In silico local structure approach: a case study on Outer Membrane Proteins

Juliette Martin, Alexandre G. de Brevern, Anne-Claude Camproux

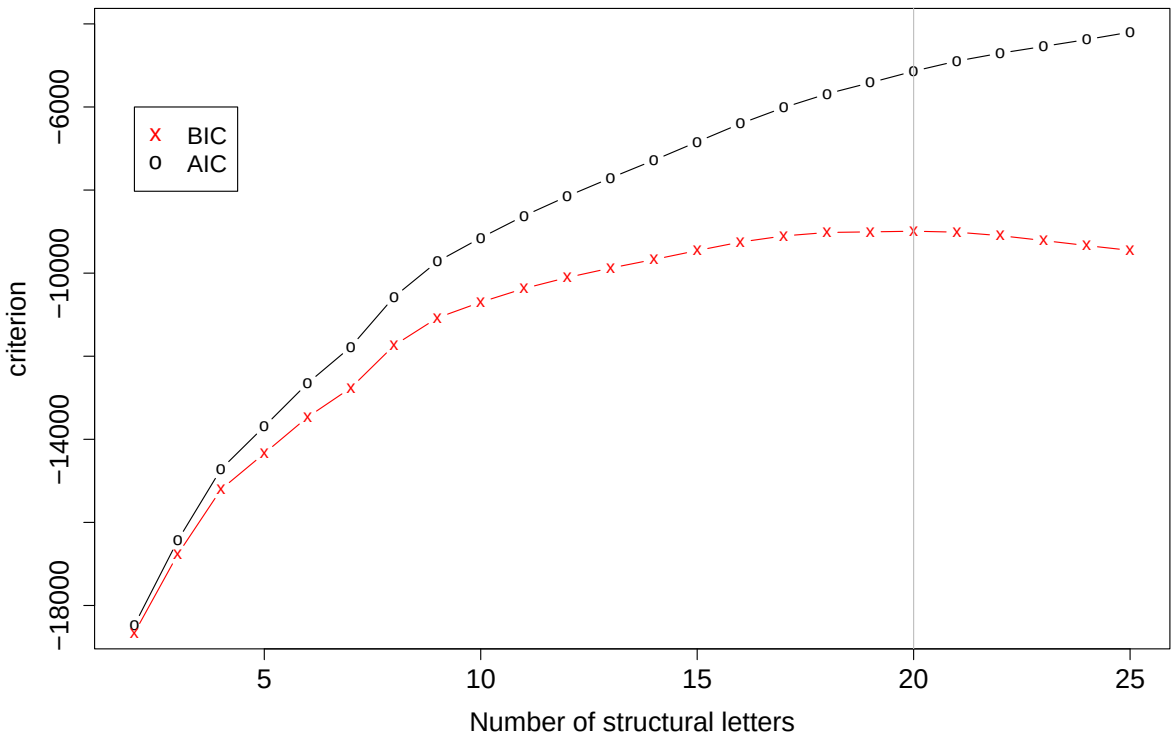

Figure 1: Evolution of BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion) with the number of structural letters in the OMP-specific alphabet. BIC is given by $BIC = L - k \times \log(N)$ and AIC is given by $AIC = L - k$ where L denotes the log-likelihood of the model, k denotes the number of independent parameters, and N denotes the amount of data. The maximum BIC is indicated by a gray line. The maximum AIC is not reached with 25 structural letters.

Table 1: Transition probabilities of SA20-OMP. Each row sum to 1.

	A	N	T	H	E	S	O	D	R	G	F	P	K	L	J	M	B	I	C	Q
A	35.5	6.6	10.2	11.6	4.7	12.0	17.0	0.0	0.0	0.0	0.0	1.4	0.0	1.2	0.0	0.0	0.0	0.0	0.0	0.0
N	4.1	3.5	2.5	17.1	6.5	25.0	31.2	0.0	0.0	0.0	0.0	8.2	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0
T	30.0	8.2	17.4	8.6	8.5	10.0	7.7	0.0	0.0	0.0	0.0	6.1	1.7	0.8	0.0	0.0	0.0	1.1	0.0	0.0
H	3.7	5.9	1.7	6.8	5.3	10.0	54.7	0.0	0.0	0.0	0.0	11.2	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0
E	0.0	0.0	0.0	0.0	0.0	0.0	0.0	19.9	22.1	18.8	8.9	0.7	4.7	1.0	2.1	5.9	0.0	0.8	7.8	7.2
S	0.0	0.4	0.0	0.0	0.0	0.0	0.0	7.6	7.1	6.5	14.9	1.6	6.0	11.7	7.3	3.7	3.5	8.0	3.6	18.0
O	5.5	8.8	5.6	0.7	4.4	5.1	3.7	4.4	2.4	0.0	11.9	5.8	4.5	13.8	7.2	0.0	12.2	4.0	0.0	0.0
D	47.1	11.2	11.9	1.3	8.2	12.8	3.9	0.0	0.0	0.0	0.0	3.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
R	4.3	27.3	2.2	1.7	0.0	2.7	1.5	1.9	1.8	1.1	6.4	2.5	4.2	7.3	4.9	8.3	6.6	8.3	2.3	4.7
G	0.0	9.9	1.4	10.4	3.1	0.5	2.8	2.8	22.5	13.9	0.0	13.2	0.0	1.8	0.0	0.0	2.9	9.7	0.8	4.2
F	0.0	9.0	0.0	0.0	0.0	1.3	0.0	12.4	5.0	0.8	14.5	8.9	11.1	19.1	3.5	0.0	7.8	6.5	0.0	0.0
P	10.8	5.6	3.8	4.0	27.3	4.4	8.2	0.0	5.5	0.0	3.5	5.2	2.5	6.2	2.8	0.0	6.2	4.0	0.0	0.0
K	9.5	2.7	6.6	2.5	20.3	10.3	5.8	0.0	0.0	0.0	0.0	35.7	0.0	4.3	0.0	0.0	0.0	2.3	0.0	0.0
L	0.0	4.8	0.0	0.0	0.0	0.0	0.0	8.5	8.3	6.1	10.4	8.2	11.5	15.3	11.7	3.6	7.0	3.8	0.0	0.7
J	0.0	1.7	0.0	0.0	0.0	0.0	0.0	0.4	0.3	0.0	0.0	4.0	5.9	9.7	42.7	0.0	28.8	6.5	0.0	0.0
M	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.3	0.4	1.0	2.4	4.2	35.0	0.0	44.4	11.5	0.0	0.0
B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.5	0.4	2.7	2.7	0.0	5.1	0.0	0.0	46.7	0.0	2.7	3.6	34.5
I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.3	12.1	2.8	9.9	0.0	2.7	0.0	0.0	3.1	0.0	6.0	48.8	12.1
C	0.0	1.4	0.0	0.0	0.0	0.0	0.0	2.0	0.2	1.2	1.4	0.0	5.1	9.5	19.1	11.8	23.9	6.1	3.2	14.9
Q	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.8	3.9	3.3	0.0	3.4	1.2	0.0	36.5	0.0	10.1	5.5	34.8

Table 2: Description of the 20 structural letters of SA20-GB. N and % are the frequencies and relative frequencies of each structural letter in the *GBset*. d_1, d_2, d_3 and d_4 denote the mean distance descriptors associated to each letter, in Å. $rmsd_w$ is the average rmsd within the fragments encoded by a given structural letter. $rmsd_b$ is the *minimum* average rmsd observed between the considered letter and all the others. rmsds are given in Å. %stride and %kaksi denote respectively the fraction of a given structural letter that correspond to a strand conformation assigned by STRIDE and KAKSI, in the *GBset*. Structural letters are sorted out by increasing d_2 value.

State	N	%	d_1	d_2	d_3	d_4	$rmsd_w$	$rmsd_b$	%stride	%kaksi	Neq	ANR
g	47	0.98	5.69	5.34	5.58	-2.27	0.40	0.87	0	0	1.98	1
q	141	2.94	5.95	5.54	5.80	1.30	0.39	0.60	3.5	1.4	6.53	1.03
a	300	6.27	5.45	5.46	5.50	3.13	0.25	0.60	1.7	0	5.05	1.74
d	167	3.49	5.74	7.38	7.09	0.34	0.39	0.80	60	50	7.56	1
m	226	4.72	6.26	7.40	5.91	-0.26	1.20	1.03	26	27	9.98	1.40
c	91	1.90	5.55	7.73	5.60	-3.32	0.35	0.63	1.1	4.4	1.93	1.01
n	208	4.34	5.76	8.07	6.83	2.70	0.51	0.78	50	36	10.94	1.02
o	134	2.80	6.77	8.19	5.46	-3.57	0.33	0.52	2.2	7.5	3.03	1
j	234	4.89	5.61	8.67	6.42	2.41	0.43	0.78	34	43	10.71	1
r	139	2.90	6.99	8.64	6.48	-2.14	0.54	0.61	41	42	13.60	1.10
i	214	4.47	6.77	8.91	6.34	-3.42	0.34	0.49	38	49	10.15	1.10
s	168	3.51	6.56	9.02	5.58	-1.87	0.40	0.59	18	29	7.52	1.03
h	194	4.05	6.27	9.32	5.88	0.47	0.37	0.64	51	47	7.10	1.06
p	305	6.37	6.51	9.37	6.43	-2.76	0.30	0.46	57	74	9.91	1.07
k	484	10.11	6.42	9.89	6.69	-1.20	0.29	0.43	81	92.1	9.02	1.12
b	319	6.66	7.04	9.84	6.52	-2.72	0.31	0.46	82	89	8.83	1
f	508	10.61	6.86	10.01	6.46	-1.37	0.31	0.44	94.5	95.5	7.36	1.09
t	156	3.26	6.60	10.21	7.15	1.17	0.33	0.50	82	85	6.50	1.02
e	321	6.70	6.40	10.11	7.01	-0.01	0.25	0.42	89	91.6	6.01	1
l	432	9.02	6.81	10.35	6.84	-0.24	0.27	0.42	94	94	7.20	1.39
total									57.5	60.8		

Figure 2: Z-scores of the 20 amino-acids in the four positions of SA20-OMP letters. Pink and blue colors denote respectively over and under-represented amino-acids. Yellow indicates that the expected frequency is two low (less than five) to compute a Z-score.

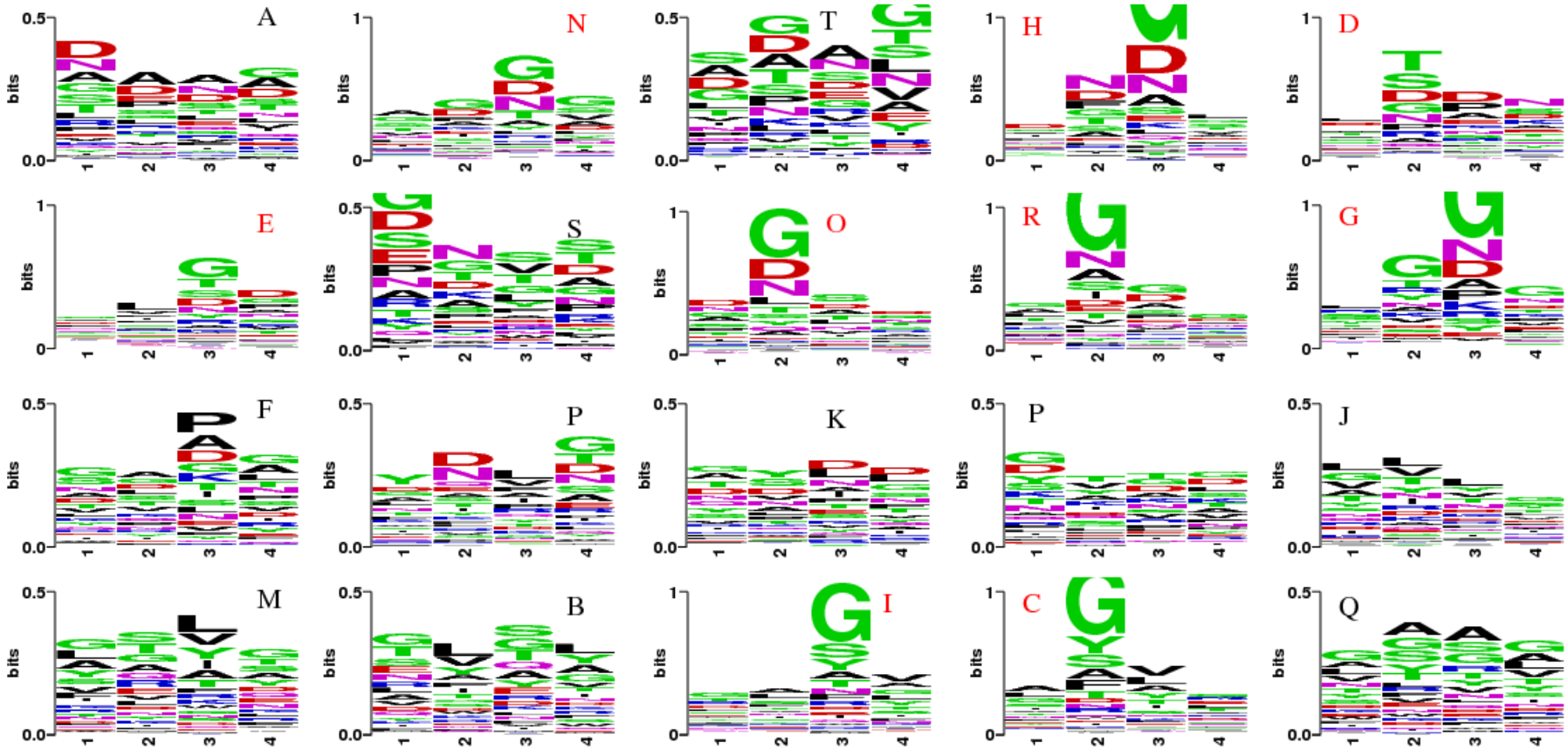

Figure 3: Sequence logos for all structural letters of SA20-OMP. Logos are generated using weblogo (WebLogo: a sequence logo generator, (2004), Crooks G. E., Hon G., Chandonia J.-M. and Brenner, S. E. Genome Res. 14(6):1188-1190.). Note that the scale is not same for all logos: logos of structural letters [$NHDORGIC$], noted in red, have their y range from 0 to 1 bit, whereas other letters have y range from 0 to 0.5

Motifs and anti-motifs in β-strands

This section presents a detailed analysis of our findings with intra and inter-strands motifs and anti-motifs identified in two previous studies:

- {1} Jackups R Jr, Cheng S, Liang J. Sequence motifs and anti-motifs in beta-barrel membrane proteins from a genome-wide analysis: the Ala-Tyr dichotomy and chaperone binding motifs. J Mol Biol. 2006 Oct 20;363(2):611-23), for intra-strand motifs and anti-motifs.

- {2} Jackups R Jr, Liang J. Interstrand pairing patterns in beta-barrel membrane proteins: the positive-outside rule, aromatic rescue, and strand registration prediction. J Mol Biol. 2005 Dec 9;354(4):979-93, for intrer-strand motifs and anti-motifs.

Since the background distribution of amino-acid in {1} and {2} is the amino-acid distribution in β-strands, we have to use the same reference. The amino-acid Zscores related to this background distribution are shown in Figure 4.

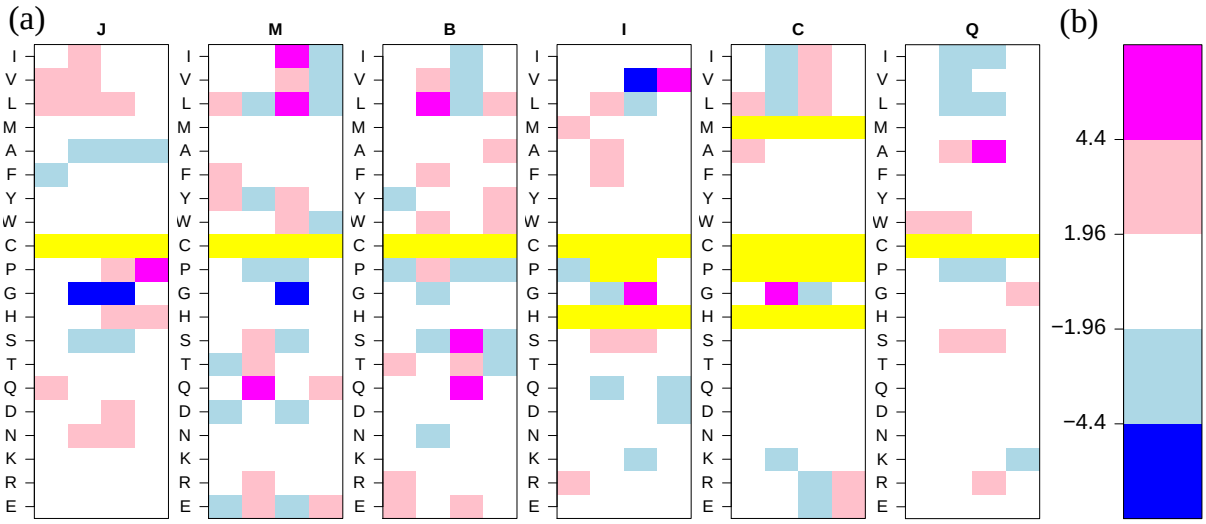

Figure 4: Z-scores of the 20 amino-acids in the four positions of β-specific SA20-OMP letters, **with the frequency of amino-acids in β-strands as a reference.** Pink and blue colors denote respectively over and under-represented amino-acids. Yellow indicates that the expected frequency is two low (less than five) to compute a Z-score.

Table 3 presents the intra-strand motifs and anti-motifs identified in {1} (Table 2, page 615), and the motifs and anti-motifs, deduced from the amino-acid Z-scores, that are in agreement with {1}. To derive sequential motifs from the structural letter presented in this study, we consider the successive over-represented amino-acids in each structural letter from Figure 4 (under-represented amino-acids in the case of anti-motifs).

Motifs identified in {1}	**GV, SY, GL**, VG, EM, **RY**, VK, **TW, LG, TV**
Structural letter	Motifs in agreement with {1}
J	
M	TV, TW, SY, RY
B	
I	LG
C	GV, GL
Q	
Anti-motifs identified in {1}	VY, **SS**, YY, DG, MT, YW, RA, TD, LV
Structural letter	Anti-Motifs in agreement with {1}
J	SS
M	
B	
I	
C	
Q	

Table 3: Intra-strand motifs and anti-motifs in OMP structures: agreement between the finding of Jackups and Liang and the sequential specificities of structural letters identified in this study.

Table 4 presents the inter-strand motifs and anti-motifs identified in {2} (Table 2, page 984) and the motifs and anti-motifs, deduced from the amino-acid Z-scores and structural pairwise preferences, that are in agreement with {2}. To derive inter-strand motifs from the structural letter presented in this study, we consider the over-represented structural pairs, and the over-represented amino-acids at their position 2 and 3. Similarly, for anti-motifs, we consider under-represented structural pairs and the over-represented amino-acids at position 2 and 3.

Motifs identified in {2}	GY, **ND**, GF, IY, KS, **LW**, **LY**, RP, **AA**, HK, **WY**, GI, **RE**, GV, **QG**, **LL**, **AV**, LP, DT, GP, EM, DP
Structural pair	Motifs in agreement with {2}
JJ	ND, LL, LP
QQ	AA
IC	LL, AV
MI	LL, AV, QG
MB	LW, LY, WY, RE, LL, LP
BC	LL, LP, QG
Anti-motifs identified in {2}	**YY**, GK, **QV**, **GY**, NL, PT, **AT**, FV
Structural pair	Anti-motifs in agreement with {2}
JQ	
JI	
MC	QV, GY
MM	YY, QV
IB	AT
BB	QV

Table 4: Inter-strand motifs and anti-motifs in OMP structures: agreement between the finding of Jackups and Liang and the sequential specificities of structural letters identified in this study.