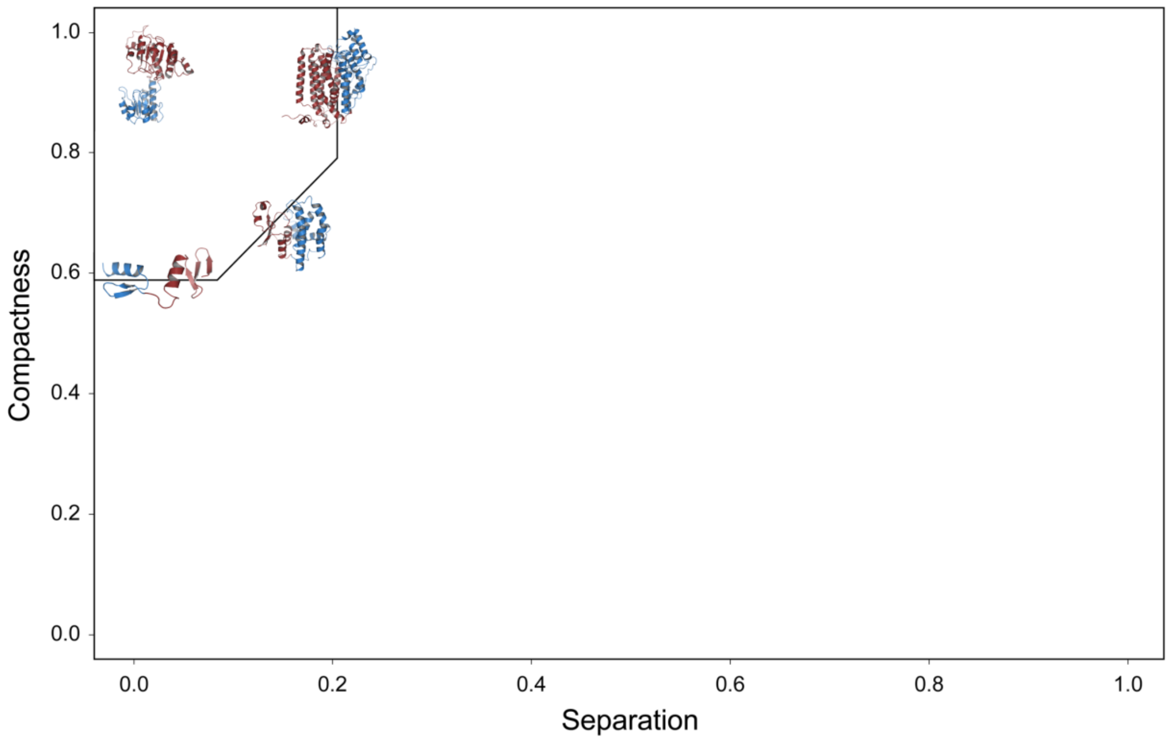


fig. S3. Representation of the domain assignment model. Three difficult cases of protein structure partitioning are illustrated on the boundaries of the acceptance region and an easy case is shown near the top left corner.

```
PDB: 1A8YA
ASSIGNMENT
#D|Min|                                                  BOUNDARIES|   AVERAGE κ|   QUALITY|
3 |102|                                       3-124 125-226 227-347|    3.564247|      ****|
ALTERNATIVES
#D|Min|                                                  BOUNDARIES|   AVERAGE κ|   QUALITY|
6 |30 |           3-38;69-124 39-68 125-193 194-226 227-292 293-347|    3.102774|         *|
6 |30 |   3-38;69-124 39-68 125-193 194-226 227-257;293-347 258-292|    3.020496|         *|
6 |30 |   3-38;69-124 39-68 125-157;194-226 158-193 227-292 293-347|    3.016253|         *|
5 |30 |                   3-38;69-124 39-68 125-226 227-292 293-347|    3.205836|         *|
5 |30 |                   3-38;69-124 39-68 125-193 194-226 227-347|    3.194637|         *|
5 |33 |                       3-124 125-193 194-226 227-292 293-347|    3.184732|         *|
4 |30 |                           3-38;69-124 39-68 125-226 227-347|    3.346431|         *|
4 |48 |                               3-124 125-226 227-292 293-347|    3.334050|         *|
4 |33 |                               3-124 125-193 194-226 227-347|    3.320050|         *|
3 |102|                                       3-124 125-226 227-347|    3.564247|      ****|
3 |30 |                           3-38;69-124;227-347 39-68 125-226|    3.359427|         *|
3 |48 |                               3-124;227-292 125-226 293-347|    3.339514|         *|
2 |102|                                       3-124;227-347 125-226|    3.622611|      ****|
2 |122|                                               3-124 125-347|    3.581932|      ****|
2 |114|                                               3-226 227-347|    3.575431|      ****|
1 |338|                                                       3-347|    0.000000|       n/a|
```

fig. S4. Domain assignments of the 1A8YA protein structure, as displayed by SWORD. The optimal partitioning is provided under 'ASSIGNMENT' and all other decompositions under 'ALTERNATIVES'. Each of these lines includes the number of domains (#D), the amino acid length of the smallest domain (Min), the sequence positions of the delimited domains (BOUNDARIES), the average compactness per domain (AVERAGE κ) and a qualitative assessment of the decomposition (QUALITY). These last two features are not applicable for 1-domain assignments (*i.e.*, no partitioning). In the 'BOUNDARIES' column, domains are separated by spaces and each part of non-contiguous domains is separated by a semicolon (;). For a given number of domains, the alternative delineations in terms of boundaries are ranked by their 'AVERAGE κ' value (the higher the better). Here, the structure of 1A8YA is optimally decomposed into 3 domains. Therefore, the next level of decomposition corresponds to the 4-domain and 2-domain assignments. Among the decompositions of this next level, the 'QUALITY' value helps to decide which number of domains is the best (in this case, it is 2).

table S1. Rate of agreement between SWORD and annotations from the five data sets of structural domains, depending on the number of assignments provided. The second column represents the decomposition levels considered, *i.e.*, the number of alternative assignments in terms of number of domains. The third column represents the number of alternative assignments in terms of domain boundaries, for a given number of domains. Thus, the first line corresponds to the optimal decompositions provided by SWORD. The fourth column is the product of the previous two columns and therefore represents the maximum number of assignments provided per query structure. The fifth column is the mean number of assignments provided per query structure. The last column is the proportion of agreement between SWORD assignments and annotations of the datasets.

Dataset	Assignments provided				Correct assignments (%)		
	Levels	Boundaries	Max.	Mean			
Jones	1	1	1	1.00	81.82		
	3	1	3	2.40	90.91		
	3	3	9	5.31	94.55		
	5	3	15	7.96	96.36		
Islam90	1	1	1	1.00	86.67		
	3	1	3	2.03	94.44		
	3	3	9	3.73	95.56		
	5	3	15	5.09	95.56		
Consensus	1	1	1	1.00	87.62		
	3	1	3	2.17	94.95		
	3	3	9	4.29	96.85		
	5	3	15	6.41	98.04		
Broad-consensus	1	1	1	1.00	93.31		
	3	1	3	2.07	96.66		
	3	3	9	3.88	97.57		
	5	3	15	5.63	98.18		
Dissensus	1	1	1	1.00	37.46[a]	33.37[b]	
	3	1	3	2.67	60.39	76.78	
	3	3	9	6.31	73.66	83.80	
	5	3	15	9.45	78.34	90.63	
Strong-dissensus	1	1	1	1.00	23.81[a]	19.05[b]	18.10[c]
	3	1	3	2.81	40.00	62.86	33.33
	3	3	9	7.03	48.57	67.62	52.38
	5	3	15	10.65	55.24	81.90	55.24

[a] CATH annotations; [b] SCOP annotations; [c] ECOD annotations

table S2. The 34 most ambiguous protein structures of the Consensus set.

2C78A	1LVAA	1DFCA	1OLZA
1ORVA	1GG3A	1N8YC	1XM9A
2AWIA	1G7SA	1VCLA	1JDHA
1GG4A	1YFSA	1A8YA	2VGLB
1WPGA	1YVRA	1Q2LA	1B3UA
1NKGA	1ZPDA	1W0PA	1B89A
2QTVA	2EZ9A	3BMVA	2BPTA
1CIYA	2Q66A	1F5NA	
1CZAN	1OYWA	1E8CA	

table S3. The *P* values of the Mann-Whitney-Wilcoxon tests comparing the A-index means of the Consensus, Dissensus, and Strong-dissensus sets.

	Consensus	**Dissensus**	**Strong-dissensus**
Consensus	-	3.369×10^{-14}	4.634×10^{-8}
Disensus	-	-	9.696×10^{-4}
Strong-dissensus	-	-	-

table S4. The *P* values of the Mann-Whitney-Wilcoxon and Pearson's χ^2 tests comparing the A-index distributions of the Consensus and Dissensus sets.

Test\Size range	**100–200**	**200–300**	**300–400**	**>400**
Mann-Whitney-Wilcoxon	$< 2.2 \times 10^{-16}$	1.367×10^{-11}	1.036×10^{-6}	1.045×10^{-4}
Pearson's chi-squared	$< 2.2 \times 10^{-16}$	1.37×10^{-9}	1.302×10^{-6}	1.216×10^{-4}

equation S1. The contact probability between two PUs.

The contact probability $p_{i,j}$ between two Protein Units i and j can be written as

$$p_{i,j} = \frac{1}{1 + \exp\left[\frac{d_{i,j} - d_0}{\Delta}\right]}$$

where $d_{i,j}$ is the Euclidean distance between the Cα of the Protein Units i and j, and the parameters d_0 and Δ are set to 6 Å and 1.5 Å, respectively (see Gelly *et al.*, 2006; PMID: 16301202).