BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 13 2005, pages 3049–3050
doi:10.1093/bioinformatics/bti439

Genome analysis

pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry

Dequan Li[1,2,*] , Yan Fu[1,2], Ruixiang Sun[1], Charles X. Ling[3], Yonggang Wei[1], Hu Zhou[4], Rong Zeng[4], Qiang Yang[5], Simin He[1] and Wen Gao[1,2]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China, [2]Graduate School of Chinese Academy of Sciences, Beijing 100039, China, [3]Department of Computer Science, The University of Western Ontario, Canada, [4]Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China and [5]Department of Computer Science, Hong Kong University of Science and Technology, Hong Kong, China

Received on November 12, 2004; revised on March 29, 2005; accepted on April 6, 2005
Advance Access publication April 7, 2005

ABSTRACT

Summary: Research in proteomics requires powerful database-searching software to automatically identify protein sequences in a complex protein mixture via tandem mass spectrometry. In this paper, we describe a novel database-searching software system called pFind (peptide/protein Finder), which employs an effective peptide-scoring algorithm that we reported earlier. The pFind server is implemented with the C++ STL, .Net and XML technologies. As a result, high speed and good usability of the software are achieved.

Availability: The pFind web server can be freely accessed through the website http://pfind.jdl.ac.cn. In this website, the compiled binary of the local version is also available.

Contact: dqli@jdl.ac.cn

INTRODUCTION

Tandem mass spectrometry (MS/MS) has become one of the most popular and successful analytical techniques for protein identification in proteomics. In MS/MS, peptides digested from complex protein mixtures are ionized and fragmented by collision-induced dissociation in the mass spectrometer to form peptide tandem mass spectra (MS/MS spectra). To identify the unknown amino acid sequences of peptides, database searching, currently the most widely used approach in high-throughput proteomics studies, is often employed. In this approach, the theoretical MS/MS spectra are predicted from the peptide sequences in a database and are compared with the experimental one to score the candidate peptide sequences. A number of top-ranked peptide sequences are returned for further identification of proteins.

We have developed a peptide/protein identification software system, pFind, using our peptide-scoring algorithm KSDP reported earlier (Fu *et al*., 2004). pFind is available in two versions, a web server or a locally installed version. The former allows worldwide users to access pFind via Internet freely. The latter, on the other hand, runs independently on the user's machine, and it is more powerful to process large-scale data at a higher speed than the web server.

As an instrument-independent software system, pFind currently is not integrated with any commercial spectrometers. It can serve as a complement or alternative to the instrument-integrated identification software in the proteomics workflow. The heart of pFind is the peptide-scoring algorithm KSDP, which makes use of correlative information among fragment ions in MS/MS spectra to reduce stochastic mismatches. In our website, new experimental results are presented to show that pFind outperforms some popular software tools, e.g. SEQUEST and Sonar MS/MS, in terms of the Top-1 (peptide sequence which gets the highest score) identification accuracy.

SYSTEM DESCRIPTION

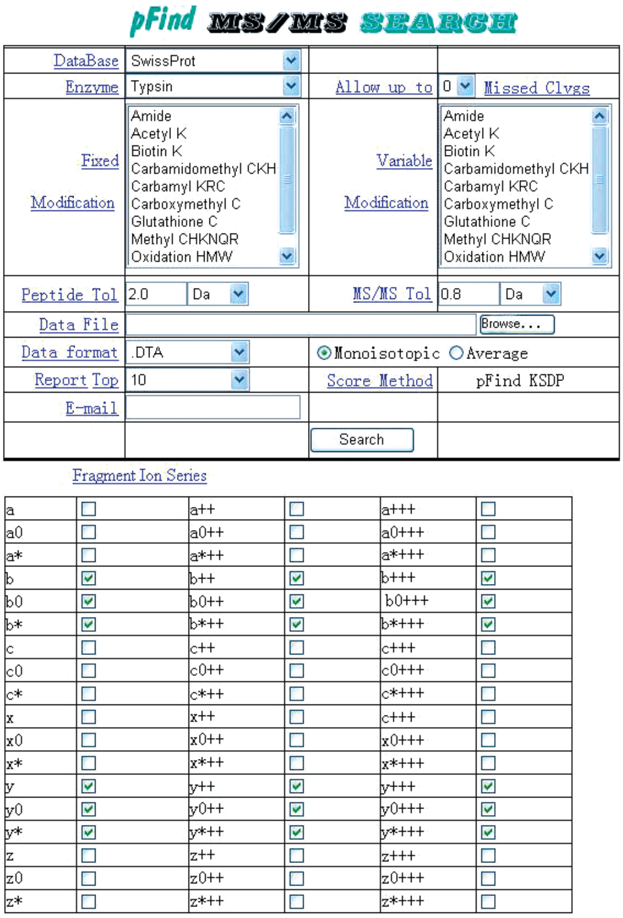
Here we briefly describe the implementation, usage and features of the pFind web server. The main differences between the server and the local version of pFind are also pointed out. Detailed descriptions can be found on the pFind website.

pFind provides a simple and user-friendly interface. The input to pFind consists of two parts: the search parameters and the MS/MS data. The search parameters are illustrated in Figure 1. The commonly used MS/MS data formats, such as DTA and PKL, are supported. In particular, the pFind web server also supports compressed archives containing multiple DTA or PKL files.

For the DTA file, pFind stores the search results in XML (eXtensible Markup Language) containing the top-ranked peptide sequences and their matching details, and returns it to the user's browser. The expect-value is utilized in pFind for significance evaluation of the search results. For the PKL file, pFind also generates the protein identification result in addition to peptide identification results and returns it to the user. For the compressed input data, the pFind web server compresses all the search result files into one archive and sends it to the e-mail box of the users.

To our knowledge, support for compressed input and output data is a unique feature of pFind in comparison with other web-based identification tools. The advantages of this feature are 2-fold: first, it can significantly reduce the time needed for data transmission through the network; second, it allows the users to operate in a batch mode.

[*]To whom correspondence should be addressed.



pFind MS/MS SEARCH

DataBase	SwissProt		
Enzyme	Typsin	Allow up to	0 Missed Clvgs
Fixed Modification	Amide Acetyl K Biotin K Carbamidomethyl CKH Carbamyl KRC Carboxymethyl C Glutathione C Methyl CHKNQR Oxidation HMW	Variable Modification	Amide Acetyl K Biotin K Carbamidomethyl CKH Carbamyl KRC Carboxymethyl C Glutathione C Methyl CHKNQR Oxidation HMW
Peptide Tol	2.0 Da	MS/MS Tol	0.8 Da
Data File	Browse...		
Data format	.DTA	● Monoisotopic ○ Average	
Report Top	10	Score Method	pFind KSDP
E-mail			
		Search	

Fragment Ion Series

a	☐	a++	☐	a+++	☐
a0	☐	a0++	☐	a0+++	☐
a*	☐	a*++	☐	a*+++	☐
b	☑	b++	☑	b+++	☑
b0	☑	b0++	☑	b0+++	☑
b*	☑	b*++	☑	b*+++	☑
c	☐	c++	☐	c+++	☐
c0	☐	c0++	☐	c0+++	☐
c*	☐	c*++	☐	c*+++	☐
x	☐	x++	☐	c+++	☐
x0	☐	x0++	☐	x0+++	☐
x*	☐	x*++	☐	x*+++	☐
y	☑	y++	☑	y+++	☑
y0	☑	y0++	☑	y0+++	☑
y*	☑	y*++	☑	y*+++	☑
z	☐	z++	☐	z+++	☐
z0	☐	z0++	☐	z0+++	☐
z*	☐	z*++	☐	z*+++	☐

Fig. 1. The interface of the pFind web server. The search parameters include the protein sequence database in the FASTA format to be searched (Database), the enzyme and the maximum number of missed cleavage sites for digesting protein sequences in the database (Enzyme and Missed Clvges), the post-translational modifications (Fixed and Variable Modification), the tolerances for mass matching of precursor ions and fragment ions (Peptide Tol and MS/MS Tol), mass accuracy (Monoisotopic or Average) and the fragment ion series used for peptide scoring (Fragment Ion Series). All the parameter titles are hyperlinked to detailed explanations.

With the .Net and XML technologies, the pFind web server can be rapidly constructed and integrated. At the server side, ASP.Net is used to receive the queries from the user's browser, transmit them to the scoring modules of pFind and return the search results in XML to the user's browser. The XSLT (eXtensible Stylesheet Language Transformations) is used to filter and sort XML documents and define the display style. Moreover, users can save the XML documents to their own machines for further data management, and they can design their own XSLT files to view the XML documents in their desired styles.

Although pFind is currently running on the MS Windows platform, the core modules of pFind are coded in C++ with standard template library (STL), which can simplify the future migration to other operating systems, such as Linux and Solaris.

NCBI BLAST and EBI IPI (Kersey *et al*., 2004) protein sequence databases are used for searching. All of these databases are well indexed and memory-mapped on the server for high-speed execution.

Compared with the pFind web server, the local version has a more powerful batch-mode function. Users can directly specify the directory that contains all the MS/MS spectra to be searched. The local version also comes with a user-friendly wizard for importing and indexing the protein databases to be searched. Besides, multiple databases installed on the local machine can be searched at the same time and several other scoring functions in addition to KSDP can be selected.

To help new users get started with pFind quickly, extensive help is provided on the website under 'For new users'.

ACKNOWLEDGEMENTS

This work was supported by National Key Basic Research & Development Program (973) of China under Grant No. 2002CB713807 and National Key Technologies R&D Programme under Grant No.2004BA711A21. We thank Yiqiang Chen, Xiaobiao Wang, Cui Zou and Haipeng Wang for the valuable discussions.

REFERENCES

Fu,Y. *et al*. (2004) Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. *Bioinformatics*, **20**, 1948–1954.

Kersey,P.J. *et al*. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, **4**, 1985–1988.