

# SRide: a server for identifying stabilizing residues in proteins

Csaba Magyar, M. Michael Gromiha<sup>1</sup>, Gerard Pujadas<sup>2</sup>, Gábor E. Tusnády<sup>1</sup> and István Simon\*

Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, H-1518 Budapest, PO Box 7, Hungary, <sup>1</sup>Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST) 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan and <sup>2</sup>Research group in 'Wine and Health', Departament de Bioquímica i Biotecnologia, Universitat Rovira i Virgili, Campus de Sant Pere Sescelades s/n, Tarragona 43007, Catalonia, Spain

Received February 11, 2005; Revised and Accepted March 21, 2005

## ABSTRACT

**Residues expected to play key roles in the stabilization of proteins [stabilizing residues (SRs)] are selected by combining several methods based mainly on the interactions of a given residue with its spatial, rather than its sequential neighborhood and by considering the evolutionary conservation of the residues. A residue is selected as a stabilizing residue if it has high surrounding hydrophobicity, high long-range order, high conservation score and if it belongs to a stabilization center. The definition of all these parameters and the thresholds used to identify the SRs are discussed in detail. The algorithm for identifying SRs was originally developed for TIM-barrel proteins [M. M. Gromiha, G. Pujadas, C. Magyar, S. Selvaraj, and I. Simon (2004), *Proteins*, 55, 316–329] and is now generalized for all proteins of known 3D structure. SRs could be applied in protein engineering and homology modeling and could also help to explain certain folds with significant stability. The SRide server is located at <http://sride.enzim.hu>.**

## INTRODUCTION

Protein structures are stabilized by numerous non-covalent interactions, e.g. hydrophobic, hydrogen bonding, electrostatic and van der Waals interactions (1,2). Hydrophobic interactions are believed to be the driving force behind protein folding and stability (3). The cooperative, non-covalent and long-range interactions between residues provide stability for resisting the local tendency to unfold (4,5). It has also been reported that the stabilizing residues (SRs) show high conservations

among protein sequences. These aspects suggest that by combining (i) surrounding hydrophobicity (6), (ii) a quantitative measure of the number of long-range residue–residue contacts [quantified by means of long-range order (LRO)] (7), (iii) stabilization centers (8,9) and (iv) conservation score (10), the SRs in protein structures could be predicted. We, therefore, developed a consensus approach for locating them in TIM-barrel proteins (11). As thermodynamic and kinetic experiments show (12,13), SRs identified by our algorithm have a significant role in the stabilization of protein structures. Thus, we believe that our definition of SRs can be a useful tool for scientists in the exploration of structural stability of proteins. For example, our TIM-barrel study suggested that the structure of most TIM-barrel proteins is stabilized by SRs, which appear in the inner core of  $\beta$ -strands and act as a skeleton of the protein. Most of the TIM-barrel proteins are enzymes or have functions that need a high level of flexibility for biochemical reactions, but at the same time they have high stability to ensure a long lifetime. By making a stable and rigid inner core and a flexible outer region of the barrel, this topology can satisfy both requirements.

In this paper, we extend this approach to all globular protein structures. We have computed surrounding hydrophobicity, LRO and involvement in stabilization center directly from 3D structures of proteins deposited in Protein Data Bank (14) and conservation score from the alignment of sequences available in Swiss-Prot database (15). Threshold values for each factor have been imposed to identify SRs. We have developed a web server for identifying SRs from protein 3D structures. Users can also select their own threshold values for each parameter to identify SRs. Using the default values, the results obtained for TIM-barrel proteins may slightly differ from the published (11) results. The difference can be accounted to the following facts. First, in the SRide server, we use more accurate values for the van der Waals atom radii (16).

\*To whom correspondence should be addressed. Tel: +36 1 4669276; Fax: +36 1 4665465, Email: [simon@enzim.hu](mailto:simon@enzim.hu)

© The Author 2005. Published by Oxford University Press. All rights reserved.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact [journals.permissions@oupjournals.org](mailto:journals.permissions@oupjournals.org)

This influences if a residue is classified as SC, or not. Second, the conservation scores calculated with our server differs slightly from the scores calculated with the ConSurf server. This could be probably due to the fact that we use a more current version of the ClustalW alignment program (17).

## IDENTIFICATION OF SRs

We checked that the four criteria mentioned above were satisfied according to the definitions given and justified in our earlier papers (7–11).

### Surrounding hydrophobicity

Surrounding hydrophobicity of a residue  $i$  [ $H_P(i)$ ] is calculated as the sum of hydrophobic indices, obtained from thermodynamic transfer experiments, of residues whose  $C_\alpha$  atoms are within the distance of 8 Å from the  $C_\alpha$  atom of residue  $i$ :

$$H_P(i) = \sum_{j=1}^{20} n_{ij} h_j$$

where  $n_{ij}$  is the total number of surrounding residues of type  $j$  around residue  $i$  of the protein, and  $h_j$  is the hydrophobic index of residue type  $j$ , in kcal/mol listed in (18).

### Long-range order

The LRO of a residue  $i$  is the number of long-range contacts of this residue counted in the following way:

$$LRO_i = \sum_{j=1}^N n_{ij} / N; n_{ij} = 1, \text{ if } |i - j| > 12; n_{ij} = 0 \text{ otherwise,}$$

where  $i$  and  $j$  are two residues, in which the  $C_\alpha$  distance between them is  $\leq 8$  Å, and  $N$  is the total number of residues in the protein.

### Stabilization center

SC residues are defined by considering the contact map of a protein. Two residues are in contact if there is at least one pair of heavy atoms with a distance less than the sum of the van der Waals radii of the two atoms plus 1.0 Å. A contact is considered long-range if it is between residues that are separated by at least 10 residues in the amino acid sequence. Two residues are SC elements if they are involved in long-range contacts and if at least one supporting residue can be found in each of the flanking tetra-peptides of these residues, in such a way that at least seven out of the possible nine interactions are formed between the two triplets (8). Stabilization centers are identified according to the definition of SC. These can also be obtained using the public server SCide (<http://www.enzim.hu/scide>) (9). If a residue is involved in a stabilization center, its SC value becomes 1, and 0 otherwise.

### Conservation scores of residues

Conservation of residues is identified by comparing the sequence of PDB (14) entries with sequences deposited in Swiss-Prot (15) using a local implementation of the public server ConSurf (10) (<http://consurf.tau.ac.il>). The ClustalW (17) aligned homologous sequences found by PSI-BLAST

(19) are used to calculate the measure of conservation by the Rate4Site algorithm (20). Residues are classified into nine categories according to their real conservation score. A score of 1 represents the most variable residues and a score of 9 represents the most conservative ones.

## THE CONSENSUS APPROACH

The SRs in the 3D structure of a protein are delineated with certain threshold values for each term (i.e. SR is the one in which the values for all these four parameters are equal to or greater than the specified threshold values). In our study of TIM-barrel proteins (11), we have used the following conditions to predict the SRs: (i)  $H_P \geq 20$  kcal/mol; (ii)  $LRO \geq 0.02$ ; (iii)  $SC \geq 1$ ; and (iv) conservation score  $\geq 6$ . The same threshold values have been used in SRide by default. The identified SRs represent a few percentages of all residues in a protein. The actual abundance varies from protein to protein. For example, our recent survey showed that in 63 TIM-barrel proteins, only 4.0% of the residues (i.e. 957 residues out of 23 968) were identified as 'stabilizing residues'. Users who prefer to apply stricter or more relaxed conditions in the definition of SRs can adjust the thresholds in the server accordingly.

## INPUT AND OUTPUT DATA OF THE SRide SERVER

The input of the SRide server is the atomic coordinate file of the protein to be analyzed. It can be specified by providing the four-letter PDB code. Alternatively, it can be any other atomic coordinate file in PDB format uploaded directly by the user. This second option is mainly for those who want to analyze structures obtained by homology modeling or other computational approaches. Calculations are carried out on the selected protein chain, and inter-chain interactions are not taken into account to calculate LRO,  $H_P$  and SC properties.

The output of the server is a list of the sequences used to calculate the conservation score and the list of the SRs, together with the  $H_P$ , LRO and conservation score values. The output is sent to the user via email because calculating the conservation score is rather time consuming (it can take several minutes).

To avoid submissions with non-existent email addresses, the user must complete a simple registration procedure. In this registration procedure, only one email address must be given and this is the address to which a registration code is sent. When the registration code is copied back into the proper field of the registration page, the email address will be enabled to place submissions.

The SRide server is located at <http://sride.enzim.hu>.

## ACKNOWLEDGEMENTS

The authors would like to thank Prof. Nir Ben-Tal (Biochemistry Department Tel Aviv University), and his former and present colleagues Dr Fabian Glaser and Dr Yossi Rosenberg for their contribution to the implementation of their Rate4Site algorithm into the SRide server, and Dr Zsuzsanna Dosztányi for her help. Financial support

from grants OTKA T-049073 and GVOP-3.1.1-2004-05-0143/3.0 is acknowledged. Funding to pay the Open Access publication charges for this article was provided by grant GVOP-3.1.1.-2004-05-0143/3.0.

*Conflict of interest statement.* None declared.

## REFERENCES

1. Dill, K.A. (1990) Dominant forces in protein folding. *Biochemistry*, **29**, 7133–7155.
2. Ponnuswamy, P.K. and Gromiha, M.M. (1994) On the conformational stability of folded proteins. *J. Theor. Biol.*, **166**, 63–74.
3. Ponnuswamy, P.K. (1993) Hydrophobic characteristics of folded proteins. *Prog. Biophys. Mol. Biol.*, **59**, 57–103.
4. Abkevich, V.I., Gutin, A.M. and Shakhnovich, E.I. (1995) Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *J. Mol. Biol.*, **252**, 460–471.
5. Gromiha, M.M. and Selvaraj, S. (2004) Inter-residue interactions in protein folding and stability. *Prog. Biophys. Mol. Biol.*, **86**, 235–277.
6. Manavalan, P. and Ponnuswamy, P.K. (1978) Hydrophobic character of amino acid residues in globular protein. *Nature*, **275**, 673–674.
7. Gromiha, M.M. and Selvaraj, S. (2001) Comparison between long-range interactions and contact order in determining the folding rates of two-state proteins: application of long-range order to folding rate prediction. *J. Mol. Biol.*, **310**, 27–32.
8. Dosztanyi, Z., Fiser, A. and Simon, I. (1997) Stabilization centers in proteins: identification, characterization and predictions. *J. Mol. Biol.*, **272**, 597–612.
9. Dosztanyi, Z., Magyar, C., Tusnady, G.E. and Simon, I. (2003) SCide: identification of stabilization centers in proteins. *Bioinformatics*, **19**, 899–900.
10. Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor, D., Martz, E. and Ben-Tal, N. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.
11. Gromiha, M.M., Pujadas, G., Magyar, C., Selvaraj, S. and Simon, I. (2004) Locating the stabilizing residues in  $(\alpha/\beta)_8$  barrel proteins based on hydrophobicity, long-range interactions, and sequence conservation. *Proteins*, **55**, 316–329.
12. Kursula, I., Partanen, S., Lambeir, A.M. and Wierenga, R.K. (2002) The importance of the conserved Arg191-Asp227 salt bridge of triosephosphate isomerase for folding, stability, and catalysis. *FEBS Lett.*, **518**, 39–42.
13. Gonzalez-Mondragon, E., Zubillaga, R.A., Saavedra, E., Chanez-Cardenas, M.E., Perez-Montfort, R. and Hernandez-Arana, A. (2004) Conserved cysteine 126 in triosephosphate isomerase is required not for enzymatic activity but for proper folding and stability. *Biochemistry*, **43**, 3255–3263.
14. 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.
15. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. and Schneider, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
16. Hubbard, S.J. and Thornton, J.M. (1993) NACCESS: Computer Program. Computer Program, Department of Biochemistry and Molecular Biology, University College, London.
17. Higgins, D., Thompson, J., Gibson, T., Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
18. Nozaki, Y. and Tanford, C. (1971) The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. *J. Biol. Chem.*, **246**, 2211–2217.
19. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Anang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
20. Pupko, T., Bell, R.E., Mayrose, I., Glaser, F. and Ben-Tal, N. (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18** (Suppl. 1), S71–S77.