

# STING Report: convenient web-based application for graphic and tabular presentations of protein sequence, structure and function descriptors from the STING database

Goran Neshich\*, Aduino L. Mancini, Michel E. B. Yamagishi, Paula R. Kuser, Renato Fileto, Ivan P. Pinto, Juliana F. Palandrani, João N. Krauchenco, Christian Baudet, Arnaldo J. Montagner and Roberto H. Higa

Núcleo de Bioinformática Estrutural, Embrapa/Informática Agropecuária, 13083-886 Campinas, Brazil

Received August 4, 2004; Revised and Accepted October 18, 2004

## ABSTRACT

The Sting Report is a versatile web-based application for extraction and presentation of detailed information about any individual amino acid of a protein structure stored in the STING Database. The extracted information is presented as a series of GIF images and tables, containing the values of up to 125 sequence/structure/function descriptors/parameters. The GIF images are generated by the Gold STING modules. The HTML page resulting from the STING Report query can be printed and, most importantly, it can be composed and visualized on a computer platform with an elementary configuration. Using the STING Report, a user can generate a collection of customized reports for amino acids of specific interest. Such a collection comes as an ideal match for a demand for the rapid and detailed consultation and documentation of data about structure/function. The inclusion of information generated with STING Report in a research report or even a textbook, allows for the increased density of its contents. STING Report is freely accessible within the Gold STING Suite at <http://www.cbi.cnptia.embrapa.br>, <http://www.es.embnet.org/SMS/>, <http://gibk26.bse.kyutech.ac.jp/SMS/> and <http://trantor.bioc.columbia.edu/SMS> (option: STING Report).

## INTRODUCTION

The ongoing evolution of projects on structural genomics is making it increasingly important to collect, access and diagrammatically present information about the sequence,

structure and function of any given macromolecule described in a Protein Data Bank (PDB) file. The systematic comparison of such information among corresponding amino acid positions in the homologous proteins (or mutants) enable us to make correlations between a sequence and its structure/function—an essential step toward the understanding of the biological role of sequence components.

The three-dimensional (3D) structures of over 27 thousand macromolecules have been deposited into the PDB (1) till date. These structures can be accessed from the RCSB/PDB server and further inspected by using additional tools provided at the same site, or alternatively, by a variety of other servers providing versatile tools and containing structure-related data. However, there is an evident limitation in the queries that one can pose to an underlying collection of complex and heterogeneous databases. These databases often lack efficient methods to integrate and report all pertinent parameters referring to a specific site in a structure. STING Report provides exactly this option to a user—extraction of a comprehensive set of integrated data related to a specific site, i.e. an individual amino acid located at the position of interest.

Gold STING (GS) and its predecessor, STING Millennium Suite (SMS) (2) are comprehensive resources for obtaining a variety of data about a selected structure. Some of the key advantages of using GS<sup>Java</sup> Protein Dossier (<sup>J</sup>PD) (3), e.g. are the volume of data that can be retrieved and the interactivity offered by this web-based Java tool. STING Report, on the other hand, provides a very simple interface which prompts the user to provide a query specification such as '1cho\_e\_195' (i.e. retrieve descriptors/parameters referring to the pdb file name '1cho.pdb', chain 'E' and the amino acid sequence number '195'). This is enough for STING Report to generate a complete summary of pertinent characteristics associated with the selected residue. The result of this query includes up to 19

\*To whom correspondence should be addressed. Tel: +55 19 3789 5774; Fax: +55 19 3289 9594; Email: neshich@cnptia.embrapa.br

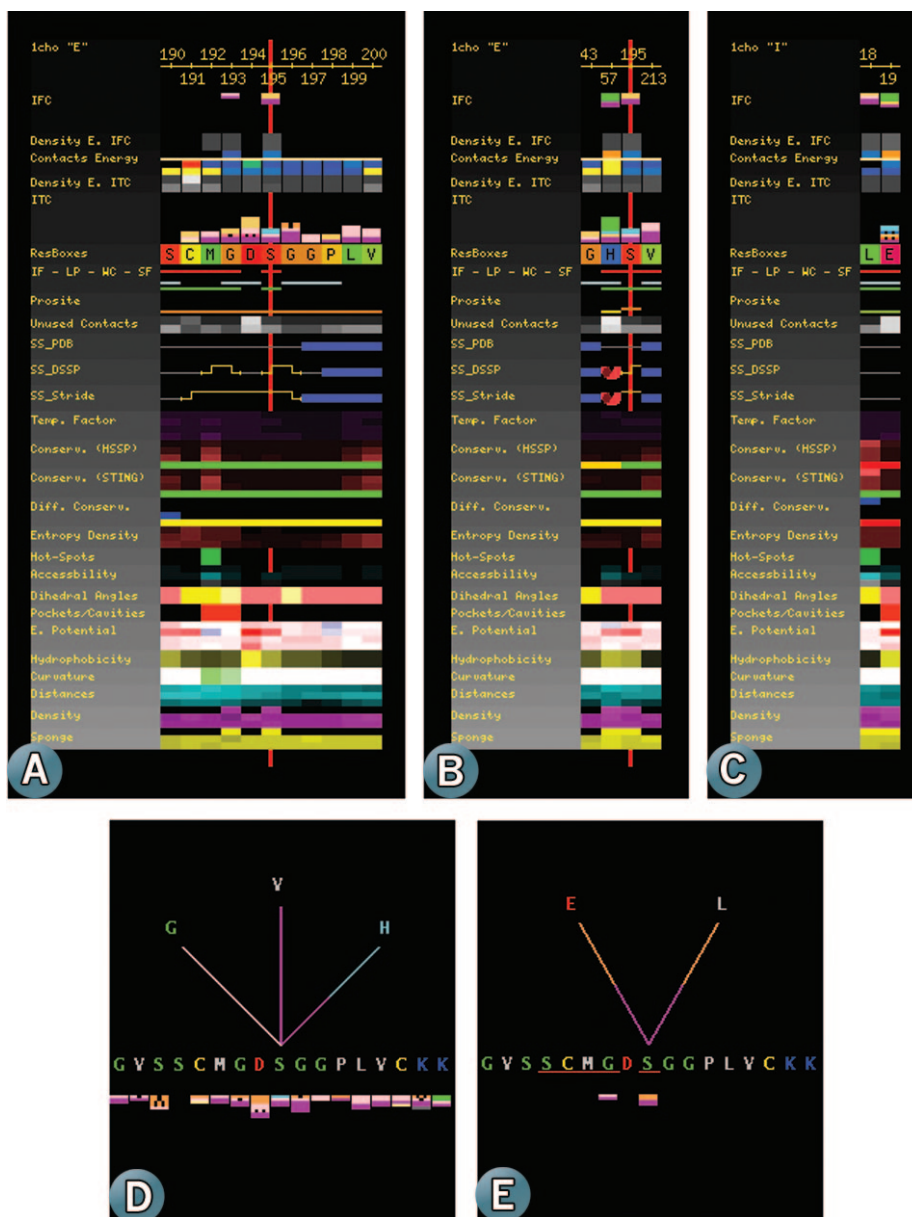
The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and 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 permissions, please contact [journals.permissions@oupjournals.org](mailto:journals.permissions@oupjournals.org).

images in the GIF format and one table (separated in five distinct, content-related areas), integrated in a single HTML document. Such output can be saved and/or printed, allowing document storing for future publication or detailed inspection.

## THE STING REPORT IMAGE AND TABLE AREAS

If a user submits a query by accepting a default option to produce all possible images and the table, the following

image areas will be shown (in the order of their appearance in the HTML window): (i) The  $^J$ PD area containing three distinct images: (a) the first image shows all parameters for a sequence of 11 residues with the selected one at the central position (see Figure 1A), (b) the second image shows the selected residue plus all residues making inter-atomic contacts with it (see Figure 1B) and (c) the third image shows all residues belonging to a chain other than the one where the selected residue is located and which make the inter-atomic



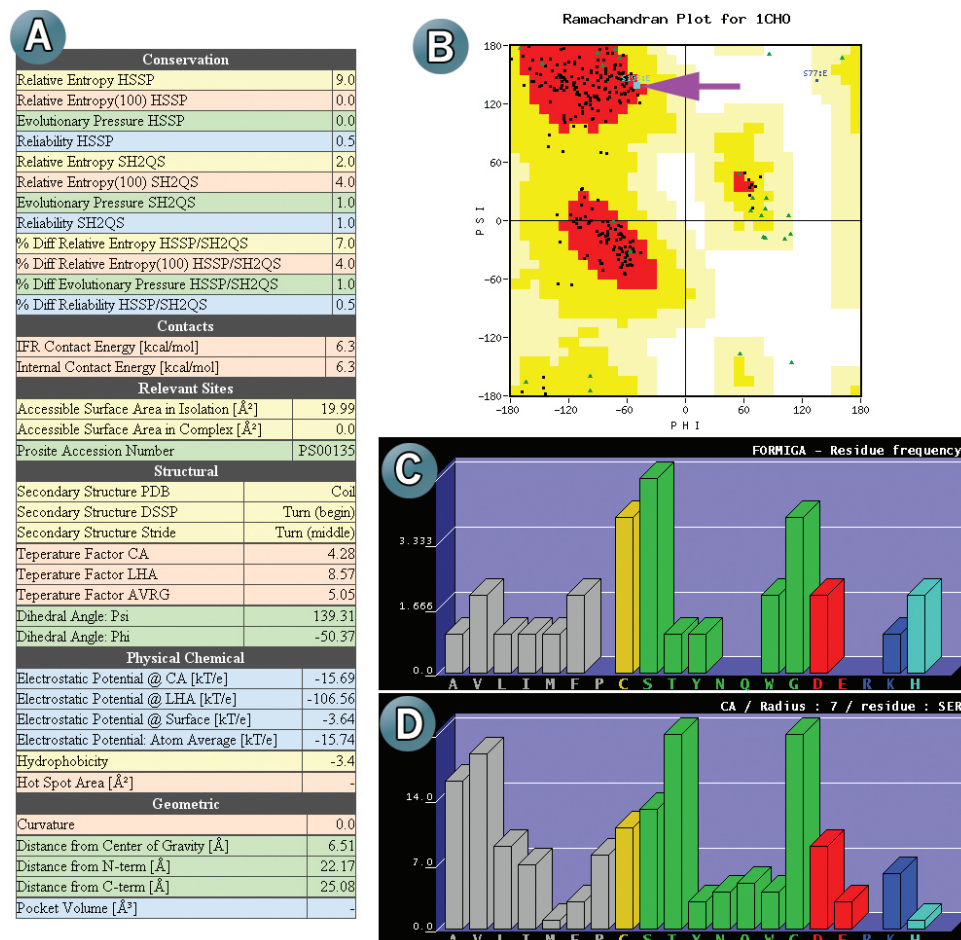
**Figure 1.** The STING Report  $^J$ PD image area contains three distinct images (data taken for the 1cho.pdb file, chain E, residue Serine\_195): (A) this image shows all parameters for a sequence of 11 residues with the selected Serine\_195, at the central position; (B) this image shows the Serine\_195 residue plus all residues of the same chain, making the inter-atomic contacts with it; and (C) this image shows all residues belonging to the 'I' chain which is different from the one where the selected residue is located (chain 'E') and which make the inter-atomic contacts across formed interface. By quick inspection of the  $^J$ PD shown parameters, one can see that the conservation is very high for a number of residues (dark boxes at the Conservation row), including Serine\_195. In addition, one can observe that the Serine\_195 makes internal contacts with three other amino acids, one of them being His\_57, also a member of the catalytic triade. The Graphical Internal (D) and Interface Forming Residue (IFR) Contacts (E) images show the fan-like representations of the amino acids inter-atomic contacts established with the Serine\_195. The color coding of the virtual contacting lines follows the established color code in STING Millennium Suite (SMS) (2). The color code practiced here is also self-evident from the Figure 2. (A) The  $^J$ Table of Contacts ( $^J$ TC) image is presented. Note that not all the red underlined residues (belonging to the interface) are actually making the IFR contacts.

contacts across the constituted interface (see Figure 1C); (ii) the <sup>Java</sup>Table of Contacts image area shows all the details of inter-atomic contacts of intra- and inter-chain type in a mixture of tabular and graphical presentation (see Figure 2A); (iii) the Graphical Internal (Figure 1D) and Interface Forming Residue (IFR) Contacts (Figure 1E)—these are fan-like representations of the amino acids inter-atomic contacts established to the selected residue; (iv) the ConSSeq image area contains two images: (a) the first image shows ConSSeq data based on the HSSP (4–6) MSA and (b) the second image shows the ConSSeq data based on the Gold STING SH<sub>2</sub>Q<sup>s</sup> (2)—both providing a wealth of information on relative entropy and sequence conservation (see Figure 2C); (v) the Ramachandran plot area (see Figure 3B) shows the dihedral angle plot (7); (vi) The X–Y plot area—this area contains five plots: (a) the Accessible Surface Area (ASA) for the amino acids belonging to a single chain which is isolated from any other chain, (b) the ASA for the amino acids within one chain but is in complex with any other chain(s) existing in a pdb file, (c) the Relative Entropy based on the HSSP data (see Figure 2B), (d) the

Relative Entropy based on the SH<sub>2</sub>Q<sup>s</sup> data and (e) the Temperature Factor; (vii) the Scorpion plot area which consist of two plots: (a) the first plot shows a frequency of occurrence for all amino acids belonging to the same chain as the selected residue and (b) the second plot shows a frequency of occurrence of the residues surrounding the selected residue and are counted if identified within a sphere of a given radius (see Figure 3D); (viii) the Formiga plot area which consists of three plots: (a) the first plot shows a frequency of occurrence for all amino acids belonging to the same chain as the selected residue and are at the interface, (b) the second plot shows a frequency of occurrence for all amino acids belonging to a chain other than the one to which the selected residue belongs and are at the interface and (c) the third plot shows a frequency of occurrence of the residues surrounding the selected residue and are counted if identified at the interface, belonging to the chain other than the one to which the selected amino acid belongs and are identified within a sphere of a given radius (see Figure 3C). The table area is divided into five content-related subtables, reporting numerical values for the sequence/



**Figure 2.** (A) The <sup>Java</sup>Table of Contacts (<sup>J</sup>TC) image area with all the details of inter-atomic contacts of intra- and inter-chain type in a mixture of tabular and graphical presentation. On the left and right extreme part of the table a user can see the data describing a Secondary Structure element and the ASA value for the chain in the isolation and in a complex with the other chain. Next column toward the center of the table shows the Relative Entropy values. The next three columns bring the residue sequence number, the residue type and the atom which engages into a given contact. The central column shows the distance between the two atoms which are establishing a particular type of the contact. In order to aid a user in quickly grasping what are the shortest distances of the contacts or which are the contacts established among the most conserved residues, the <sup>J</sup>TC bar color-coded graph accompanies the numerical values. (B) This graph represents only one (out of five available) X–Y plot: the Relative Entropy based on the HSSP data. (C) The ConSSeq image based on the Gold STING SH<sub>2</sub>Q<sup>s</sup> (16) data. Both the logo and bars above the sequence indicate that the Serine<sub>195</sub> is a very conserved residue.



**Figure 3.** (A) A part of the table reporting the numerical values for the structure parameters is shown. (B) The Ramachandran plot area shows the dihedral angle plot with the selected residue clearly labeled. (C) This graph shows one (out of three available) Formiga plot: the frequency of occurrence of the residues surrounding the selected residue is presented. These residues are counted in if identified at the interface and if they belong to the chain other than the one to which the selected amino acid belongs, and if they are identified within a sphere of a given radius. This graph also shows the preferred local environment residue constitution at the interface which belongs to the chain 'I'. (D) This graph shows the Scorpion plot with the frequency of occurrence of the residues surrounding the selected residue. Counter is incremented if the presence of the residue is identified within a sphere of a given radius. This graph also shows the preferred 3D local environment for all present serine residues in the chain 'E'.

structure/function parameters presented above in the graphic images.

Let us analyze, for illustration purposes, just one of the possible components of a document produced by querying the STING Database with the STING Report—the <sup>J</sup>PD image area. This component presents the values for each descriptor/parameter along an amino acid sequence on a residue-by-residue basis by using the color and shape codes. Such a graphic presentation and the large amount of information presented are distinctive characteristics of STING components in comparison with other software packages (8,9). This diagram is intuitive and easily understood by a typical STING Report user; the details describing all parameters and the procedures to calculate them are available on online help and in our earlier publications (10,11). In addition, the legend pointer at the STING Report HTML page contains the graphical presentation of all parameter classes which are clickable and which leads a user to the information describing a parameter definition and the procedure applied to calculate it.

As far as the appearance is concerned, the STING Report <sup>J</sup>PD component provides a graphical summary of

several important structural characteristics for a chosen protein. It allows a user to make very knowledgeable decisions about the possible roles of a specific amino acid in defining the function of the protein. It also helps to analyze what effect a specific mutation will possibly have on the structure and function of the protein, specifically by observing the changes in intra- and interface contact signatures, the sum of energy values for established contacts, the electrostatic potential at the surface and the conservation.

### STING REPORT STRUCTURE/FUNCTION PARAMETERS

Table 1 lists the categories of structure/function parameters that STING Report offers to a user (in parentheses are given references relating to the method used for calculation of those parameters).

Many of the structure/function descriptors provided by STING Report are calculated with a variety of conditions

**Table 1.** The categories of structure/function parameters that STING Report offers to a user

AA sequence & Prosite (12,13)	Density of Contact Energy
Contact Energy	Sponge (to be published)
Interface area (14)	Water contacting residues
Surface forming residues (14)	Ligand Pocket forming residues
Unused contacts	Secondary Structure elements (15,16)
Multiple Occupancy	Temperature Factor
AA Conservation (4–6,17)	Entropy density
Hot Spots (18)	Solvent accessibility (14)
Dihedral angles (7)	Pockets/Cavities (19)
Electrostatic potential (20,21)	Hydrophobicity (18)
Curvature (22)	Distance Geometry
Density (manuscript in preparation)	Inter-atomic Contacts (internal and across the interface)

References relating to the method used for calculation of those parameters are given in parentheses.

(e.g. variable volume of the probing sphere, atom at which the center of the probing sphere is placed, the variable size of the sliding window), leading to a total number of descriptors that can be reported to be 125. The detailed description of the parameters and procedures employed to obtain these parameters can also be found in the <sup>J</sup>PD Help—an interactive relative of the STING Report. The procedure we recommend to a user is to first use <sup>J</sup>PD for selecting the residues of interest by interactive parameter filtration and then employ the STING Report to obtain desired set of images/tables for the selected residues.

## SELECTING DOCUMENT COMPONENTS WITH STING REPORT

STING Report is completely integrated with Gold STING (3), the new edition of STING Millennium Suite (2). The STING Report user is able to select the images, graphs and tables that will be included in a resulting document. This is convenient if the focus of a user is already determined. The different components to be inserted in the resulting document are built by distinct modules of Gold STING, such as <sup>J</sup>PD (3), Contacts, Ramachandran Plot, ConSSeq (10), Scorpion and Formiga (11). STING Report coordinates the execution of these modules, by submitting to them the query that specifies the residue to be recovered and eventual conditions to calculate the parameters and generate the outputs. STING Report automatically activates the modules for which results are selected as required by the STING Report user, and assembles the results in a single HTML document.

Figures 1–3 show some fragments of an HTML document produced with the STING Report for an individual amino acid—the Serine<sub>195</sub> in the chain E of the 1cho.pdb. This particular amino acid was chosen because it belongs to the chymotrypsine active site and is one of the catalytic triade residues as well. The user can infer many valuable conclusions about how important this particular amino acid might be for the protein stability, function and binding to an inhibitor or a substrate. The legends of the figures describe each of these GIF components of the STING Report document.

## IMPLEMENTATION AND FUTURE DEVELOPMENTS

The STING Report package was implemented using the PERL language to process web requests through a CGI, submit them to the requested modules and then assemble the partial results in a single HTML report. The PERL script was also used to implement many of the modules that generate GIF images and HTML tables to compose the final report. Some specific modules were implemented in Java, but they rely on PERL libraries to load data from the STING database.

The STING DB is updated weekly and synchronized with the update of the PDB. However, recalculating some of the parameters such as the relative entropy, evolutionary pressure, MSA and phylogenetic trees requires a complete STING DB renovation. Such renovation is currently being performed every two/three months. Depending on the availability of the CPU time in the future, we will be striving to achieve weekly updates for all parameters/descriptors found in the STING database.

We have decided to disable the local file processing option on the STING Report entry web page as this facility can easily overload the Gold STING server if a number of requests for processing become high (as anticipated). However, in a future STING release—the Diamond STING—this option will be enabled. We are considering the addition of a new server and possibly, the transference of code to the client side, in the form of plugins, to efficiently process users' local files.

## CONCLUSIONS

STING Report is described here as a simple though content-rich information resource describing in full detail the participation of any given amino acid ('structural bit') in maintaining a protein structure stability and functional specificity. A compilation of the images and tables having a wealth of data about given structural spot allows a user to inspect and compare the key parameters/descriptors for the residues of interest to him/her (such as amino acids belonging to catalytic site, amino acids important for ligand binding, etc.).

The goal of STING Report is to facilitate the extraction of customized collections of data about specific amino acids. The main advantages of using STING Report are as follows: (i) getting the most complete set of descriptors/parameters about one residue in a single step; (ii) easy to use mechanisms to pose queries and specify the information to be returned; (iii) presentation of the results of a query in one HTML document that can be easily displayed or stored for future use; and (iv) ability to access all the Gold STING components using a browser with the most elementary configuration.

The documents containing the STING Report generated images have an increased information density and therefore might be very suitable for didactic purposes or as a summary of research findings.

## ACKNOWLEDGEMENTS

This work was supported in part by following grants: FAPESP 01/08895-0, FINEP 1945/01 and CNPq 521093/2001-5 (NV).

## REFERENCES

- Berman, H.H., 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.
- Neshich, G., Togawa, R., Mancini, A.L., Kuser, P.R., Yamagishi, M.E.B., Pappas, G., Jr, 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.
- Neshich, G., Rocchia, W., Mancini, A.L., Yamagishi, M.E.B., Kuser, P.R., Fileto, R., Baudet, C., Pinto, I.P., Montagner, A.J., Palandrani, J.F. *et al.* (2004) <sup>Java</sup>Protein Dossier: a novel web-based data visualization tool for comprehensive analysis of protein structure. *Nucleic Acids Res.*, **32**, W595–W601.
- Sander, C. and Schneider, R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
- Schneider, R., de Daruvar, A. and Sander, C. (1997) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, **25**, 226–230.
- Schneider, R. and Sander, C. (1996) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, **24**, 201–205.
- Ramachandran, G.N., Ramakrishnan, C. and Sasisekharan, V. (1963) Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, **7**, 95–99.
- Laskowski, R.A., Hutchinson, E.G., Michie, A.D., Wallace, A.C., Jones, M.L. and Thornton, J.M. (1997) PDBsum: a web-based database of summaries and analyses of all PDB structures. *Trends Biochem. Sci.*, **22**, 488–490.
- Hogue, C.W. (1997) Cn3D: a new generation of three-dimensional molecular structure viewer. *Trends Biochem. Sci.*, **22**, 314–316.
- Higa, R.H., Montagner, A.J., Togawa, R.C., Kuser, P.R., Yamagishi, M.E.B., Mancini, A.L., Pappas, G., Jr, Miura, R.T., Horita, L.G. and Neshich, G. (2004) ConSSeq: a web-based application for analysis of amino acid conservation based on HSSP database and within context of structure. *Bioinformatics*, **20**, 1983–1985.
- Higa, R.H., Oliveira, A.G., Horita, L.G., Miura, R.T., Inoue, M.K., Kuser, P.R., Mancini, A.L., Yamagishi, M.E., Togawa, R. and Neshich, G. (2004) Defining 3D residue environment in protein structures using SCORPION and FORMIGA. *Bioinformatics*, **20**, 1989–1991.
- Bucher, P. and Bairoch, A. (1994) A generalized profile syntax for biomolecular sequences motifs and its function in automatic sequence interpretation. In Altman, R., Brutlag, D., Karp, P., Lathrop, R. and Searls, D. (eds), *Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology (ISMB-94)*. AAAI Press, Menlo Park, CA, pp. 53–61.
- Gattiker, A., Gasteiger, E. and Bairoch, A. (2002) ScanProsite: a reference implementation of a PROSITE scanning tool. *Appl. Bioinformatics*, **1**, 107–108.
- Sridharan, S., Nicholls, A. and Honig, B. (1992) A new vertex algorithm to calculate solvent accessible surface areas. *Biophys. J.*, **61**, A174.
- Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometric features. *Biopolymers*, **22**, 2577–2637.
- Frishman, D. and Argos, P. (1995) Knowledge-based protein secondary structure assignment. *Proteins*, **23**, 566–679.
- 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**, S71–S77.
- Radzicka, A. and Wolfenden, R. (1988) Comparing the polarities of the amino-acids—side-chain distribution coefficients between the vapor-phase, cyclohexane, 1-octanol, and neutral aqueous-solution. *Biochemistry*, **27**, 1664–1670.
- Nicholls, A., Sharp, K. and Honig, B. (1991) Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins*, **11**, 281–296.
- Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P.V. and Subramaniam, S. (1998) Analytical shape computation of macromolecules: II. Inaccessible cavities in proteins. *Proteins*, **33**, 18–29.
- Rocchia, W., Sridharan, S., Nicholls, A., Alexov, E., Chiabrera, A. and Honig, B. (2002) Rapid grid-based construction of the molecular surface for both molecules and geometric objects: applications to the finite difference Poisson–Boltzmann method. *J. Comput. Chem.*, **23**, 128–137.
- Tsodikov, O.V., Record, M.T., Jr and Sergeev, Y.V. (2002) A novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. *J. Comput. Chem.*, **23**, 600–609.