# Visualization in Bioinformatics: Protein Structures with Physicochemical and Biological Annotations

Marco Wiltgen and Andreas Holzinger

Institute for Medical Informatics, Statistics, and Documentation, Medical University of Graz, Austria
(marco.wiltgen@meduni-graz.at, andreas.holzinger@meduni-graz.at)

**Abstract**

*In this paper we concentrate on the visualization of protein structures with annotations of non-structural properties. The visualization of protein structures is of importance for the understanding of biological processes and diseases at a molecular level. The appropriate representation of structural features reveals the nature of the protein function and behaviour. Important is the interactive visualization of non-structure-based functional annotations in protein 3D structures. We present a visualization method which includes information of three different databases: a structure database, a sequence pattern database and an electron density map database. The detected sequence pattern (a so called motif) is responsible for the biological function of the protein. Because the functional specificity of proteins is linked to the structure it is necessary to visualize the sequence pattern in its correct geometrical arrangement. The additional visualization of the electron densities, which are responsible for the chemical properties of residues, allows insights into the dynamics of the active site. Additionally, by use of interactive windows, further results (solvent accessibility, electric charges, geometrical distances between the residues in the motif etc.) from automated calculations can be annotated to the structure.*

Categories and Subject Descriptors (according to ACM CCS): J.3, I.3.3, H5.1 [Life and Medical Sciences, Computer Graphics, Visualization]: Visualizaion of protein structures

## 1. Introduction

Through history, scientists of all fields used visualizations to represent their data and information [GFG*94]. Visualization take on the complexity of biomedical computing, improving its utility to scientists and clinicians alike [JMPW04]. Earliest computer representations of macromolecular structures were made in the sixties [Lev66]. First at the MIT a system, based on an oscilloscope and displaying rotating wire frame representations of macromolecular structures, was developed [LB68]. At the same time, a program to produce stereoscopic drawings of molecular structures with a pen plotter was developed [Joh65]. In the mid 1970’s a protein structure was visualized entirely with computers for the first time [BRR77]. During the 1980’s computer systems showing wire frame renderings of the amino acid chain, which could be rotated in real times become very popular for crystallographers [RR94]. In the 1990’s first programs running on Macintosh and PC brought molecular visualization to a large number of scientists, educators and students [SMW95]. Nowadays a number of protein viewers are freely available from referenced Web sites and run on most computer platforms and operating systems including Microsoft Windows, Macintosh and UNIX X-Windows [MK04]. Such programs convert the atomic coordinates into a view of the molecule and allow manipulation of the molecule [Hog97], [GP97], and [Wal97]. Some viewers are used as Web browser plug-ins to display and manipulate structures inside a Web page.

Whereas the problems concerning the representation of proteins (for example: wire frame representation) and the virtual 3D effects (for example: the illusion of depth) were worked out in the 1980’s and 1990’s, the research activities now mainly deal with the mapping of physicochemical properties and biological annotations onto the protein structure [GLW03]. Web access to gene and protein databases with a lot of annotations has increased dramatically the availability of biological information [GOT04]. To investigate the structure and function of a protein, researchers create increasingly sophisticated methods for transforming genetic and functional data into comprehensive information at the protein level to better understand biological processes. Several research groups have developed software for visualization and analysis of protein structures, by taking into account different properties and aspects. Current research encompass: Visualization of protein structures with annotations [NRM04], considering sequence-structure relationships [Old04], macromolecular interfaces [FMB05], relationships between genes and protein structures [VTR03] and molecular surface analysis [GWW03].

## 2. Protein structure and function

Proteins are the molecules used by the cell for performing and controlling cellular processes, including: degradation and biosynthesis of molecules, physiological signalling, energy storage and conversion, formation of cellular structures etc.

Proteins are built up as polymers of the amino acids, whereas 20 different amino acids (residues) are involved as elements in protein sequences. The amino acids are building up a polypeptide backbone and the different residues differ by their side chains. Once a protein is synthesized in a cell it folds together to a well defined 3D structure. It can be differentiated between the primary structure (the sequence of the residues), the secondary structure ( α-helices, β-sheets and loops) and the tertiary structure (folding of the secondary structure elements into a three dimensional structure).

The functional specificity of a protein is linked to its structure. Due to the folding structure each of the critical residues, responsible for interactions of a protein with other molecules, are brought into a precise geometric arrangement.

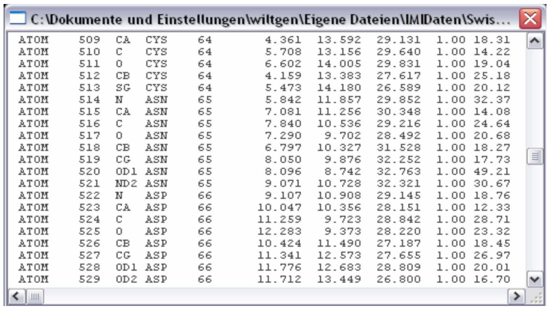
Then interactions of a protein with other molecules are determined by residues, which are close in 3D space, but may be very distant in the amino acid sequence. An active site of a protein is a localized combination of amino acids within the tertiary structure that acts with other molecules and provides the protein with biological activity. The active site is then often found only in a small part of the structure and the rest of the protein structure is mainly necessary to enable and maintain the correct spatial position between the amino acids on the active site.

## 3. Protein structure database

Protein structures are determined with crystallographic methods or by nuclear magnetic resonance spectroscopy. Once the atomic coordinates of the protein structure have been determined, a table of these coordinates is deposited into the protein database (PDB), an international repository for 3D structure files [BWF*00]. The PDB database is handled by the RCSB (Research Collaboratory for Structural Biology) at the Rutgers University and UC San Diego. PDB is the most important source for protein structures. At the moment PDB contains more than 26.000 protein structures. Before a new structure of a protein is added, a careful examination of the data must be carried out to guarantee the quality of the structure.

PDB: `http://www.rcsb.org/pdb/`

The PDB data file contains, among others, the coordinates of all the atoms of the protein (Figure 1). It is important to notice that these data is determined by crystallographic methods or by nuclear magnetic resonance spectroscopy. PDB files are used as input for protein visualization and they offer the necessary information for the calculation of protein properties, manipulation and representation the structures. From the PDB database the structure (coordinate) files can be downloaded and the protein visualized at the local place.


C:\Dokumente und Einstellungen\wiltgen\Eigene Dateien\IMIDaten\Swis...

```
ATOM     509  CA   CYS     64        4.361  13.592  29.131  1.00 18.31
ATOM     510  C    CYS     64        5.708  13.156  29.640  1.00 14.22
ATOM     511  O    CYS     64        6.602  14.005  29.831  1.00 19.04
ATOM     512  CB   CYS     64        4.159  13.383  27.617  1.00 25.18
ATOM     513  SG   CYS     64        5.473  14.180  26.589  1.00 20.12
ATOM     514  N    ASN     65        5.842  11.857  29.852  1.00 32.37
ATOM     515  CA   ASN     65        7.081  11.256  30.348  1.00 14.08
ATOM     516  C    ASN     65        7.840  10.536  29.216  1.00 24.64
ATOM     517  O    ASN     65        7.290   9.702  28.492  1.00 20.68
ATOM     518  CB   ASN     65        6.797  10.327  31.528  1.00 18.27
ATOM     519  CG   ASN     65        8.050   9.876  32.252  1.00 17.73
ATOM     520  OD1  ASN     65        8.096   8.742  32.763  1.00 49.21
ATOM     521  ND2  ASN     65        9.071  10.728  32.321  1.00 30.67
ATOM     522  N    ASP     66        9.107  10.908  29.145  1.00 18.76
ATOM     523  CA   ASP     66       10.047  10.356  28.151  1.00 12.33
ATOM     524  C    ASP     66       11.259   9.723  28.842  1.00 28.71
ATOM     525  O    ASP     66       12.283   9.373  28.220  1.00 23.32
ATOM     526  CB   ASP     66       10.424  11.490  27.187  1.00 18.45
ATOM     527  CG   ASP     66       11.341  12.573  27.655  1.00 26.97
ATOM     528  OD1  ASP     66       11.776  12.683  28.809  1.00 20.01
ATOM     529  OD2  ASP     66       11.712  13.449  26.800  1.00 16.70
```

**Figure 1:** *The input data file for protein visualization contains real data (among others the atomic coordinates) from protein structures determined by crystallographic methods or by nuclear magnetic resonance spectroscopy.*

Protein visualization at interactive rates has become more and more important, as the number of proteins in the protein data bank is increasing very fast. The ability to visualize the 3D structure of these proteins is critical in various areas such as drug design or protein modelling, because the function, which means the possible interactions with other molecules, of a protein is closely connected to its 3D structure.

## 4. Visualization tool

For our experiments we used the Swiss-PDB Viewer which is actually one of the most sophisticated tools [GP97]. The viewer is of special interest because it enables advanced protein structure visualization and complements the services offered by Swiss Institute of Bioinformatics.

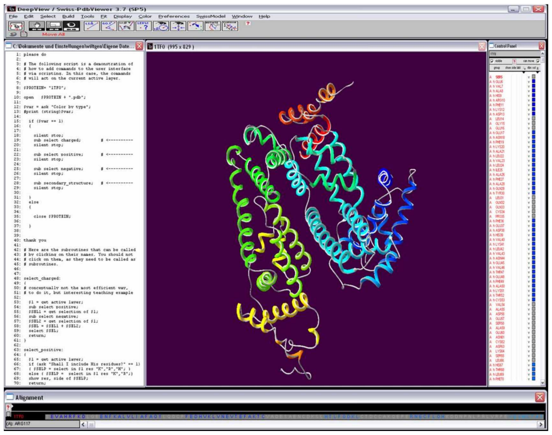
Swiss-Prot: `http://au.expasy.org/sprot/`

The viewer offers functions like superimposition of structures which can be used to prepare input for the prediction of unknown structures (via homology modelling) using the

Swiss-Model web server. The server is accessed directly by the Swiss-PDB Viewer.

This program converts the atomic coordinates into a view of the protein. The viewer provides ways to manipulate the protein by rotating and zooming the molecule and enables features like distance measurements, calculation and display of H-bonds, analysis of torsions angles etc. Additionally an illusion of depth is possible by creating two images that provide a stereo view.

The viewer program runs on a PC with Microsoft Windows. The workspace of the Swiss-PDB Viewer is divided into several windows (Figure 2). The main window is used for manipulation, measurements, etc. of protein structures. At the display window the protein structure is visualized. The control panel is used to select residues for display.



```
please do

# The following script is a demonstration of
# how to add commands to the user interface
# via scripting. In this case, the commands
# will act on the current active layer.

$PROTEIN= "1TF0";

open   $PROTEIN + ".pdb";

$var = ask "Color by type";
#print (string)$var;

  if ($var == 1)
  {

    silent stop;
    sub select charged;         # <----------
    silent stop;

    sub select positive;        # <----------
    silent stop;

    sub select negative;        # <----------
    silent stop;

    sub secondary_structure;    # <----------
    silent stop;

  }
  else
  {

    close $PROTEIN;

  }


thank you

# Here are the subroutines that can be called
# by clicking on their names. You should not
# click on them, as they need to be called as
# subroutines.


select_charged:
{
# conceptually not the most efficient way,
# to do it, but interesting teaching example

  $1 = get active layer;
  sub select positive;
  $SEL1 = get selection of $1;
  sub select negative;
  $SEL2 = get selection of $1;
  $SEL = $SEL1 + $SEL2;
  select $SEL;
  return;
}

select_positive:
{
  $1 = get active layer;
  if (ask "Shall I include His residues?" == 1)
  { $SELP = select in $1 res "K","R","H"; }
  else { $SELP =  select in $1 res "K","R";}
  show res, side of $SELP;
  return;
```

**Figure 2:** *The Swiss-PDB Viewer: The main window (top) enables the interactive manipulation of the protein structure in the display window (middle). In the control panel window (right side) individual amino acid residues can be selected for display. In the script window (left side), proprietary programs are displayed, allowing problem based interactive manipulations.*

For special problem representations, proprietary programs (written in the Deep View scripting language, a kind of PERL derivate) can be read into the script windows, enabling complex and time consuming calculations on the protein structure. The scripting language supports variables, conditional branching, loops, arrays and file access. There exists the possibility to stop the scripts at defined break points, enabling the user to interact with the graphical interface before resuming the operation. Subroutines are used for jump tables of functions that are executed by clicking with the mouse on their name in the script window. At the end of the program run, the routines can be called by the program and displayed in the script window allowing further manipulation of the results interactively by the user (see Figure 2). The routines are initiated by mouse clicking on their names in the script window. Interacting windows enables a relationship between the output data of calculations and the corresponding parts of the protein structure. This offers the possibility to interactively connect annotations with structural properties.

## 5. Protein visualization

Being able to“see” the 3D-structure of a protein and analyze its shape is of crucial importance for understanding protein properties and interaction. Looking at the protein structure means: locate different types of amino acids, visualize specific regions of the protein, visualize secondary structure elements, determine residues in the score or solvent accessible residues on the surface of the protein, determine binding sites, visualize physicochemical properties at active sites etc. But a protein can not be seen, for example by a microscope (with X-rays focussing lenses)! Therefore no real image of a protein exists (like a microscopic view from cell). Instead a model, resulting from the best fitting into the experimental data (determined by X-ray crystallography or nuclear magnetic resonance spectroscopy) must be used. This model of a structure is a 3D-representation of a protein that reflects the experimental data in a consistent way by providing information about the spatial arrangements of groups of atoms.

### 5.1. Representation of protein structures

Protein structure data are stored in the PDB file as a collection of Cartesian coordinates, with labeling information. For the visualization, the connectivity between atoms has to be taken into account. The protein is then visible as a 3D graphic, which can be rotated in the display window. The

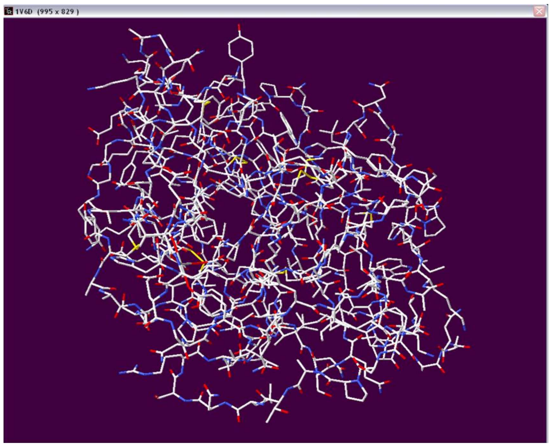


**Figure 3:** *Wire frame representation of the backbone and side chains of the trypsin protein. The atoms are coloured according to the CPK colour scheme; the chemical bonds are represented by sticks.*

proteins are so complex, that the 3D representations are difficult to interpret. The number of amino acids in proteins ranges from 50 to 2000 residues. (For example: the human serum albumin protein contains 4902 a toms). The first problem in the visualization and analysis of protein structure is the appropriate representation. The human eye can interpret 3D solids but has difficulties with topologically complex 3D data sets (Figure 3).

Because of the complex protein structure, special simplified representations are necessary. There are a number of conventionally simplified representations of protein structure that allow the visualization of overall topology of the protein structure, without the confusion of atomic details. Although the representations are constructed, they are based on experimental data and therefore they represent real aspects of proteins. A simplification of the representation and visualization of proteins is based on ribbons and helices, allowing the user to visualize essential features of the protein topology like secondary structure elements (Figure 4).

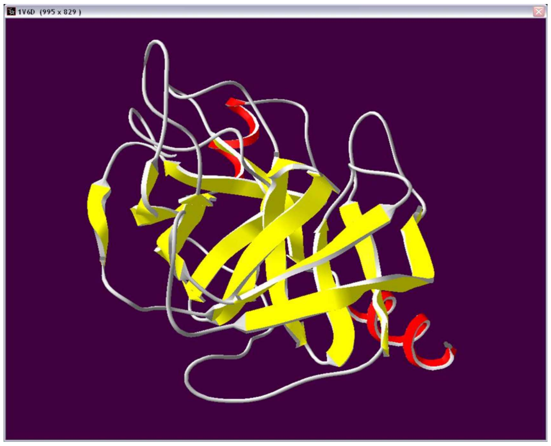


**Figure 4:** *Visualization of secondary structure elements like: α-helices,β-sheets by ribbons which are connected by loops. The polypeptide backbone lies in the ribbon of the α-helices and β-sheets, whereas the side chains are oriented outside the ribbon. Represented is again the trypsin protein.*

Molecular surface shows the overall shape of the protein but smoothes out the atomic details. The molecular (contact) surface of a protein is defined by the van der Waals radii of the atoms. The surface can considered as the boundary of that volume within any probe sphere (representing, for example, a water molecule) of a given radius sharing no volume with the hard sphere atoms which make up the protein molecule (Figure 5). The molecular surface is helpful for the analysis and understanding of protein-protein or protein-substrate interactions.

Molecular shapes are useful when protein-protein interactions are studied and wire frame representations when functions at atomic level are studied.

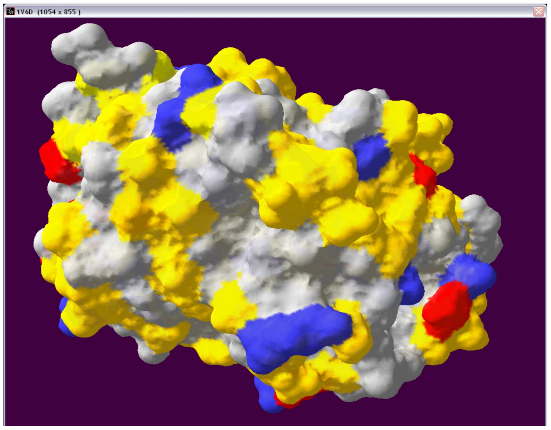


**Figure 5:** *Molecular surface of the trypsin protein. The surface is coloured according to the type of the amino acids located at the surface. Non-polar amino acids are grey, polar amino acids are yellow, basic (positive) amino acids are blue and acidic (negative) amino acids are red.*

### 5.2. Special problem based representations

The complexity of protein structures makes it necessary to choose problem based representations, allowing the investigator to concentrate on the specific features of interest without being overwhelmed. Visualizations of structural properties can reveal problems leading to new compelling questions by showing existing information in a new way. Adequate representation of the structure together with the mapping of physicochemical properties and annotations helps to understand proteins visually showing context and connection.

Special procedures are required for exploring structure-function relationships, aiding scientists in fast automated analysis of the functional properties of proteins. Sequence-based functional annotations need to be mapped to the corresponding part of the protein structure. Beside the PDB structure database we used two additional databases with different types of information:

PROSITE: http://au.expasy.org/prosite/

with biological significant sequence patterns and the Electron Density Server at Uppsala University [BB94], [KHZ*04]

EDS: http://fsrv1.bmc.uu.se/eds/

A motif is a locally conserved region in a sequence or a short sequence pattern shared by a set of sequences. Often motifs are localized in active sites or binding sites for substrates and coenzymes. PROSITE is a database which contains biologically significant patterns of residues responsible for the function of a protein family. These motifs are described by the PROSITE syntax:

$$C-x(3)-C-x(2)-[LMF]-x(3)-[DEN]-[LI]-x(5)-C$$

This pattern represents a common motif in the lysozyme family. The amino acids are represented in the one letter code. $x(n)$ means an pattern of $n$ arbitrary residues. Letters between brackets (for example: $[LMF]$) mean that one of the involved amino acid (represented in the one letter code: $L$ Leucin, $M$ Methionin, $F$ Phenylalanin, $C$ Cystein) must be present at the specific position. One letter (e.g. $C$) in the pattern means that exactly this amino acid and no other must be present at the specific position.

This motif pattern was retrieved from the PROSITE database and used to search (in a script program) for the corresponding residues, which are responsible for the biological function, in the protein. The positions and orientations of the detected residues in the protein fold are then visualized (Figure 6). It can be seen, that residues in the motif pattern that are separated in the sequence are close in 3D space due to the protein folding. The rest of the protein fold structure is represented as a ribbon visualizing secondary structure elements.

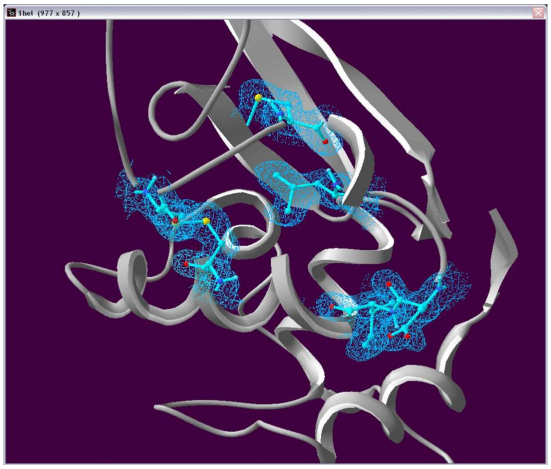


**Figure 6:** *Visualization of a PROSITE motif in wire frame representation superimposed with the corresponding part of the electron density map (resulting from the Electron Density Server, University of Uppsala) of lysozyme.*

Electron densities, responsible for the chemical properties of residues, allow insights into the dynamics of the active site. The visualization was done by superimposing the selected residues with the corresponding electron density map, retrieved from the Electron Density Server. The electron density maps result from X-ray crystallography of the molecule; the X-rays are scattered by the electrons surrounding the atoms. Electron density maps reflect real experimental data. The plotted map around the atoms and bindings in the representation defines surfaces with constant electron densities or in other words: the distribution of the electrons in space (Figure 6). Where there are many electrons the density is higher than in places there are few electrons.

From the structure file several properties of the protein (solvent accessibility, electric charges, geometrical distances between the residues in the motif etc.) are calculated automatically during the visualization. The results of the calculations are displayed in an output window. A relationship between the output data and corresponding parts of the protein structure in the display window was established by interacting windows.

## 6. Conclusion

In principle the DNA string holds the template for human development, physiology and certain diseases. The transformation of such information into understanding, prevention and treatment of diseases is a goal of bioinformatics. In future it will be routine to tailor medical treatments to the protein structure resulting from individual genetic profiles of patients, their pathogens etc. Genetic variations among patients, resulting in structural variability of the corresponding proteins, must be taken into account. Therefore the visualization of fundamental aspects of the protein fold structure is of special importance. The choice of an appropriate protein representation facilitates the analysis and visualization of specific protein properties and function.

## 7. Future outlook

In the future we will work towards adapted software tools for specific problems. The Swiss-PDB Viewer supports stereo hardware. In order to carry out further experiments, we plan to use a graphic card that enables QuadBuffered Stereo in OpenGL (where LCD shutter glasses (e.g. CrystalEyes) and emitters are needed). The visualization of the complex protein molecules reveals from a wealth of information the required contexts and connections. Methods in clinical practice fall basically into two categories: performed automatically or by interactive visualization. Due to the fact that biological research demands the ability to manipulate molecules in three dimensions, we are aiming towards the development of end user centered augmented reality applications, exactly suited to the needs of clinicians and biologists. We concentrate future work on the practical implications of the specific adapted viewer for the end-users. How can the end-users gain (complementary) information from different visualization variants, how can they benefit in better workflows. The first step is to apply Usability Engineering methods [Hol05], in order to gain a better understanding of biologists' behavior, their experience, tasks, and work contexts. Empirical studies, involving all aspects of biologists' experience and their interaction, are necessary to determine how the tools are ideally implemented.

## 8. Acknowledgement

The authors wish to thank Dr. Gerhard Bachmaier for putting his expertise at out disposal and preparing the layout of this paper using LaTeX.

## References

[BB94] Bucher P., Bairoch A.: A generalized profile syntax for biomolecular sequences motifs. In *roceedings of Intelligent Systems for Molecular Biology (ISMB 94)* (1994), AAAI Press, pp. 53–61.

[BRR77] Beem K. M., Richardson D. C., Rajagopalan K. V.: K.v. metal sites of copper-zinc superoxide dismutase. *Biochemistry 16*, 9 (19977), 1930–1936.

[BWF*00] Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N., Bourne P. E.: The protein data bank. *Nucleic Acids Research 28* (2000), 235–242.

[FMB05] Finn R. D., Marshall M., Bateman A.: ipfam: visualization of protein-protein interactions in pdb. *Bioinformatics 21*, 3 (2005), 410–412.

[GFG*94] Gershon N. D., Friedhoff R. M., Gass J., Langridge R., Meinzer H.-P., Pearlman J. D.: Is visualization really necessary. In *Proceedings of 21st Annual Conference on Computer Graphics* (1994), pp. 499–500.

[GLW03] Gabdoulline R. R., Leitner R. H. F., Wade R. C.: Prosat: functional annotation of protein 3d structures. *Bioinformatics 19*, 13 (2003), 1723–1725.

[GOT04] Golovin A., Oldfield T. J., Tate J. G.: E-msd: an integrated data resource for bioinformatics. *Nucleic Acids Research 32* (2004), 211–216.

[GP97] Guex N., Peitsch M. C.: The swiss-pdb viewer: An environment for comparative protein modelling. *Electrophoresis 18* (1997), 2714–2723.

[GWW03] Gabdoulline R. R., Wade R. C., Walther D.: Molsurfer: A macromolecular interface navigator. *Nucleic Acids Research 31*, 13 (2003), 3349–3351.

[Hog97] Hogue C. W.: Cn3d: A new generation of three-dimensional molecular structure viewer. *Trends in Biochemical Science 22* (1997), 314–316.

[Hol05] Holzinger A.: Usability engineering for software developers. *Communications of the ACM 48*, 1 (2005), 71–74.

[JMPW04] Johnson C. R., MacLeod R., Parker S. G., Weinstein D.: Biomedical computing. *Communications of the ACM 47*, 11 (2004), 64–71.

[Joh65] Johnson C. K.: *OR TEP: A FORTRAN Thermal-Ellipsoid Plot Program for Crystal Structure Illustrations*. 1965.

[KHZ*04] Kleywegt G. J., Harris M. R., Zou J. Y., Taylor T. C., Wählby A., Jones T. A.: The uppsala electron-density server. *Acta Crystallography D60* (2004), 2240–2249.

[LB68] Levinthal C., Barry C. D.: *Computer Graphics in Macromolecular Chemistry*. Benjamin, New York, 1968, pp. 231–253.

[Lev66] Levinthal C.: Molecular model-building by computer. *Scientific American 214*, 6 (1966), 42–52.

[MK04] Martz E., Kramer T. D.: Molecular visualization resources; online at: http://molvis.sdsc.edu, 2004.

[NRM04] Neshich G., Rocchia W., Mancini A. L.: Java protein dossier: a novel web based data visualization tool for comprehensive analysis of protein structure. *Nucleic Acids Research 32* (2004), 595–601.

[Old04] Oldfield T. J.: A java applet for multiple linked visualization of protein structure and sequence. *Journal of Computer Aided Molecular Research 18*, 4 (2004), 225–234.

[RR94] Richardson D. C., Richardson J. S.: Kinemages Ű simple macromolecular graphics for interactive teaching and publication. *Trends in Biochemical Science 19* (1994), 135–138.

[SMW95] Sayle R. A., Milner-White E. J.: Rasmol: Biomolecular graphics for all. *Trends in Biochemical Science 20* (1995), 374.

[VTR03] Vivek G., Tan T. W., Ranganathan S.: Xdomview: protein domain and exon position visualization. *Bioinformatics 19*, 1 (2003), 159–160.

[Wal97] Walther D.: Webmol - a java based pdb viewer. *Trends in Biochemical Science 22* (1997), 274–275.