iMolTalk: an interactive, internet-based protein structure analysis server

Alexander V. Diemand* and Holger Scheib[1]

University of Lausanne and Swiss Institute of Bioinformatics, Chemin des Boveresses 155, 1066 Epalinges s/Lausanne, Switzerland and [1]University of Geneva and Swiss Institute of Bioinformatics, Centre Médicale Universitaire, 1 rue Michel-Servet, 1211 Geneva 4, Switzerland

ABSTRACT

iMolTalk (http://i.moltalk.org) is a new and interactive web server for protein structure analysis. It addresses the need to identify and highlight biochemically important regions in protein structures. As input, the server requires only the four-digit Protein Data Bank (PDB) identifier, of an experimentally determined structure or a structure file in PDB format stemming e.g. from comparative modelling. iMolTalk offers a wide range of implemented tools (i) to extract general information from PDB files, such as generic header information or the sequence derived from three-dimensional co-ordinates; (ii) to map corresponding residues from sequence to structure; (iii) to search for contacts of residues (amino or nucleic acids) or heterogeneous groups to the protein, present cofactors and substrates; and (iv) to identify protein–protein interfaces between chains in a structure. The server provides results as user-friendly two-dimensional graphical representations and in textual format, ideal for further processing. At any time during the analysis, the user can choose, for the following step, from the set of implemented tools or submit his/her own script to the server to extend the functionality of iMolTalk.

INTRODUCTION

Today, numerous complete genomes are in our hands, ready to be deciphered. Sequencing genomes has become a standard protocol, producing an ever-larger amount of nucleic acids and, eventually, protein sequences. In contrast to the many protein sequences available today, structure elucidation still happens at a slower pace. Yet, structural information deposited in the publicly accessible Protein Data Bank (PDB) (1) also increases at a high rate thanks to the recent structural genomics initiatives (2) (for details see http://targetdb.rutgers.edu) and significant improvements in structure determination methods in general. To further narrow the gap between known sequences and structures, three-dimensional model structures can be predicted by virtue of evolutionary homology (3,4). Therefore, molecular biologists interested in protein function and structure can benefit greatly from the growing amount of structural information. Typical questions concern the relationship between sequence and co-ordinates, or the spatial organization of residues in active sites, as well as their interactions with bound ligands, inhibitors, cofactors and metal ions. Or else they are concerned with identifying the residues that are located at the interface between chains in a structure. However, all too often, protein structure analysis remains an expert task. One way to analyse 3D structures today is by using molecular modelling and visualization tools, e.g. MolMol (5), Rasmol (6), VMD (7) and SPDBV (8). But, to use them, data have to be stored locally and software installed, an increasingly embarrassing task on local area networks. In many cases, the quality of structural analyses can benefit greatly from installing additional hardware, e.g. shutter glasses for three-dimensional representation. In addition to these system prerequisites, many of these tools are not intuitive at first for non-expert users, who require time and effort to become adequately trained.

iMolTalk takes a different approach. It does not require local installation either of soft- or hardware or of data. All computations are carried out on the server and results are presented in the browser of the user. The methods available can be applied to structure, chain or residue. They are organized in so-called toolchains, which represent a logical sequence of steps to gather the necessary input for particular algorithms. From each toolchain, any other toolchain can be accessed with the current result as input for further analyses on the level of structure, chain and residue. Furthermore, the functionality of the server is not limited to the implemented methods but can be extended by users providing their own structural data or tailor-made scripts, which are then executed on the server.

*To whom correspondence should be addressed. Tel: +41 0 79 213 0571; Fax: +41 0 21 692 5945; Email: alexander.diemand@isb-sib.ch

Table 1. The implemented toolchains available on iMolTalk with required input, computed output and potential usage

Toolchain	Input	Output	Usage
PDB file information	PDB or model code	Detailed information about a structure and its chains	Header information of the structure file, co-ordinate-derived sequence from structure file
Secondary structure assignment	(i) PDB or model code (ii) Chain identifier	Secondary structure is assigned applying STRIDE (11) and mapped to each residue in the sequence	Secondary structure states of residues (α-helix, β-strand, turn)
Sequence to structure alignment	(i) PDB or model code (ii) Chain identifier (iii) Sequence or Swiss-Prot AC/ID identifier	Pairwise, global alignment of the co-ordinate-derived sequence of the structure to the protein sequence	Correspondence between residues in structure and sequence
Ramachandran plot	(i) PDB or model code (ii) Chain identifier	A Ramachandran-plot (12) is computed from the dihedral angles phi and psi of the protein backbone atoms. The graph is divided into core, allowed, generous and disallowed regions (13)	Quality assessment of experimental or model structure
Distance matrix	(i) PDB or model code (ii) Chain identifier (iii) Distance threshold	Pairwise internal distances between Cα atoms are plotted in a 2D representation (14). The background is colour-shaded according to secondary structure	Interactions of secondary structure elements in a protein structure
Residue contacts	(i) PDB or model code (ii) Chain identifier (iii) Identifier of residue (amino or nucleic acid) or heterogeneous group (iv) Distance threshold	For a standard residue or heterogeneous group, contacts to other residues/groups are searched. For each pairwise contact, the bond type is inferred (H-bond or salt-bridge) (15)	To determine amino acids, nucleic acids or heterogeneous groups in spatial proximity to a residue or heterogeneous group of interest
Interface finder	(i) PDB or model code (ii) Two chain identifiers (iii) Distance threshold	For any two chains in a protein structure, the protein–protein interface is revealed. For each pairwise contact at the interface, the bond type is inferred (H-bond or salt-bridge) (15)	To identify amino acids, nucleic acids or heterogeneous groups at the interface between two chains in the same structure file
Scripting editor	MolTalk Script	Results are reported in the browser	To program and execute individual scripts on the server

In an example analysis, we used iMolTalk to relate the annotation of a protein at the residue level from sequence to structure. For the covalently bound cofactor, we determined its contacts to the protein environment, in both the apo and holo forms, all within minutes of connecting to the server.

IMPLEMENTATION

iMolTalk was implemented as a web server using CGI (common gateway interface) for communication with standard Internet browsers. Computations are server-centric, i.e. all data and programs are available on the server and do not require local installation by the user. Nevertheless, users are able to upload their own structure files in PDB format and crafted scripts to extend the functionality of the server. A weekly mirror of the PDB was installed on the server to guarantee the availability of the most recently released structures to iMolTalk.

The underlying software was implemented in Objective-C using the MolTalk library libmoltalk (http://www.moltalk.org, submitted) or was directly programmed in the MolTalk scripting language. MolTalk is a computational environment which maps PDB files to an object-oriented representation.

iMolTalk provides computation on three types of structural objects: structure, chain and residue. A structure is the representation of a PDB file containing single or multiple chains, which themselves hold a list of residues. A residue is either an amino acid or a nucleic acid or can be a heterogeneous group of atoms. The scripting language included (related to the programming language Smalltalk) is inherently object-oriented and allows access to all objects (structure, chain and residue) and implemented algorithms (e.g. structure superposition and geometric hashing of residue co-ordinates).

TOOLCHAINS

The services provided by the iMolTalk server are organized into predefined logical sequences of mandatory user input termed toolchains (Table 1, Figure 1). During the last step of a toolchain, a result is computed and reported back to the user. Objects of type structure, chain and residue in the report are turned into active links. These links lead to characteristic pop-up menus, which, for each object, provide direct access to the results of other toolchains (Table 2, Figure 2B). Within a toolchain, one can go back and forth to change input parameters and to re-compute results (Figure 1).

EXAMPLE ANALYSIS

Some possible ways to analyse protein structures with iMolTalk are presented for the analysis of the mitochondrial aspartate aminotransferase (Swiss-Prot identifier AATM_ CHICK) and its corresponding structures in open (PDB code 7AAT) and closed (PDB code 1AMA) form (9,10). The family of aspartate aminotransferases exists as two
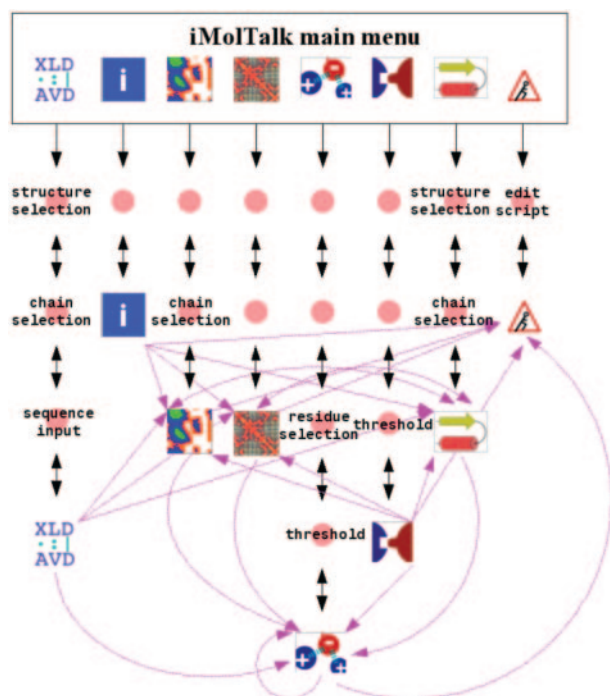
isozymes: one located in the cytosol and the other in the mitochondria. The enzyme catalyses the reversible transfer of an amino group with the help of PLP (pyridoxal-5′-phosphate or vitamin B_6) as a cofactor. The homo-dimer of two subunits forms the active enzyme with two independent active sites.

First, the correspondence of the residues in the protein sequence to the residues present in the structure was established (Figure 2B). Often, residues in a structure cannot be identified by their number in the sequence, and vice versa, because the numbering schemes differ. A pairwise global alignment of the two sequences can reveal such a correspondence, assuming a reasonably high homology between the two sequences. For the mitochondrial aspartate aminotransferases, the alignment showed that the structure lacks the N-terminal target sequence. The shift in numbering could be detected easily. As an example, K272, which in Swiss-Prot was annotated to bind 'pyridoxal phosphate', corresponds to K258 in the open-form structure. In the alignment, the sequence of the structure is coloured according to the secondary structure assignment based on STRIDE (11). Each residue in the structure sequence represents an active link to a pop-up menu. As shown for K258, the pop-up menu allows direct access to the results of the toolchains 'Residue contacts' and 'Scripting editor' for this residue.

Second, the report of the toolchain 'Residue contacts' (Figure 2C; detail in Figure 3A) showed that the cofactor (PLP258) is covalently bound to the terminal ammonium group (atom 'NZ') of K258. Moreover, a specific H-bond to Y70 of the other subunit (chain 'B') in the dimer was highlighted. This contact of the active site in chain 'A' to Y70 of chain 'B' might be an important functional feature of the dimer. In the closed form, the cofactor is covalently bound to the substrate in exchange for K258. The report of the toolchain 'Residue contacts' for PLP258 in the structure of the protein in open conformation revealed that the terminal ammonium group of K258 now formed a hydrogen bond to the phosphate group of the cofactor–substrate complex (Figure 3B).

The Cα-distance map highlights contacts between secondary structure elements in a structure. A parallel β-sheet shows up as a diagonal line parallel to the main diagonal of the graph; an

Figure 1. Schematic representation of the toolchains available in iMolTalk. Navigation is possible either within the predefined sequence of a toolchain (vertical arrows) or between different toolchains (broken lines). The icons represent, from left to right, the toolchains 'Sequence to structure alignment', 'PDB file information', 'Ramachandran plot', 'Distance matrix', 'Residue contacts', 'Interface finder', 'Secondary structure assignment', and 'Scripting editor'.

Table 2. The context and navigation options of the characteristic pop-up menus for the objects structure, chain and residue

Object	Name	Toolchain	Action
Structure	PDB File Information	PDB file information	A modifiable script in MolTalk extracts information and presents the results in the browser window
	Run Script	Scripting editor	Runs a modifiable script directly on this structure
	Download PDB File		Downloads and opens structure with MIME-type application/x-pdb
Chain	Secondary Structure	Secondary structure assignment	Assigns secondary structure on this chain
	Cα-distance map	Distance matrix	Computes graphical representation on this chain
	Ramachandran plot	Ramachandran plot	Computes graphical representation on this chain
	Run Script	Scripting editor	Runs a modifiable script directly on this chain
	Download PDB File		Downloads and opens this chain only with MIME-type application/x-pdb
Residue	Contacts	Residue contacts	Lists contacts to neighbouring residues or heterogeneous groups
	Run Script	Scripting editor	Runs a modifiable script on this residue

Figure 2. Representation of the iMolTalk entry page and the output of three toolchains. (**A**) The entry page provides access to each of the eight toolchains as well as to the Help pages, the Wiki discussion forum and the MolTalk project. (**B**) Alignment of the Swiss-Prot sequence AATM_CHICK and chain 'A' of structure 7AAT. The structure sequence is coloured according to its secondary structure and each residue is actively linked to a pop-up menu. Also, the characteristic pop-up menus for the structure, the chain and residue K258 are displayed. (**C**) Report of the toolchain 'Residue contacts', showing contacts for the heterogeneous group PLP258 in 7AAT. (**D**) In the background, the Cα-distance map is shown for PDB structure 7AAT (chain 'A', residues 50–270). The background of the map is coloured according to the secondary structure: red for α-helices, yellow for β-strands and cyan for turns. Rectangles indicate pairwise contacts between residues and are in yellow for contacts closer than half of a user-defined threshold (6 Å), and are in red for contacts within this threshold (12 Å). Active links in the reports are coloured in red.

Figure 3. Co-ordination of the terminal ammonium group of K258 in both the closed- and open-form structures of mitochondrial aspartate aminotransferase. Screenshots show excerpts of the report from the toolchain 'Residue contacts'. (**A**) In the open form (PDB code 7AAT, chain 'A'), the cofactor is covalently bound to the atom Nζ of K258 at a distance of 1.29 Å, forming a Schiff base. (**B**) In the closed form (PDB code 1AMA) with bound substrate, the same atom of K258 forms a hydrogen bond with the cofactor–substrate complex.

anti-parallel β-sheet is perpendicular to the main diagonal. Owing to their local contacts, α-helices appear thicker along the main diagonal. For chain 'A' of the PDB structure 7AAT, the typical pattern of helix–helix contacts and parallel β-sheets at the C-terminal end of the sequence can be displayed (Figure 2D). In the Cα-distance map provided by iMolTalk, rows and columns represent the residues in a protein and are coloured according to their secondary structure. Rows are active links to the residue-specific pop-up menu (Figure 2B, Table 2).

CONCLUSION

With iMolTalk we provide a web server for protein structure analysis, e.g. to map annotation from sequence to structure or to investigate atom contacts and structural interfaces in a highly interactive manner. Results are represented in a user-friendly format and can be readily used in further analyses. As input, the server requires only a PDB identifier or a file in PDB format. The functionality of the server can be extended by user-provided scripts.

ACKNOWLEDGEMENTS

REFERENCES

1. 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.
2. Stevens,R.C., Yokoyama,S. and Wilson,I.A. (2001) Global efforts in structural genomics. *Science*, **294**, 89–92.
3. Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.
4. Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.
5. Koradi,R., Billeter,M. and Wuthrich,K. (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.*, **14**, 51–55.
6. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.
7. Humphrey,W., Dalke,A. and Schulten,K. (1996) VMD: visual molecular dynamics. *J. Mol. Graph.*, **14**, 27–28.
8. Guex,N. and Peitsch,M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.
9. 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.
10. McPhalen,C.A., Vincent,M.G., Picot,D., Jansonius,J.N., Lesk,A.M. and Chothia,C. (1992) Domain closure in mitochondrial aspartate aminotransferase. *J. Mol. Biol.*, **227**, 197–213.
11. Frishman,D. and Argos,P. (1995) Knowledge-based protein secondary structure assignment. *Proteins*, **23**, 566–579.
12. Ramachandran,G.N., Ramakrishnan,C. and Sasisekharan,V. (1963) Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, **7**, 95–99.
13. Morris,A.L., MacArthur,M.W., Hutchinson,E.G. and Thornton,J.M. (1992) Stereochemical quality of protein structure coordinates. *Proteins*, **12**, 345–364.
14. Richardson,J.S. (1981) The anatomy and taxonomy of protein structure. *Adv. Protein Chem.*, **34**, 167–339.
15. Stickle,D.F., Presta,L.G., Dill,K.A. and Rose,G.D. (1992) Hydrogen bonding in globular proteins. *J. Mol. Biol.*, **226**, 1143–1159.