3dLOGO: a web server for the identification, analysis and use of conserved protein substructures

Allegra Via1,*, Daniele Peluso1, Pier Federico Gherardini1, Emanuele de Rinaldis1,2, Teresa Colombo3,4, Gabriele Ausiello1 and Manuela Helmer-Citterich1

1Centre for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata, 00133 Rome, Italy, 2Bioinformatics Group, I.R.B.M. P. Angeletti, MRL-Rome, Via Pontina Km, 30600 Pomezia, Italy, 3Center for Comparative Functional Genomics, Department of Biology, New York University, NY 10003, USA and 4Systems Biology Group - Max-Delbrück-Centrum für Molekulare Medizin, Berlin

Received December 21, 2006; Revised March 15, 2007; Accepted March 28, 2007

ABSTRACT
3dLOGO is a web server for the identification and analysis of conserved protein 3D substructures. Given a set of residues in a PDB (Protein Data Bank) chain, the server detects the matching substructure(s) in a set of user-provided protein structures, generates a multiple structure alignment centered on the input substructures and highlights other residues whose structural conservation becomes evident after the defined superposition. Conserved residues are proposed to the user for highlighting functional areas, deriving refined structural motifs or building sequence patterns. Residue structural conservation can be visualized through an expressly designed Java application, 3dProLogo, which is a 3D implementation of a sequence logo. The 3dLOGO server, with related documentation, is available at http://3dlogo.uniroma2.it/

INTRODUCTION
The superimposition of a set of related protein structures is a straightforward methodology for identifying structurally conserved residues that cannot be detected using sequence alignments (1,2). Structurally conserved regions are usually of functional significance and can also improve our understanding of protein evolution. Despite the fact that a variety of multiple structure alignment algorithms have been reported over the years [(3–9); for a comprehensive review see (10)], methods for automatic identification of variable and conserved regions within a multiple structure alignment of proteins have not been fully exploited. The approach of identifying structurally conserved residues in a set of protein structures has been already used (11) to improve the sensitivity and/or specificity of poorly performing PROSITE patterns (12).

METHODS
Overview
The 3dLOGO web server combines three procedures. First, a multiple alignment of protein structures is built on the substructures common to a set of
user-selected structures. Then the alignment is analyzed to identify conserved residues; the structural alignment can be further improved by iterating the whole procedure and using the newly identified conserved residues as input. Finally, a sequence consensus can be built with the spatially conserved residues that are co-linear in the corresponding protein sequences.

A complete description of the method and usage technical details can be found on the documentation web pages.

The input data

Users must supply \( n \) \((n \geq 2)\) protein chains, which can be specified using PDB codes (13) or whose coordinates can be uploaded. For reasons of speed, the maximum number of input structures is 25. For each given structure, a single chain ID must be provided. The user must also supply the number \( \text{ID} \) of at least three amino acids of at least one input protein chain. This minimal substructure (the ‘object’ substructure) cannot comprise more than ten residues and is needed in order to localize the protein area to be studied. For instance, the minimal substructure can be a region known to bind a ligand (e.g. the solvent exposed residues of an SH3 domain or a set of residues binding ATP or GTP), to have a specific function (e.g. the catalytic triad of a serine protease), to match a known functional motif (e.g. PROSITE) or it can be any other protein area of interest. The server uses the Query3d application (14) for the detection of local similarities, in order to identify a structural match of the ‘object’ substructure (query) in the other \( n - 1 \) input structures. A match is valid if \( \text{a}\) it is made up of at least three residues, \( \text{b}\) pairs of matching residues are physico-chemically similar and display an averaged \( \text{rmsd} <2 \text{ Å} \). If no local similarities are found in one of the \( n - 1 \) input structures, this last is skipped and the subsequent steps are carried out only for the remaining \( n - 2 \) structures. The ‘residue similarity options’ in the input page, makes it possible to set the degree of physico-chemical similarity under which 3D conserved positions are identified, e.g. by setting ‘identical’, only positions made up of identical residues will be shown.

Multiple alignment of structures

All structures are superimposed using two points for each given residue, namely the C-alpha and a pseudo-atom, calculated as the average of the coordinates of the residue side-chain atoms. In the case of \( n \) input structures, the first structure is assumed to be the reference (or ‘target’) structure for the structural superimposition(s). The remaining \( n - 1 \) structures are aligned pairwise to the reference structure.

The output data: identification of conserved residues

After superimposing the input structures on the given residues, all the nearby conserved residues are identified. Two amino acids belonging to different structures are assumed to be conserved if they are spatially well superimposed and display physico-chemical similarities, according to a substitution matrix [see (15) and the documentation web page for more details]. The user can make a selection from the input page if only identical, very similar, similar or all the spatially well-superimposed residues have to be displayed. Identification of the conserved residues is carried out using the 3D profiles method (1). The multi-alignment output (Figure 1) is displayed in a table containing a column for each input structure and a row for each conserved position. User-supplied residues are highlighted by green dots, whereas yellow marks pinpoint newly identified well-superimposed residues.

In the score column, the similarity measure derived from the substitution matrix is shown. Its value is maximum if the row’s residue types are identical. The ‘\( \text{rmsd} \)’ column reports the average \( \text{rmsd} \) between the residues belonging to a row of the table with respect to the reference structure. Only positions whose residues display an average \( \text{rmsd} <2.5 \text{ Å} \) are shown in the table.

The structural alignment can be visualized with the Astex Viewer\textsuperscript{TM} applet (16) or downloaded in text format (Figure 1). Furthermore, the user can visualize the variable or conserved positions of the structural multiple alignment through a specifically designed Java application, 3dProLogo, which is a 3D implementation of a sequence logo (17).

The 3dProLogo Java application

A sequence logo is a graphical representation of an amino acid or nucleic acid multiple sequence alignment (MSA) which provides an instant way of visualizing variable and conserved positions of the MSA (17). In the sequence logo, the logo is constructed by calculating the information content of each position of the aligned sequences, and then displaying the characters representing the amino acids stacked on top of each other. The height of each letter is made proportional to its frequency. The height of each stack is then adjusted according to the information content at that position, as detailed in (17). Similarly, 3dProLogo is a 3D graphical representation of the residues in a multiple structural alignment. The total information for each position in the structural alignment is calculated as in (17) and used to derive the height of each residue type’s letter present in the 3D position. The resulting representation (3D logo) can be seen in Figure 1, where the conserved positions in P-loop containing proteins 3D multiple alignment (see the last section) are displayed. The strong conservation of the residues involved in the nucleotide binding can be detected immediately as can be their arrangement in space. 3D logos can be rotated according to the users need.

Sequence consensus building

In order to build a sequence consensus, the user is allowed to manually select the positions in the MSTA that better fit with his/her requirements. When the user is satisfied with the proposed structural alignment, a sequence consensus (pattern) can be generated from the matched residues. By using all the identified conserved residues, the sequence pattern is constructed as a regular expression for searching sequence databases. The format of the regular expression
Iterated alignments

The search for structurally conserved residues can be repeated by superimposing the structures on a new choice of conserved residues, including the newly detected ones. The whole process can be iterated until the identification of an ‘extended’ structure or sequence ‘consensus’. The user can obtain a new alignment by starting from a manual selection of the conserved positions listed in the structural alignment output table and using these as a new input of the procedure.

ONE EXAMPLE: P-LOOP CONTAINING PROTEINS

The P-loop is a phosphate-binding loop commonly found in ATP and GTP-binding proteins (2,19). The ‘consensus’ sequence of this loop is \[AG\]XXXXGK\[TS\] (20).

A typical example of a P-loop-containing GTP-binding protein is the small GTPase c-H-ras p21, belonging to the Ras family (SwissProt: P01112). Guanylate kinase (SwissProt: Q8UGD7) is an ATP-binding protein belonging to the guanylate kinases family. The two proteins display 11% sequence identity, and the SSM (21) structure alignment of the two proteins finds only a few highly similar matching residues (concentrated around the P-loop), and an overall lack of structural similarity. The P-loop in the c-H-ras p21 structure (PDB 5P21) matches residues 10–18, whereas in the guanylate kinase structure (PDB: 1gky) the P-loop is found at residues 8–16 (2). We considered 5p21 and 1gky plus a set of four different P-loop-containing proteins (PDB IDs: 1grn, 1gua, 1tad, 1ega) extracted from (2) to perform a 3dLOGO query. By using these six structures as 3dLOGO input, all the residues belonging to the P-loop were found, as well as several new well-superimposed residues which are close to the P-loop in space, but very distant along the sequence (Figure 1). In particular we retrieved the ras N116 (correctly superimposed to the kinase N168), which belongs to the well-known G protein conserved motif NKXD, whose residues are involved in the interaction with the nucleotide.

More examples are described in detail and graphically displayed in the 3dLOGO examples Web page (http://3dlogo.uniroma2.it/3dLOGO/examples.html).
CONCLUSIONS

The 3dLOGO web server is designed for the local study of protein structures. It represents a new useful tool for the analysis of conserved structural regions and for the derivation of more specific structure and sequence patterns. The 3dLOGO tool was also implemented to produce a novel 3D representation of the conservation of residues in a set of protein structures. Representative examples are reported in the Results and in the 3dLOGO Web pages.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of Telethon (GGP04273), a PNR 2001-2003 (FIRB art.8) and a PNR 2003-2007 (FIRB art.8). We thank Memmo Federici for his essential contribution to the 3dProLogo applet. Funding to pay the Open access publication charges for this article was provided by AIRC.

Conflict of interest statement. None declared

REFERENCES