# 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.

th: I:\	database\ipi.HUMAN.v3.55.fasta		fasta 👱	🖉 Open Fil
splay Q	uery Protein Substitute Create Statistic	al		
ID	AC	Tt 🔊	edetail information of the selected protein:	
0	IPI:IPI00000001.2 SWISS-PROT:095793-1	TR .		
1	IPI:IPI00000005.1 SWISS-PROT:P01111 TR	E., A	IPI:IPI00000017.1[TREMBL:Q16191	
2	IPI:IPI00000006.1 SWISS-PROT:P01112 EN	ISE		
3	IPI:IPI00000012.4 SWISS-PROT:Q6XR72-1	E., DE	Tax_Id=9606 Gene_Symbol=- Orf1 5' to PD-	1
4	IPI:IPI00000013.1 SWISS-PROT:060911 TF	RE.	ECGF/TP protein	
5	IPI:IPI00000015.2 SWISS-PROT:008170 TF	RE.		142
6	IPI:IPI00000017.1 TREMBL:Q16191		1	1.00
7	IPI:IPI0000020.1 TREMBL:Q16192			
8	IPI:IPI00000021.5 TREMBL:A1L1A8	50	MGLGAGRPDANSDAPREREGHDPCGRAPPPSPSA	RASPR S
9	IPI:IPI00000023.4 SWISS-PROT:P18507 TR	E	RRRAAPGQATWCPLA	
10	IPI:IPI00000024.2[SWISS-PROT:008174-1]	TR		
11	IPI:IPI00000026.5[SWISS-PROT:Q96NX5-1]	T		
12	IPI:IPI00000027.2 SWISS-PROT:P18509 TR	E		
13	IPI:IPI00000030.1  SWISS-PROT:Q14738-1	TR		
14	IPI:IPI00000033.4 SWISS-PROT:P42568 TR	E.,		
15	IPI:IPI00000035.1 SWISS-PROT:Q9BXE9 H-	IN		
16	IPI:IPI00000041.1 SWISS-PROT:P62745 TR	E.,		
17	IPI:IPI00000044.1 SWISS-PROT:P01127 TR	E		
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 Ef	NS.		1
22	IPI:IPI00000049.3 TREMBL:A6NL36 ENSEMB	3L:.		
23	IPI:IPI00000051.4 SWISS-PROT:060925 TF	RE.		
24	IPI:IPI00000057.1 SWISS-PROT:Q14746 TF	RE.	Longth of Sequences	
25	IPI:IPI00000058.1 SWISS-PROT:060928 TF	RE.	Length of Sequence:	55
26	IPI:IPI00000059.2 TREMBL:Q14748;Q16393	3 E	Marca Marca	01 071500
27	IPI:IPI0000060.1 ENSEMBL:ENSP00000361	1. 🥃		01.0/1500
20	TRUTTOTOTOTOTO 1 SEWTES DE OT DOT 1 20170		Protoin TD: C Search Bu TD:	
<u> </u>			Protein ID: 6 Search By ID:	6
	76664		and the second second	-
otal Prote	ein Num: <u>73554</u>			

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.

and a							Transaction of the			
th:	I:\database\ipi.H	HUMAN.v3.5	5.fasta				fasta	🖌 🗌 Open Fil		
play	Query Protein	Substitute	Create	Statistical						
0	C. March				Result:					
Quer.	y conditions				ID	AC		~		
AC (	ACcession numbe	r):			0	TPT TPT00000001 2	ISWISS-PROT	095793-1		
1					1	IPI: IPI0000005.1	SWISS-PROT	P01111  TR.		
					2	IPI: IPI0000006.1	SWISS-PROT	P01112 EN.		
					3	IPI: IPI00000012.4	SWISS-PROT	06XR72-1		
					4	IPI:IPI00000013.1 SWISS-PROT:060911 TR.				
DE ÍT	Freerintion				5	IPI:IPI00000015.2	SWISS-PROT	Q08170 TR.		
DE (L	iscription).				6	IPI: IPI00000017.1	TREMBL: Q16!	191		
					7	IFI: IPI0000020.1	TREMBL: Q16!	192		
					9	IPI: IPI0000023.4	SWISS-PROT	: P18507  TR		
					10	IPI: IPI0000024.2	SWISS-PROT	Q08174-1		
					11	IPI: IPI00000026.5  SWISS-PROT: Q96NX5-1				
SQ (S	SeQuence Data):				12	IPI: IPI0000027.2	SWISS-PROT	P18509 TR.		
					13	IPI:IPI0000030.1	SWISS-PROT	Q14738-1		
^ <u>M</u>				14	IPI:IPI0000033.4	SWISS-PROT	P42568 TR.			
					15	IPI:IPI0000035.1	SWISS-PROT	Q9BXE9 H		
					16	IPI:IPI00000041.1	SWISS-PROT	:P62745 TR.		
					17	IPI:IPI0000044.1	SWISS-PROT	: P01127  TR		
6 3	ranazio da permite				18	IPI:IPI0000045.1	SWISS-PROT	P18510-1		
Leng	th Range: 0		>> 0		19	IPI:IPI0000046.1	SWISS-PROT:	Q9H478 H		
				13	20	IPI:IPI0000047.2	SWISS-PROT:	Q5TGY1-2		
					21	IPI:IPI0000048.2	SWISS-PROT	Q9NYY1 EN.		
					22	IPI:IPI00000049.3	TREMBL: AGNI	J6 ENSEMB.		
					23	IPI:IPI00000051.4	SWISS-PROT	:060925 TR		
	2008 B	1.5 52		1.1	24	IPI:IPI00000057.1	SWISS-PROT	Q14746 TR		
Mass	: Range: 0>> 0				25	IPI:IPI0000058.1	SWISS-PROT	.060928  TR		
		18	1	15	26	IPI: IPI00000059.2	TREMBL: Q14	748;Q16393		
					27	IPI: IPI0000060. 1	ENSEMBL: ENS	SP00000361		
					28	1P1:1P10000070.1	SWISS-PROT	POILIGO TR.		
					29	TP1: TP100000073.1	SWISS-PRUT	POILOT TR.		
					30	IFI:IF10000075.3	SWISS-PRUT	FOIIST IK.		
					<		and the second second	<b>N</b>		
Tota	1 Num: <u>69</u> 45-	4/75554			Same		-	Some		
	water a second state of the second state of th				Search	a		Save		

**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.

ath:	I:\database\ipi.H	(:\database\ipi.HUMAN.v3.55.fasta						
		Culeatituite		(accurate the				
splay	Query Protein	Substitute	Create	Statistical				
Subs	stitute Protein	Information	(Regular	Expression)				
Fi	nd AC:							
Su	ıbstitute As:	-						
Fi Su	nd DE: ibstitute As:			DScan 区 Substitute file successfully!				
			L					
Fi	nd SQ:	040(.)(3)	r (c)					
Su	ubstitute As:	J\$2\$3						

**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.

🖥 pScan		
Path: I:\database\ipi.I	HUMAN.v3.55.fasta	fasta 👻 🛛 Open File
Display Query Protein	Substitute Create Statistical	
Target_Reversed	Decoy DB without the target sequences:the order of the le target protein sequence is changed from end to beginning	etters in the in the decoy.
Target-Reversed	Composite target-decoy database:the target and reversed s are merged into a composite decoy database.	sequences
Target_Shuffled		
Shuffle	Decoy DB without the target sequences:each letter from the target protein sequence is put to a randomly chosen position in the decoy protein sequence.	For each of the original sequence, please input the number of the decoy sequence to generate:
Target-Shuffled	Composite target-decoy database:the target and shuffled sequences are merged into a composite decoy database.	
Union&Split		
Split Fasta	Split by file size(MB)	0
Sub		
Sub Fasta	N-terminal of SQ:	0
Total Num:		

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.

<b>S</b> pScan					
Path:				fasta 👻	Open File
Display Query Protein Si	ubstitute Create St	atistical.			-
Create peptides Length: 4 ~ 14 Mass: 400 ~ 64 Image: 400 Average Amino Acid Ratio	00 Missed cleav 000 Cleave metho e Reverse (")	age: 2	Enzyme list Trypsin Arg-C Asp-W ambic Chymotrypsin CKBr Formic_acid Lys-C Lys-C/P Ponrinð	Extr	Create act Peptide
Calculate Modification Patterns	The ratio of diges Ratio	ted peptides with sp	ecific amino acids	s to all peptide	sequences.
Calculate	(e.g.,'NXS/T/C' in oxidation, and 'C'	glycosylation, 'S/T, in carbamidomethyla	/Y' in phosphoryls tion) to all pepti	ation, 'M' in ide sequences.	
Mass distribution of a	all peptides candidat Modification:	tes Carbamidomethyl Oxidation Phospho	Mass: 9 Bindhum: 1	50 ~ 1 00	050

Fig. 5.



Database Statistical Characterization and The Robust Framework.

Fig. 6.