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.

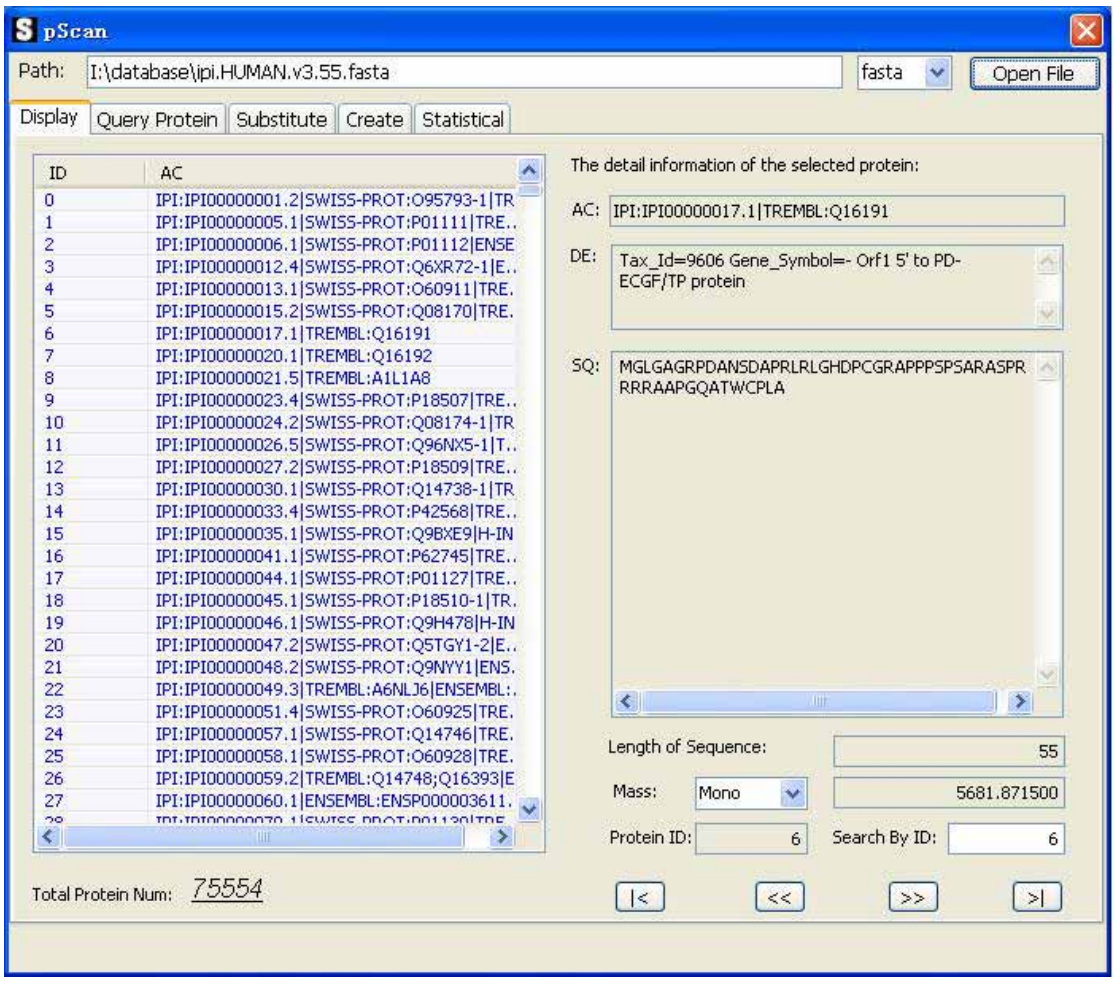


Fig. 1.

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.

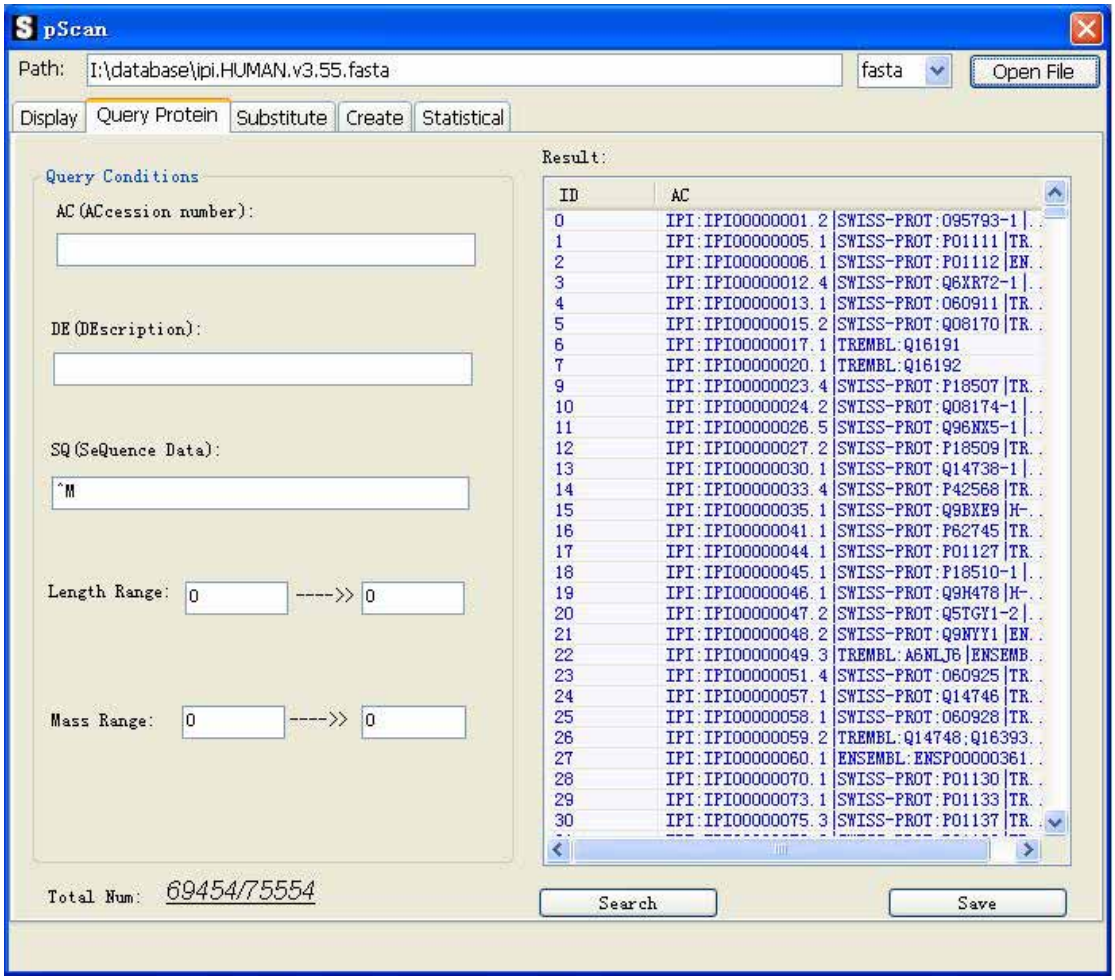


Fig. 2. To calculate the number of proteins that begin with the amino acid of ‘M’, biologists can use the regular expression ‘^M’ to search by pScan.

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

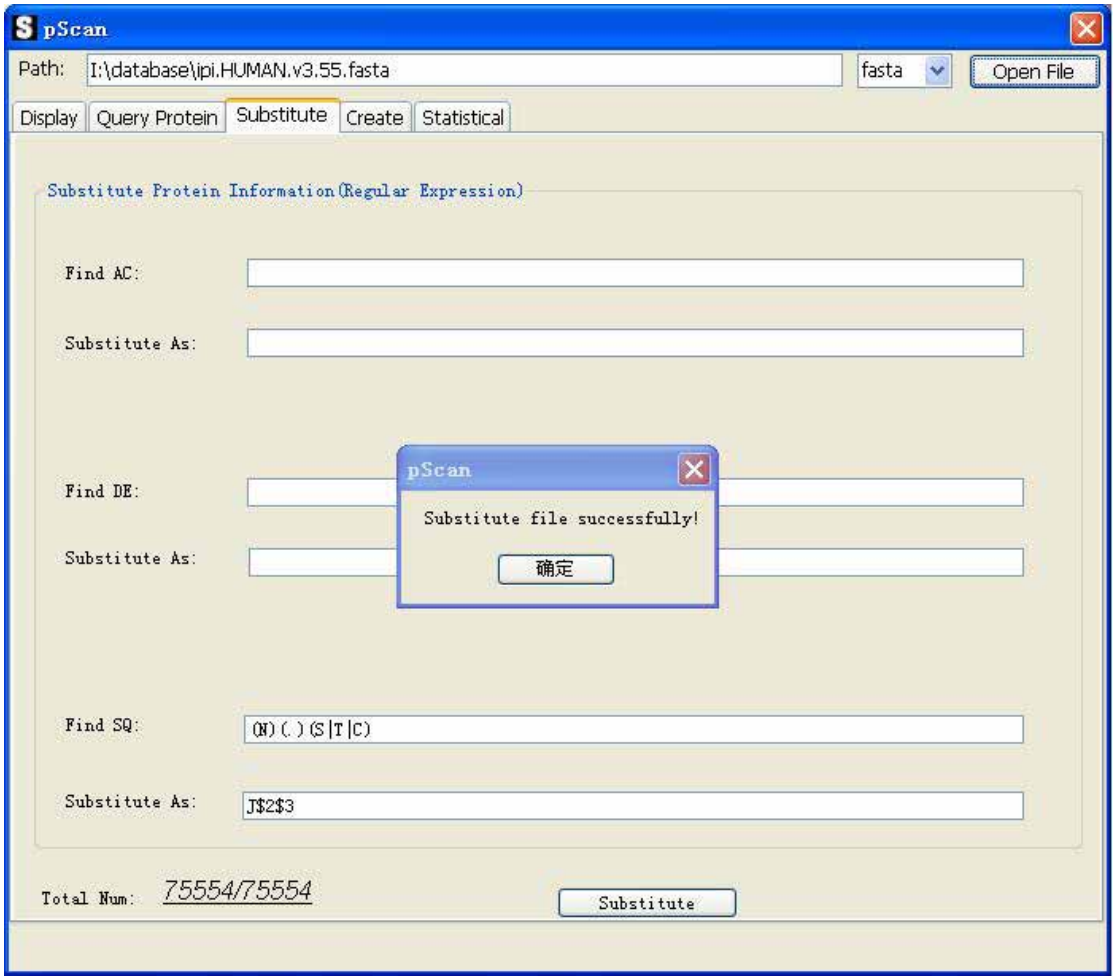


Fig. 3. The character 'N' in the glycosylation sequon 'NXS/T/C' is substituted as 'J' by submitting the old pattern 'N*(S|T|C)' and the new pattern 'J*(S|T|C)' to pScan, which is a commonly used method for mass spectral identification of N-linked glycopeptides.

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.

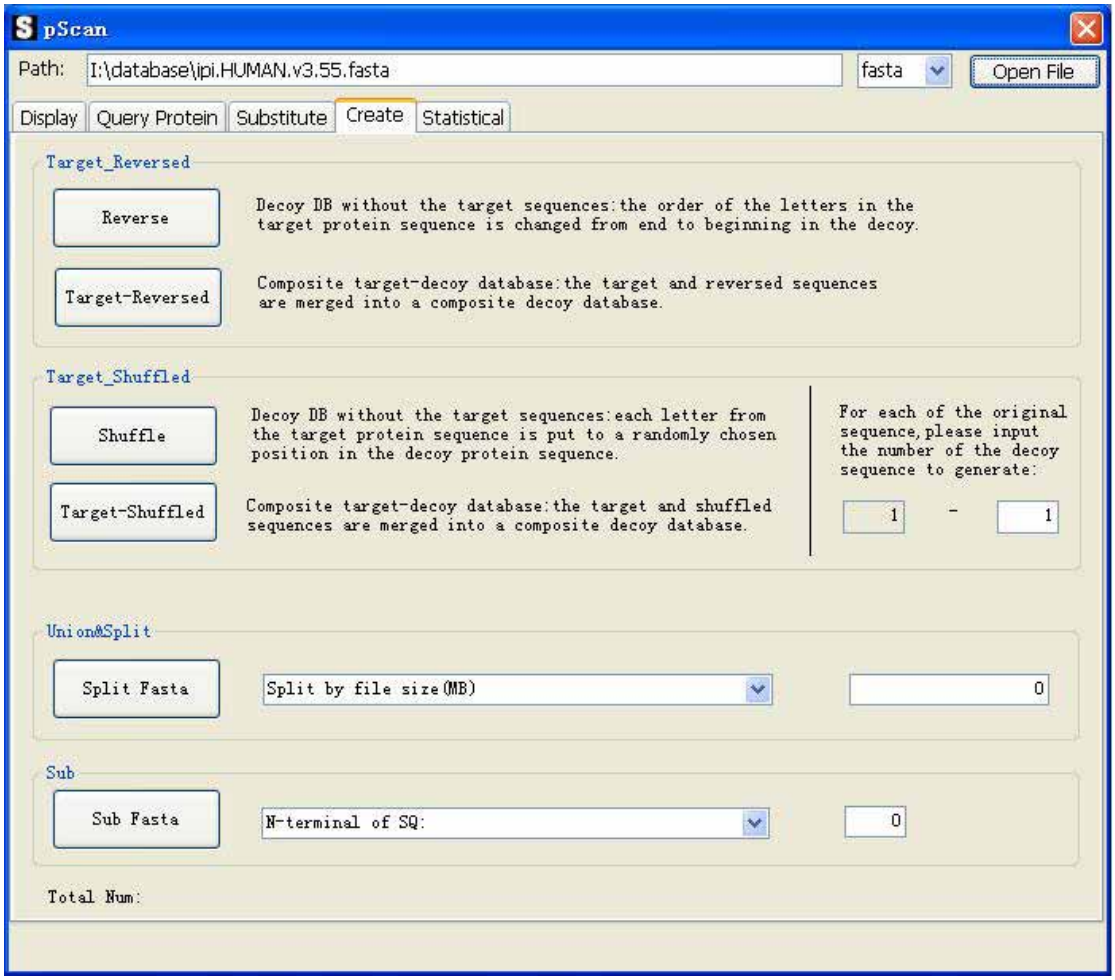


Fig. 4.

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.

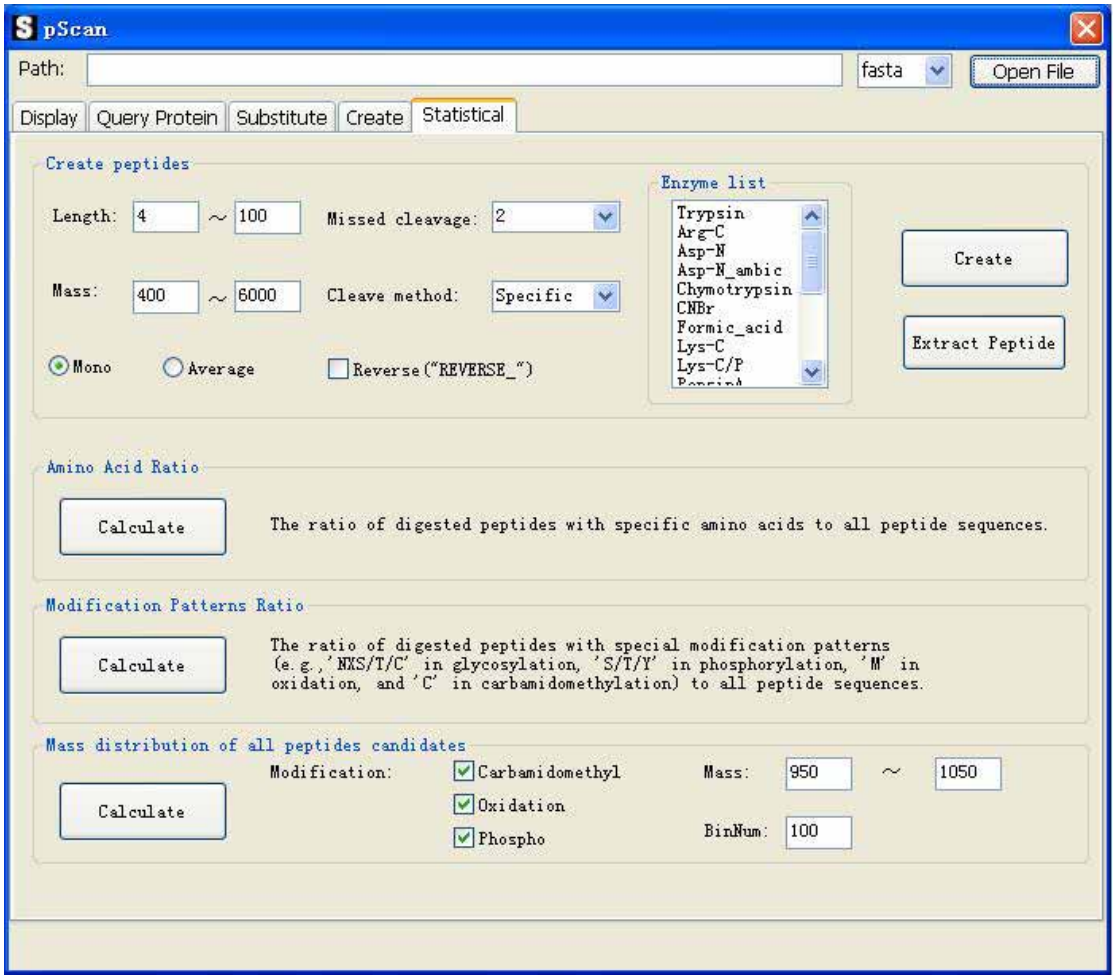
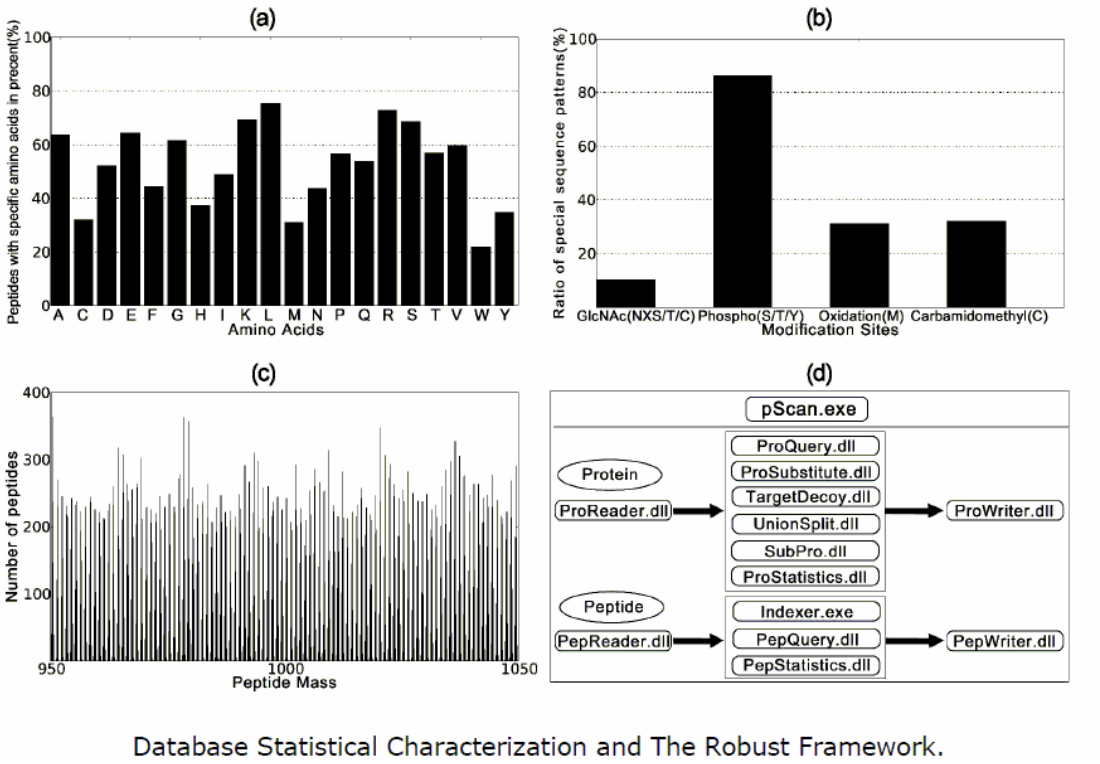


Fig. 5.



Database Statistical Characterization and The Robust Framework.

Fig. 6.