Supplementary Table S1. The structure and function of segment-swapped proteins as annotated in various databases

PDB chains (central protein listed first)	Name or function	SCOP	CATH	main function from GO or literature	Summary
1rf6A	5-enolpyruvyl-shikimate-3-phosphate synthase	α+β; 6 repeats of IF3 fold organized into two RPTC like doms	α-β prism	enzyme, transferase	α-β prism (6 repeats of IF3 fold organized into two RPTC (RNA 3'-terminal phosphate cyclase)-like domains)
2yvwA	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	-	α-β prism	enzyme, transferase	
1g6sA	EPSP synthase	same	α-β prism	same	
1ejdA	UDP-N-acetylglucosamine enolpyruvyltransferase	same	α-β prism	same	
2pqcA	EPSP synthase	-	α-β prism	same	
2o0bA	EPSP synthase	-	-	same	
2rl1A	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	-	α-β prism	same	
2ql3A	LysR transcriptional regulator, C-term dom	-	3-layer αβα sandwich, maltodextrin binding protein	transcription regulator	3-layer αβα sandwich, periplasmic binding protein-like II
3hhfA	CrgA reg dom (LysR type transcr. reg)	-	-	transcription regulator	
1ixcA	CbnR (LysR type transcr. reg)	α/β; Periplasmic binding protein-like II	same	transcription regulator	
1yavA	B. subtilis, unknown function	α+β; CBS domain pair	-	adenosyl binding	α+β; CBS domain pair
3hf7A	Uncharacterized CBS domain pair	-	-	adenosyl binding	
2emqA	G. caustophilus, unknown function	-	-	adenosyl binding	
2r58A	Polycomb protein, sex comb on midleg (SCM) protein	-	domain 1: SH3 type barrel	transcription regulator	all-β, SH3-like barrel, two MBT repeats
2bivA	Sex comb on midleg like protein 2 (Scml2)	all-β; SH3-like barrel, two MBT repeats	same	transcription regulator	
1oz2A*	Lethal (3) malignant brain tumor (MBT)-like protein	same	same	cell cycle regulator, transcription regulator	
1wcwA	Uroporphyrinogen III synthase (T. thermo.)	-	3-layer αβα sandwich, Rossmann fold	enzyme, lyase	3-layer αβα sandwich, Rossmann, HemD-like
1jr2A	Uroporphyrinogen III synthase (human)	α/β; HemD-like	same	enzyme, lyase	
1n00A	Annexin GH1 (cotton)	all-α, Annexin	4 domains, mainly-α, orthog. bundle, Annexin V domain 1	membrane binding	all-α, orthogonal bundle, annexin
1dk5A	Annexin 24(ca32)	all-α, Annexin	same	phospholipid binding	

3d3aA	β-galactosidase	-	-	enzyme, hydrolase	2-layer β sandwich
2vqaA	SLL1358 protein MncA, periplasmic metal binding protein, cupin	-	-	metal binding	Double-stranded β helix with three short α helices added
2qqrA	Jumonji domain containing histone demethylation protein 3A	all-β, SH3-like barrel, tandem repeat of two segment-swapped Tudor domains	-	enzyme, oxidoreductase (demethylase)	SH3-like, Jumonji domain, two segment-swapped Tudor domains
2q5cA	NtrC family transcriptional regulator	α/β; Chelatase-like, PrpR receptor domain-like	dom 1: 3-layer αβα sandwich, Rossmann; dom 2: PrpR receptor domain like	transcriptional regulator	3-layer αβα sandwich, Rossmann? PrpR receptor domain-like
2q0tA	Putative gamma-carboxymuconolactone decarboxylase	all-α, AhpD-like, duplication: subunits form a helix-swapped trimer	mainly α, up-down bundle, AhpD-like	enzyme, lyase, decarboxylase	all-α up-down bundle AhpD-like. Note: Swapping is also present between subunits!
2jh3A	DR2241, ribosomal protein S2 related protein, CbiX	-	-	unknown, may be related to cobalamin biosynthesis	3-layer αβα sandwich, Rossmann?
2hcrA	phosphoribosyl pyrophosphate synthetase I	-	domain 1: 3-layer αβα sandwich, Rossmann fold;	enzyme, transferase	3-layer αβα sandwich, Rossmann?
2h9fA	Unknown function, DUF453	α+β; Diaminopimelate epimerase-like, PA0793-like	ab, Roll, Diaminopimelate epimerase	unknown, may be enzyme, isomerase (PrpF)	ab roll, diaminopimelate epimerase like
2et6A	Hydroxyacyl-CoA dehydrogenase domain of multifunctional β-oxidation protein	-	-	enzyme, oxidoreductase	3-layer αβα sandwich, NAD(P)-binding Rossmann fold
2b5iD	Interleukin-2 receptor α chain	small disulphide-rich all-β; segment-swapped SCR/Sushi domains	-	receptor	all-β, disulphide-rich, complement control module, two segment-swapped SCR domains
2a90A	WWE domain of Deltex protein (Drosophila)	α+β; WWE domain fold	-	Signaling, protein-protein interaction	α+β, WWE domain
1y3tA	YxaG, quercetin 2,3-dioxygenase	all-β; Double-stranded β-helix, RmlC-like cupins, Quercetin 2,3-dioxygenase-like	-	enzyme, oxidoreductase	double-stranded β helix, RmlC-like cupins, Quercetin 2,3-dioxygenase-like

Supplementary Table S2. The 12 central proteins of the permissively defined segment-swapped proteins along with their structure and function as annotated in various databases. A full listing of the families can be found in Supplementary Dataset S1. According to the permissive definition, a protein was considered as segment-swapped when TM(BA,M) > 0.5 (i.e. Domain 1 with its segments swapped was found similar to Domain 2) without making sure that segment-wise structural similarity also applies; i.e. alignment coverages were not checked, and TM-scores were normalized by the length of the shorter chain, thus allowing for large insertions and deletions. This definition is intentionally loose in order to identify proteins that might have originated by segment-swapping but where the domains have diverged too much for a confident detection of segment-swapping. Using these criteria, 452 potential segment-swapped proteins were identified in ReprPDB; these could be clustered into 57 clusters. The central proteins of these clusters were compared to those found using the stricter definition. Out of the 57 central proteins, 12 had global structural similarity (TM-score >0.5) to a segment-swapped protein found by the stricter definition, and 7 had structural similarity limited to one domain. From the remaining 38 central proteins, 26 were found to be false hits (β-propellers, $(\beta\alpha)_8$-barrels, αα-barrels, etc.). Finally, 12 central proteins (representing 84 proteins) appear to be valid potential segment-swapped proteins that had not been identified by the stricter definition. A visual inspection of the structures confirmed that segment-swapping may have occurred in these cases, but a distortion of the domain structure, or large insertions (e.g. in 2nydA), deletions, or the addition or removal of large terminal segments (e.g. in 3i04A or 2gagC) made the domains too dissimilar to unambiguously recognize these cases as segment-swapped proteins. In some less convincing cases, the extent of the swap is fairly small (e.g. only a single β-strand is "swapped" between the domains, like in 3gc3A or the 5-domain 3hrzA).

PDB chains	Name or function	SCOP	CATH	main function from GO or literature	Summary
3ladA	lipoamide dehydrogenase	3-layer ββα, FAD/NAD(P) binding domain	3-layer ββα, FAD/NAD(P) binding domain	oxidoreductase	3-layer ββα sandwich
3h8lA	sulfide:quinine oxidoreductase	-	-	oxidoreductase	
2zbwA	thioredoxin reductase like	-	3-layer ββα, FAD/NAD(P) binding domain	?	
3i04A	bifunctional carbon monoxide dehydrogenase / acetyl-CoA synthase	-	-	enzyme	3-layer αβα sandwich
2rkbA	serine dehydratase like-1	-	3-layer αβα, Rossmann fold	?	
2nydA	unknown	-	-	?	
3delB	arginine binding protein	-	-	transport	
3gc3A	clathrin heavy chain 1	-	-	scaffold	2-layer β-sandwich
2f68X	collagen adhesin	-	-	binding	
3hrzA	cobra venom factor	-	-	binding	
2gagC	heterotetrameric sarcosine oxidase gamma subunit	-	-	metalloenzyme	αβ sandwich
2aprA	aspartic proteinase	β-barrel, pepsin-like	β-barrel	enzyme	complex fold with β-barrels and sheets

Supplementary Table S3. Comparisons of fold ages of the two domains (Domain 1, AB-type; Domain 2, BA-type) of the SSPs listed in Table 1. The numbers of domain analogs (pooled for each family) and their source organisms in ReprPDB and in 22 genomes, respectively, are listed for each protein family. Higher numbers suggest older folds. The last column shows the phylogenetic position corresponding to the first appearance of each fold as determined in a phylogenetic tree of the 22 organisms. The position is the height of the node of earliest appearance of the fold as counted from the root of the tree; thus, higher numbers indicate a younger fold. Families where any of the domain folds was found in less than 2 species were omitted.

Group	Central protein	In ReprPDB			In 22 genomes		
		All domain analogs AB-type / BA-type	Domain analogs with <30% sequence identity[a] AB-type / BA-type	Source organisms of all domain analogs AB-type / BA-type	Proteins containing domain analogs AB-type / BA-type	Source organisms of domain analogs AB-type / BA-type	Phylogenetic position of appearance AB-type / BA-type
Mainly-α	1n00A	0 / 0	0 / 0	0 / 0	0 / 1	0 / 1	
	2q0tA	1 / 0	1 / 0	1 / 0	21 / 1	11 / 1	
Mainly-β	2r58A	0 / 4	0 / 4	0 / 1	3 / 0	2 / 0	
	2qqrA	189 / 0	54 / 0	43 / 0	67 / 0	6 / 0	
	3d3aA	78 / 13	41 / 11	37 / 11	0 / 0	0 / 0	
	2b5iD	43 / 0	8 / 0	4 / 0	32 / 5	4 / 3	1 / 20
	1y3tA	2 / 0	2 / 0	2 / 0	13 / 0	5 / 0	
3-layer αβα sandwich	2ql3A	15 / 1	5 / 1	13 / 1	11 / 4	9 / 4	0 / 1
	1wcwA	415 / 200	196 / 110	137 / 98	212 / 13	21 / 10	0 / 2
	2q5cA	88 / 40	48 / 24	43 / 24	31 / 9	20 / 8	0 / 1
	2et6A	0 / 0	0 / 0	0 / 0	12 / 1	6 / 1	
	2hcrA	295 / 136	115 / 75	105 / 66	10 / 5	9 / 5	0 / 0
	2jh3A	143 / 12	49 / 10	67 / 10	5 / 11	4 / 10	0 / 0
Other αβ	1rf6A	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	
	1yavA	0 / 1	0 / 1	0 / 1	68 / 1	20 / 1	
	2vqaA	0 / 0	0 / 0	0 / 0	35 / 8	2 / 6	1 / 1
	2h9fA	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	
	2a90A	0 / 0	0 / 0	0 / 0	7 / 2	2 / 1	

[a] The number of analogs after reducing the set to <30% pairwise sequence identity.

Figure S1. A full gallery of segment-swapped proteins

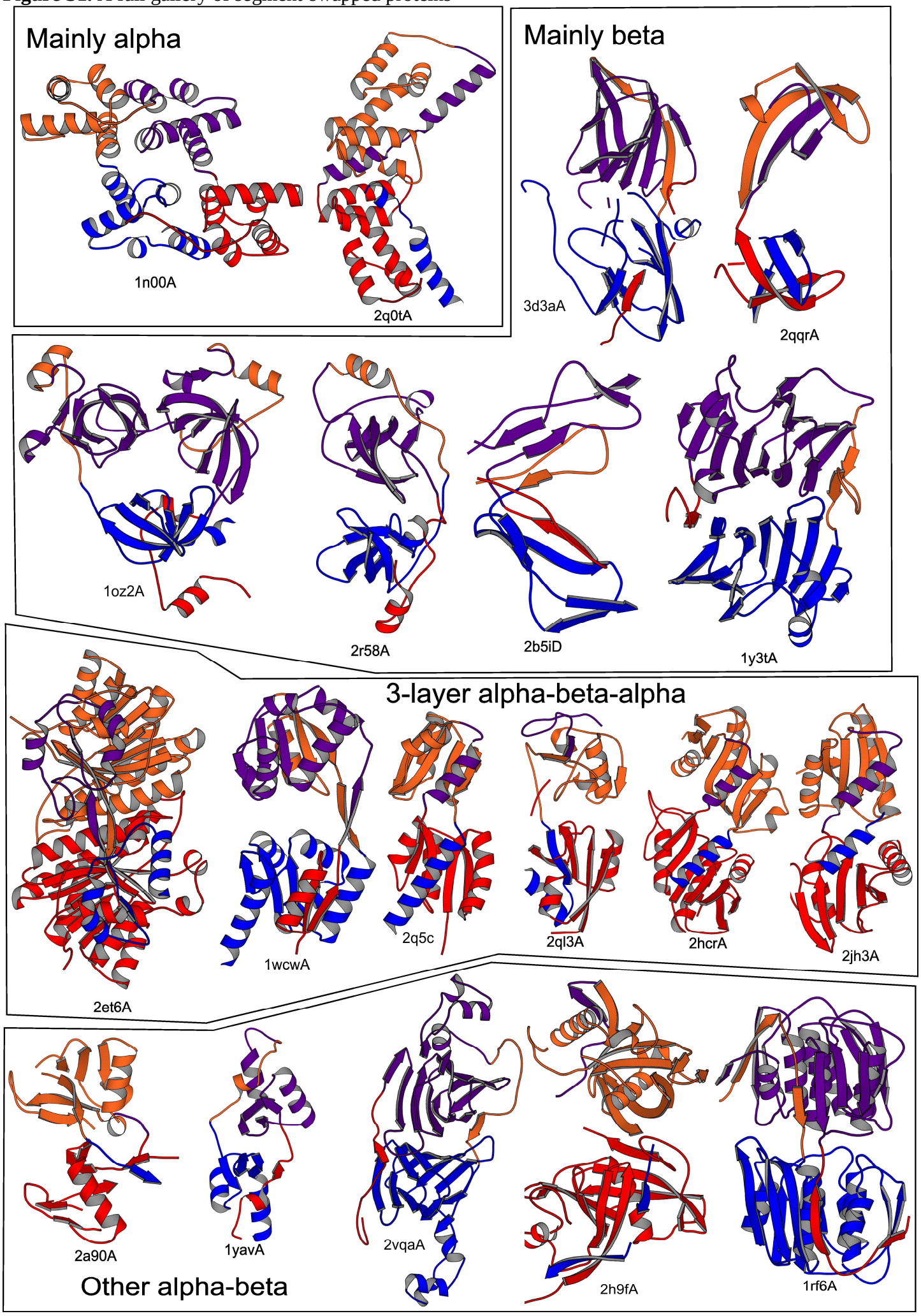

Figure S2. Major circular permutants of Domain 2 of 1wcwA. We scanned ReprPDB for circular permutants of the domain fold of 1wcwA, and found 291 hits. These were grouped by the amount of shift relative to 1wcwA, and those with a shift value difference <5 were unified. Almost all loops in the domain fold are potential sites for circular permutant generation. The topology cartoon, drawn in TOPS [60] style, represents the structure of Domain 2 of 1wcwA (triangular symbols indicate β-strands and circular ones helices). Scissors icons mark the approximate locations of the sites where the N- (and C-) termini of several identified circular permutants (indicated with their PDB codes) align by structural alignment.

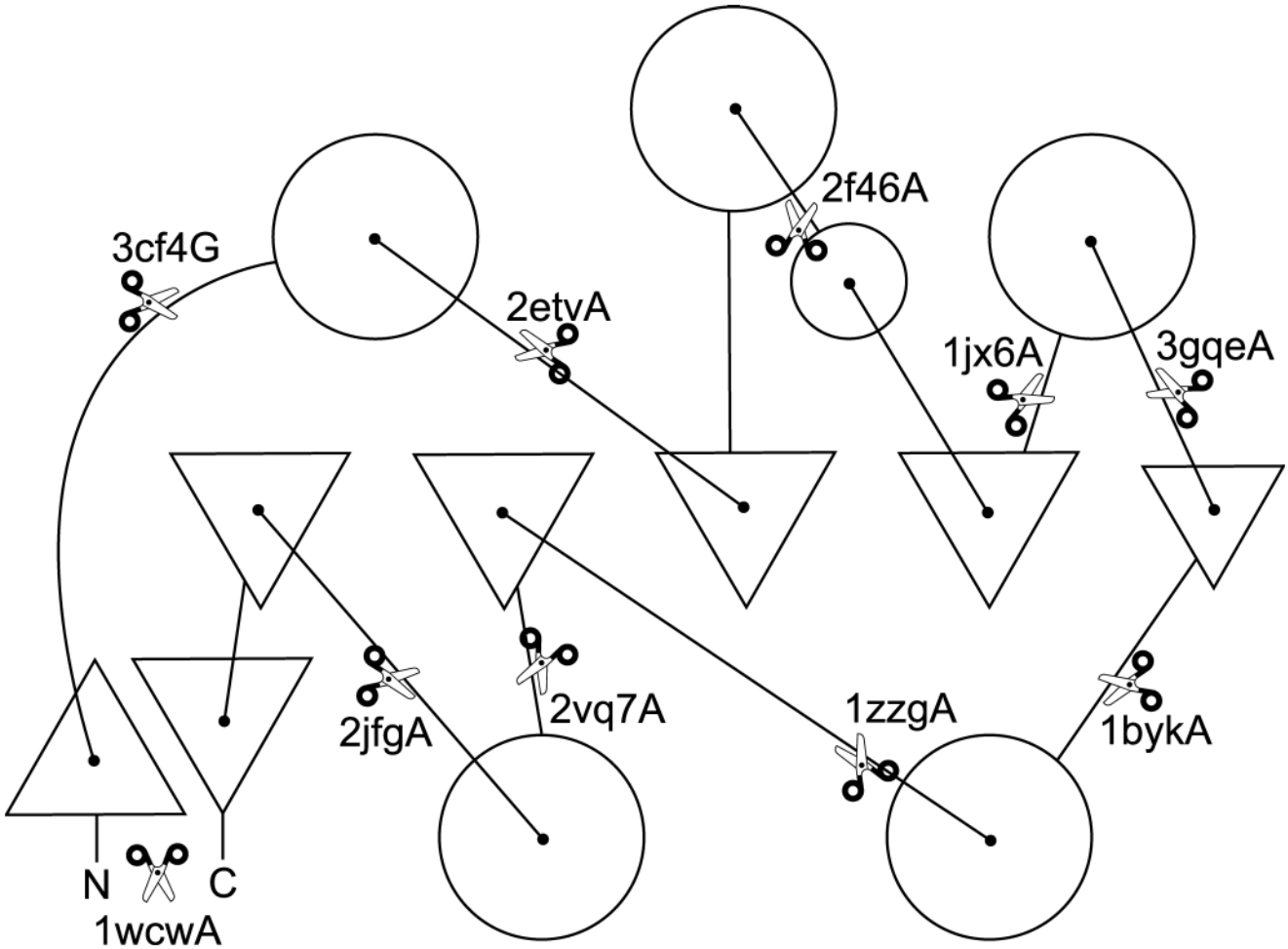

Figure S3. Schematic representation of SSP variants generated from the same base fold by different mechanisms. (a) Variants generated by the DSF mechanism (depending on where the domains open up to get rearranged) have different continuous domains while the continuous domains are similar to each other apart from the site where the continuous domain is inserted into them. (b) Variants generated by the CP mechanism (depending on where the circularized chain is cut to produce the two termini) have different discontinuous domains while the continuous domains are similar to each other.