Vol. 20 no. 12 2004, pages 1983–1985
doi:10.1093/bioinformatics/bth185

ConSSeq: a web-based application for analysis of amino acid conservation based on HSSP database and within context of structure

R. H. Higa[1], A. J. Montagner[1], R. C. Togawa[2], P. R. Kuser[1], M. E. B. Yamagishi[1], A. L. Mancini[1], G. Pappas Jr.[3], R. T. Miura[2], L. G. Horita[2] and G. Neshich[1,*]

[1]Núcleo de Bioinformática, Centro Nacional de Pesquisa Agropecuária, Empresa Brasileira de Pesquisa Agropecuária, Campinas, SP, Brazil, [2]Laboratório de Bioinformática, Embrapa/Recursos Genéticos e Biotecnologia and [3]Laboratório de Bioinformática, Universidade Católica de Brasília, DF-70770-9001, Brasilia

Received on September 2, 2003; revised on January 24, 2004; accepted on February 1, 2004
Advance Access publication March 25, 2004

ABSTRACT

Summary: A web-based application to analyze protein amino acids conservation—Consensus Sequence (ConSSeq) is presented. ConSSeq graphically represents information about amino acid conservation based on sequence alignments reported in homology-derived structures of proteins. Beyond the relative entropy for each position in the alignment, ConSSeq also presents the consensus sequence and information about the amino acids, which are predominant at each position of the alignment. ConSSeq is part of the STING Millennium Suite and is implemented as a Java Applet.

Availability: http://sms.cbi.cnptia.embrapa.br/SMS/STINGm/consseq/, http://trantor.bioc.columbia.edu/SMS/STINGm/consseq/, http://mirrors.rcsb.org//SMS/STINGm/consseq/, http://www.es.embnet.org/SMS/STINGm/consseq/ and http://www.ar.embnet.org/SMS/STINGm/consseq/

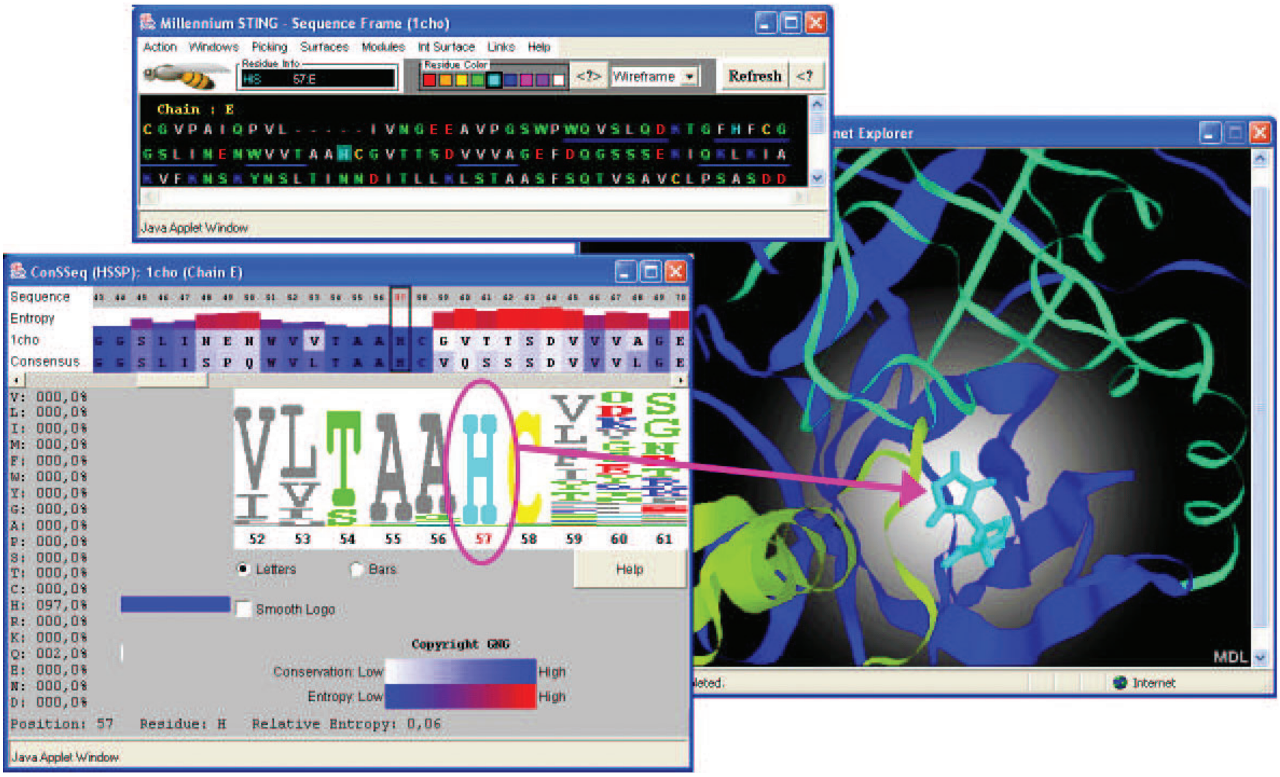
Contact: neshich@cbi.cnptia.embrapa.br

Consensus Sequence (ConSSeq) is a web-based application that allows the user to visualize graphically how each position in the sequence alignment provided by homology-derived structures of proteins (HSSP) (Sander and Schneider, 1991; Dodge *et al*., 1998) varies in terms of occupancy by 20 amino acids. ConSSeq is part of the STING Millennium Suite (SMS) (Neshich *et al*., 2003), but can also be accessed as an independent application. ConSSeq uses different graphical resources to provide the user with all the wealth of information contained in HSSP about residue conservation. The program is interactive and user friendly. It was implemented as a Java Applet (Campione *et al*., 2001) and can be executed through Netscape and Internet Explorer web-browsers.

Homology-derived structures of proteins is a derived database merging information on three-dimensional (3D) protein structures and sequences of homologous proteins. For each protein structure in the Protein Data Bank (PDB) (Berman *et al*., 2000), the database reports an alignment containing a list of putative homologs selected from SWISS-PROT (Bairoch and Apweiler, 2000). This alignment is built by using an iterative position-weight dynamic programming method and a sequence profile alignment (MaxHom). A well-tested threshold for structural homology (Sander and Schneider, 1991) is also used in order to decide if a sequence is a putative homolog or not. In particular, HSSP is useful for analyzing residue conservation in structural context. For each position in the multiple sequence alignment, HSSP provides two measures of variability: conservation weight and relative entropy (Sander and Schneider, 1991). Over evolutionary time, conserved amino acids may be taken as evidence of differential selective pressure in mutational events. In light of a 3D protein structure, the conservation can reflect the importance of individual residues for the architecture of protein-fold and/or for the protein function (Sander and Schneider, 1991).

In order to use ConSSeq, a user needs to supply a four-character PDB-ID, consequently indicating a corresponding HSSP entry available at EBI. In Figure 1, the ConSSeq output for alpha chymotrypsin extracted from bovine pancreas (PDB-ID: 1CHO, chain E) is presented. At the top the ConSSeq window, the query sequence is presented aligned to the HSSP calculated consensus sequence. Above the two sequences, the ConSSeq window presents a bar chart colored according to a scale of colors that reflects the degree of amino acid conservation. The bar height reflects the corresponding relative entropy at that position of the alignment. On the left-hand side of the ConSSeq panel, the information

[*]To whom correspondence should be addressed.



Millennium STING - Sequence Frame (1cho)

Residue Info: HIS 57:E

Chain : E

```
C G V P A I Q P V L - - - - - I V N G E E A V P G S W P W Q V S L Q D K T G F H F C G
G S L I N E N W V V T A A H C G V T T S D V V V A G E F D Q G S S S E K I Q K L K I A
K V F K N S K Y N S L T I N N D I T L L K L S T A A S F S Q T V S A V C L P S A S D D
```

ConSSeq (HSSP): 1cho (Chain E)

```
Sequence   43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70
1cho       G  G  S  L  I  N  E  N  W  V  V  T  A  A  H  C  G  V  T  T  S  D  V  V  V  A  G  E
Consensus  G  G  S  L  I  S  P  Q  W  V  L  T  A  A  H  C  V  Q  S  S  S  D  V  V  V  L  G  E
```

```
V: 000,0%
L: 000,0%
I: 000,0%
M: 000,0%
F: 000,0%
W: 000,0%
Y: 000,0%
G: 000,0%
A: 000,0%
P: 000,0%
S: 000,0%
T: 000,0%
C: 000,0%
H: 097,0%
R: 000,0%
K: 000,0%
Q: 002,0%
E: 000,0%
N: 000,0%
D: 000,0%
Position: 57   Residue: H   Relative Entropy: 0,06
```

● Letters ○ Bars

□ Smooth Logo

Fig. 1. Example of ConSSeq output: alpha chymotrypsin extracted from bovine pancreas (PDB-ID: 1CHO, chain: E). A logo was generated showing His:57 at the central position. The central amino acid is indicated in the structure SMS window (inset on the right) and also shown in the sequence SMS window (inset above the ConSSeq panel).

about frequencies of residues found in other homologous sequences is shown. For fast visualization, ConSSeq generates a sequence logo in two flavors, one of them according to the logo introduced by Schneider and Stephens (1990). In both logos, the single letter code for each residue is colored according to the STINGpaint standard (Neshich *et al.*, 2003). The logo has a width of 10 amino acids and the user can choose the central one by sliding horizontal bar below the sequence presented at the top of the window. This allows the user to observe the degree of the conservation as well as the proportion of occurrence of each amino acid at any given position of the alignment.

ConSSeq can be used free of charge by the researchers interested in analyzing the proteins whose structures have been resolved as well as by the students interested in learning more about the relationship between conserved residues and their importance to protein folding and function.

FUTURE DEVELOPMENTS

ConSSeq will be including more information starting from SMS new release: Gold STING. Namely, our laboratory is making final tests on home-made HSSP data base: SHQ2S. Our own SHQ2S will allow for the immediate response to the inquiry on new sequences in terms of conservation and would offer additional adjustable threshold for calculating relative entropy from MSA data. ConSSeq will include data from other methods estimating conservation like the one from ConSurf (Armon *et al.*, 2002) and will also allow visualization and editing of multiple sequence alignment as well as display of phylogenetic trees generated from these alignments.

ACKNOWLEDGEMENTS

Work supported by FAPESP, Fundação de Amparo a Pesquisa do Estado de São Paulo—Project no. 1945/01, CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico—Project no. 521093/2001-5 (NV) and FINEP, Financiadora de Estudos e Projetos—Project no. 01/08895-0.

REFERENCES

Armon,A., Graur,D. and Ben-Tal,N. (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.*, **307**, 447–463

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Campione,M., Walrath,K. and Huml,A. (2001) *The Java(TM) tutorial: A Short Course on the Basics*, 3rd edn. Addison-Wesley.

Dodge,C., Schneider,R. and Sander,C. (1998) The HSSP database of protein structure–sequence alignments and family profiles. N*ucleic Acids Res.*, **26**, 313–315.

Neshich,G., Togawa,R., Mancini,A.L., Kuser,P.R., Yamagishi,M.E.B., Pappas Jr.,G., Torres,W.V., Campos,T.F., Ferreira,L.L., Luna,F.M. *et al*. (2003) STING millennium: a web based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. *Nucleic Acids Res.*, **31**, 3386–3392.

Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Prot. Struct. Func. Genet.*, **9**, 56–68.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acid Res.*, **18**, 6097–6100.