BIOINFORMATICS APPLICATIONS NOTE

Vol. 16 no. 3 2000
Pages 286–287

MPSA: integrated system for multiple protein sequence analysis with client/server capabilities

Christophe Blanchet, Christophe Combet, Christophe Geourjon and Gilbert Deléage

Institut de Biologie et Chimie des Protéines, BCP-CNRS UPR 412, Laboratoire de conformation des Protéines, 7, Passage du Vercors, 69 367, Lyon Cedex 07, France

Received on July 26, 1999; accepted on November 23, 1999

Abstract

Summary: *MPSA is a stand-alone software intended to protein sequence analysis with a high integration level and Web clients/server capabilities. It provides many methods and tools, which are integrated into an interactive graphical user interface. It is available for most Unix/Linux and non-Unix systems. MPSA is able to connect to a Web server (e.g. http://pbil.ibcp.fr/NPSA) in order to perform large-scale sequence comparison on up-to-date databanks.*

Availability: *Free to academic http://www.ibcp.fr/mpsa/*

Contact: *c.blanchet@ibcp.fr*

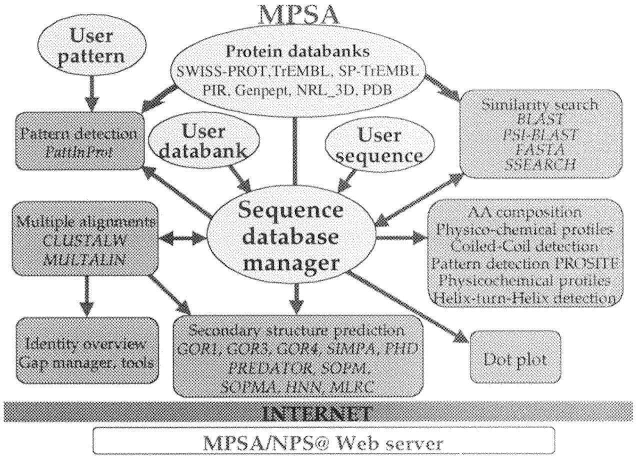


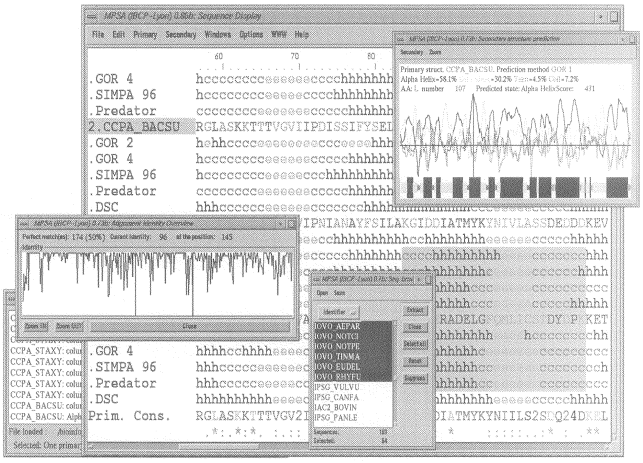
Fig. 1. General architecture of MPSA. All methods are coupled together and invoked from a common general GUI.

The numerous projects to determine the complete sequence of whole organisms yields data to the community with increasingly high flow rate. Thus, the theoretical treatment of ever-growing sets of protein sequences is required. In this context, we develop the MPSA package which is an integrated solution (i) including most of the individual methods in a single GUI (Graphic User Interface), (ii) able to run on almost all platforms, (iii) capable of submitting biological analysis jobs on a remote server and (iv) able to retrieve data from a remote Web server. MPSA allows the incorporation of secondary structure predictions within the multiple alignment and full interactive editing of huge alignments.

The software is written in C and thus fully supports dynamic memory allocation. It uses the VIBRANT GUI library (J. Kans, NCBI Software Toolkit) and a *biolcp* library of bioinformatic methods developed in our lab. The program has been successfully tested on UNIX/Linux and Macintosh. Biological Web client/server facilities are available as a small set of Perl v5 scripts (MPSAweb) that should be first installed on the remote Web server, as we did on our Web server NPS@. Once MPSAweb server is configured, MPSA is able to ask the remote server for remotely available databanks and to submit a query to analyse user data with remote methods. This Web request uses a new MIME type we defined as ‘chemical/x-mpsa’. This MIME type can be set up in the user browser so as to define MPSA as a helper application allowing the user to directly download data generated by NPS@ within MPSA.

MPSA automatically recognizes the file formats, which can be amino acids sequence, multiple alignment, secondary structure or physico-chemical profile files. MPSA uses the ≪ ReadSeq ≫ utility (D. Gilbert, Biology Dept, Indiana University) as sequence formats (Pearson/FASTA, EMBL, NBRF/PIR, . . .). Moreover, the sequence browser can be considered as a database manager/converter/extractor since it is able to handle very large lists of protein sequences.

MPSA integrates more than 25 different methods or algorithms for protein sequence analysis (Figure 1) with numerous coupling tools. The methods are available either in local mode from within the package or through a remote server specified by the user. The search for homologous proteins in sequence databanks like SWISS-PROT, trEMBL, or NRL3D can be performed by using FASTA, SSEARCH algorithms (Pearson and Lipman,



MPSA (IBCP-Lyon) 0.86b: Sequence Display
File
Edit
Primary
Secondary
Windows
Options
WWW
Help
60
70
80
MPSA (IBCP-Lyon) 0.73b: Secondary structure prediction
Secondary
Zoom
MPSA (IBCP-Lyon) 0.73b: Alignment Identity Overview
Identity
Zoom IN
Zoom OUT
Close
MPSA (IBCP-Lyon) 0.7b: Seq. brow
Open
Save
Identifier
Extract
Close
Select all
Reset
Suppress

```
.GOR 4         hcccccccceeeeeecccchhhhhhh
.SIMPA 96      hccccccccceeeeecccchhhhhhhh
.Predator      cccccccceeeeeeecccchhhhhh
2.CCPA_BACSU   RGLASKKTTTVGVIIPDISSIFYSEL
.GOR 2         hehhcccceeeeeeeccceeeehhh
.GOR 4         hcccccccceeeeeecccchhhhhh
.SIMPA 96      hccccccccceeeeecccchhhhhhhh
.Predator      cccccccceeeeeeecccchhhhhh
.DSC           hhhccccceeeeeecccccchhhhhhhhhhhhhhhccceeeeccccchhhhh
               VIPNIANAYFSILAKGIDDIATMYKYNIVLASSDEDDDKEV
               eecchhhhhhhhhcchhhhhhhheeeeeccccchhhhh
               ecccchhhhhhhccchhhhhhhhh      ccccccchhh
                                  hhhhhhhc     cccccchhhh
                                  hhhhccc      ccccccccc
                                  hhhhccc      cccchhhhh
                                  RADELG       DY P KET
                                  hhhhhhhh     hccccccchh
.GOR 4         hhhhcchhhheeeec    hhhhhhc      ccccccccc
.SIMPA 96      hhhccccceeeeeec    hhhhhhcc     cccccccch
.Predator      hhhcccccceeeeecc   hhhhhhcc     ccccccccc
.DSC           hhhhhhhhhheeeec    hhhhhhhheeeeeccccchhhhh
Prim. Cons.    RGLASKKTTTVGV2I    DIATMYKYNIILS2SDQ24DKEL
```

Primary struct. CCPA_BACSU. Prediction method GOR 1

Alpha Helix=58.1% Beta Sheet=30.2% Turn=4.5% Coil=7.2%

AA: L number 107 Predicted state: Alpha HelixScore: 431

Perfect match(es): 174 (50%) Current identity: 96 at the position: 145

IOVO_AEPAR
IOVO_NOTCI
IOVO_NOTPE
IOVO_TINMA
IOVO_EUDEL
IOVO_RHYFU
IPSG_VULVU
IPSG_CANFA
IAC2_BOVIN
IPSG_PANLE

Sequences: 169

Selected: 84

CCPA_STAXY: colu
CCPA_STAXY: colu
CCPA_STAXY: colu
CCPA_STAXY: colu
CCPA_STAXY: colu
CCPA_BACSU: colu
CCPA_BACSU: Alph

File loaded : /bioinfo

Selected: One primar

Fig. 2. The MPSA main windows.

1988) and BLAST (Altschul *et al.*, 1997). The alignment of homologous protein subset can be performed with CLUSTALW (Thompson *et al.*, 1994) either locally or remotely on the Web server. Secondary structure either predicted or user-defined can be included automatically in the alignment (see a review of Rost and O'Donoghue, 1997). These structures can be obtained either by predictive methods provided by MPSA/NPS@ or from known structures (PDB data through DSSP tool). Other tools for analysing protein families such as PROSITE (Bairoch *et al.*, 1996) scan and the physico-chemical profiles like hydrophibicity, antigenicity, flexibility and solvent accessibility can be included into the multiple alignment display. The data displayed in MPSA are fully accessible through the mouse or graphic tools. For example, information is returned by clicking onto the alignment (positions in gapped and ungapped sequence, amino acid type and sequence number). Selection of sequences and blocks can be made to generate a consensus sequence which can be directly used in our method PattInProt to scan protein sequence databanks. Alignment edition can be achieved through a gap manager and 'drag and drop' facility. The alignment quality can be assessed by identity and homology overviews shown as graphs or by colour coding systems. For example, MPSA is able to colour the multiple alignment as a function of the identity level, as a function of the similarity level as defined in CLUSTALW 1.7 or with a user-defined function. In order to simply show the relationship between the protein sequences and their secondary structures (predicted or measured) these data can be coloured and included in the multiple alignment (Figure 2). Many types of consensus can be derived from the alignment (primary or secondary, partial or global).

MPSA is fully customizable for colour codes, input/output/program pathnames and externally callable methods, font size, and for the Web server to query. The graphic windows are fully interactive with cursors, scroll, zoom, resize and selection functions capabilities. At the printing level, MPSA can generate PostScript and RTF (Rich Text Format) files.

The main goal in making MPSA was to provide the biological user with a powerful integrated and open system to combine multiple alignment with different secondary structure prediction methods and many other useful bioinformatic methods. As a conclusion, MPSA and its Web facilities is an efficient integrated system that could be a complete solution for Intranet protein sequence analysis for universities, biological research institutes or companies.

Acknowledgements

This work is supported by MESR (ACC-SV13), Région Rhône-Alpes and CNRS (COMI). CB is a recipient of a fellowship from MENESR. Thanks are due to NCBI Vibrant staff and to all biocomputing centres for providing up-to-date sequences databanks and powerful algorithms.

References

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.*, **17**, 3389–3402.

Bairoch,A., Bucher,P. and Hofmann,K. (1996) The PROSITE database, its status in 1995. *Nucl. Acids Res.*, **24**, 189–196.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence analysis. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Rost,B. and O'Donoghue,S. (1997) Sisyphus and prediction of protein structure. *Comput. Appl. Biosci.*, **13**, 345–356.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucl. Acids Res.*, **22**, 4673–4680.