

# pScan User Guide

## **Introduction:**

pScan is a flexible tool that helps biologists to preprocess protein sequence databases in proteomics research. Besides the commonly used functions, such as sequence pattern-matching, building decoy databases, and converting protein sequence databases to peptide sequence databases, pScan also supports querying and substituting of protein entries based on the regular expression, creating customized databases, and conducting statistical characterization of the databases. pScan can greatly help biologists to improve the design of proteomics experiments and to facilitate the database search and analysis by making full use of the information content contained in the sequence databases.

## **FUNCTIONALITIES**

### **Display, Query and Substitute Sequences:**

pScan allows biologists to edit, query and substitute the accession ID, the description information and the sequence for each entry in the FASTA file, collectively or separately, which are based on various types of regular expressions.

### **Create Customized Databases:**

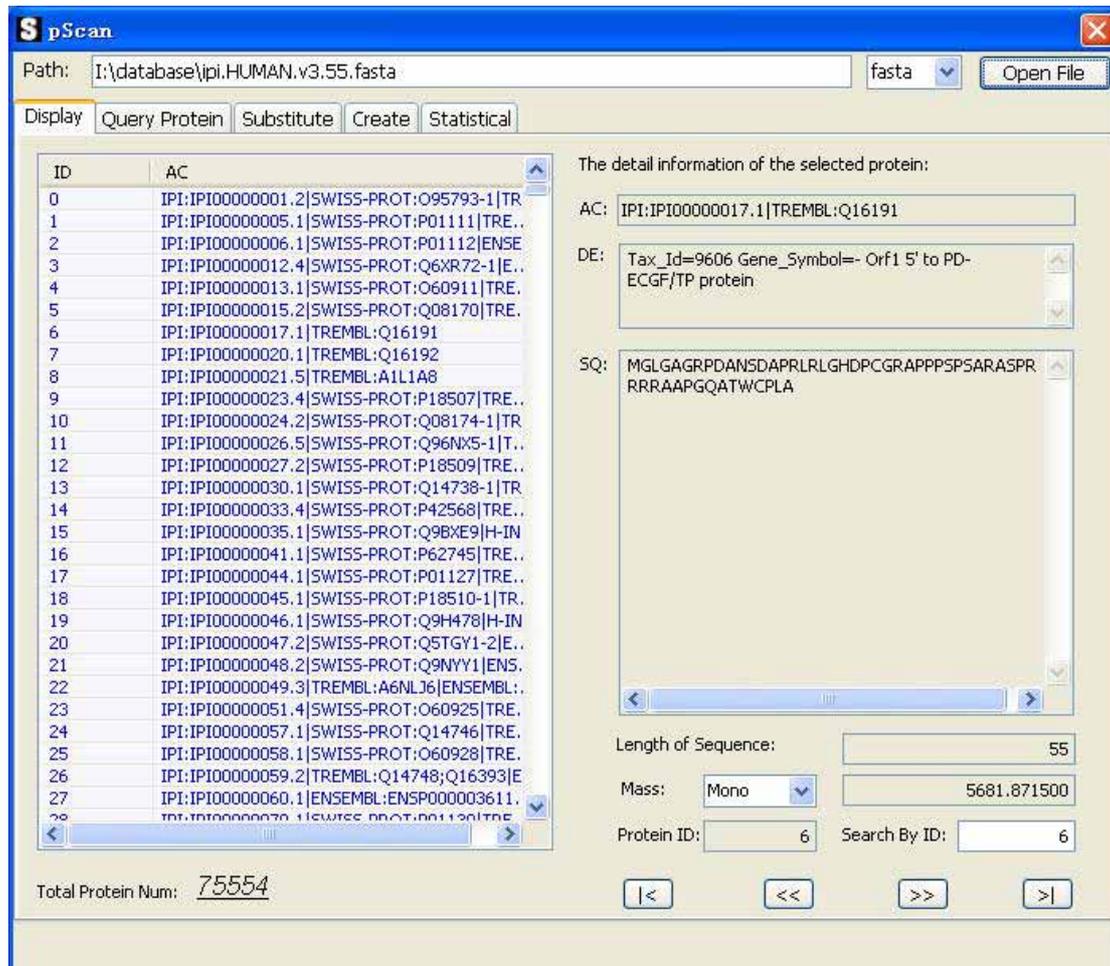
pScan can be used to create some customized databases, e.g., sub-species databases, N- and C-terminal sequence databases, and target-decoy databases with different decoy strategies, which are very helpful for peptide identification in database search engines, such as pFind (<http://pfind.ict.ac.cn>), SEQUEST and Mascot.

### **Conduct Statistical Characterization:**

pScan also supports the statistical characterization of the protein sequence databases, for example, the ratio of digested peptides with a specific amino acid to all peptide sequences, the ratio of digested peptides with special modification patterns (e.g., 'NXS/T/C' in glycosylation and 'S/T/Y' in phosphorylation) to all peptide sequences, and the distribution of mass values of all peptides (with or without modifications) obtained from digestion of the proteins.

## Display, Query and Substitute Sequences

**Displaying** protein's detail information (accession ID, the description information and the sequence for each entry in the FASTA file), collectively or separately, counting proteins and mass calculating of sequence are performed in pScan.



The screenshot shows the pScan software interface. The main window has a title bar with the 'pScan' logo and a close button. Below the title bar, there is a 'Path' field containing 'I:\database\ipi.HUMAN.v3.55.fasta' and a file type dropdown set to 'fasta'. An 'Open File' button is located to the right of the path field. Below the path field, there are five tabs: 'Display', 'Query Protein', 'Substitute', 'Create', and 'Statistical'. The 'Display' tab is currently selected.

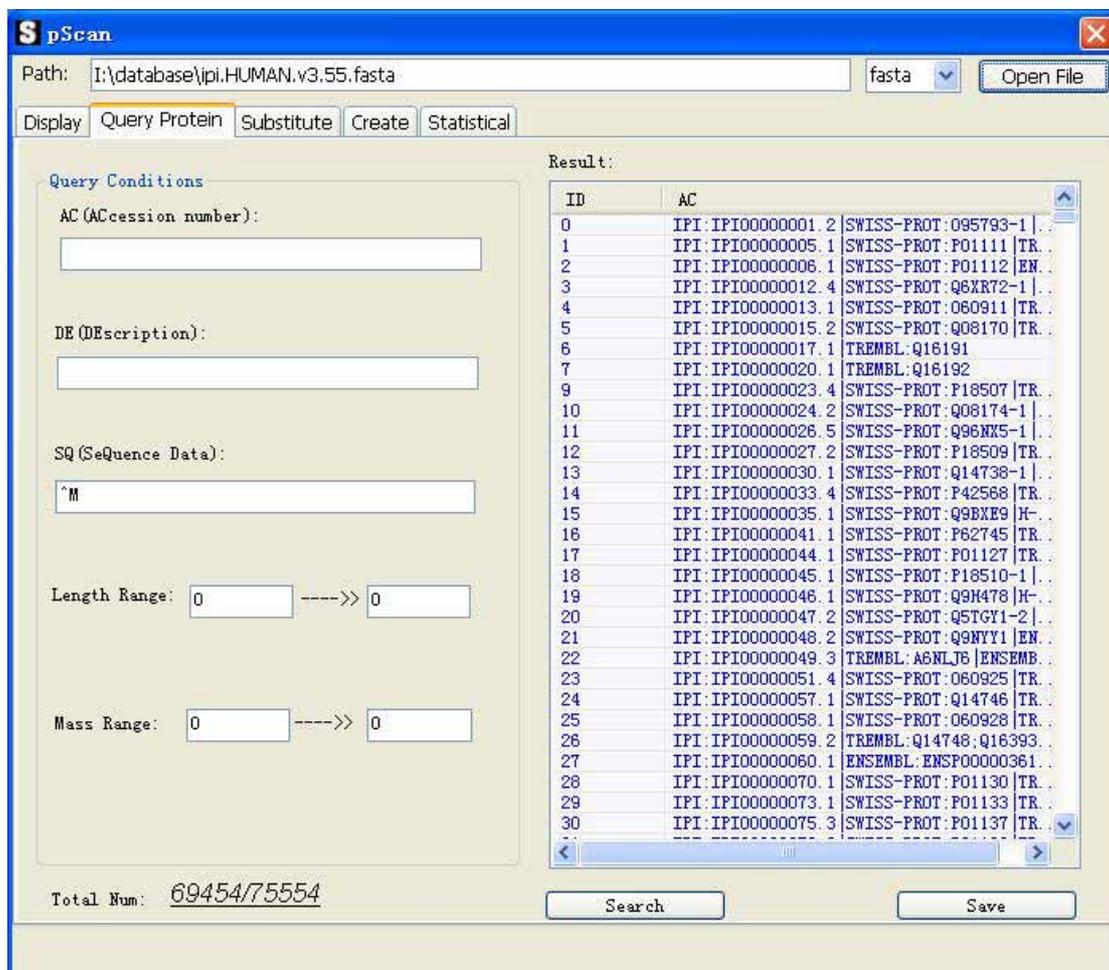
The 'Display' tab shows a list of protein entries with columns for 'ID' and 'AC'. The list contains 28 entries, with the 6th entry (ID 6) selected. The 'AC' column contains accession numbers and database identifiers, such as 'IPI:IPI00000017.1|TREMBL:Q16191'.

To the right of the list, there is a section titled 'The detail information of the selected protein:'. This section contains three text boxes: 'AC' (IPI:IPI00000017.1|TREMBL:Q16191), 'DE' (Tax\_Id=9606 Gene\_Symbol=- Orf1 5' to PD-ECGF/TP protein), and 'SQ' (MGLGAGRPDANSDAPRLRLGHDPGCRAPPPSPSARASPRRRRAAPGQATWCPLA). Below these text boxes, there are three input fields: 'Length of Sequence' (55), 'Mass' (Mono, 5681.871500), and 'Protein ID' (6). A 'Search By ID' field also contains the value 6. At the bottom of the interface, there are four navigation buttons: '<|', '<<', '>>', and '>|'. The 'Total Protein Num:' is displayed as 75554.

| ID | AC   |
|----|--|
| 0  | IPI:IPI00000001.2 SWISS-PROT:O95793-1 TR   |
| 1  | IPI:IPI00000005.1 SWISS-PROT:P01111 TRE..  |
| 2  | IPI:IPI00000006.1 SWISS-PROT:P01112 ENSE   |
| 3  | IPI:IPI00000012.4 SWISS-PROT:Q6XR72-1 E..  |
| 4  | IPI:IPI00000013.1 SWISS-PROT:O60911 TRE..  |
| 5  | IPI:IPI00000015.2 SWISS-PROT:Q08170 TRE..  |
| 6  | IPI:IPI00000017.1 TREMBL:Q16191            |
| 7  | IPI:IPI00000020.1 TREMBL:Q16192            |
| 8  | IPI:IPI00000021.5 TREMBL:A1L1A8            |
| 9  | IPI:IPI00000023.4 SWISS-PROT:P18507 TRE..  |
| 10 | IPI:IPI00000024.2 SWISS-PROT:Q08174-1 TR   |
| 11 | IPI:IPI00000026.5 SWISS-PROT:Q96NX5-1 T..  |
| 12 | IPI:IPI00000027.2 SWISS-PROT:P18509 TRE..  |
| 13 | IPI:IPI00000030.1 SWISS-PROT:Q14738-1 TR   |
| 14 | IPI:IPI00000033.4 SWISS-PROT:P42568 TRE..  |
| 15 | IPI:IPI00000035.1 SWISS-PROT:Q9BXE9 H-IN   |
| 16 | IPI:IPI00000041.1 SWISS-PROT:P62745 TRE..  |
| 17 | IPI:IPI00000044.1 SWISS-PROT:P01127 TRE..  |
| 18 | IPI:IPI00000045.1 SWISS-PROT:P18510-1 TR   |
| 19 | IPI:IPI00000046.1 SWISS-PROT:Q9H478 H-IN   |
| 20 | IPI:IPI00000047.2 SWISS-PROT:Q5TGY1-2 E..  |
| 21 | IPI:IPI00000048.2 SWISS-PROT:Q9NYY1 ENS.   |
| 22 | IPI:IPI00000049.3 TREMBL:A6NLJ6 ENSEMBL:.. |
| 23 | IPI:IPI00000051.4 SWISS-PROT:O60925 TRE..  |
| 24 | IPI:IPI00000057.1 SWISS-PROT:Q14746 TRE..  |
| 25 | IPI:IPI00000058.1 SWISS-PROT:O60928 TRE..  |
| 26 | IPI:IPI00000059.2 TREMBL:Q14748;Q16393 E   |
| 27 | IPI:IPI00000060.1 ENSEMBL:ENSP000003611..  |
| 28 | IPI:IPI00000070.1 SWISS-PROT:P01130 TR     |

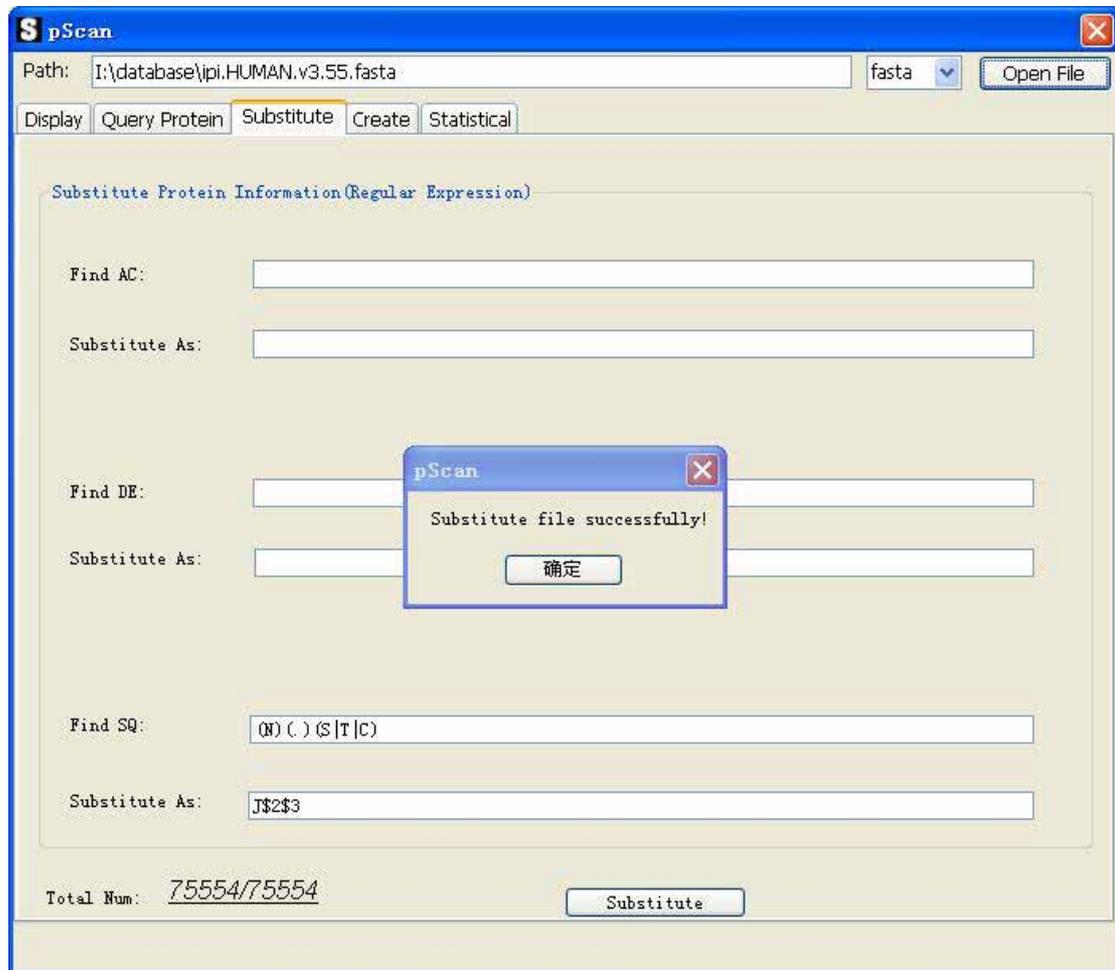
Fig. 1.

**Querying** species-specific proteins, restricted to the conditions of the accession ID, the description information, the sequence, the length range of sequence, and the mass range of sequence, collectively or separately.



**Fig. 2.** To calculate the number of proteins that begin with the amino acid of 'M', biologists can use the regular expression '^M' to search by pScan.

**Substituting** protein's detail information (accession ID, the description information and the sequence for each entry in the FASTA file), collectively or separately.



**Fig. 3.** The character 'N' in the glycosylation sequon 'NXS/T/C' is substituted as 'J' by submitting the old pattern 'N\*(S|T|C)' and the new pattern 'J\*(S|T|C)' to pScan, which is a commonly used method for mass spectral identification of N-linked glycopeptides.

## Create Customized Databases

pScan can be used to create, sub-species databases, N- and C-terminal sequence databases, and target-decoy databases with different decoy strategies.

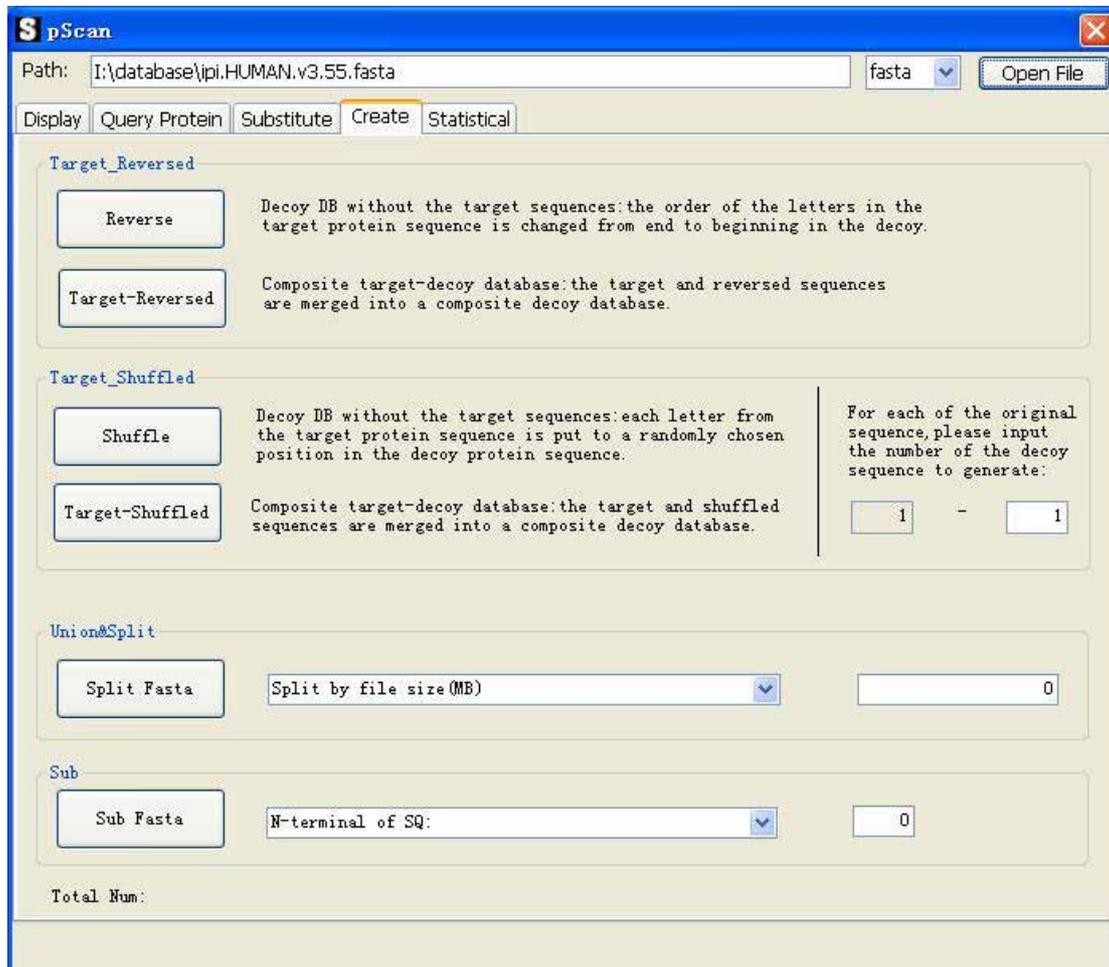
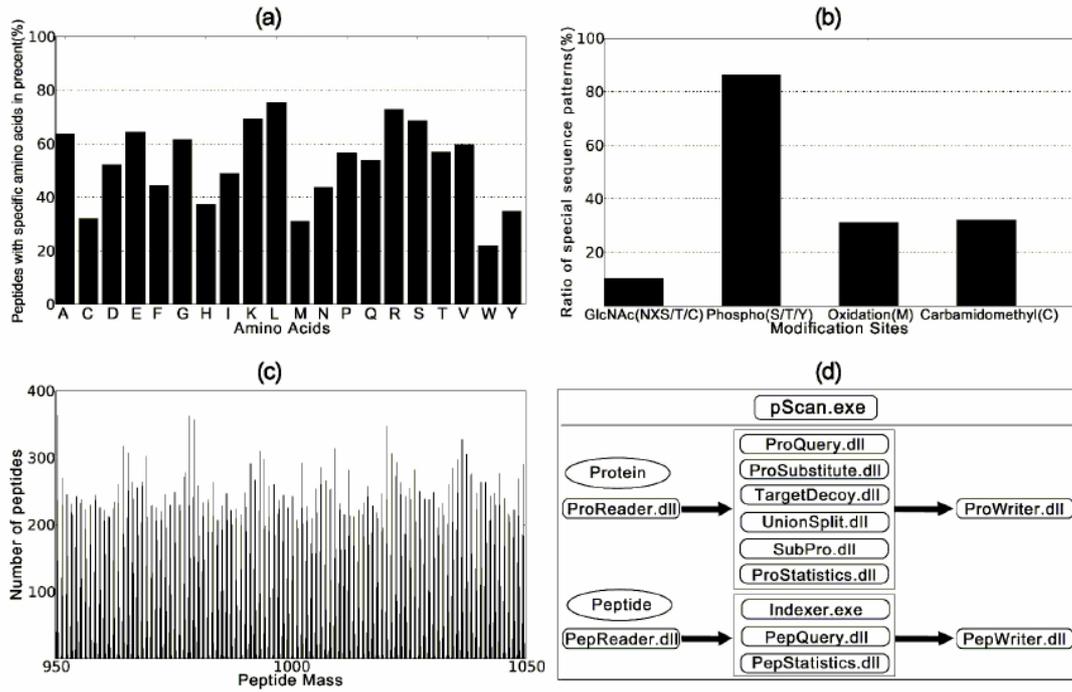


Fig. 4.

## Conduct Statistical Characterization

Currently, pScan supports the statistical characterization of the ratio of digested peptides with a specific amino acid to all peptide sequences, the ratio of digested peptides with special modification patterns (e.g., 'NXS/T/C' in glycosylation and 'S/T/Y' in phosphorylation) to all peptide sequences, and the distribution of mass values of all peptides (with or without modifications) obtained from digestion of the proteins.

Fig. 5.



Database Statistical Characterization and The Robust Framework.

Fig. 6.