BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 9 2001
Pages 847–848

InterProScan – an integration platform for the signature-recognition methods in InterPro

Evgeni M. Zdobnov and Rolf Apweiler*

EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received on August 2, 2000; revised on January 12, 2001; accepted on May 30, 2001

ABSTRACT

Summary: InterProScan is a tool that scans given protein sequences against the protein signatures of the InterPro member databases, currently – PROSITE, PRINTS, Pfam, ProDom and SMART. The number of signature databases and their associated scanning tools as well as the further refinement procedures make the problem complex. InterProScan is designed to be a scalable and extensible system with a robust internal architecture.

Availability: The Perl-based InterProScan implementation is available from the EBI ftp server (ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/) and the SRS-based InterProScan is available upon request. We provide the public web interface (http://www.ebi.ac.uk/interpro/scan.html) as well as email submission server (interproscan@ebi.ac.uk).

Contact: Evgueni.Zdobnov@EBI.ac.uk

INTRODUCTION

Databases of protein domains and functional sites have become vital resources for the prediction of protein functions. During the last decade, several signature-recognition methods have evolved to address different sequence analysis problems, resulting in rather different and, for the most part, independent databases. Diagnostically, these resources have different areas of optimum application owing to the different strengths and weaknesses of their underlying analysis methods. Thus, for best results, search strategies should ideally combine all of them. InterPro (The InterPro Consortium, 2001) is a collaborative project aimed at providing an integrated layer on top of the most commonly used signature databases by creating a unique, non-redundant characterization of a given protein family, domain or functional site. The InterPro database integrates PROSITE (Hofmann *et al.*, 1999), PRINTS (Attwood *et al.*, 2000), Pfam (Bateman *et al.*, 2000), ProDom (Corpet *et al.*, 1999) and SMART (Schultz *et al.*, 2000) databases and the addition of others is scheduled.

*To whom correspondence should be addressed.

INTERPRO MEMBER DATABASES AND SCANNING METHODS

Legend: ❖ denotes a database and ➢ denotes the associated scanning tools.

❖**PROSITE patterns.** Some biologically significant amino acid patterns can be summarized in the form of regular expressions.

➢ *ScanRegExp* (by Wolfgang.Fleischmann@ebi.ac.uk).

❖**PROSITE profiles** based on weight matrices (also known as profiles) are more sensitive in detection of divergent protein families.

➢ *pfscan* from the *Pftools* package (by Philipp.Bucher@isrec.unil.ch).

❖**PRINTS** database houses a collection of protein family fingerprints. These are groups of motifs that together are diagnostically more potent than single motifs by making use of the biological context inherent in a multiple-motif method.

➢ *FingerPRINTScan* (Scordis *et al.*, 1999).

❖**Pfam** is a database of protein domain families. Pfam contains curated multiple sequence alignments for each family and corresponding profile hidden Markov models (HMMs).

➢ *hmmpfam* from the *HMMER2.1* package (by Sean Eddy, eddy@genetics.wustl.edu, http://hmmer.wustl.edu), DeCypher™ (TimeLogic) HMM search.

❖**ProDom** families are built by an automated process based on a recursive use of *PSI-BLAST* homology searches.

➢*BlastProDom.pl* (by Florence Servant, fservant@toulouse.inra.fr) a filter on top of the *Blast* package (Altschul *et al.*, 1997).

❖**SMART** domains are extensively annotated with respect to phyletic distributions, functional class, tertiary structures and functionally important residues. SMART alignments are optimised manually and following construction of corresponding Hidden Markov Models (HMMs).

➢ *hmmpfam* from the *HMMER2.1* package, DeCypher™ (TimeLogic) HMM search.

INTERPROSCAN

InterProScan is a tool that combines different protein signature recognition methods into one resource. The number of signature databases and their associated scanning tools as well as the further refinement procedures increase the complexity of the problem. InterProScan is more than a simple wrapping of sequence analysis applications since it requires performing a considerable data look-up from some databases and program outputs. The need for production scale efficiency and an easy extensibility require a robust and efficient (parallel) internal architecture that can benefit from network distributed computing with the support of UNIX queueing systems.

We developed an SRS-based InterProScan suite as well as the stand-alone Perl-based InterProScan package.

Nowadays SRS (Etzold *et al.*, 1996) has become an integration system for both data retrieval and applications for data analysis that is ideally suited to resolve the data flow complexity in InterProScan. Firstly, InterProScan was implemented using the introduced technique of joining some of the SRS integrated applications into one virtual application that can organize the execution of the underlying steps in an efficient (parallel) manner. Later, we developed a client web interface using the SRS Perl API that is a compromise between the SRS inter-database linking integrity and the simplicity of the user interface, providing 'one-click-away' results.

While the SRS-based InterProScan has several benefits from the close integration with other databases it requires some SRS expertise and is bound to the licensed SRS distribution. To overcome these limitations we decided to develop a stand-alone InterProScan version based on the popular scripting language Perl. The Perl-based InterProScan was intended as an extensible and scalable system optimised to cope with bulk data processing. In the package a Perl-based simple data retrieval system was introduced to provide the required data look-up efficiency and easy extensibility. The system has a modular structure and is designed in an SRS-like fashion. Each of the data description modules defines the data schema of the source text data and the parsing rules. The corresponding Perl module provides an object-oriented interface to the underlying entry attributes. The parsing of the source data into the memory objects happens only once and is done upon request, implementing so-called lazy-parsing. Hierarchical parsing rules are implemented using the recursive-descent approach (Parse-RecDescent package). Fast data retrieval is implemented using the Perl native B-trees indexing (DB_File.pm, based on Berkeley DB). The simple 'one Perl module per data source' organisation makes it possible to reuse the modules in other stand-alone *ad-hoc* solutions. The Perl-based InterProScan is capable of providing post-processed, integrated results in several formats and it could be used as a simple retrieval system for the underlying data.

The tool has become popular in the bioinformatics community. The EBI public web interface serves more than 10 000 interactive requests a month. There are more than 60 installations worldwide of the Perl-based InterProScan package that has already been used to analyse the complete genomes on a production scale.

ACKNOWLEDGEMENTS

We would like to thank Rodrigo Lopez for general support and the mailserver backend as well as Thure Etzold and Henning Hermjakob for useful discussions and ideas.

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. *Nucleic Acids Res.*, **25**, 3389–3402.

Attwood,T.K., Croning,M.D., Flower,D.R., Lewis,A.P., Mabey,J.E., Scordis,P., Selley,J.N. and Wright,W. (2000) PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.*, **28**, 225–227.

Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.

Corpet,F., Gouzy,J. and Kahn,D. (1999) Recent improvements of the ProDom database of protein domain families. *Nucleic Acids Res.*, **27**, 263–267.

Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Meth. Enzymol.*, **266**, 114–128.

Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.

Schultz,J., Copley,R.R., Doerks,T., Ponting,C.P. and Bork,P. (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.*, **28**, 231–234.

Scordis,P., Flower,D.R. and Attwood,T.K. (1999) FingerPRINTScan: intelligent searching of the PRINTS motif database. *Bioinformatics*, **15**, 799–806.

The InterPro Consortium (Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D., Durbin,R., Falquet,L., Fleischmann,W., Gouzy,J., Hermjakob,H., Hulo,N., Jonassen,I., Kahn,D., Kanapin,A., Karavidopoulou,Y., Lopez,R., Marx,B., Mulder,N.J., Oinn,T.M., Pagni,M., Servant,F., Sigrist,C.J. and Zdobnov,E.M.) (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.