STING Millennium: a web-based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence


Nucl. Sin., Brasília, Brazil, 1Laboratório de Bioinformática, Embrapa/Recursos Genéticos e Biotecnologia and 2Laboratório de Bioinformática, Universidade Católica de Brasília, Brazil

Received February 14, 2003; Revised and Accepted April 2, 2003

ABSTRACT

STING Millennium Suite (SMS) is a new web-based suite of programs and databases providing visualization and a complex analysis of molecular sequence and structure for the data deposited at the Protein Data Bank (PDB). SMS operates with a collection of both publicly available data (PDB, HSSP, Prosite) and its own data (contacts, interface contacts, surface accessibility). Biologists find SMS useful because it provides a variety of algorithms and validated data, wrapped-up in a user friendly web interface. Using SMS it is now possible to analyze sequence to structure relationships, the quality of the structure, nature and volume of atomic contacts of intra and inter chain type, relative conservation of amino acids at the specific sequence position based on multiple sequence alignment, indications of folding essential residue (FER) based on the relationship of the residue conservation to the intra-chain contacts and Ca–Ca and Cβ–Cβ distance geometry. Specific emphasis in SMS is given to interface forming residues (IFR)—amino acids that define the interactive portion of the protein surfaces. SMS may simultaneously display and analyze previously superimposed structures. PDB updates trigger SMS updates in a synchronized fashion. SMS is freely accessible for public data at http://www.cbi.cnptia.embrapa.br, http://mirrors.rcsb.org/SMS and http://trantor.biocolumbia.edu/SMS.

INTRODUCTION

SMS was designed in order to meet exciting new challenges originated specifically from the recent rise in biologists’ interest for describing and quantifying the relationship between sequence, structure and the function of proteins. This sudden rise in interest occurs in parallel to a disproportional growth between sequence and structure databases. Knowing that well organized information about macromolecular sequence, structure and function is an important platform for understanding processes in living organisms and diseases, as well as for designing new drugs, biologists search for appropriate tools to crosslink and analyze (often heterogeneous) data that have already been collected. Such tools should help elucidating patterns and underlying associations hidden in original raw data. To date, the structure of over 20 000 macromolecules and their complexes has been determined and deposited at the Protein Data Bank (PDB) (1). By calculating sequence–structure–function descriptors from experimental databases we may establish a subtle link between key sequence patterns and structure/function attributes. In addition, by identifying similarities in folds from alignments of protein chains we might be able to find a common base that defines function.

*To whom correspondence should be addressed. Tel: +55 19 3789 5774; Fax: +55 19 3789 5711; Email: neshich@cnptia.embrapa.br
Protein function determination and identification of possible function modifiers, is in fact the most desirable goal of many genome project efforts. To achieve such goal, bioinformatics counts with some basic activities, such as:

1. sequence homology search and position specific sequence conservation;
2. sequence alignments;
3. structural alignments;
4. search for sequence to structure relationship;
5. definition of structural parameters like intra and inter chain contacts;
6. structure modeling with evaluation of the quality of obtained models;
7. protein interfaces and identification of active and ligand binding sites;
8. cumulative statistics on protein family characteristics.

All these activities require an integrated environment with computational tools that can visually and schematically map 3D information together with parameters and the patterns extracted from a number of databases, on the protein sequence. In some other software packages, for multi-parameter presentation, the key role is trusted to simple structure display. Although the molecular structure display and manipulation is becoming an everyday task not only to structural biologists, but to an increasing audience of molecular biologists and biochemists as well, multi-parameter mapping on the structure itself is intrinsically limited. This is because any point on the structure/surface might be painted (colored) only once, following the color code designed for given amino acid characteristics. Amino acid sequence on the other hand, can be mapped by number of parameters, organized and displayed in underlying rows. Appropriate visualization of numerical value for any given parameter and for any amino acid in the sequence, might be easily achieved by color coding. This is the crucial difference between existing software packages (2,3) on one side and our SMS component—Protein Dossier.

The constant increase in accumulation of sequence data prompts for a more complete functional characterization of the gene product. This necessarily leads to insights into the protein’s three-dimensional (3D) structure. Given this scenario, it is of utmost importance that the programs designed to provide most complete and versatile information through user friendly interface, are as intuitive as possible. Such programs ought to save the end-user from dealing with a complicated sequence of programming commands typically required in order to achieve complex data cross-referencing. This task is only possible at the expense of an elaborate user interface. STING Millennium Suite is the mature product, the results of our attempts to meet such challenging requirements.

### STING Millennium Suite INTRINSICS

STING is an acronym of Sequence To and withIN Graphics. The original STING program was developed to enable bi-directional coupling between sequence and structural information and also to offer a simple and easy way to map a single amino acid (or nucleotide) letter code to its position in 3D, and vice versa. This work was originally published elsewhere (4).

STING Millennium was developed from the original STING concept. However, we have completely rewritten all code and made the STING concept much more versatile.

SMS version 2.2 is now available from the network of mirrors: http://www.cbi.cnptia.embrapa.br, at EMBRAPA/CNPTIA, Campinas-SP, Brazil, http://mirrors.rcsb.org/SMS, at PDB and http://trantor.bioc.columbia.edu/SMS, at Columbia University in Barry Honig’s laboratory. Other SMS mirror sites are anticipated and some are in the final phase of implementation. SMS entry page clearly indicates current SMS version, SMS host server (city and country) and three access options: (i) SMS web page illustrated with art images and which is best seen if using a fast internet connection and high resolution monitors; (ii) SMS access web page similar to the one described above, but with alternative molecular image illustrations; and (iii) SMS simple web page designed for access by users having low resolution monitors and slow internet connections. SMS web page menu options have been designed and written in such a way so that a user may quickly find critical information about each SMS component and possible interpretation of presented data.

In terms of the system design the package is implemented using a variety of programming languages, each one better suited to a specific task: (i) the Java™ language is used for the SMS Sequence window, which allows a complete control of the Chime plugin as well as initiation of any of the SMS components; (ii) the JavaScript is used to make the communication between Java Sequence window and Chime-plugin; (iii) the Perl language is used as a wrapper for processing web requests through CGI; (iv) the C++ language is used for more intensive programming tasks in structure parameter calculations; and (v) Chime molecular visualization plugin (MDL inc.) is used to provide the molecular rendering.

### SMS ORGANIZATION

SMS is organized in two logical layers (Fig. 1): SMS server and SMS client. The server side is responsible for updating regularly all relevant public domain databases used by SMS: PDB, HSSP (5–7) and PROSITE (8). At the same time, the SMS server is also responsible for calculation of a number of macromolecular properties for each PDB structure: electrostatic potential is calculated using GRASP (9) software, curvature is calculated using SurfRace (10) software, solvent accessible area for each protein chain and for the whole molecular complex is calculated using Surf v (11) software adapted to our own requirements, secondary structure identification is calculated according to DSSP (12) and STRIDE (13), intra and inter chain amino acid contacts as well as protein/dna interaction are calculated by our own software: contacts, hydrophobicity is assigned according to Radzicka and Wolfenden (14), dihedral angles are calculated by our own ‘Ramachan’ program and PROSITE patterns are identified using Ps_Scan (15) software. SMS components written in italic are accessible in SMS 3.0, now available on our beta test site, and SMS components written in italic and underlined are to be available in SMS 3.1 version. Server side is also
Resulting information presentation at the client side is done by structural parameters mapped to each residue in the sequence. Data and a user will receive a cartoon containing several protein structure. SMS search engine will collect corresponding Protein Dossier among the available SMS components. The selected Protein Dossier is then passed to the user. For instance, if a user has input a query, the server will provide a response containing information about the selected PDB file. The SMS version 2.2 components are listed within black border boxes, SMS version 3.0 components (now available at beta test site) are within red border boxes, SMS version 3.1 components (now at alpha test site) are within blue border boxes, and finally SMS version 4.0 component (in final phase of development) is shown within yellow border box. SMS_MSA stands for the new SMS component: multiple sequence alignment. Those sequences are extracted both from HSSP and from our own SH2Qs (Sequences Homologue to the Query [Structure-having] Sequence). The latter one is designed so that we can benefit from the possibility of making a choice of the number of sequences that we want to consider in the relative entropy calculation. SMS_MSA will use standard MSA viewer package: SeaView (20). SMS_P3 stands for SMS phylogenetic tree component. Phylogenetic tree is generated starting from either HSSP or SH2Q alignments. TreeView (21) is used to present to a user calculated phylogenetic trees. Finally, GPC is a new SMS component and it stands for global protein cartoon. This component displays, on a global sequence scale, a summary of a number of parameters which are presented in detail both in Protein Dossier and Java Protein Dossier.

**STING Millennium MODES**

STING Millennium has several working modes permitting its use with: (i) a single PDB structure (either from PDB or a single modeled structure in PDB format); (ii) a structural alignment of several protein structures [structures previously aligned with CE (22) or PRISM (23) software]; (iii) a PDB structure along with a sequence alignment, where the first sequence corresponds to the one with the known 3D structure. This SMS mode offers a way to visualize sequence conservation/variation and correlate it with structural regions in the target protein. The structure can also have its residues colored by an index that indicates sequence conservation (sequence relative entropy) promoting a rapid identification of the sites subject to differential (selective) evolutionary pressure.

**STING Millennium MODULES**

Several analytical programs developed by our group are listed as SMS modules and are available under the modules menu in the SMS sequence frame (Fig. 2A). These SMS modules (components) perform a number of tasks providing a powerful combination of tools to analyze a given PDB structure. These modules (some are to be described in separate publications) are:

1. Ramachandran plot [SMS_RP] (24): a fully interactive java plot displaying the main-chain dihedral angles ($\phi$ and $\psi$). SMS_RP identifies the allowed/disallowed regions for the angles (Fig. 3F) and permits selection of residues that can be displayed in the SMS structure window.

2. PDB file TXT: lists the original PDB entry in text format.
3. Hornet: this SMS component does calculation and visualization of hydrogen bonds between two chains (protein and/or nucleic acid) with special attention given to water molecules (trapped between two chains) participating in H-bond formation.
4. Scorpion: this tool is designed to calculate the frequency of amino acid presence for a given (list of) PDB file(s) as well as frequency of amino acid neighbors in 3D.

5. Formiga: this SMS component calculates the frequency of amino acid neighbors in 3D, only from the interfacial areas between two or more protein chains, for a given (list of) PDB file(s) (Fig. 3G).

6. Graphical contacts and graphical IFR contacts: these two tools calculate atomic contacts between residues in a PDB entry. The complete profile of the interactions can be analyzed in the graphical contacts window and graphical IFR contacts window (Fig. 3D and E, respectively). Each of those windows displays the primary sequence of a chain and below it a histogram containing the number of contacts or IFR contacts made by the respective residue. Such contacts are divided into number of classes: hydrogen bonds (main-chain/main-chain, side-chain/main-chain, side-chain/side-chain), hydrogen bonds with intermediary water molecules, hydrophobic contacts, aromatic ring stacking contacts, electrostatic (attractive and repulsive) contacts and finally disulphide bridges. Each contact type is attributed with specific color both in histogram and in virtual contact lines presentation. Virtual contact lines connect residues of interest in a fan fashion. This presentation is very intuitive and the user can easily grasp contact parameters along the protein sequence. In the case of IFR contacts, IFRs are underlined with a red line for easy identification along the sequence. Both internal and IFR contacts are presented in tabular form as well (Fig. 3C).

7. StingPaint: visual aid to display multiple sequence alignments with the amino acids colored by physico-chemical properties. STINGpaint now supports the following sequence and Multiple Sequence Alignment (MSA) formats: (i) sequence of any PDB entry; (ii) sequence in FASTA format; (iii) MSA in CE and PRISM output format; (iv) MSA in ClustalW (25), PSI-BLAST (26) and GCG (27) output format.

8. PDB cartoon: image of the amino acid sequence along with the secondary structure elements rendered as cartoons.

9. PDB_Mining: this SMS component is designed to retrieve specific data from the PDB. This tool is an off-spin product, resulting from our compilation of scripts made to extract relevant data about PDB while constructing and specifically testing SMS components. Therefore, PDB_Mining is highly recommended for software developers while dealing with some of not best structured data from PDB. PDB_Mining provides following data: total residue occurrence in whole PDB, keyword search, PDB files sorted by file size, PDB files sorted by number of protein chains, PDB files sorted by X-ray resolution of 3D data, PDB files sorted by deposition date, PDB files classified with respect to number of DNA chains, protein families classification etc.

10. ConSSeq: displays the amino acid sequence of a given protein structure along with the consensus sequence (as found in HSSP), color coded with respect to the calculated degree of conservation. ConSSeq also displays color coded graphic bars of relative entropy as well as information about residues present in other homologous sequences, with their respective frequency. For fast visualization, this program also generates the sequence logo within very effective user interface (Fig. 3B).

11. Cα–Cα and Cβ–Cβ distance java plot: the Java Cα–Cα (Cβ–Cβ) distance plot is a diagram where the distances between the alpha (beta) carbon of one residue and all alpha (beta) carbon atoms of other residues, within a single chain of the PDB file, are represented by colored squares in a symmetrical plot (Fig. 3A). A user defined color code of those squares follows the scale of calculated distances among Cα–Cα or Cβ–Cβ atoms.

12. Protein Dossier (Fig. 2B): this SMS component provides a graphical summary of several important structural characteristics for a chosen protein. Protein Dossier displays the amino acid sequence cartoon, accompanied with two histograms representing the atomic contacts (as in the Graphical contacts module), as well as IFR residue identification and IFR contacts. Below the sequence cartoon there is a Prosite pattern information, followed by three different secondary structure indicators (PDB, DSSP, STRIDE). Another five parameters are displayed and annotated with color coded scales, representing for each amino acid the corresponding: temperature factor, sequence conservation in a multiple alignment (relative sequence entropy), solvent accessibility of the protein chain in isolation and in complex with other protein chain present in PDB file, and finally, dihedral angles. Protein Dossier presentation is designed so that it can justify the very name of this module. Protein Dossier is a tool that a user can employ to browse through valuable information about the protein structure. The collection of information presented (growing with each new SMS versions release) is allowing a user to make very knowledgeable decisions about the possible role of specific amino acids in defining the function of the protein. It also helps in deciphering what effect the specific mutation will have on the structure and function of the protein, specifically by observing changes in intra and interface contact signatures.

STING Millennium FEATURES

The main aspect of the STING Millennium is its ability to combine data delivery through the web with the structural analysis tools, in order to provide a self-contained instrument for macromolecular studies. More than a simple front-end to the Chime plugin, SMS offers analytical services which we will only briefly describe here, counting that users will refer to extensive on-line help for further details.

STING Millennium is composed of two main windows. The SMS sequence window (Fig. 2A) displays sequence and contains the general menus with the commands and the SMS structure window (Fig. 2C) that renders the macromolecular
3D structure. In general terms STING Millennium provides the following services:

1. Ability to easily select residues in the sequence, select elements of secondary structure, as well as offer a wide variety of methods for rendering and coloring a molecule (mostly available through the ACTION menu).
2. Defining 3D neighbors to arbitrary selected residues.
3. Building surfaces of whole molecule or just of the IFR part of it.
4. Interactive Ramachandran plot, permitting rapid identification of residues in the disallowed regions and display of selected residues in the structure window.
5. Calculation of residue frequency within selected chain or on interface, as well as frequency of those residues filtered through chosen contact parameters.
6. Hydrogen bond net calculation with special attention given to the participation of water molecules.
7. Contacts definition and calculation for the whole molecule and/or interfaces.
8. Convenient 2D graphical presentation of parameters extracted from the 3D structure.
9. Display of sequence neighbors and calculation of relative sequence conservation for the family of homologous proteins.

SMS provides a great number of links to other web databases and programs. Such extensive cross-referencing and linking allows a user to have a most complete and integrated environment for sequence/structure/function analyses.

**EXAMPLE OF STING Millennium APPLICATION**

In Figures 2 and 3 we show a collage of snapshots produced by different SMS modules during a session that analyzed alphachymotrypsin (E chain) in complex with turkey ovomucoid third domain (I chain) of structure available in the PDB file 1cho.pdb. Specific attention in this example is given to the charged His residue (residue number 57) in the E chain. Analyzing information collected by SMS components and displayed in Figures 2 and 3, a user can infer many valuable conclusions about how important this amino acid is for protein stability, for protein function and for binding to inhibitor/substrate:

1. **SMS Sequence and control window** (Fig. 2A): sequence window displays linear protein and/or DNA sequence. Amino acids are color coded according to hydrophobicity and charged groups. The SMS sequence window also shows the numbering of the residues in the sequence, gaps in the PDB sequence, a chain identifier and ranges for all secondary structure elements as annotated in the original PDB file. Each residue in the sequence window is 'clickable', resulting in a presentation of its position in the SMS structure window (Fig. 2C). There are seven different display options in the menu bar situated just left from the Refresh button on SMS control panel: **Wireframe, WS = Wireframe + spacefill, CPK, Ribbon, Backbone, Strands** and **Cartoon**. Blue and red lines below the sequence are also 'clickable' resulting in a graphical ribbon presentation of the specified sequence region (red lines indicate helical region and blue lines indicate extended sheets region). Circled in magenta in Figure 2A is His_57 of the E chain.

2. Cβ–Cβ distance plot shown in Figure 3A, displays I chain data. In this plot a user is able to immediately identify glycine residues as they do not have Cβ atoms and their presence is evident from the existence of a light gray color vertical and horizontal lines (missing distance values due to non-existing Cβ atoms).

3. Figure 3B shows SMS ConSSeq output centered at His_57. A user may confirm both from ‘logo’ and histogram presentation, that the conservation of His_57 residue in the protein sequence is very high (even larger than for the neighboring Cys_58, which makes disulphide bond with Cys_42).

4. In Figure 3C and D, we show tabular information on amino acid IFR contacts and fan-like virtual internal-contact lines coming out from residue His_57 of chain E, respectively. Those virtual contact lines point to an amino acid that makes a particular contact (identified by the color of the line connecting it) with His_57. Information about the distance, atom partners and type of contact is provided in text format and if a particular residue is selected with the mouse click, zoomed image is displayed in the SMS structure window, showing the contacting 3D environment in details. The histogram along sequence of the chain E aids in rapid localization of critical residues, defined as those having larger than average number of contacts. Also, by color differentiation, a user may spot those amino acids having more energy-valuable contacts (e.g. electrostatic interactions).

5. Interface Forming Residues (IFR) Graphical Contacts are shown in Figure 3E. This inset demonstrates the same sequence of the E chain, but now a user can see IFR as red underlined sequence regions. Crucial difference from Figure 3D is that here contacts are counted only between residues belonging to different chains. A user can easily spot that chosen residue His_57 of chain E, make a couple of different contacts with the I chain residues E and T, and within its own E chain, with residues D, S and A. All details about IFR contacts are given in the table shown in Figure 3C, and in graphical form in Figure 2C.

6. Ramachandran plot is displaying the main-chain dihedral angles (φ and ψ), in Figure 3F. Green arrow points to the values of His_57, a position indicative of alpha helical secondary structure.

7. **SMS Formiga provided graphs** in Figure 3G. This graphical presentation shows which amino acids are found within 7 Å radius sphere, centered at the Last Heavy Atom (LHA) of the histidine residues. LHAs are defined as non-hydrogen side chain atoms which are most distant from the Cz atom. This information may also be transformed in cumulative frequencies for hydrophobic, polar and charged groups (one click away from demonstrated graph). Differences in frequencies are visually much easier to analyze than otherwise. Clearly, this information is valuable to define local environment for any amino acid and within any type of secondary structure. Variation in the size of the probing sphere allows for fine calibration of the potential function describing local residue environments.

8. Finally, Figure 2B shows a summary of the data calculated by a number of SMS programs and collected into SMS_DB. Now a user may analyze all pertinent parameters within a single visual area in Protein Dossier generated image.
This is a very powerful tool which enables a quick estimate of the level of engagement for each amino acid within its own protein chain and functionally more importantly, in the mechanism of binding to substrate and/or inhibitor. Consequently, a molecular biologist is able now to make intelligent predictions about effects of mutations (previously judged as necessary) on both protein structure and protein function.

CONCLUSIONS

STING Millennium Suite is described here in terms of a solution that brings together a number of protein analysis tools at a single web server. Furthermore, a user is now able to download and install a whole package on its own platform and have network independent version running. We hope that for many laboratories, STING Millennium Suite will be able to decipher some of the hidden 3D parameters and aid in research. For some others, we hope that STING Millennium Suite will show itself as a valuable didactic tool. In number of research. For some others, we hope that STING Millennium Suite is described here in terms of a continuous interest in this project.

ACKNOWLEDGEMENTS

This work was supported in part by the following grants: FAPESP 01/08895-0, FINEP 1945/01 and CNPq 521093/2001-5 (NV) and by NSF grant #DBI-9904841 to Barry Honig. The authors wish to express special thanks to Phil Bourne and Wolfgang Bluhm for their collaboration in testing and maintaining SMS at the PDB site. Special thanks are expressed to Barry Honig for his valuable insights and continuous interest in this project.

REFERENCES