Integrating public databases into an existing protein visualization and modeling program – BRAGI

Guido Dieterich[1], Marsel Kvesic[2], Dietmar Schomburg[2], Dirk W. Heinz[1], Joachim Reichelt[1*]

[1] Division of Structural Biology, German Research Center for Biotechnology (GBF), Mascheroder Weg 1, D-38124 Braunschweig, Germany
[2] University of Cologne, Institute for Biotechnology, Zülpicher Straße 47, D-50674 Köln, Germany
Guido.Dieterich@gbf.de
marsel.kvesic@gmx.de
D.Schomburg@uni-koeln.de
Dirk.Heinz@gbf.de
Reichelt@gbf.de

Abstract: BRAGI offers an efficient visual access to sequence alignment information, 3D alignments and annotated protein structure-function correlations. As a new feature we have mapped information from SWISS-PROT and InterPro to individual entries of the PDB. 3D structural alignments from DALI database were converted to XML files for easy access in BRAGI. BRAGI provides interactive access to NCBI-Blast and the DALI server. Linking and visualizing different types of information hopefully allow the structure function of proteins to be appreciated more intuitively.

Availability: http://bragi.gbf.de
Cost: Freeware, but you have to sign a license

Introduction

The relationship between sequence, structure and function of biomacromolecules is of central interest in biology. Structural data on proteins and nucleic acids are stored in the Protein Data Bank [Be02]. Its archives contain atomic 3D coordinates, bibliographic citations, as well as information on the sequence and secondary structure. To better visualize these relationships, a large number of new features has been added to BRAGI [SR88], a package for viewing and modeling of proteins. It operates on hardware accelerated with OpenGL on ordinary Windows, Linux and SGI computers. The user interface provides an intuitive, standard "look and feel" of the computer used. Information from public databases such as SWISS-PROT [Bo03] or InterPro [Mu03] to PDB entries have been integrated, and can be displayed graphically while links to additional data are available via standard browsers.

Integrating SWISS-PROT

SWISS-PROT is a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases. PDB entries contain DBREF records that provide cross-references to various databases. Such references are, however, not systematically available for all PDB entries and direct chain cross-reference to a SWISS-PROT entry is frequently entirely absent. SWISS-PROT itself does provide such links to PDB entries but not to individual protein chains. Based on a list of these entries, we have extracted the protein sequences of ATOM records from the corresponding PDB entries using Perl with the Bioperl-toolkit [St02]. Using the bl2seq routine of the NCBI Blast program [TM99], the PDB entry chain with the highest score was assigned to the SWISS-PROT entry. The annotations of this SWISS-PROT entry are found in the feature table. We were able to match residues listed in this feature table to residues of the chain. Using the Bioperl toolkit, we extracted all functions ignoring items concerning the secondary structure (HELIX, TURN, STRAND key), the VARSPLIC and CONFLICT key.

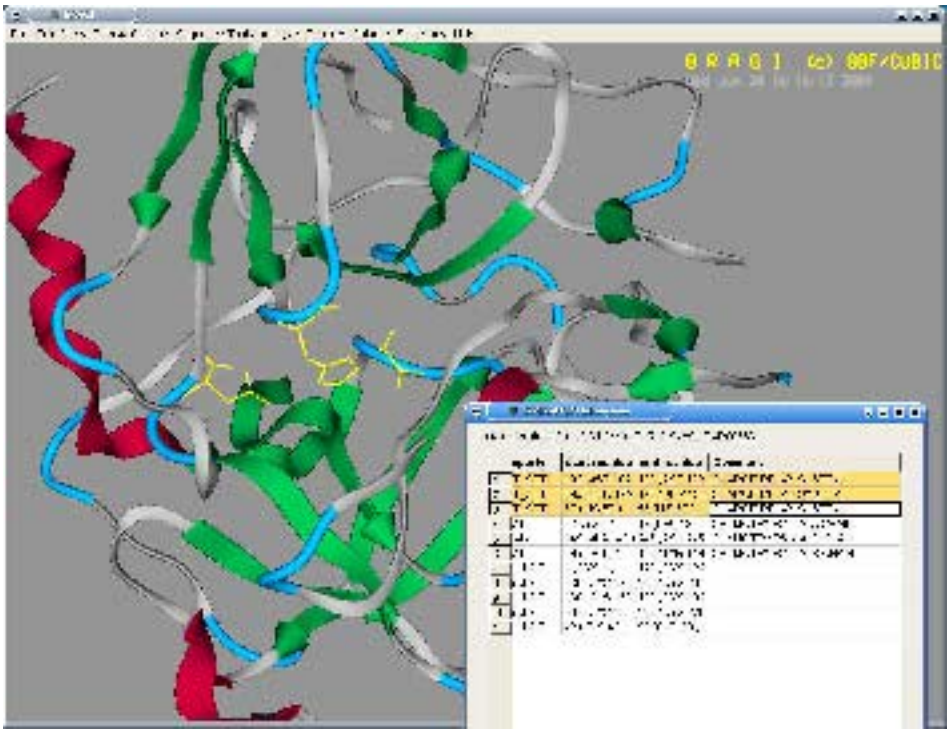

Figure 1: Screenshot of a session with BRAGI. A protein structure 'Alpha-Chymotrypsinogen (1CGJ)' is entered by the user and three functional properties are chosen from the 'SWISS-PROT Information'-display resulting in a region (in yellow colored ball and sticks representation) of the residues comprising this chain on the 3D structure (main window). These residues are building the 'charge relay system'. The functional annotations as extracted from SWISS-PROT are shown in the 'SWISS-PROT Information' display.

The amino acid sequence corresponding to each feature was then matched to the chain sequence through the use of regular expressions. In ambiguous cases, such as more than one match or a single residue being involved, the search pattern was extended by two residues at either end. Thereby we could match more features such as disulfide bonds to a PDB entry than the set resulting from ProSat [Ga03]. Information related to entries and found in data collections (e.g. Ensembl, EMBL, GenBank, PDB and many others) other than SWISS-PROT from the DR (database cross-reference) lines were extracted, thus identifying corresponding entries, for example in PDB.

Integrating InterPro

InterPro is a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences. The same list of PDB entries described above was compared to the InterPro database using a local version of InterProScan. The resulting domain descriptions for each PDB entry are stored in XML files.

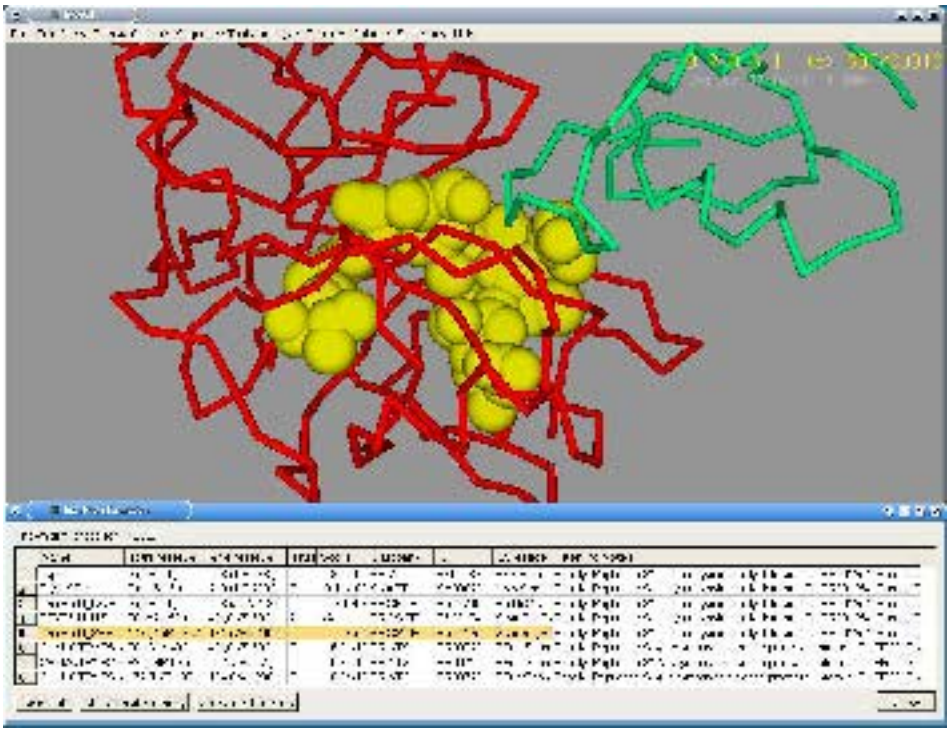

Figure 2: Screenshot of a session with BRAGI. A protein structure 'Trypsinogen Complex (1TGS)' is entered by the user and a functional domain (line 5,TRYPSIN_SER) is chosen from the 'InterPro Information'-display resulting in a region of the residues - highlighted in yellow CPK representation - comprising this domain on the 3D structure in the main window. This domain is the serine active site of serine proteases of the trypsin family. The domains as evaluated from the InterProScan tool are shown in the 'InterPro Information' display.

Integrating OMIM

To link PDB entries to information inherent in the OMIM catalog of of human genes and genetic disorders related to inherited diseases [Mc00] we used a database generated from OCA [4] - a browser-database for structure/function relations - , listing all PDB entries relating to diseases. We combined the PDB ID, the corresponding SWISS-PROT IDs and OMIM IDs into one file. The mutation loci were referenced to the PDB chain sequence and deposited into a XML file.

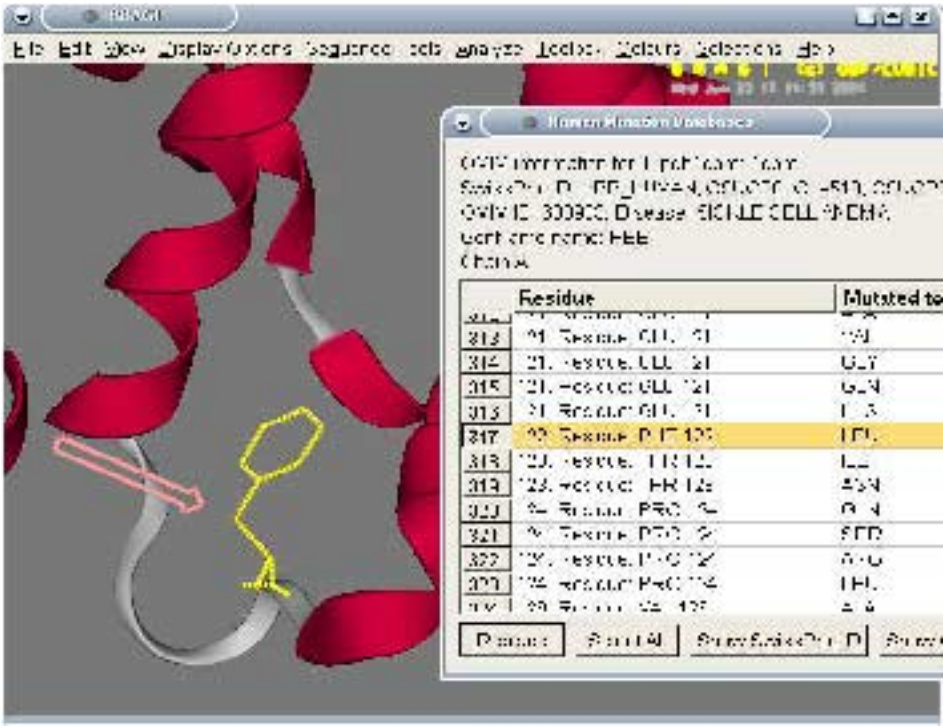

Figure 3: Screenshot of a session with BRAGI. The protein structure 'Hemoglobin (1CBM)' is entered by the user. In the 'Human Mutation Databases' window containing all known mutation of this structure from OMIM one mutation PHE122LEU (No. 317) - jointly responsible for sickle cell anemia – is selected. This PHE residue is highlighted in the 3D viewer and marked with an arrow.

Integration of DALI-3D structure comparison

A comparison of protein structures in 3D using DALI [HH00] provides a multiple alignment of structural homologues of a query structure. The dali_dccp file [Da00] lists a one-line summary per matched structure. These data and the aligned residues (dali_fragments file) were merged into one XML-file per query structure. BRAGI offers all alignments for seamless download and automatic 3D alignment without further manual intervention. By selecting parts of the sequence alignment BRAGI immediately highlights

this reference. To get a list of structural neighbours of a structure, BRAGI sends the co-ordinates to the DALI server in PDB format. BRAGI displays the multiple 3D alignment information of the reply email without previous manipulations in contrast to e.g. VMD [HDS02]

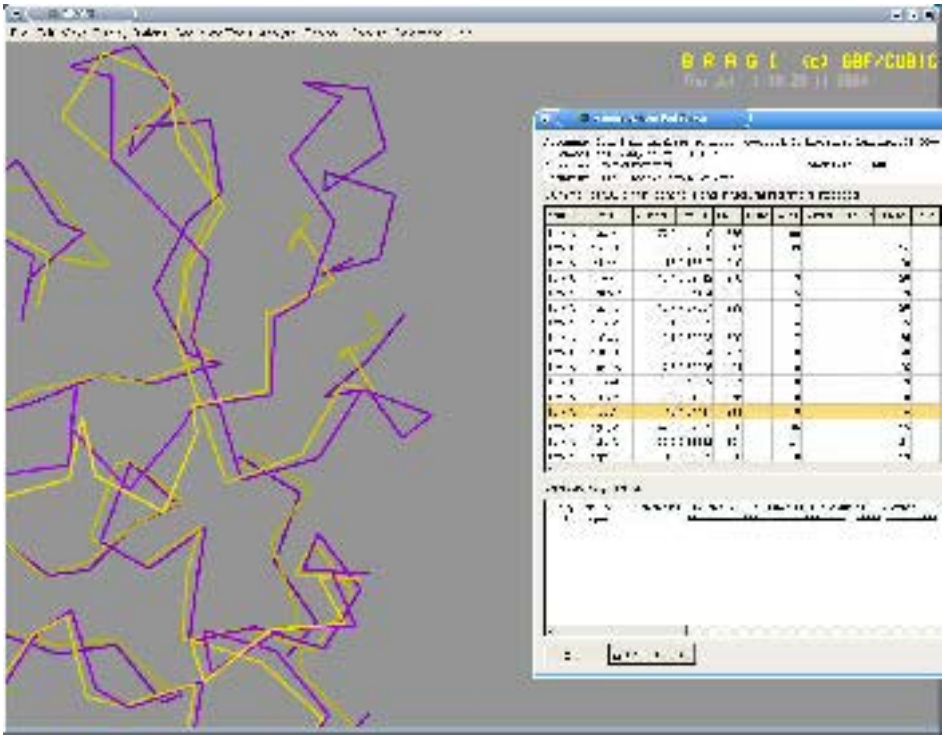

Figure 4: Screenshot of a session with BRAGI. The protein structure 'Oxidoreductase - Old Yellow Enzyme' (1BWL) is entered by the user. In the 'Alignment from Dali Server' window containing all 3D related structures found using DALI, the structure 1ICP - 12-Oxophytodienoate Reductase 1 from Tomato - chain A is selected. BRAGI has loaded this structure and shows the Calpha traces of both proteins (pink: 1BWL, yellow: 1ICP) aligned as given by DALI

Integrating NCBI Blast

BRAGI offers a Blast search of the sequence belonging to any structure via direct HTTP-encoded requests to the NCBI web server to allow sequence comparisons. By default BRAGI searches the amino acid sequences derived from the 3D structure records from the PDB databank. Using Blast BRAGI offers the possibility to use the build in modeling tools to predict a homologous structure.

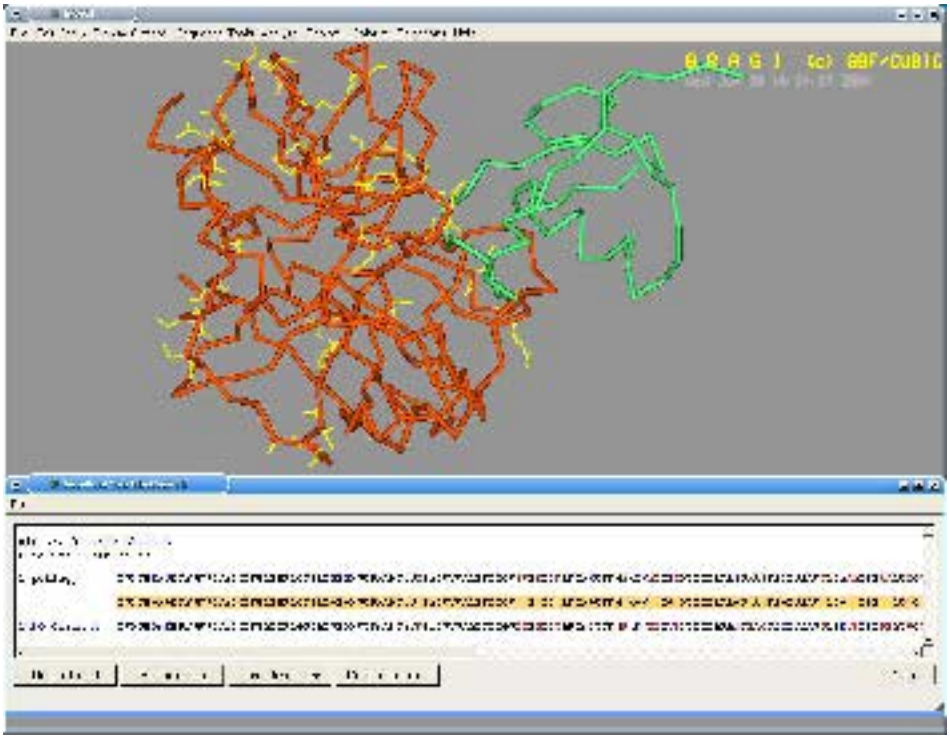

Figure 5: Screenshot of a session with BRAGI. A protein structure 'Alpha-Chymotrypsinogen (1CGJ)' is entered by the user. In the 'Results of Your Blast Search' window the consensus sequence of an aligned sequence from a different PDB entry is highlighted. In the 3D structure view all those residues are highlighted in yellow that are different in these proteins. This was done pressing the button 'Mark Differences'

Discussion and Conclusion

Publicly available powerful 3D viewers such as PyMOL [De02] or YASARA [KNV03] currently do not provide the visualization of information available from various public databases, a precondition to a better understanding of protein function. The combination of a proven modelling and visualization tool, as established in BRAGI, and the linkage of information from public databases harbors an enormous simplification for the analysis of proteins structures and rational protein design.

Acknowledgements

Part of this project was supported by grant from the Deutsche Forschungsgemeinschaft (DFG) to D. W. Heinz (He 1852/6-1).

Literature

[Be02] Berman,H. M., Battistuz,T., Bhat,T. N., Bluhm,W. F., Bourne,P. E., Burkhardt,K., Feng,Z., Gilliland,G. L., Iype,L., Jain,S., Fagan,P., Marvin,J., Padilla,D., Ravichandran,V., Schneider,B., Thanki,N., Weissig,H., Westbrook,J. D., Zardecki,C., 2002, The Protein Data Bank., Acta Crystallogr D Biol Crystallogr, 58, 899-907

[Bo03] 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., Schneider, M., 2003, The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. , Nucleic Acids Res , 31, 365-70

[Da00] //www.bioinfo.biocenter.helsinki.fi:8080/dali/download.

[De02] DeLano,W.L. The PyMOL Molecular Graphics System (2002) DeLano Scientific, San Carlos, CA, USA. http://www.pymol.org

[Ga03] Gabdoulline,R. R., Hoffmann,R., Leitner,F., Wade,R. C., 2003, ProSAT: functional annotation of protein 3D structures. , Bioinformatics, 19, 1723-5

[HDS96] Humphrey, W., Dalke, A. and Schulten, K., "VMD - Visual Molecular Dynamics", J. Molec. Graphics, 1996, vol. 14, pp. 33-38.; http://www.ks.uiuc.edu/Research/vmd/

[HH00] Heger,A., Holm,L., 2000, Towards a covering set of protein family profiles., Prog Biophys Mol Biol , 73, 321-37

[KNV03] Krieger,E. ,Nabuurs,S. B., Vriend,G., 2003, Homology modeling, *Methods Biochem Anal,* **44**, 509-23

[Mc00] McKusick-Nathans,V. A., Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Cente, Online Mendelian Inheritance in Man, O, 2000, http://www.ncbi.nlm.nih.gov/omim/

[Mu03] Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R. R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S. E., Pagni, M., Peyruc, D., Ponting, C. P., Selengut, J. D., Servant, F., Sigrist, C. J., Vaughan, R., Zdobnov, E. M., 2003, The InterPro Database, 2003 brings increased coverage and new features. , Nucleic Acids Res , 31, 315-8

[SR88] Schomburg, Dietmar, Reichelt, Joachim, 1988, BRAGI: A comprehensive protein modeling program system , J. Mol. Graphics , 6, 161-165

[St02] Stajich,J. E., Block,D., Boulez,K., Brenner,S. E., Chervitz,S. A., Dagdigian,C., Fuellen,G., Gilbert,J. G., Korf,I., Lapp,H., Lehvaslaiho,H., Matsalla,C., Mungall,C. J., Osborne,B. I., Pocock,M. R., Schattner,P., Senger,M., Stein,L. D., Stupka,E., Wilkinson,M. D., Birney,E., 2002, The Bioperl toolkit: Perl modules for the life sciences. , Genome Res , 12, 1611-8

[TM99] Tatusova,T. A., Madden,T. L., 1999, BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. , FEMS Microbiol Lett , 174, 247-50