BIOINFORMATICS APPLICATIONS NOTE Vol. 16 no. 3 2000
Pages 290–291

PIR: a new resource for bioinformatics

Peter B. McGarvey, Hongzhan Huang, Winona C. Barker, Bruce C. Orcutt, John S. Garavelli, Geetha Y. Srinivasarao, Lai-Su L. Yeh, Chunlin Xiao and Cathy H. Wu

Protein Information Resource, National Biomedical Research Foundation, 3900 Reservoir Road, NW, Washington, DC 20007, USA

Received on August 24, 1999; revised on October 22, 1999; accepted on October 26, 1999

Abstract

Summary: *The Protein Information Resource (PIR) has greatly expanded its Web site and developed a set of interactive search and analysis tools to facilitate the analysis, annotation, and functional identification of proteins. New search engines have been implemented to combine sequence similarity search results with database annotation information. The new PIR search systems have proved very useful in providing enriched functional annotation of protein sequences, determining protein superfamily-domain relationships, and detecting annotation errors in genomic database archives.*

Availability: *http://pir.georgetown.edu/.*

Contact: *mcgarvey@nbrf.georgetown.edu*

Genome sequencing projects have greatly increased the volume and complexity of available molecular data. Fast and easy means for retrieval using all relevant information have become essential for a database search system to be truly useful to the average scientist. With these points in mind, PIR has recently redesigned its Web site to provide a number of new user-friendly search and analysis tools, with the goal of making the search, analysis, and identification of protein sequences quicker and easier.

The Protein Information Resource (PIR), along with the Munich Information Center for Protein Sequences (MIPS), and the Japanese International Protein Sequence Database (JIPID), maintains the PIR-International Protein Sequence Database (PIR-PSD), a comprehensive, annotated, and non-redundant protein sequence database in which entries are classified into family groups (Barker *et al.*, 1999).

The PIR search and analysis system

The PIR search and analysis system consists of a set of search engines of three types:

(1) standard sequence search and analysis engines, which include Pattern Match, BLAST, FASTA, and Multiple Alignment;
(2) interactive text-based search engines;
(3) advanced search engines, which combine sequence similarity with annotation searches as detailed below.

The PIR Domain Similarity Search engine uses FASTA (Pearson and Lipman, 1988) to search domain databases created from domains annotated in the PIR-PSD. The output display includes the domain database searched, the PIR entry containing the annotation, the name and the length of the domain, the overlap with the query sequence, a graphical representation of the region of similarity, and an alignment of the similar regions. Any combination of complete sequences and domains retrieved by the search can be selected and viewed in a multiple alignment generated by CLUSTALW (Thompson *et al.*, 1994) and displayed using MView (Brown *et al.*, 1998). The PIR Global and Domain Similarity Search uses BLAST (Altschul *et al.*, 1997) to search for global similarity to the query sequence and FASTA to search all the PIR domain databases for local similarity. The results are presented in order of highest global alignment score, with all domain hits listed in order directly below. Again, any of the resulting hits can be viewed in a multiple alignment. Figure 1(A) shows a sample output from this search engine. The PIR Annotation-Sorted Similarity Search provides a capability that combines sequence similarity searches with annotation. This program searches the PIR-PSD and displays the best search matches to the query sequence along with the superfamily classification, percent identity, and the overlap length between the query and each hit. The search results and the annotation are presented in a table and users can then select different annotation and sort the table by the selected annotation. Figure 1(B) shows sample output from this search engine. The Integrated Environment for Sequence Analysis provides a new interface that accesses all of the sequence search options described above, and, in addition, provides text search options, complete links to the PDB, COUG, and KEGG databases, and a precompiled FASTA database of the PIR-PSD.

A) PIR Global and Domain Similarity Search

ID	Region	Title	e-value	#aa	%idn	Ov.lap	
T24918	Global	hypothetical protein T14G10.1 - Caenorhabditis ele...	0.0	652			☐
JC4383	Global	3'-phosphoadenosine-5'-phosphosulfate synthetase -...	0.0	610			☐
	0037-0200	ASK adenylylsulfate kin	8e-51	164	66	164	☐
	0211-0605	SAT sulfate adenylyltra	2.5e-102	395	58	419	☐
JW0087	Global	3'-phosphoadenosine-5'-phosphosulfate synthetase -...	0.0	624			☐
	0052-0215	ASK adenylylsulfate kin	1.4e-49	164	64	164	☐
	0226-0620	SAT sulfate adenylyltra	1.8e-98	395	57	410	☐
S44079	Global	sulfate adenylyltransferase (EC 2.7.7.4) met3-1 - ...	e-120	424			☐
	0011-0409	SAT sulfate adenylyltra	1.5e-93	399	55	416	☐

B) PIR Annotation-Sorted Similarity Search

ID	Superfamily	MIPSfamily	Species	Tax. group	Keywords	Title	Score	e-value	#aa	%idn	Ovlap
T24918	--------	--------	Caenorhabditis elegans	Euk/animal	-------	hypothetical protein T14G10.1 - Caenorhabditis ele...	1260	0.0	652	100	652
JC4383	SF1552.0	FAM20020	Urechis caupo	Euk/animal	multifunctional enzyme; nucleotidyltransferase; P-loop; phosphoprotein; phosphotransferase	3'-phosphoadenosine-5'-phosphosulfate synthetase -...	709	0.0	610	59	624
JW0087	SF1552.0	FAM20020	Homo sapiens (man)	Euk/animal	multifunctional enzyme; nucleotidyltransferase; P-loop; phosphoprotein; phosphotransferase	3'-phosphoadenosine-5'-phosphosulfate synthetase -...	690	0.0	624	57	626
S44079	SF1553.0	FAM20021	Solanum tuberosum (potato)	Euk/plant	nucleotidyltransferase	sulfate adenylyltransferase (EC 2.7.7.4) met3-1 - ...	430	e-120	424	54	432

Fig. 1. Sample output from the PIR Global and Domain Similarity Search (A) and the PIR Annotation- Sorted Similarity Search (B) with the option to display all annotation selected. Both searches used the same protein from the *C. elegans* genome (see text).

Example

By combining sequence and annotation searches, the PIR search system has proved very useful in rapidly determining protein superfamily-domain relationships, annotating the database and correcting annotation from genomic database archives. An example is shown in Figure 1, where a protein from the *Caenorhabditis elegans* genome was analyzed. The *C. elegans* protein entered the PIR database entitled 'hypothetical' because no unambiguous product or function was annotated in the original EMBL submission. The Global and Domain Search shown in Figure 1(A) shows at a glance that the protein has high end-to-end similarity to multifunctional enzymes from *Urechis caupo* (spoonworm) and *Homo sapiens* and contains functional domains for both adenylylsulfate kinase activity and sulfate adenylyltransferase activity. It also shows the similarity to the single function sulfate adenylyltransferase PIR:S44079. An Annotation-Sorted Similarity Search of the same sequence reveals additional classification and annotation information, including superfamily, family, and keywords. Thus, two quick searches have revealed the following information to be added to entry PIR:T24918, a correct title, superfamily, family, keywords, and homology domains.

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.*, **25**, 3389–3402.

Barker,W.C., Garavelli,J.S., McGarvey,P., Marzec,C.R., Orcutt,B.C., Srinivasarao,G.Y., Yeh,L.-S.L., Ledley,R.S., Mewes,H.-W., Pfeiffer,F., Tsugita,A. and Wu,C. (1999) The PIR-International Protein Sequence Database. *Nucl. Acids Res.*, **27**, 39–43.

Brown,N.P., Leroy,C. and Sander,C. (1998) MView: A web compatible database search or multiple alignment viewer. *Bioinformatics*, **14**, 380–381.

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

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