PERGAMON

Computers in Biology and Medicine 31 (2001) 259–267

Computers in Biology and Medicine

www.elsevier.com/locate/compbiomed

ANTHEPROT: An integrated protein sequence analysis software with client/server capabilities ☆

G. Deléage *, C. Combet, C. Blanchet, C. Geourjon

Pôle Bioinformatique Lyonnais, Institut de Biologie et Chimie des Protéines, CNRS UMR 5086 Bioinformatique et RMN structurales, 7, Passage du Vercors, 69 367 Lyon, Cedex 07, France

Received 26 May 2000; accepted 16 January 2001

Abstract

Programs devoted to the analysis of protein sequences exist either as stand-alone programs or as Web servers. However, stand-alone programs can hardly accommodate for the analysis that involves comparisons on databanks, which require regular updates. Moreover, Web servers cannot be as efficient as stand-alone programs when dealing with real-time graphic display. We describe here a stand-alone software program called ANTHEPROT, which is intended to perform protein sequence analysis with a high integration level and clients/server capabilities. It is an interactive program with a graphical user interface that allows handling of protein sequence and data in a very interactive and convenient manner. It provides many methods and tools, which are integrated into a graphical user interface. ANTHEPROT is available for Windows-based systems. It is able to connect to a Web server in order to perform large-scale sequence comparison on up-to-date databanks. ANTHEPROT is freely available to academic users and may be downloaded at http://pbil.ibcp.fr/ANTHEPROT.

Keywords: Protein sequence analysis; Multiple alignment; Secondary structure prediction; Web server

☆ All the authors of this paper belong to the PBIL ("Pôle BioInformatique Lyonnais") and are in charge of the molecular modelling of protein and structural aspects of the PBIL. Dr. Gilbert Deléage (Professor of BioInformatics at the Claude Bernard University of Lyon) published one of the first package (ANTHEPROT for DOS and Apple) for protein sequence analysis in 1988. He is the author of ANTHEPROT for Windows. Christophe Combet (Ph.D. student) is the author of NPS@ web server (http ://pbil.ibcp.fr/NPSA). He contributes to the client/server interface of ANTHEPROT with the server side as a specialist of Perl cgiscripting. Dr. Christophe Blanchet is the author of MPSA (http ://pbil.ibcp.fr/mpsa) under Unix and Macintosh systems. He contributes at the system level of the client/server mode. MPSA also supports a client/server mode. Dr. Christophe Geourjon is the author of the IBM rs6000 version of the ANTHEPROT software and is mainly involved in the development molecular modelling tools.

* Corresponding author.

E-mail address: g.deleage@ibcp.fr (G. Deléage).

0010-4825/01/$ - see front matter
PII: S0010-4825(01)00008-7

1. Introduction

A large number of projects to determine the complete sequence of whole organisms yield data to the community with increasingly high rate. Thus, the need for the theoretical treatment of an ever-growing set of protein sequences is required. Many local programs have been recently developed for managing and/or editing multiple alignments (Refs. [1–4]). Other viewers or editors of multiple alignments are also available on the Web such as CINEMA [http://www.bioinf.man.ac.uk/dbbrowser/] Ref. [5] or NPS@ (Ref. [6]). However, few comprehensive packages specially developed for protein sequence analysis are available and most of them are commercial (GCG, Protean) or no longer supported (PC/Gene). In the past, we developed ANTHEPROT for DOS-based PC computers, one of the first packages devoted to protein sequence analysis Refs. [7,8], and set it as a freely available package (ftp.ibcp.fr). This program was able to interactively couple multiple alignments with secondary structure predictions (Ref. [9]) but was only available on the IBM rs6000 workstation with AIX operating system. Moreover, it did not allow the edition of alignments. In this paper, we describe a completely new version of ANTHEPROT for PC computers that allows the incorporation of secondary structure predictions within multiple alignment and full interactive editing of alignments, graphic windows and client/server capabilities. Indeed, this package is an integrated solution: (i) including most of the individual methods in a single interface, (ii) able to submit tasks on a remote server, and (iii) able to retrieve data from a remote Web server.

2. Systems requirements

ANTHEPROT is a program written with Microsoft Visual Basic development interfaces. VB3 and VB6 have been used for Windows 3.1 and for 32 bits Windows operating systems, respectively. The program has been successfully tested on all windows systems (16 bits for 3.1 and 32 bits for Win95, Win98 or NT). The program is available as a complete ready-to-run archive, by anonymous ftp to ftp.ibcp.fr in the pub/ANTHEPROT/WINDOWS directory. The suitable machine for ANTHEPROT is a PC (with at least a 600 MHz processor) running under Windows 32 bits operating system (Win 98 or NT) equipped with at least 64 Mb RAM (128 Mb is better), at least 5 GB hard drive space (depending upon the use of the program), and a modem or an ethernet card for network access. A wide screen should be used (19 in is perfect) due to the numerous graphical windows generated by the program. Ink jet printers are fully supported for colour printing of multiple alignment and graphic curves. The client/server mode uses socket functionalities for TCP/IP connections under the HTTP 1.1 protocol with POST method for data transfer.

ANTHEPROT Web server facilities are available as a small set of Perl v5 scripts that should be first installed on the remote Web server. A perl package is available to set up the pathways for remote programs and databanks.

3. General features

The general interface keeps the spirit of the previous versions of ANTHEPROT. This means that most tools for protein sequence analysis should be available in a single graphical user interface.

This is a particularly important point for interactive graphic displays. ANTHEPROT also provides network facilities (client mode and internal Web browser). The program is fully customisable for colour codes, font size, and for the Web server to query. Real-time movable cursors, scroll, zoom, resize and selection functions capabilities are supplied in all graphic windows. The interface has been completely rewritten in order to offer printing outputs, cut and paste functions, interactive image resizing facilities, network access, graphics data exchange (cursor position). The program features are so numerous that they cannot be easily described in details. Indeed, ANTHEPROT is easier to use than to describe.

4. File formats and inputs/outputs

The program automatically recognises standard biological file formats like EMBL, Pearson/FASTA, NBRF/PIR, CLUSTALW, Multalin, PHD, PREDATOR, GOR, etc. An input layer automatically detects the file formats that are accepted by the program. These files can contain a single amino acid sequence, a multiple alignment, or secondary structure predictions. The input layer is mainly constituted of a powerful sequence manager and editor. This manager uses an internal syntax analyser as a filtering converter and it is able to read sequences in the most commonly used sequence formats (Pearson/FASTA, EMBL, NBRF/PIR,...). The sequences can be manipulated as one-sequence file or sequence database (multiple sequence) file. The graphical output layer is able to directly write on the printer device, copy data into the clipboard and save pictures as BMP files. Alternatively, RTF (Rich Text Format) files can be generated from within the program for multiple alignments. The RTF format file can be further loaded with any program that accepts RTF as input format and the multiple alignment will appear as colour-coded text.

5. Protein sequence analysis methods

ANTHEPROT integrates 25 different methods or algorithms for protein sequence analysis and numerous coupling tools. It is an ever-growing platform with an increasing number of methods. The general organisation of ANTHEPROT is given in Fig. 1. All individual methods have been implemented or incorporated as originally described by their authors. The most time-consuming jobs (local alignment, local fasta or prosite scans) are launched in a DOS window as a background job, thus allowing the user to maintain a rather good interactivity during calculations.

To start a sequence analysis, a good strategy is to search for homologous proteins in databases such as SWISS-PROT, trEMBL, or NRL3D in order to collect all proteins that may belong to a given family. These homologous proteins can be detected by using FASTA (Ref. [10]), BLAST (Ref. [11]) or PattInProt (Blanchet et al., in preparation) algorithms. Thereafter, homologous proteins can be aligned together with CLUSTALW (Ref. [12]) or Multalin (Ref. [13]) either on the local machine or on the remote Web server. Secondary structure predictions performed by several methods can then be included automatically in the alignment. This strategy permits to validate the multiple alignment procedure on the basis of predicted secondary structures. The methods for secondary structure prediction are available either in local mode from within the package or through a remote server (specified by the user). Local methods are nearest-neighbour methods (Ref. [14]), directional

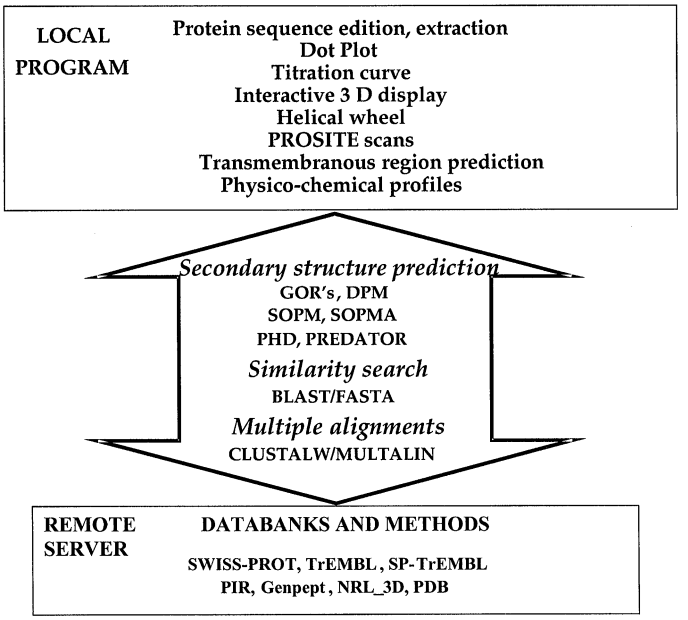

Fig. 1. Methods available in the ANTHEPROT program. All methods are coupled together and invoked from a common general GUI. Italicised methods correspond to remotely available algorithms whereas normal ones are for locally installed methods.

information methods (GOR Refs. [15,16]), Double Prediction Method (Ref. [17]). Remote methods are for example Self Optimised Prediction Method from Alignments (Ref. [18]), and the PHD method of Rost and Sander [19].

6. Sites/signatures detection

The site/signature detection using the PROSITE dictionary (Refs. [9,20]) is a powerful way to identify protein functions, which can be performed in a stand-alone mode provided that PROSITE.DAT and PROSITE.DOC files are available on the local system. For the PROSITE scan, the user may allow some errors as in the case of the pattern search program PattInProt. Even in local mode, the cross-reference system between PROSITE.DAT and PROSITE.DOC allows the generation of hypertext links in the result file to display the documentation in another window.

7. Physico-chemical profiles

Protein families can be analysed with the help of physico-chemical profiles such as hydrophobicity (Ref. [21]), antigenicity (Ref. [22]), flexibility (Ref. [23]), and solvent accessibility (Ref. [24]). These tools can be useful to design peptides that would lead to antibodies (after immunisation) able to recognise the entire protein.

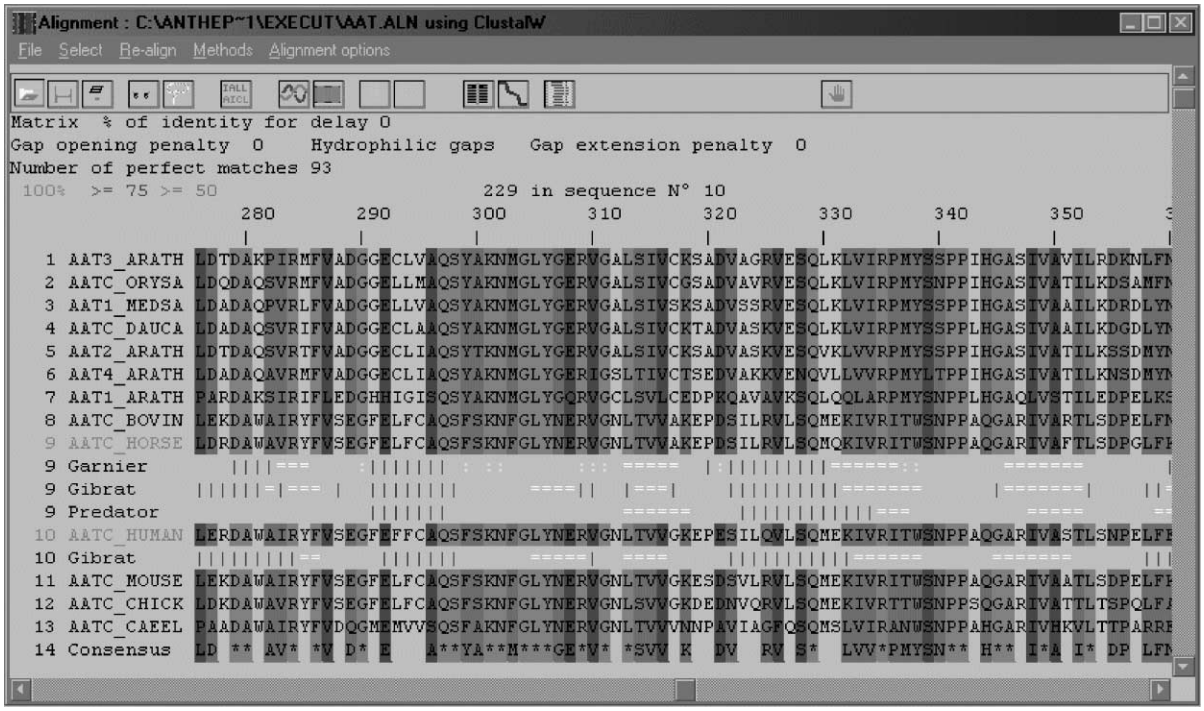

Fig. 2. The multiple alignment main window. Sequences alignment was performed from within the ANTHEPROT program by using the CLUSTALW (1.8) package after extraction (through the HTTP server) of the 13 most similar proteins from SWISSPROT. Three secondary structure predictions have been inserted within the multiple alignment window for AATC_HORSE and one for AATC_HUMAN.

8. Multiple alignment

Alignment can be directly loaded or generated from within the program (after selection of the sequences) by using CLUSTALW 1.8 program. A dialog window allows the user to set all CLUSTALW parameters for both pairwise and multiple alignment steps just like CLUSTALX (Ref. [3]). A typical multiple alignment in boxshade mode is shown in Fig. 2. The alignment is clickable and the cursor position returns the amino acid type, its position in the ungapped sequence, and the sequence number. Sequences can be selected by clicking onto the protein sequence names for further analysis. A part of the alignment can be selected in a box for further analysis. This tool allows the generation of a consensus sequence with the PROSITE syntax so as to scan protein sequence databases such as SWISS-PROT with the PattInProt method and look for the presence of the pattern. This tool is helpful to detect particular motifs that are typical of a protein family. The multiple alignment editor provides gap insertion/deletion capabilities with interactive update of the coloured alignment display. The modified alignment can be saved in CLUSTAL format. The alignment can be coloured as a function of the similarity level as defined since CLUSTALW 1.7 or with a user-defined colour. A primary consensus line reports the most represented amino acid type in a given position. Three different display modes are available: all residues, identical residues or different residues. Changing

the mode is helpful when looking at highly conserved regions. The secondary structure of protein sequences can be displayed after computation by several methods within the multiple alignment window.

9. Secondary structures

Some of the numerous graphic capabilities of ANTHEPROT are illustrated in Fig. 3. A user-defined secondary structure obtained either by another predictive method or from known structures (PDB data through DSSP tool) can be inserted into the multiple alignment window. The display system is versatile enough to allow the selection of the sequence for which the method has to be shown. A

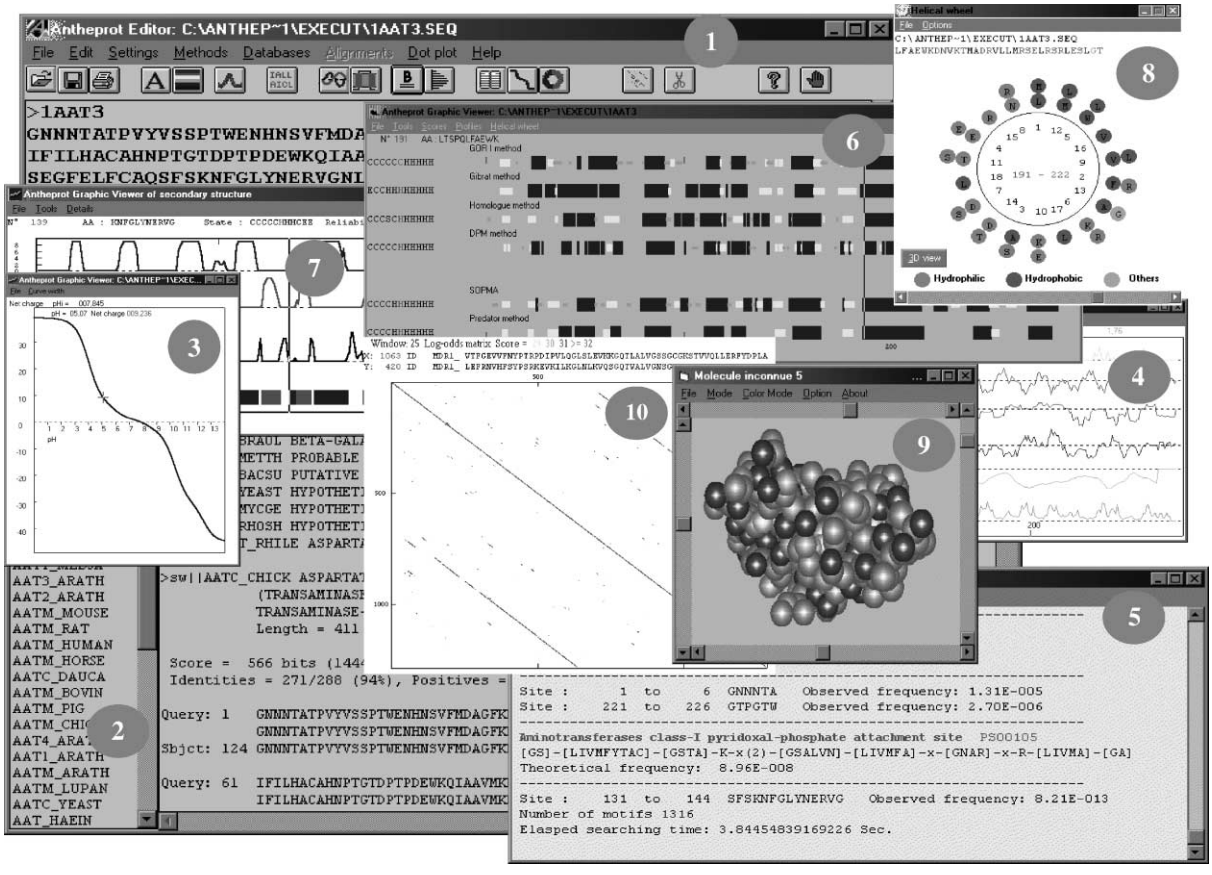

Fig. 3. Overview of the program graphic windows. The main sequence editor window (1) contains the sequence for chicken Aspartate amino transferase (SWISSPROT code AATC-CHICK). Result form BLAST analysis (remote mode onto SWISSPROT databank) window (2). Titration curve (3). Physico-chemical profiles (4). PROSITE scan (5). Secondary structure predictions (6). Graph of conformational confidence index for the PHD method (7). Helical wheel projection for predicted helical region (8). The 3D structure of crambin (PDB code 1 crn) in space filling mode (9). The dot-matrix plot of MDR1_HUMAN versus itself (10).

distinction between highly probable secondary structures and less probable ones is therefore possible. A double click on the name of the prediction method leads to the graphical display (in another window) of the scores for the different conformational states and to a schematic view of predicted secondary structure. In graphics, a cursor indicates amino acid position as well as the corresponding value (physico-chemical profiles or conformational scores). When moved in a window, the cursor position is updated in all windows. The insertion of predicted secondary structures within a multiple alignment is one of the most useful and original tools of ANTHEPROT.

10. Interaction with network protein sequence analysis (NPS@) remote server

ANTHEPROT is able to submit jobs for BLAST/FASTA, multiple alignment, secondary structure prediction, pattern searches on a Web server such as the NPS@ server (http://pbil.ibcp.fr/NPSA). In this context, no databanks are needed on the user local machine thus avoiding the crucial problem of maintaining up-to-date databanks on local machines. Update is performed daily on our remote server by mirroring the original servers for protein sequence databanks (SWISS-PROT, TrEMBL and Nrl_3D). Mirror sites for distant countries may also be installed upon request. A list of biological Web servers compatible with the software will be maintained. Results generated by the NPS@ server (Ref. [6]) can be directly loaded into the program by defining the "chemical/x-antheprot" MIME type in the Web browser. So far, the ANTHEPROT package has been retrieved by more than 6000 users from our anonymous ftp site.

11. Help

On line documentation and help are available through the main Web page (http://pbil.ibcp.fr/). ANTHEPROT includes an internal Web browser that is useful to get the help page from within the program. Users may subscribe to the mailing list on the form of the main web page. Current version is 5.0 and users are notified of new releases or updates through the mailing list.

12. Discussion

The main goal in developing this program was to provide biological users with a powerful integrated system to combine multiple alignment with different secondary structure prediction methods and other useful methods. In order to investigate the conformational scores for each state (helix, sheet, turn and coil), many graphical views can be displayed. This system is able to insert and manage predicted or user-defined secondary structures for individual aligned sequence. A 3D module is also available to interactively manipulate protein 3D structure.

The system combines the advantages of both stand-alone and connected programs into a single platform. The client/server operating mode of the program is fully transparent for the user and no particular knowledge about network is required to use the system. It has the advantage of stand-alone programs since it handles multiple graphic windows that are all interactive and linked together. ANTHEPROT, as a client program, makes use the power of a Web server with up-to-date protein

databanks. In the future, it will be improved with the possibility to generate Rasmol scripts directly from sequence analysis tools. The program is a Windows-based platform able to incorporate many other methods and to interact with a server on which time-consuming programs and large size databanks are installed. We are open to include other methods or to integrate them into the program as external procedures. In conclusion, this program is an useful integrated system that could be a complete solution for protein sequence analysis for universities, research institutes or private companies.

13. Summary

In this paper, we describe a stand-alone software called ANTHEPROT intended for protein sequence analysis with a high level of integration and clients/server capabilities. The main goal in creating this program was to provide biological users with a powerful integrated system that combines multiple alignment with different secondary structure prediction methods and other useful methods. In order to investigate the conformational scores for each state (helix, sheet, turn and coil), several graphical views can be displayed. ANTHEPROT integrates 25 different methods or algorithms for protein sequence analysis and many coupling tools in an ever-growing platform with an increasing number of methods. It is an interactive graphic program that allows handling of protein sequences and data in a very interactive and convenient manner. It provides many methods and tools, which are integrated into a graphical user interface.

The system combines the advantages of both stand-alone and connected programs into a single product. The client/server operating mode of the program is fully transparent for the user and no particular knowledge about network is required to use the system. It has the advantage of stand-alone programs since it handles multiple graphic windows that are all interactive and connected together. On the other hand, it offers, as a connected program, the power of a Web server with up-to-date protein databanks. The program is a Windows based platform able to welcome many other methods and to interact with a server on which all time-consuming programs and large size databanks will be installed. As a conclusion, this program is an efficient integrated system that could be a complete solution for Intranet protein sequence analysis for universities, biological research institutes or biomedical companies. ANTHEPROT is available for Windows based systems. It can be downloaded by anonymous ftp or from within its Web page (http://pbil.ibcp.fr/ANTHEPROT).

Acknowledgements

This work is supported by MESR (ACC-SV13) and CNRS (IMABIO). Thanks are due to all biocomputing centres and people for providing up-to-date sequences databanks and powerful algorithms. The authors would like to thank Dr. L. Baggetto for English improvement.

References

[1] N. Galtier, M. Gouy, C. Gautier, SEAVIEW and PHYLO_WIN: two graphics tools for sequence alignment and molecular phylogeny, Comp. Appl. Biosci. 12 (1996) 543–548.

[2] A.S. Frolov, I.S. Pika, A.M. Reoshkin, ProMSED: protein multiple sequence editor for Windows 3.11/95, Comp. Appl. Biosci. 13 (1997) 243–248.
[3] J.D. Thompson, T.J. Gibson, F. Plewniak, F. Jeanmougin, D.G. Higgins, The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, Nucl. Acids Res. 25 (1997) 4876–4882.
[4] P. Gouet, E. Courcelle, D.I. Stuart, F. Métoz, ESPript: analysis of multiple sequence alignments in PostScript, Bioinformatics 15 (1999) 305–308.
[5] D.J. Parry-Smith, A.W.R. Payne, A.D. Michie, T.K. Attwood, CINEMA—A novel Colour INteractive Editor for Multiple Alignments, Gene 221 (1998) 57–63.
[6] C. Combet, C. Blanchet, C. Geourjon, G. Deléage, NPSat: Network protein sequence analysis, TIBS 291 (2000) 147–150.
[7] G. Deléage, F.F. Clerc, B. Roux, D.C. Gautheron, ANTHEPROT: a package for protein sequence analysis using a microcomputer, Comp. Appl. Biosci. 4 (1988) 351–356.
[8] C. Geourjon G. Deléage, ANTHEPROT 2.0: A three dimensional module fully coupled with protein sequence analysis method, J. Mol. Graph. 13 (1995) 209–212.
[9] C. Geourjon, G. Deléage, Interactive and graphic coupling between multiple alignments, secondary structure predictions and motif/pattern scanning into proteins, Comp. Appl. Biosci. 9 (1993) 87–91.
[10] W.R. Pearson, D.J. Lipman, Improved tools for biological sequence analysis, Proc. Natl. Acad. Sci. 85 (1988) 2444–2448.
[11] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucl. Acids Res. 17 (1997) 3389–3402.
[12] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice, Nucl. Acids Res. 22 (1994) 4673–4680.
[13] F. Corpet, Multiple sequence alignment with hierarchical clustering, Nucl. Acids Res. 16 (1988) 10 881–10 890.
[14] J.M. Levin, B. Robson, J. Garnier, An algorithm for secondary structure determination in proteins based on sequence similarity, FEBS Lett. 205 (1986) 303–308.
[15] J. Garnier, D.J. Osguthorpe, B. Robson, Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, J. Mol. Biol. 120 (1978) 97–120.
[16] J.F. Gibrat, J. Garnier, B. Robson, Further developments of protein secondary structure prediction using information theory, J. Mol. Biol. 198 (1987) 425–443.
[17] G. Deléage, B. Roux, An algorithm for protein secondary structure prediction based on class prediction, Prot. Engng. 1 (1987) 289–294.
[18] C. Geourjon, G. Deléage, SOPMA: Significant improvement in protein secondary structure prediction by consensus prediction from multiple alignments, Comp. Appl. Biosci. 11 (1995) 681–684.
[19] B. Rost, C. Sander, Prediction of protein secondary structure at better than 70% accuracy, J. Mol. Biol. 232 (1993) 584–599.
[20] A. Bairoch, P. Bucher, K. Hofmann, The PROSITE database, its status in 1995, Nucl. Acids Res. 24 (1996) 189–196.
[21] J. Kyte, R.F. Doolittle, A simple method for displaying the hydrophobic character of a protein, J. Mol. Biol. 157 (1982) 105–132.
[22] T.P. Hopp, K.R. Woods, Prediction of protein antigenic determinants from amino acid sequences, Proc. Natl. Acad. Sci. USA 78 (1981) 3824–3828.
[23] P.A. Karplus, G.E. Schulz, Prediction of chain flexibility in proteins, Naturwissenschaften 72 (1985) 212–213.
[24] J. Boger, E.A. Emini, A. Schmidt, Surface probability profile. An heuristic approach to the selection of synthetic peptide antigens, Reports on the Sixth International Congress in Immunology, Toronto, 1986, p. 250.