Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W495–W498***
doi:10.1093/nar/gkm406

Sequence harmony: detecting functional specificity from alignments

K. Anton Feenstra[1], Walter Pirovano[1], Klaas Krab[2] and Jaap Heringa[2,*]

[1]Centre for Integrative Bioinformatics VU (IBIVU) and [2]Institute of Molecular Cell Biology, Vrije Universiteit Amsterdam, De Boelelaan 1081A, 1081 HV Amsterdam, The Netherlands

Received January 31, 2007; Revised April 5, 2007; Accepted May 3, 2007

ABSTRACT

Multiple sequence alignments are often used for the identification of key specificity-determining residues within protein families. We present a web server implementation of the Sequence Harmony (SH) method previously introduced. SH accurately detects subfamily specific positions from a multiple alignment by scoring compositional differences between subfamilies, without imposing conservation. The SH web server allows a quick selection of subtype specific sites from a multiple alignment given a subfamily grouping. In addition, it allows the predicted sites to be directly mapped onto a protein structure and displayed. We demonstrate the use of the SH server using the family of plant mitochondrial alternative oxidases (AOX). In addition, we illustrate the usefulness of combining sequence and structural information by showing that the predicted sites are clustered into a few distinct regions in an AOX homology model. The SH web server can be accessed at www.ibi.vu.nl/programs/seqharmwww.

INTRODUCTION

During the past years there has been a wide interest in studies of specificity-determining residues within protein subfamilies (1). Consequently, an increasing number of methods and web applications has become available that offer functional analyses of subtype specificity from multiple alignments (2–5).

Previously we evaluated advantages and limitations of several of the state-of-the-art methods and introduced a new method called Sequence Harmony (*SH*) (6). *SH* accurately detects positions within an alignment that are responsible for functional differences between two protein subfamilies. To further facilitate the use of *SH* for a broad audience, we have implemented a comprehensive web server. The *SH* web server offers a fast, one-step analysis with a number of options, yielding results that can be interpreted easily.

In this article we will guide the user through all the steps of the *SH* web-application by means of a biologically relevant example. We will look for subtype specific sites for the two subfamilies of the alternative oxidases (AOX) protein family of plant alternative oxidases. The subtype specific sites found are the best candidates to explain the functional differences. Other relevant applications of this method include pathway specificity, ligand specificity, host-specific viral infection, viral disease progression differences and viral drug resistance.

METHODS

Implementation

The SH method, as described previously (6), is currently implemented as an awk program. The main steps are as follows. Sequences are read from the alignment and separated into two user-specified groups. For each group separately and combined entropies (E) are calculated. SH values are calculated as $SH = \frac{1}{2}(E_{A+B} - E_A - E_B)$, with $E_{A+B} = -\sum(p_A + p_B)\log(p_A + p_B)$, i.e. using the sum of the normalized frequencies of groups A and B. SH values range from zero for completely non-overlapping residue compositions, to one for identical compositions. Next, sites are selected that have a SH score below a cutoff. Stretches of neighbouring selected sites are identified and the size of each of these stretches is assigned to the sites as the 'Rank'. Finally, selected sites are sorted on i) increasing SH, ii) decreasing Rank and iii) increasing entropy. This sorted list of selected sites is the primary result of the *SH* algorithm. Currently alignments of up to about 5 million residues (including gaps) can be processed with runtimes on the order of tens of seconds, excluding the generation of the high-quality output PyMol image.

The separate chains of the optional protein structure (PDB file) are aligned with the input alignment using the 'profile' option of Muscle (7). Subsequently, the chains with the highest average SH score with respect to the input

*To whom correspondence should be addressed. Tel: +31 20 598 7649; Fax: +31 20 598 7653; Email: heringa@few.vu.nl

alignment are selected for graphical display as an image in an interactive Jmol applet.

Use of the Server

The *SH* web server contains basic and advanced options. Sanity of the input options is checked, e.g. whether an input alignment is uploaded or pasted, and not both, whether it is proper FASTA format. Problems are reported with an error message and offending input fields are highlighted. An example screenshot of the main page and input form is shown in Figure 1. The main input is a multiple sequence alignment of the protein family and a subdivision into two groups. Sequence data in FASTA format is case-insensitive and may be split over multiple lines; gaps should be represented with a dash ('–'); and lines beginning with a semi-colon (';') are ignored as comment. From the alignment the program automatically generates a list of sequence IDs. The user defines the two input groups by selecting the identifier of the first sequence of the second group. By default, a cutoff of 0.2 for the *SH* score is used.

Advanced features allow more control over the analysis and output, but the default settings usually suffice for a basic analysis. The *SH* cutoff can be adjusted between zero (allowing no compositional overlap) and one (allowing full overlap). A higher cutoff will lead to the selection of a larger number of sites. A reference sequence can be selected from the alignment and a starting position can be set to provide a reference numbering in the output tables. Additionally, a PDB identifier can be specified by its four-letter code, to visualize the selected sites in the protein structure. Alternatively, a PDB file can be uploaded. The PDB file is automatically aligned with the multiple alignment, or manually when the reference position is set.

The output is a sorted table and optional graphical representations of the selected sites as an image (generated by PyMol, www.pymol.org) and as an interactive Jmol applet (8). The Jmol applet includes buttons to set the SH cutoff for displaying selected sites, and to show or hide non-aligned chains in the PDB file. An example screenshot of the main output page is shown in Figure 2. An additional, non-sorted, table is available that lists all sites in the alignment. The tables are coloured from red (predicted subtype specific sites) to blue (sites not predicted).

RESULTS

The AOX family of alternative plant oxidases

The cyanide-insensitive plant alternative oxidase (AOX) is a mitochondrial non-haeme quinol oxidase (9). An established function of AOX is thermogenesis in the spadices of *Aroid* lilies, but little is known about AOX function in other tissues (10). AOX is encoded in two discrete gene subfamilies (11). AOX1 is found in monocot and dicot plants and is induced by stress stimuli (11,12). AOX2 is usually constitutive and has at present not been found in monocot plants (11).

Figure 1. Input screen of the Sequence Harmony server, showing part of the AOX alignment and model input. (**A**) Overview of the Sequence Harmony web page; and (**B**) the Sequence Harmony input options and advanced features.

Sequence Harmony for 'AOX_PSIPral-AliSorted.fasta'

Generated on Tue 27 Mar 2007 15:08 by SeqHarm version 1.1
from the Sequence Harmony Webserver at http://www.ibi.vu.nl/programs/seqharmwww/
Please cite:
Walter Pirovano*, K. Anton Feenstra* and Jaap Heringa
"*Sequence Comparison by Sequence Harmony Identifies Subtype Specific Functional Sites*"
Nucl. Acids Res., 2006, Vol. 34, No. 22 6540-6548
* joint first authors.
Found reference sequence 'Aox1_Saur_gu/1-349', offset=1.
Group A: 24 sequences, length 366
Sequences: Aox1_Glyc_ma/1-321, Aox1_Nico_at/1-353, Aox1_Saur_gu/1-349, Aox1_Sola_tu/1-34
Group B: 7 sequences, length 366
Sequences: Aox2_Arab_th/1-353, Aox2_Cucu_sa/1-346, Aox2_Mang_in/1-274, Aox2a_Glyc_ma/1-
Cutoff: 0.2.

Archive of all output files

Results for AOX_PSIPral-AliSorted.fasta

Aligned chains from AOX_SwissMODEL_em.pdb or together with whole alignment

Selected 17 positions below cutoff (0.2)

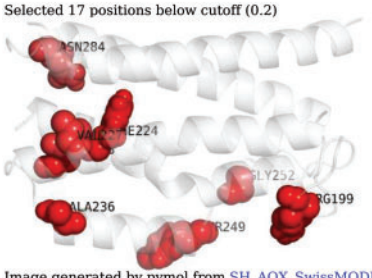
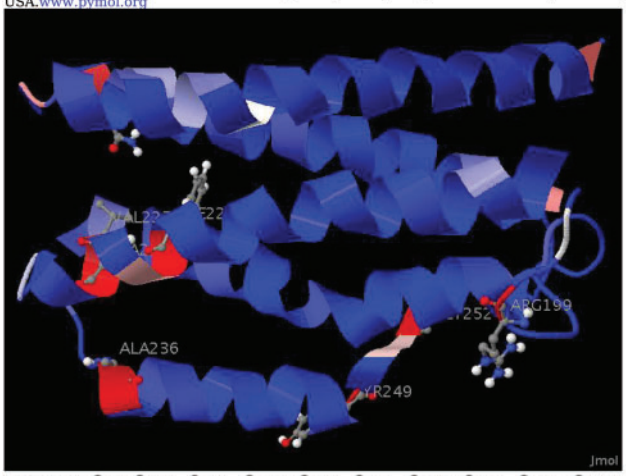


Image generated by pymol from SH_AOX_SwissMODEL_em.pdb pdb file with '1-SH' as B-factors and using display.pml pymol script:
DeLano, W.L. The PyMOL Molecular Graphics System (2002) DeLano Scientific, Palo Alto, CA, USA.www.pymol.org



SH cutoff ○ 0 ○ 0.1 ◉ 0.2 ○ 0.3 ○ 0.4 ○ 0.5 ○ 0.6 ○ 0.7 ○ 0.8 ○ 0.9 ○ 1
Show ○ Whole structure ◉ Only Aligned

Raw Table

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

Figure 2. Output screen of the Sequence Harmony server with results of the AOX analysis. Included are a summary of input parameters, a PyMol rendered image, the Jmol applet and controls, and the output table. Links to the raw table with all alignment positions and to additional output files are also provided.

We retrieved 31 AOX sequences from the NCBI non-redundant database (www.ncbi.nih.gov); 24 of the AOX1 and 7 of the AOX2 subfamily. Sequences were aligned using PSI-Praline (13,14) using secondary structure prediction information from PSIPRED (15). The alignment is available as Supplementary Data. Lacking an experimental AOX structure, we constructed a homology model using of the catalytic iron-binding N-terminal domain. Swiss-Model (16) was used to align the *Sauromatum guttatum* AOX1 sequence with the template structure 1AFR (17), following the description of Andersson and Nordlund (18). The PDB file is provided as Supplementary Data.

Sample input

The multiple alignment obtained was uploaded in the appropriate box of the *SH* web server and the first AOX2 sequence (AOX2_Arab_th) was selected as the first sequence of the second group (Figure 1). The default *SH* cutoff of 0.2 was used, meaning that partial overlap in residue composition is allowed. We chose AOX1 from *Sauromatum guttatum* as reference sequence so that the subtype specific sites obtained will be numbered accordingly (18). The PDB file of the model structure was uploaded to the *SH* web server.

Sample output

The results page of the SH server (Figure 2) shows at the top a summary of the input parameters, and optionally the graphical representations of the protein structure. Below that is the table containing all sites with a *SH* score (SH) below the cutoff value (default 0.2). These are sorted first on increasing *SH* score, then on decreasing rank (Rnk, number of neighbouring sites below the cutoff) and finally on increasing total Entropy (AB) of the alignment position, as described previously (6). Position Ali and Ref show the sequence position in the alignment and reference sequence, respectively. The 'Consensus' columns give all residue types present in groups A and B, respectively, in order of decreasing frequency and in lowercase when the frequency is less than half of the highest. In addition, a link is provided to the raw table, that holds all information of the complete alignment without selection or ranking. The raw table is available as Supplementary Data.

For the AOX family we found nine sites with a *SH* value of zero (bright red in Figure 2). An additional eight sites have SH values between zero and 0.2 (from light red to white). In the 'Rnk' column, a value of 2 signifies a stretch of two consecutive sites with SH values below the cutoff (alignment positions 244–245 and 121–122). Interestingly, out of nine sites with zero harmony, four have non-zero entropy in at least one group, i.e. they are not totally conserved within both groups (alignment positions 136, 245, 269, 354). This is also reflected in the 'Consensus' columns, which show multiple residue types occurring.

The graphical representations of the protein structure show the predicted sites in our homology model of the C-terminal domain (positions 182–352), see Figure 3 and also Figure 2. This domain contains two sites of zero

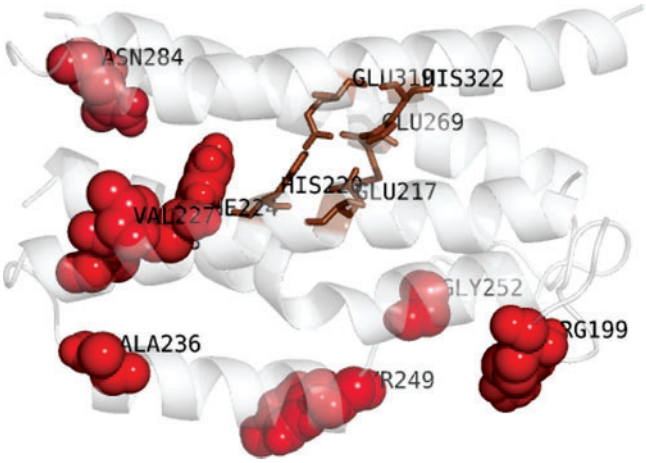


Figure 3. Model of the AOX iron-binding domain with low-harmony sites shown in spheres and the iron-binding Glu and His residues in sticks. Figure created using PyMol (www.pymol.org).

harmony and six of low harmony. Four low-harmony sites on the upper-left (Phe224, Val227, Ala228 and Asn284) are close to the Glu and His iron binding residues (18) (Figure 3), and could conceivably play a role in ligand binding or activation mechanisms. Three other low-harmony sites, Arg199, Tyr249 and Gly252 are located around a putative membrane-binding region (18), and could conceivably play a role in modulating the protein-membrane interactions or substrate access. The functional prediction of *SH* from the multiple alignment seems to be consistent with the structural prediction using homology modelling, although a different structural model could lead to different conclusions.

CONCLUSION

Our SH web server identifies putative subtype specific sites based on an alignment and selection of two groups as input only. In addition, sites can be easily linked to structural information using a structure from the PDB directly, or, as shown in our example, using a homology model of the protein. Using this structural information, selected subtype specific sites can be grouped into spatial clusters of sites that are likely to share functional relationships.

The combination of the SH algorithm and protein structural information has yielded a useful tool to interpret multiple sequence alignments, and to guide subsequent selection of interesting sites for experimental investigation.

SUPPLEMENTARY DATA

The AOX input alignment, raw output table, 'coloured' PDB and pymol script are available as Supplementary Data. In addition, the output is available at www.ibi.vu.nl/programs/seqharmwww/showcase_aox.

ACKNOWLEDGEMENTS

K.A.F. and W.P. acknowledge financial support from the Netherlands Bioinformatics Centre, BioRange Bioinformatics research programme (SP 2.3.1 and SP 3.2.2) and the EC Network of Excellence 'ENFIN' (LSHG-CT-2005-518254). The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Whisstock,J.C. and Lesk,A.M. (2003) Prediction of protein function from protein sequence and structure. *Quart. Rev. Biophys.*, **36**, 307–340.
2. Mirny,L.A. and Gelfand,M.S. (2002) Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J. Mol. Biol.*, **321**, 7–20.
3. Kalinina,O.V., Mironov,A.A., Gelfand,M.S. and Rakhmaninova,A.B. (2004) Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Prot. Sci.*, **13**, 443–456.
4. Donald,J.E. and Shakhnovich,E.I. (2005) Predicting specificity-determining residues in two large eukaryotic transcription factor families. *Nucleic Acids Res.*, **33**, 4455–4465.
5. Ye,K., Lameijer,E., Beukers,M. and IJzerman,A. (2006) A two-entropies analysis to identify functional positions in the transmembrane region of class A G protein-coupled receptors. *Proteins*, **63**, 1018–1030.
6. Pirovano,W., Feenstra,K.A. and Heringa,J. (2006) Sequence comparison by sequence harmony identifies subtype specific functional sites. *Nucleic Acids Res.*, **34**, 6540–6548.
7. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
8. Herráez, A. (2006) Biomolecules in the computer: Jmol to the rescue. *Biochem. Educ.*, **34**, 255–261.
9. Affourtit,C., Albury,M., Crichton,P. and Moore,A. (2002) Exploring the molecular nature of alternative oxidase regulation and catalysis. *FEBS Lett.*, **510**, 121–126.
10. Wagner,A. and Moore,A. (1997) Structure and function of the plant alternative oxidase: its putative role in the oxygen defence mechanism. *Biosci. Reports*, **17**, 319–333.
11. Considine,M., Holzapffel,R., Day,D., Whelan,J. and Millar,A. (2002) Molecular distinction between alternative oxidase from monocots and dicots. *Plant Physiol.*, **129**, 949–953.
12. Vanlerberghe,G. and McIntosh,L. (1997) Alternative oxidase: from gene to function. *Annu. Rev. Plant Physiol. Plant Mol. Biol.*, **48**, 703–734.
13. Simossis,V.A. and Heringa,J. (2003) The PRALINE online server: optimising progressive multiple alignment on the web. *Comput. Biol. Chem.*, **27**, 511–519.
14. Simossis,V.A. and Heringa,J. (2005) PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Res.*, **33**, W289–W294.
15. Jones,D. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
16. Guex,N. and Peitsch,M. (1997) Swiss-model and the swiss-pdbviewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.
17. Lindqvist,Y., Huang,W., Schneider,G. and Shanklin,J. (1996) Crystal structure of Δ^9 stearoyl-acyl carrier protein desaturase from castor seed and its relationship to other di-iron proteins. *EMBO J.*, **15**, 4081–4092.
18. Andersson,M. and Nordlund,P. (1999) A revised model of the active site of alternative oxidase. *FEBS Lett.*, **449**, 17–22.