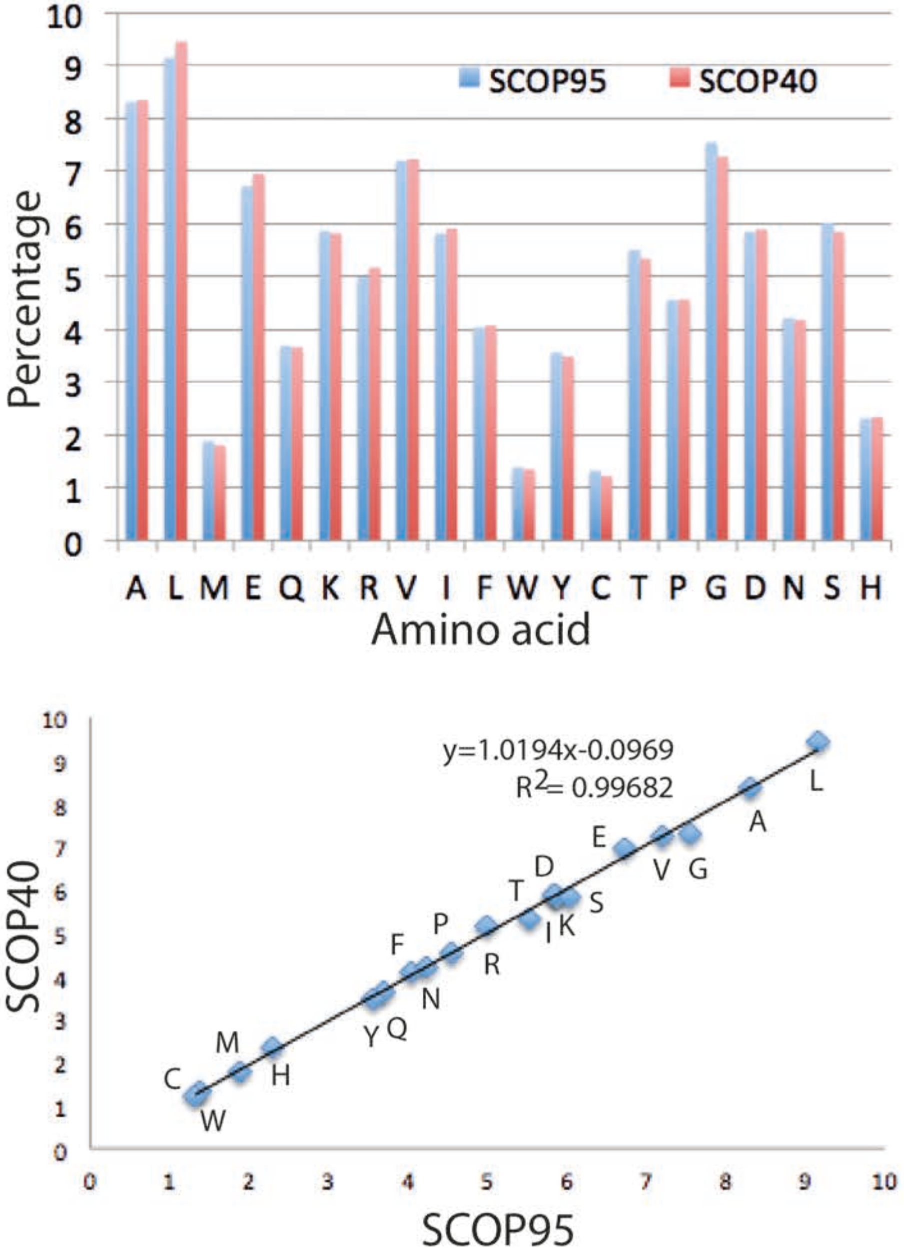
THE RESPECTIVE ROLES OF POLAR/NON POLAR BINARY PATTERNS AND AMINO ACID COMPOSITION IN PROTEIN REGULAR SECONDARY STRUCTURES EXPLORED EXHAUSIVELY USING HYDROPHOBIC CLUSTER ANALYSIS

JOSEPH REBEHMED, FLAVIEN QUINTUS, JEAN-PAUL MORNON, ISABELLE CALLEBAUT*

SUPPLEMENTARY DATA

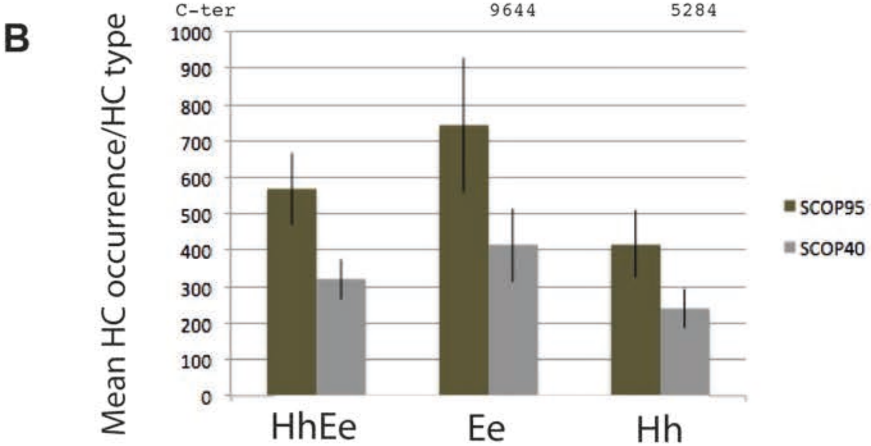
Amino acid	Occurrence		Percentage	
	SCOP95	SCOP40	SCOP95	SCOP40
A	238680	135229	8,31	8,36
L	262990	153007	9,15	9,46
M	54278	29119	1,89	1,80
E	193196	112554	6,72	6,96
Q	105953	59210	3,69	3,66
K	168634	94287	5,87	5,83
R	143503	83766	4,99	5,18
V	207080	117065	7,21	7,23
I	167333	95717	5,82	5,92
F	116118	65981	4,04	4,08
W	39869	21829	1,39	1,35
Y	102418	56469	3,56	3,49
C	38072	19799	1,33	1,22
T	158260	86525	5,51	5,35
P	130867	73897	4,55	4,57
G	216913	117813	7,55	7,28
D	168220	95432	5,86	5,90
N	121191	67749	4,22	4,19
S	173149	94741	6,03	5,86
H	66362	37912	2,31	2,34
Total	2873086	1618101	100,00	100,00

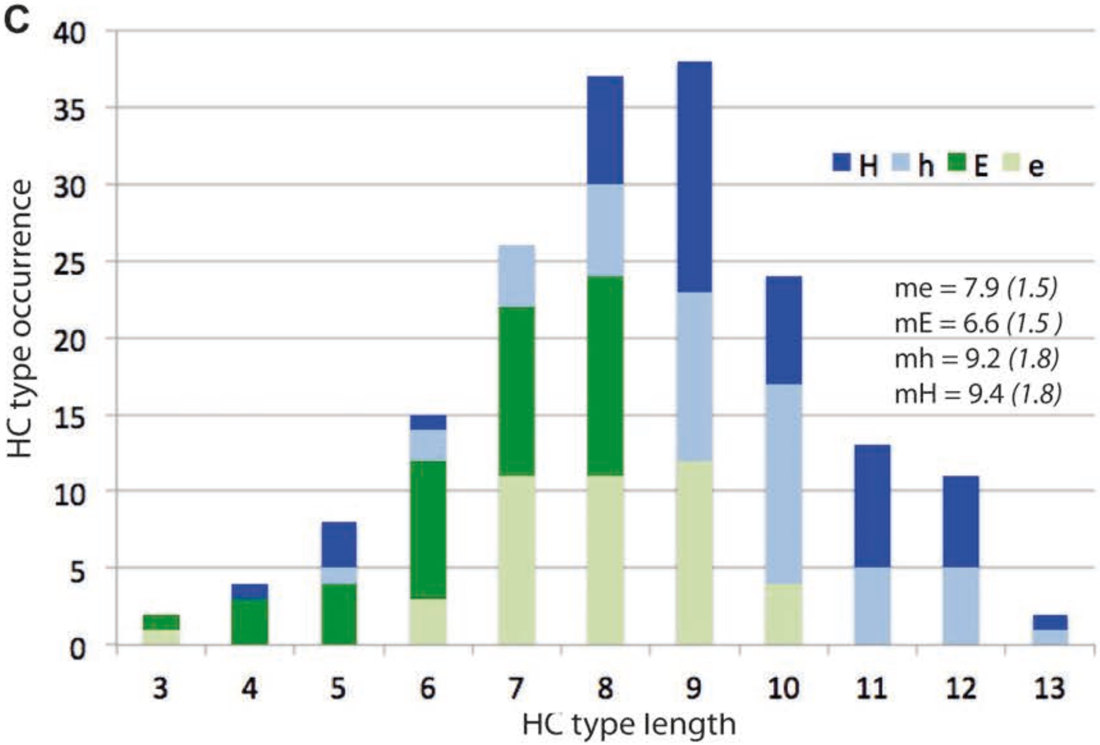


Supplementary Data 1: *Occurrences and percentages of the 20 amino acids in the SCOP95- and SCOP40-derived datasets and the correlation between the two datasets.*

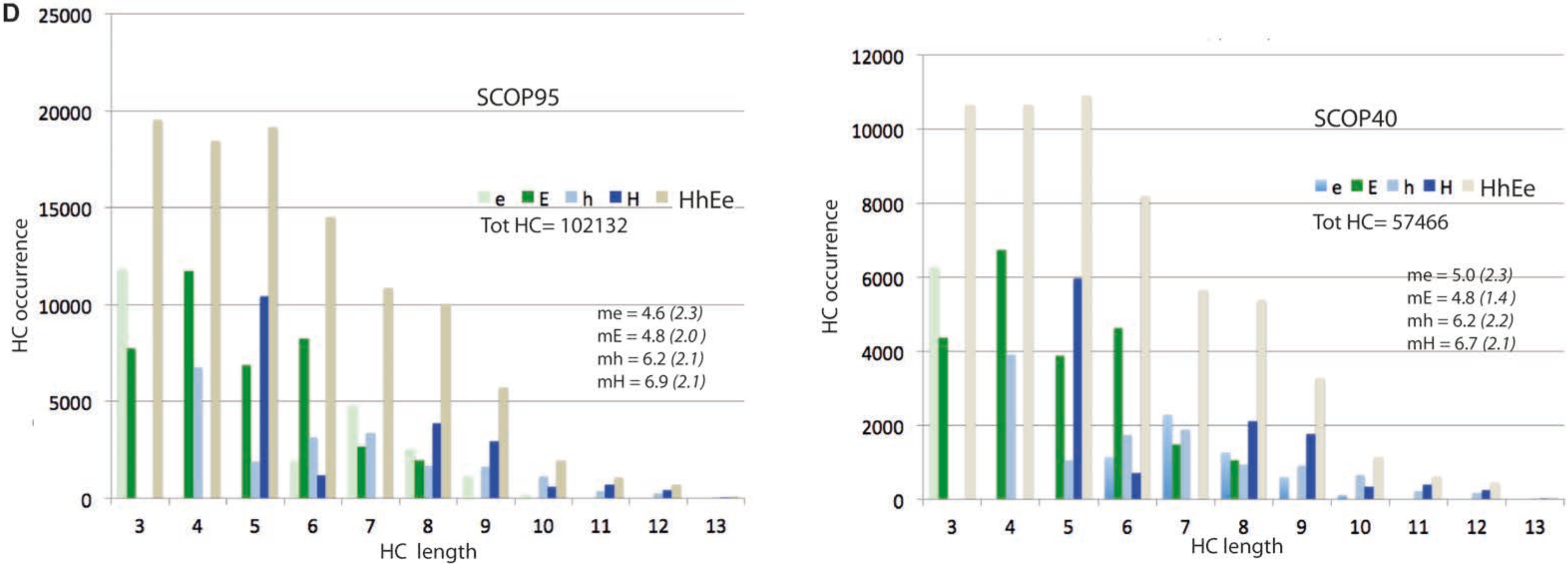
A

HC classes	SCOP95	SCOP40
Helix affinity (97)	40434	23076
H (48)	20213	11544
h(49)	20221	11532
Strand affinity (83)	61698	34390
E (41)	39158	22205
e (42)	22540	12185
Multiple affinity (112)	11828	6566
Not in dictionary:		
HC	41752	23844
HC types	18962	13584
Not relevant		
with gaps	3928	2338
N-ter	8919	4933
C-ter	9644	5284





Supplementary Data 2 (A-B-C) : see below



Supplementary Data 2: *HC databases, built from the SCOP95 and SCOP40-derived datasets. A) Global statistics of the HC occurrences. B) Mean HC occurrence per HC type for HC with propensities for α-helices and β-strands. C) Occurrence of HC types as a function of HC type length. D) Occurrence of HC as a function of HC length. Mean lengths (standard deviations) are indicated for the SCOP95 and SCOP40 databases.*

A

P-153 10011001 (Expected secondary structure Helix)

a) Concordant

```
     (≥ 80% H)                      (≥ 60% H)

1zkj MRAYYRQW                   1x1n IRGWWDTL
     HHHHHHHC    H 7/8              HHHHHHCC    H 6/8

2z01 IAAIVKGV                   1r9f LHRVLGSW
     HHHHHHHH    H 8/8              HHHHHCCC    H 5/8
```

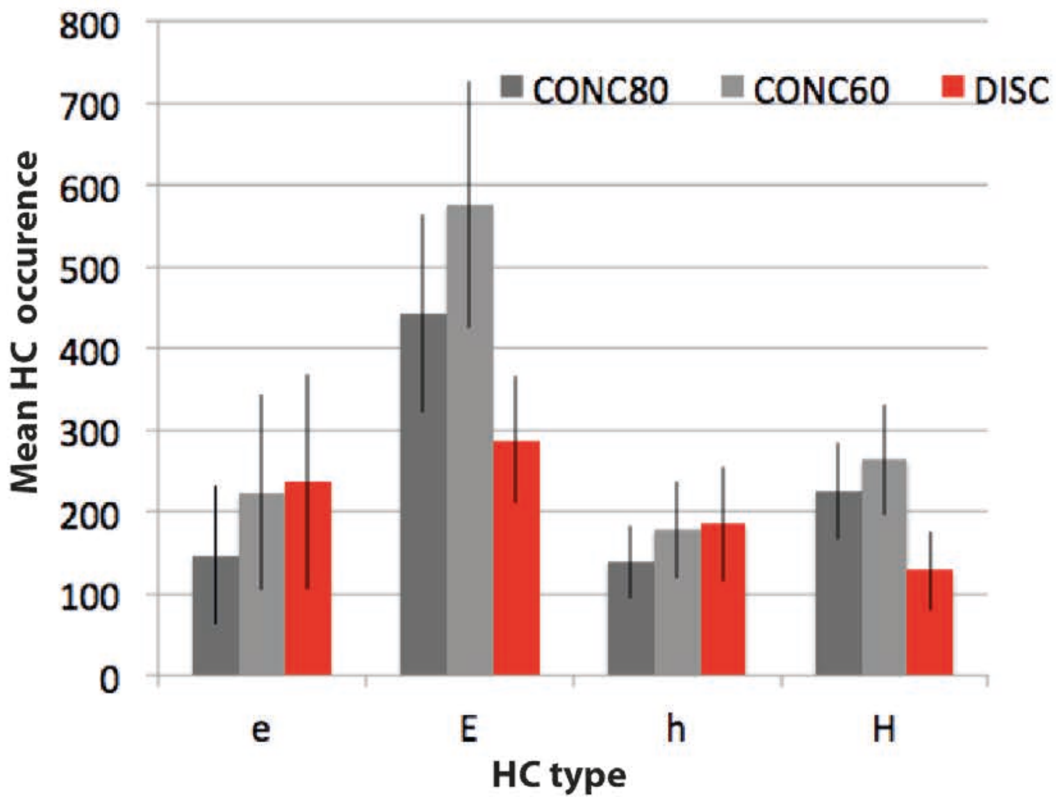
---

b) Discordant

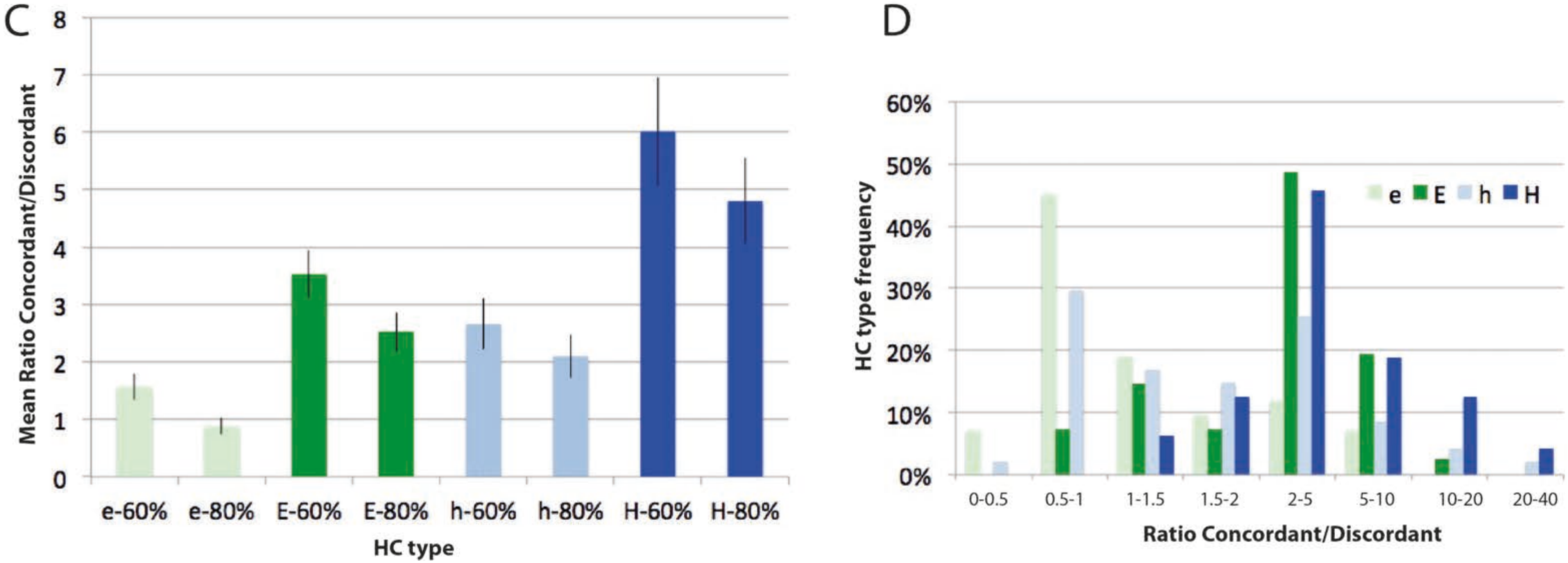
```
     (≤ 20% H)                      (≥ 80% E)

1sh5 ISDLYEDL                   2xik FGEVYKGI
     CCCCCCCC    H 0/8              CEEEEEEE    E 7/8

2epj LEAVFTGL                   2s6n VQGILRGF
     CCCEECCC    H 0/8              EEEEEEEE    E 8/8
```

B

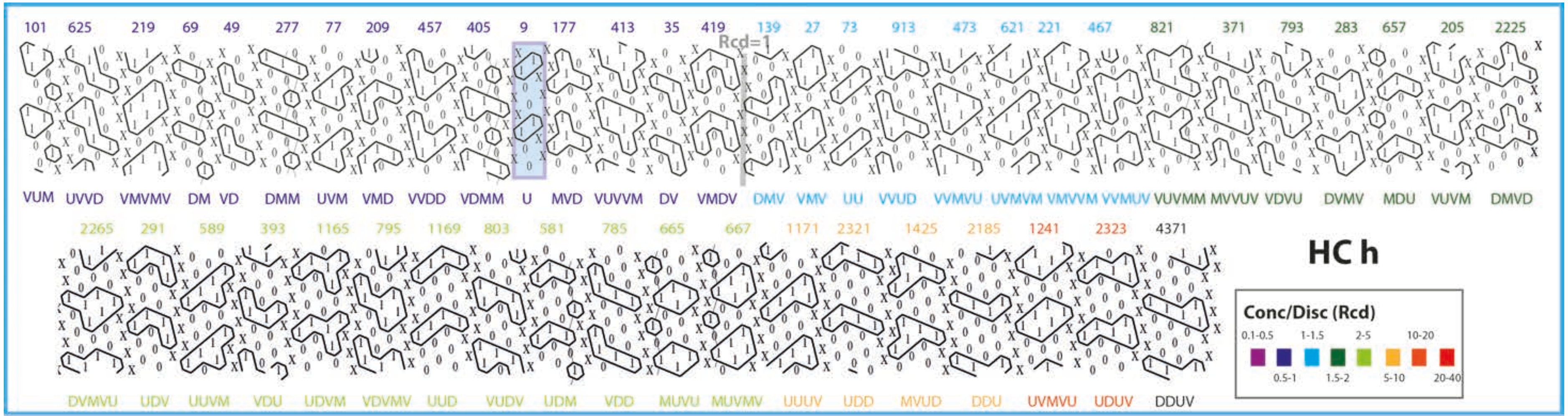
	SCOP95	SCOP40
Helix affinity (H,h)	Total = 40434 Conc = 17586 (97) 21353 (97) Disc = 14963 (96) 14963 (96) Int = 7885 (97) 4118 (95)	Total = 23076 Conc = 10231 (97) 12455 (97) Disc = 8287 (95) 8287 (95) Int = 4558 (97) 2334 (95)
Strand affinity (E,e)	Total = 61698 Conc = 24306 (83) 32929 (83) Disc = 21722 (83) 21722 (83) Int = 15670 (83) 7047 (83)	Total = 34390 Conc = 13569 (83) 18265 (83) Disc = 12344 (83) 12344 (83) Intermediate = 8477 (83) 3781 (83)

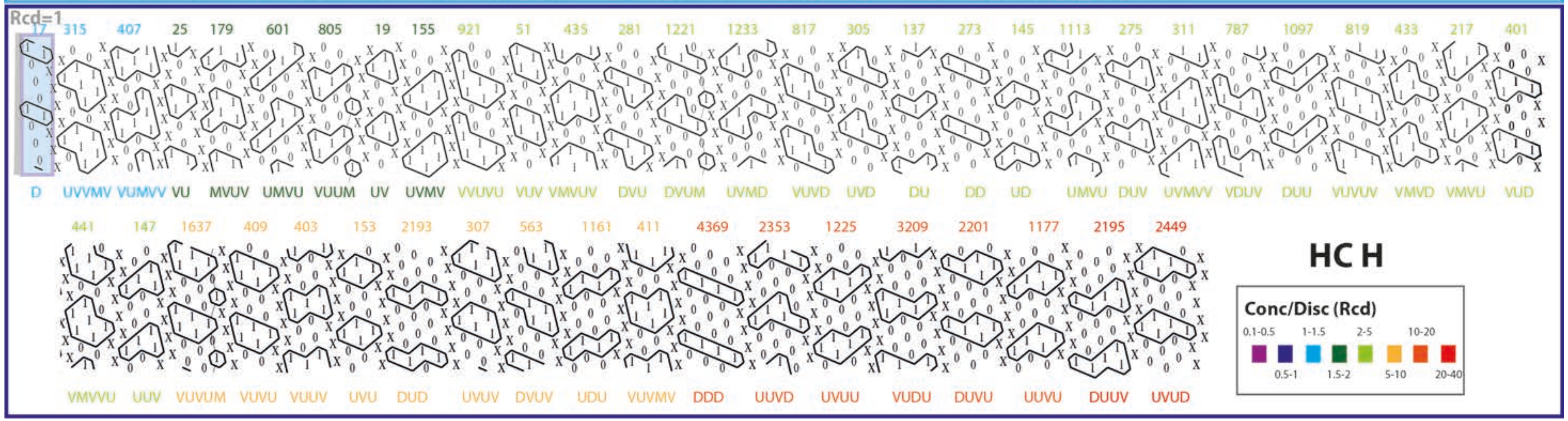


Supplementary Data 3 (A-B) : see below

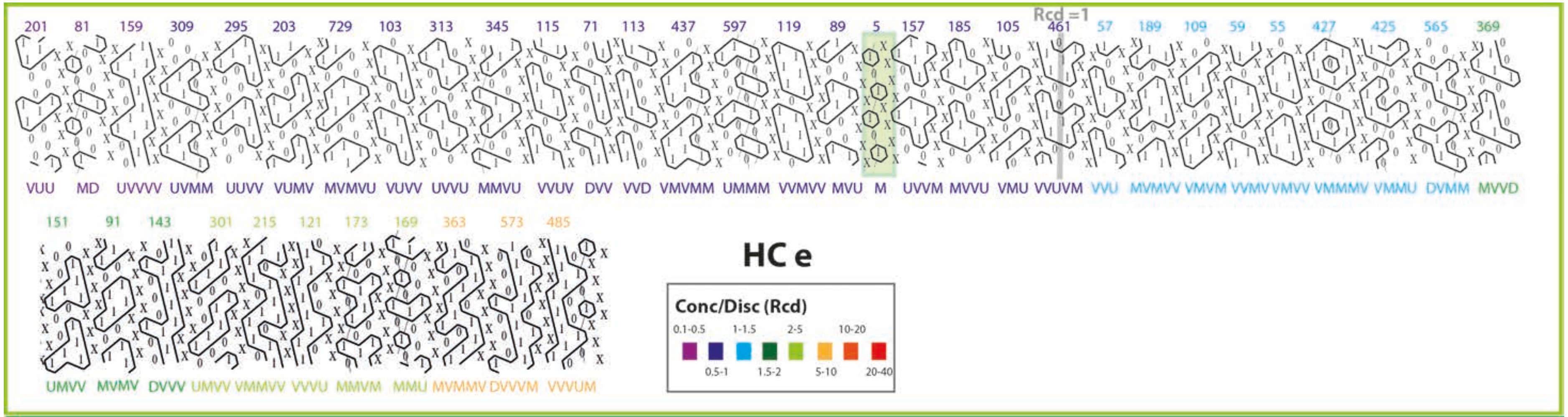


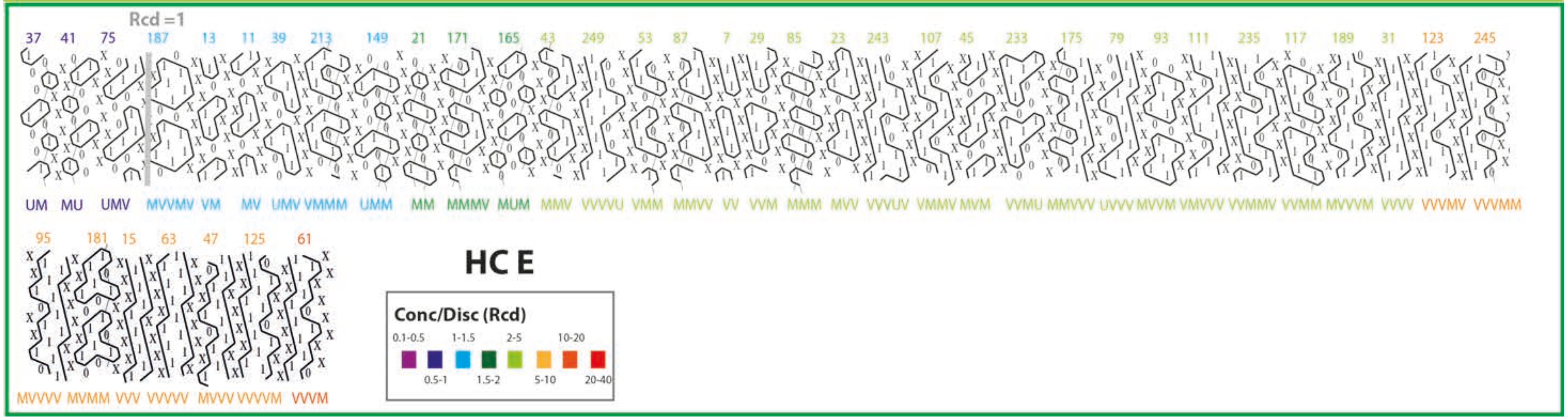
Supplementary Data 3: *Concordant and Discordant HC. A) Definition of concordant/discordant states relative to the behavior expected from the dictionary. The occurrences of concordant, discordant and intermediate HC are counted according to the following rules: for HC with helix and strand affinities, HC is described as Concordant if more than 80% or 60 % of the HC respects the predicted RSS in the dictionary; Discordant if the HC presents less than 20% of the predicted RSS in the dictionary and Intermediate for the other cases. B) Occurrences of concordant, discordant and intermediate HC in the SCOP95 and SCOP40 datasets. The two numbers designate the 80- and 60-levels, whereas values within brackets refer to the number of concerned HC types. Mean HC occurrences per HC type are shown in the bottom part of panel B. C) Mean of the ratios calculated per HC type between concordant and discordant clusters (mean Rcd). D) Distribution (frequencies) of the ratios, calculated per HC type, between concordant and discordant clusters (Rcd).*





Supplementary Data 4 (see below)


HC e


HC E

Supplementary Data 4 (see below)

Supplementary Data 4: *HCA plots of the 180 HC types analyzed in this study, classified by SSR affinities (h,H,e,E) and as a function of the increasing ratio of the number of concordant and discordant HC (Rcd) in the SCOP95 database (60 % level for definition of the concordance/discordance states). HC are represented by both their P-codes (top line) and Q-codes (bottom line). The colors indicate the range of Rcd values. Bar indicates the Rcd values of 1, before which there are more discordant than concordant HC per HC type. Note that the HC type with P-code 4371 (Q-code DDUV) has no discordant representative in our database. HC types corresponding to the Q-codes M, U and D are shaded in colors. The Q-code V is not present in this list as it is mainly associated with coils.*

Cluster			SCOP 95					SCOP 40				
P-code	**Q-code**	**Binary**	**Conc-80**	**Conc-60**	**Disc**	**Int-80**	**Int-60**	**Conc-80**	**Conc-60**	**Disc**	**Int-80**	**Int-60**
P5-e	M	101	3573	5102	5600	2650	1121	1843	2616	3100	1335	562
P7-E	VV	111	4249	5216	2097	1390	423	2388	2932	1216	770	226
P9-h	U	1001	2003	2711	3162	1578	870	1160	1566	1861	901	495
P11-E	MV	1011	1469	2189	1730	1374	654	885	1317	983	761	329
P13-E	VM	1101	1522	2127	1716	1097	492	895	1256	953	655	294
P15-E	VVV	1111	1930	2325	325	559	164	1094	1328	192	319	85
P17-H	D	10001	1630	1950	1793	524	204	949	1123	947	294	120
P19-H	UV	10011	1859	2022	1123	267	104	1073	1177	649	162	58
P21-E	MM	10101	1368	1643	929	441	166	749	904	544	251	96
P23-E	MVV	10111	1026	1245	496	280	61	590	699	280	140	31
P25-H	VU	11001	1605	1848	1206	431	188	956	1094	674	257	119
P27-h	VMV	11011	924	974	848	107	57	520	551	469	64	33
P29-E	VVM	11101	1022	1143	459	182	61	632	706	252	108	34
P31-E	VVVV	11111	509	531	109	27	5	265	277	65	14	2
P35-h	DV	100011	509	642	670	287	154	304	388	371	170	86
P37-E	UM	100101	370	515	712	391	246	220	315	378	226	131
P39-E	UVV	100111	299	525	398	373	147	179	298	226	218	99
P41-E	MU	101001	424	591	753	459	292	236	324	424	208	120
P43-E	MMV	101011	562	682	338	198	78	298	367	176	114	45
P45-E	MVM	101101	468	599	219	217	86	295	378	129	130	47
P47-E	MVVV	101111	273	381	47	139	31	160	227	29	83	16
P49-h	VD	110001	480	591	863	321	210	294	342	448	154	106
P51-H	VUV	110011	725	779	332	122	68	441	479	182	75	37
P53-E	VMM	110101	481	594	265	212	99	284	348	129	115	51
P55-e	VMVV	110111	219	243	185	44	20	100	118	116	29	11
P57-e	VVU	111001	317	449	416	266	134	191	272	235	170	89
P59-e	VVMV	111011	230	279	219	79	30	133	156	132	33	10
P61-E	VVVM	111101	368	443	35	93	18	213	250	21	51	14
P63-E	VVVVV	111111	115	124	16	12	3	59	65	12	9	3
P69-h	DM	1000101	199	313	506	274	160	107	168	282	146	85
P71-e	DVV	1000111	82	123	158	203	162	41	72	93	122	91
P73-h	UU	1001001	330	409	331	316	237	186	233	189	170	123

P75-E	UMV	1001011	101	159	170	180	122	65	100	108	90	55
P77-h	UVM	1001101	206	260	343	141	87	125	159	182	86	52
P79-E	UVVV	1001111	81	142	48	116	55	42	75	28	62	29
P81-e	MD	1010001	138	180	445	236	194	76	109	270	162	129
P85-E	MMM	1010101	224	282	113	174	116	127	162	64	106	71
P87-E	MMVV	1010111	110	128	55	61	43	60	75	32	42	27
P89-e	MVU	1011001	200	286	319	197	111	86	132	213	110	64
P91-e	MVMV	1011011	114	157	96	86	43	37	56	63	38	19
P93-E	MVVM	1011101	154	195	65	59	18	62	89	37	38	11
P95-E	MVVVV	1011111	49	58	9	16	7	35	43	4	11	3
P101-h	VUM	1100101	119	179	362	233	173	72	111	184	148	109
P103-e	VUVV	1100111	72	82	124	75	65	46	52	76	44	38
P105-e	VMU	1101001	131	202	215	304	233	83	120	132	117	80
P107-E	VMMV	1101011	127	189	72	121	59	68	94	39	60	34
P109-e	VMVM	1101101	123	160	126	70	33	53	78	80	43	18
P111-E	VMVVV	1101111	37	45	14	15	7	24	30	11	11	5
P113-e	VVD	1110001	110	174	217	249	185	56	92	120	141	105
P115-e	VVUV	1110011	82	113	150	95	64	48	65	84	57	40
P117-E	VVMM	1110101	162	196	47	83	49	88	107	23	42	23
P119-e	VVMVV	1110111	33	37	42	8	4	20	24	29	8	4
P121-e	VVVU	1111001	92	155	52	177	114	53	95	39	104	62
P123-E	VVVMV	1111011	54	72	13	33	15	27	38	7	19	8
P125-E	VVVVM	1111101	74	96	11	30	8	31	44	7	20	7
P137-H	DU	10001001	367	482	164	198	83	217	281	93	111	47
P139-h	DMV	10001011	178	212	188	93	59	106	127	91	54	33
P143-e	DVVV	10001111	30	73	38	85	42	15	44	15	54	25
P145-H	UD	10010001	417	508	159	191	100	225	282	96	110	53
P147-H	UUV	10010011	359	419	86	118	58	201	239	51	67	29
P149-E	UMM	10010101	86	178	120	161	69	48	99	73	89	38
P151-e	UMVV	10010111	48	90	58	79	37	25	49	29	45	21
P153-H	UVU	10011001	473	573	87	137	37	262	321	49	85	26
P155-H	UVMV	10011011	245	270	137	38	13	134	151	34	24	7
P157-e	UVVM	10011101	38	94	103	80	24	18	52	45	48	14
P159-e	UVVVV	10011111	11	26	59	19	4	7	17	19	11	1
P165-E	MUM	10100101	78	162	84	153	69	45	93	56	90	42
P169-e	MMU	10101001	66	423	121	421	64	41	100	66	94	35

P171-E	MMMV	10101011	46	89	48	70	27	24	55	28	44	13
P173-e	MMVM	10101101	56	90	27	65	31	36	58	12	47	25
P175-E	MMVVV	10101111	17	35	12	21	3	10	19	9	11	2
P177-h	MVD	10110001	157	181	209	54	30	80	94	114	29	15
P179-H	MVUV	10110011	174	192	121	32	14	92	101	52	17	8
P181-E	MVMM	10110101	57	112	16	94	39	31	66	10	52	17
P183-e	MVMVV	10110111	32	53	45	33	12	10	15	28	12	7
P185-e	MVVU	10111001	58	106	114	78	30	26	56	66	53	23
P187-E	MVVMV	10111011	25	47	39	31	9	11	23	23	13	1
P189-E	MVVVM	10111101	31	62	13	37	6	13	32	11	24	5
P201-e	VUU	11001001	20	58	437	93	55	9	28	244	54	35
P203-e	VUMV	11001011	34	54	90	65	45	17	28	50	34	23
P205-h	VUVM	11001101	71	119	69	71	23	51	78	49	40	13
P209-h	VMD	11010001	114	161	199	103	56	72	101	106	60	31
P213-E	VMMM	11010101	43	70	50	47	20	32	50	32	33	15
P215-e	VMMVV	11010111	23	32	12	20	11	17	24	7	14	7
P217-H	VMVU	11011001	207	258	71	64	13	114	147	38	40	7
P219-h	VMVMV	11011011	34	42	69	15	7	20	26	29	11	5
P221-h	VMVVM	11011101	32	37	27	11	6	13	17	15	5	1
P233-E	VVMU	11101001	17	157	55	181	41	8	38	35	57	27
P235-E	VVMMV	11101011	32	41	10	15	6	18	23	7	9	4
P243-E	VVVUV	11110011	33	58	23	43	18	21	30	13	21	12
P245-E	VVVMM	11110101	26	59	10	40	7	11	38	7	32	5
P249-E	VVVVU	11111001	13	47	21	39	5	8	26	12	21	3
P273-H	DD	100010001	181	244	77	134	71	98	140	35	79	37
P275-H	DUV	100010011	198	222	68	69	45	127	146	35	48	29
P277-h	DMM	100010101	78	84	116	47	41	45	49	64	29	25
P281-H	DVU	100011001	96	146	60	80	30	68	100	33	50	18
P283-h	DVMV	100011011	83	99	59	44	28	52	64	30	22	10
P291-h	UDV	100100011	106	135	62	81	52	59	75	33	37	21
P295-e	UUVV	100100111	31	49	83	45	27	11	21	45	26	16
P301-e	UMVM	100101101	25	55	22	60	30	14	38	17	44	20
P305-H	UVD	100110001	162	207	72	90	45	104	132	33	54	26
P307-H	UVUV	100110011	165	185	27	30	10	105	116	13	18	7
P309-e	UVMM	100110101	10	40	72	78	48	7	21	40	25	11
P311-H	UVMVV	100110111	33	33	10	3	3	19	19	8	3	3

P313-e	UVVU	100111001	16	47	71	50	19	5	23	38	31	13
P315-H	UVVMV	100111011	29	34	27	9	4	15	18	13	4	1
P345-e	MMVU	101011001	17	37	52	51	31	12	21	40	26	17
P363-e	MVMMV	101101011	25	38	7	24	11	13	19	3	13	7
P369-e	MVVD	101110001	14	46	30	58	26	6	21	18	31	16
P371-h	MVVUV	101110011	31	36	22	9	4	20	22	14	3	1
P393-h	VDU	110001001	111	139	62	83	55	69	82	35	45	32
P401-H	VUD	110010001	227	253	62	67	41	141	159	40	47	29
P403-H	VUUV	110010011	154	168	26	47	33	83	93	10	26	16
P405-h	VUMM	110010101	36	43	51	39	32	21	26	33	22	17
P407-H	VUMVV	110010111	24	25	19	19	18	13	14	12	14	13
P409-H	VUVU	110011001	150	184	29	49	15	95	118	16	35	12
P411-H	VUVMV	110011011	74	82	9	13	5	41	45	4	7	3
P413-h	VUVVM	110011101	28	31	33	7	4	12	13	16	4	3
P419-h	VMDV	110100011	47	53	55	31	25	23	28	39	19	14
P425-e	VMMU	110101001	12	39	28	53	26	10	27	17	28	11
P427-e	VMMMV	110101011	8	23	17	32	17	6	14	11	15	7
P433-H	VMVD	110110001	101	119	34	41	23	53	64	16	25	14
P435-H	VMVUV	110110011	83	90	37	17	10	49	56	17	12	5
P437-e	VMVMM	110110101	7	14	17	31	24	5	7	6	8	6
P441-H	VMVVU	110111001	44	49	12	7	2	23	27	7	6	2
P457-h	VVUU	111001001	33	50	61	41	24	23	37	35	29	15
P461-e	VVUVM	111001101	13	18	18	26	21	10	12	13	11	9
P467-h	VVMUV	111010011	27	28	20	17	16	17	17	16	7	7
P473-h	VVMVU	111011001	42	53	40	12	1	20	23	20	4	1
P485-e	VVVUM	111100101	22	32	5	25	15	13	22	3	16	7
P563-H	DVUV	1000110011	80	97	12	23	6	45	55	8	13	3
P565-e	DVMM	1000110101	12	29	20	37	20	7	21	15	27	13
P573-e	DVVVM	1000111101	16	38	6	35	13	4	15	3	16	5
P581-h	UDM	1001000101	59	99	31	61	21	38	64	16	38	12
P589-h	UUVM	1001001101	39	55	25	32	16	20	33	18	22	9
P597-e	UMMM	1001010101	8	16	19	14	6	6	10	11	8	4
P601-H	UMVU	1001011001	36	45	28	19	10	15	18	19	10	7
P621-h	UVMVM	1001101101	22	27	20	12	7	12	14	12	5	3
P625-h	UVVD	1001110001	8	23	39	31	16	3	13	24	19	9
P657-h	MUD	1010010001	66	80	47	26	12	33	43	24	20	10

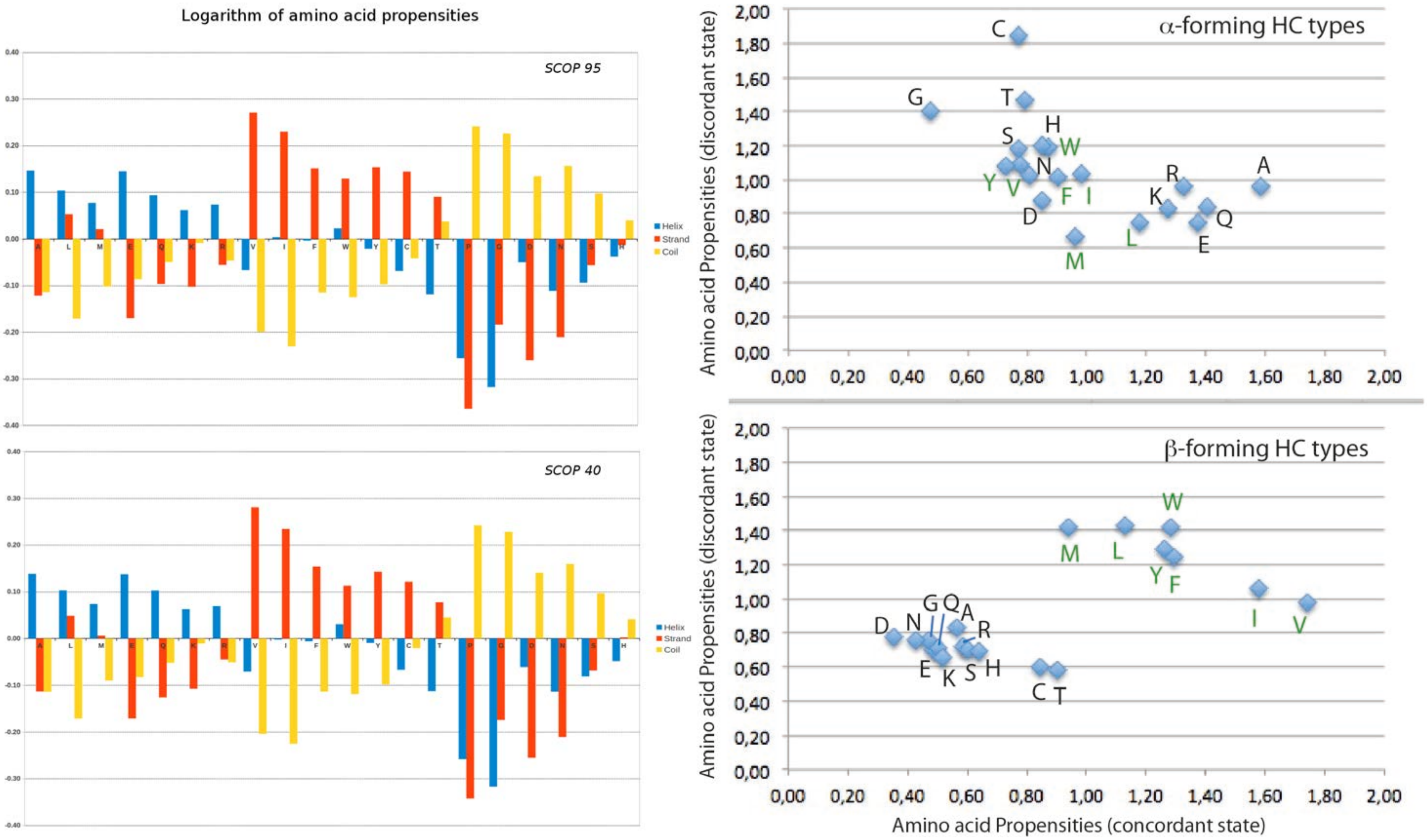
P665-h	MUVU	1010011001	64	80	20	23	7	36	47	13	14	3
P667-h	MUVMV	1010011011	20	28	6	12	4	12	17	3	7	2
P729-e	MVMVU	1011011001	6	15	25	10	1	2	5	16	4	1
P785-h	VDD	1100010001	55	76	22	49	28	36	44	12	22	14
P787-H	VDUV	1100010011	57	64	19	22	15	36	42	11	15	9
P793-h	VDVU	1100011001	29	52	31	31	8	22	33	20	13	2
P795-h	VDVMV	1100011011	28	34	12	13	7	15	18	8	6	3
P803-h	VUDV	1100100011	39	46	15	18	11	22	27	8	11	6
P805-H	VUUM	1100100101	37	67	39	43	13	20	36	22	27	11
P817-H	VUVD	1100110001	44	71	25	33	6	22	40	12	21	3
P819-H	VUVUV	1100110011	34	38	11	4		15	18	7	3	
P821-h	VUVMM	1100110101	16	25	16	10	1	11	18	7	8	1
P913-h	VVUD	1110010001	33	43	34	19	9	19	28	21	16	7
P921-H	VVUVU	1110011001	25	27	13	5	3	20	21	7	3	2
P1097-H	DUU	10001001001	44	64	19	40	20	29	41	10	24	12
P1113-H	DMVU	10001011001	39	45	14	22	16	21	22	8	6	5
P1161-H	UDU	10010001001	89	105	12	32	16	47	56	7	22	13
P1165-h	UDVM	10010001101	21	30	11	16	7	14	18	10	8	4
P1169-h	UUD	10010010001	56	76	26	39	19	35	50	13	29	14
P1171-h	UUUV	10010010011	26	46	9	30	10	14	25	6	18	7
P1177-H	UUVU	10010011001	56	83	5	32	5	29	41	3	15	3
P1221-H	UVDM	10011000101	35	49	18	26	12	18	27	8	17	8
P1225-H	UVUU	10011001001	65	86	7	33	12	39	47	4	14	6
P1233-H	UVMD	10011010001	40	47	17	25	18	20	23	11	10	7
P1241-h	UVMVU	10011011001	37	44	3	7		25	29	3	4	
P1425-h	MVUD	10110010001	33	46	7	24	11	18	29	3	19	8
P1637-H	VUVUM	11001100101	22	29	5	10	3	14	18	2	6	2
P2185-h	DDU	100010001001	37	50	6	25	12	20	27	6	16	9
P2193-H	DUD	100010010001	52	66	10	24	10	33	42	6	15	6
P2195-H	DUUV	100010010011	36	45	2	16	7	25	31		12	6
P2201-H	DUVU	100010011001	45	58	4	18	5	29	35	1	9	3
P2225-h	DMVD	100010110001	26	28	14	8	6	15	17	10	7	5
P2265-h	DVMVU	100011011001	20	21	10	9	8	13	14	5	4	3
P2321-h	UDD	100100010001	37	47	9	31	21	18	26	7	21	13
P2323-h	UDUV	100100010011	34	37	2	10	7	18	21	2	8	5
P2353-H	UUVD	100100110001	30	44	4	28	14	25	33	3	15	7

P2449-H	UVUD	100110010001	63	77	2	27	13	43	52	2	12	3
P3209-H	VUDU	110010001001	41	56	4	21	6	20	26	1	10	4
P4369-H	DDD	1000100010001	21	32	3	19	8	12	17	2	9	4
P4371-h	DDUV	1000100010011	30	41		15	4	16	20		6	2

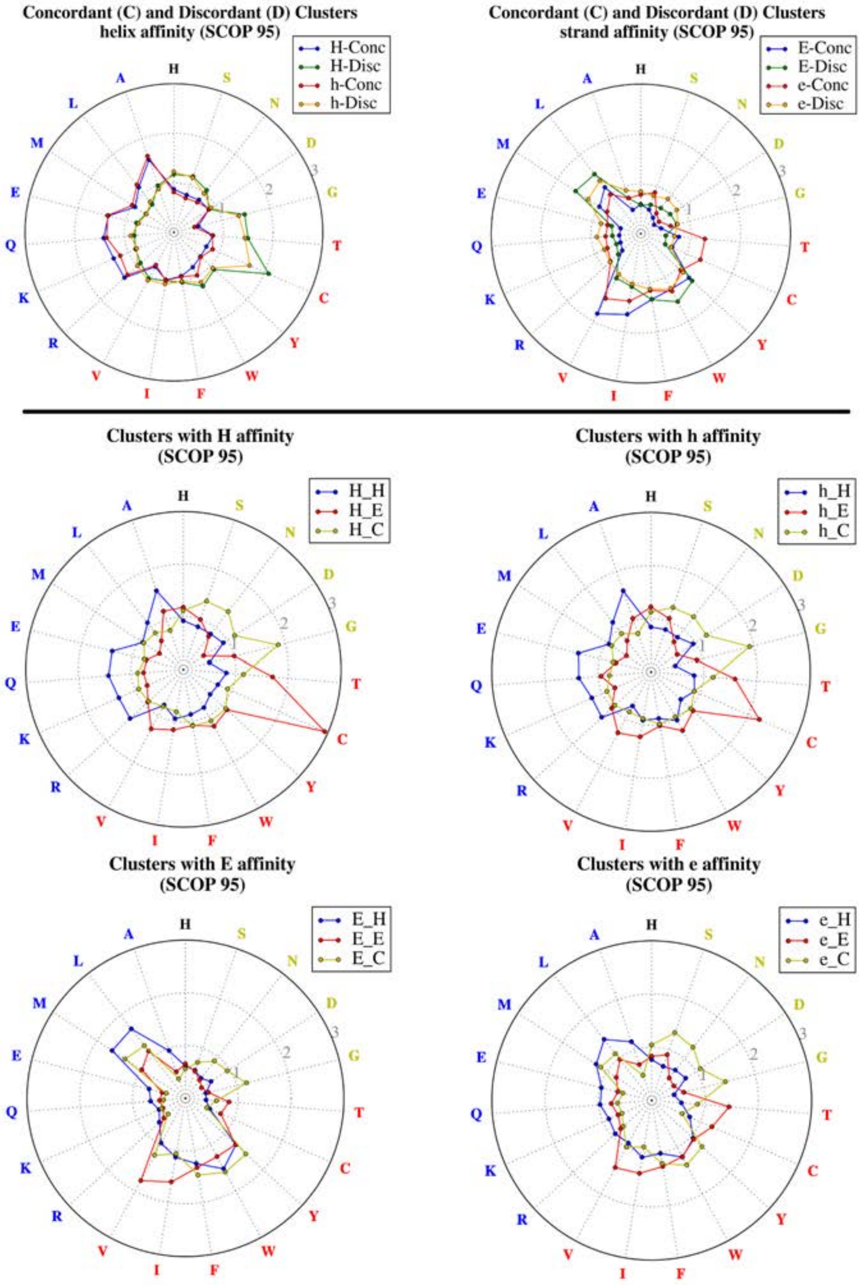
Supplementary Data 5: *Occurrences of HC with affinities for the helix (H,h) and strand (E,e) states, distinguishing between the concordant, discordant and intermediate states. Upper and lower cases indicate strong and weak affinities for RSS, respectively. These data were calculated for the SCOP95- and SCOP40-derived databases.*

Two levels are considered for the definition of the concordant state: 80-level (Conc-80) and 60-level (Conc-60) if more than 80% and 60% of the HC respects the RSS affinity predicted from the dictionary, respectively. The discordant state (Disc) is assigned if the HC presents less than 20% of the RSS affinity predicted from the dictionary, and the Intermediate state (Int-80 and Int-60) for the other cases.

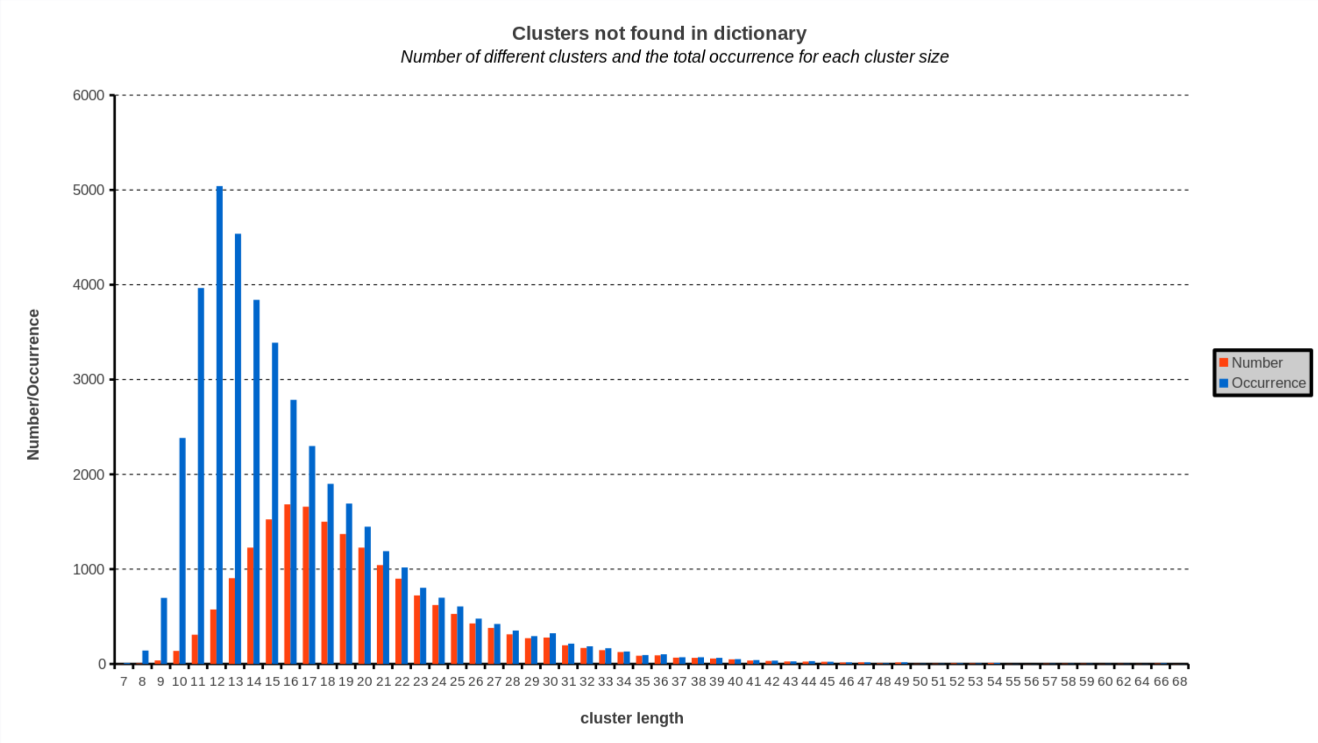
Supplementary Data 6: *WebLogo representations, for the 180 HC types considered here, of the observed secondary structure and amino acid composition (each of the 20 amino acids considered and amino acids grouped according to their RSS affinities). Data can be found http://www.impmc.upmc.fr/~callebau/SD6_WL, including a README file for detailed explanations.*



***Supplementary Data 7: A** (left) Logarithms of amino acid propensities for the three secondary structure states (Helix, Strand, Coil) calculated for the SCOP95- and SCOP40-derived databases (whole protein sequences), **B** (right) Couples of amino acid propensities for the concordant and discordant states for HC with affinities for α-helices (top) and β-strands (bottom).*



Supplementary Data 8*: Propensities of amino acids for the Concordant/Discordant states, distinguishing HC with strong (H, E) and weak (h,e) affinities for RSS. Top panel: propensities for HC with helix (left) and strand (right) propensities, regardless the secondary structure which is actually observed. Bottom panel: propensities analyzed according to the secondary structure which is actually observed.*



Supplementary Data 9: *Distribution, according to the HC length, of the number of HC types (colored in red; each HC types being defined by a unique binary pattern) that are not listed in the previously established dictionary (Eudes et al. 2007), due to a too low occurrence. The total number of HC (occurrence) for a specific length is reported in blue.*

Reference :

Eudes R, Le Tuan K, Delettre J, Mornon JP, Callebaut I. A generalized analysis of hydrophobic and loop clusters within globular protein sequences. BMC Struct Biol 2007;7:2.