Sequence Harmony: Detecting Functional Specificity from Alignments

K. Anton Feenstra, Walter Pirovano and Jaap Heringa
Centre for Integrative Bioinformatics VU (IBIVU), Vrije Universiteit Amsterdam

Sequence Harmony

We present the *Sequence Harmony* (*SH*) method and web server introduced previously [1,2]. *SH* allows the quick selection of subtype-specific sites from a multiple alignment given a subfamily grouping, by scoring compositional differences, without imposing conservation. In addition, it maps the predicted sites directly onto a protein structure for display and manipulation.
The web server can be accessed from:
`www.ibi.vu.nl/programs/SeqHarmWWW/`.

HIV Sequence Divergence and Disease Progression

Divergence between HIV virus populations in infected patients is important for predicting AIDS progression. Figure 3 shows selected sites in the crystal structure of the HIV capsid protein. The 26 sites selected by *SH* contain all 7 known B57 restricted CTL escape mutants. Comparison with non-B57 patients in early- and late-stage of disease progression, and those that show long-term non-progression, revealed 18 sites of which 12 are known mutation sites that correlate with HIV replication rates.

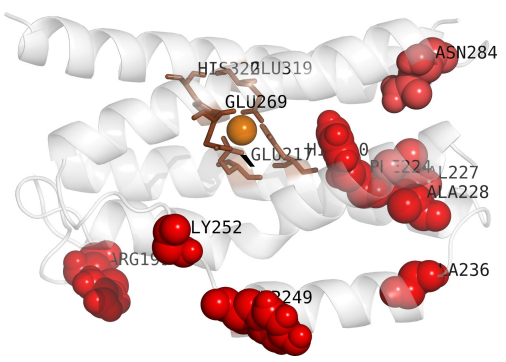

Figure 1 | **Low-harmony sites (red balls) for AOX1 vs. AOX2 in an AOX homology model. The putative active site iron (orange ball) and binding residues (sticks) are shown as well.**

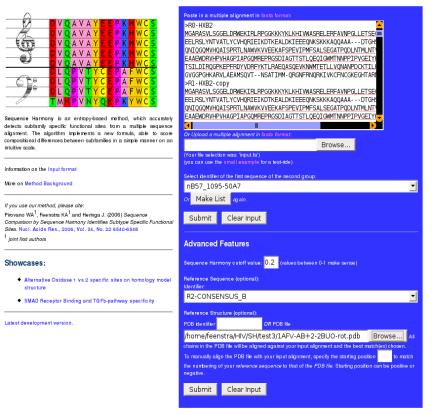

Figure 2 | **Input web-form of the SH Web Server, showing input and settings for analysis of the HIV alignment.**

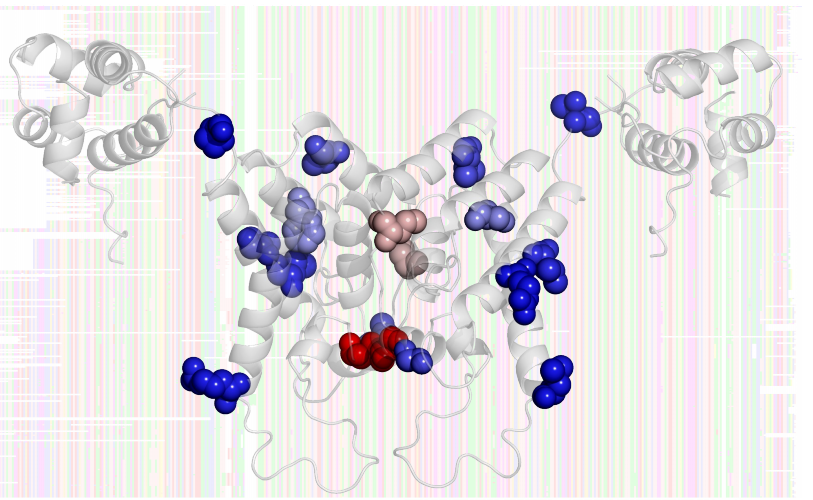

Figure 3 | **Low-harmony sites for non-B57 *vs*. B57 in the HIV capsid protein (1AFV, 2BU0). The backdrop shows the GAG region alignment where the capsid is located.**

Alternative Oxidase

The mitochondrial plant alternative oxidase (AOX) is encoded in two discrete gene subfamilies. AOX1 is found in monocot and dicot plants and is induced by stress stimuli. AOX2 is usually constitutive and has at present not been 55 found in monocot plants. Of the nine non-harmonious (SH=0) residues (Table 1), 5 are located at the substrate binding site as indicated in Figure 1.

Table 1 | **Output table for the Alternative Oxidase.**

Selected 17 positions below cutoff (0.2)

SH: 0.00 | 0.04 | 0.08 | 0.12 | 0.16 | 0.20 | 0.36 | 0.52 | 0.68 | 0.84 | 1.00

Position		Entropy				SH	Rnk	Consensus	
Ali	Ref	A	B	AB	rel.			A	B
245	A228	1.06	0.00	1.59	1.26	0.00	2	Astm	V
216	R199	0.00	0.00	0.77	1.26	0.00	1	R	K
241	F224	0.00	0.00	0.77	1.26	0.00	1	F	M
141	R124	0.00	0.00	0.77	1.26	0.00	1	R	M
253	A236	0.00	0.00	0.77	1.26	0.00	1	A	L
266	Y249	0.00	0.00	0.77	1.26	0.00	1	Y	F
136	R119	0.74	0.00	1.34	1.26	0.00	1	Kr	P
269	G252	1.06	1.38	1.90	1.26	0.00	1	Gat	Lfc
354	L337	2.05	0.00	2.36	1.26	0.00	1	Mqrhl	K
122	P107	1.36	0.59	1.89	1.17	0.07	2	Pqat-	Rt
244	V227	0.25	0.00	0.82	1.10	0.13	2	Vl	L
170	R153	0.90	0.00	1.33	1.10	0.13	1	Wlry	R
301	N284	1.67	1.15	2.20	1.07	0.15	1	AKnd	Sen
55	Q48	3.13	2.52	3.64	1.07	0.15	1	V-armlpqstw	Fcghy-
147	Q130	2.03	1.45	2.50	1.04	0.17	1	KTesqr	HRq
101	V86	2.73	1.38	3.02	1.03	0.18	1	AESvg-kt	Knt
121	P106	2.67	1.66	3.07	1.02	0.20	2	EPQdkav-	Sety

Method

The input of *SH* is a multiple sequence alignment of the protein family and a subdivision into two groups. Advanced features optionally allow more control over the analysis and output (see also Figure 2). The Sequence Harmony between two groups (A and B) at position *i* is defined as follows:

$$SH_i^{AB} = \sum_x p_{i,x}^A \log \frac{p_{i,x}^A}{p_{i,x}^A + p_{i,x}^B}$$

This can also be seen as a form of 'relative entropy' of group A relative to the sum of the probabilities of both groups (pA + pB). This function becomes zero for maximally different sites and one for sites with identical distributions. The final *SH* score used in the method was the average of SH^{AB} and SH^{BA}.

References

[1] Pirovano, W, KA Feenstra, & J Heringa "*Sequence Comparison by Sequence Harmony Identifies Subtype Specific Sites Functional Sites*" Nucl. Acid. Res., 34: 6540 2006.
[2] Feenstra, KA, W Pirovano, K Krab & J Heringa "*Sequence Harmony: Detecting Functional Specificity from Alignments*", Nucl. Acid. Res., 35: W495 2007.
[3] HIV sequence data was kindly provided by M Navis & N Kootstra, Clinical Viro-Immunology, Sanquin Research Amsterdam.

ENFIN Enabling Systems Biology
CENTRE FOR INTEGRATIVE BIOINFORMATICS VU
nbic Netherlands Bioinformatics Centre