

GPS Manual

Group-based Prediction System

Version 5.0 20/7/2019

Author: Chenwei Wang, Haodong Xu and Yu Xue *

Contact: Chenwei Wang, <u>wangchenwei@hust.edu.cn</u>, Yu Xue, <u>xueyu@hust.edu.cn</u> The software is only free for academic research. The latest version of GPS software is available from <u>http://gps.biocuckoo.cn/download.php</u>. Copyright (c) 2004-2019. The CUCKOO Workgroup. All right reserved.

Index

INDEX	1
STATEMENT	2
INTRODUCTION	3
DOWNLOAD & INSTALLATION	4
PREDICTION OF KINASE-SPECIFIC PHOSPHORYLATION SITES	6
DIRECT PREDICTION	6
BATCH PREDICTION	8
PREDICTION OF SPECIES-SPECIFIC KINASE-SPECIFIC PHOSPHORYLATIC	ON SITES . 12
DIRECT PREDICTION	
BATCH PREDICTION	16
EVALUATION OF PREDICTION PERFORMANCES	17
REFERENCES	
RELEASE NOTE	24

Statement

1. **Implementation**. The softwares of the CUCKOO Workgroup are implemented in JAVA (J2SE). Usually, both of online service and local stand-alone packages will be provided.

2. **Availability**. Our softwares are freely available for academic researches. For nonprofit users, you can copy, distribute and use the softwares for your scientific studies. Our softwares are not free for commercial usage.

3. **GPS**. Previously, we used the GPS to denote our Group-based Phosphorylation Scoring algorithm. Currently, we are developing an integrated computational platform for post-translational modifications (PTMs) of proteins. We re-denote the GPS as Group-based Prediction Systems. This software is an indispensable part of GPS.

4. **Usage**. Our softwares are designed in an easy-to-use manner. Also, we invite you to read the manual before using the softwares.

5. **Updation**. Our softwares will be updated routinely based on users' suggestions and advices. Thus, your feedback is greatly important for our future updation. Please do not hesitate to contact with us if you have any concerns.

6. **Citation**. Usually, the latest published articles will be shown on the software websites. We wish you could cite the article if the software has been helpful for your work.

7. Acknowledgements. Funding for open access charge: Special Project on Precision Medicine under the National Key R&D Program [2017YFC0906600, 2018YFC0910500]; Natural Science Foundation of China [31671360, 81701567, 31801095]; National Program for Support of Top-Notch Young Professionals; Changjiang Scholars Program of China; The program for HUST Academic Frontier Youth Team; China Postdoctoral Science Foundation [2018M642816, 2018M632870].

Introduction

Identification of phosphorylation sites with their cognate protein kinases (PKs) is the foundation for understanding the functional dynamics and plasticity of various cellular processes. Although nearly 10 kinase-specific predictors were developed, numerous PKs were casually classified into sub-groups without a standard rule. And for large-scale predictions, the false positive rate (FPR) was also never addressed. Here we adopted a well-established rule to classify PKs with their verified sites into a hierarchical structure with four levels, including group, family, subfamily and single PK⁻¹. Then we constructed the **GPS** (**Group-based Prediction System**, ver 2.0) software, with a modified version of GPS (Group-based Phosphorylation Scoring) algorithm^{2, 3}. As the first stand-alone software for computational phosphorylation, GPS 2.0 was implemented in JAVA and could predict kinase-specific phosphorylation sites for **408** human PKs in hierarchy.

Currently, as more and more phosphorylation sites have been experimentally validated, a more accurate and reliable kinases-specific phosphorylation sites prediction tool is in urgent need. At the same time, several advanced GPS algorithm have been developed and achieved higher performance, including GPS 2.1 and GPS 2.2. To improve the performance of phosphorylation sites prediction, we applied the GPS 2.2 algorithm and enlarged the training set. In addition, we included the phosphorylation sites prediction in 161 species. For dul-specific kinases, we also provided the prediction for these kinases. Finally, we constructed the **GPS (Group-based Prediction System**, ver 5.0) software package.

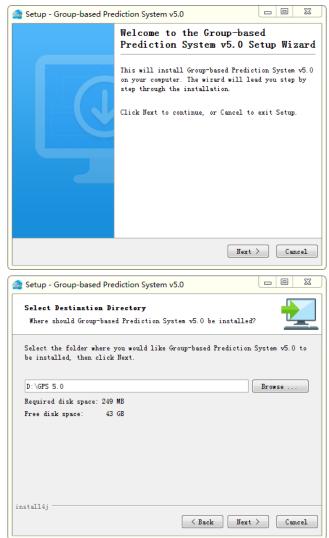
💩 GPS 5.0						
File Tools Help						
PKs	Predicted Sites					
 Phosphorylation Serine/Threonine Kinase G.AGC C.KI C.KI C.KI C.KGC P.KL RGC STE T.KL Alypical Dual Tyrosine Kinase T.K Dual 	Position	Code (s) in FASTA form	Kinase	Peptide	Score	Cutoff
	Threshold			Console		
	🔾 High 🛛 🖲	Medium 🔾 Lo	w 🔾 All	Example	Clear	Submit

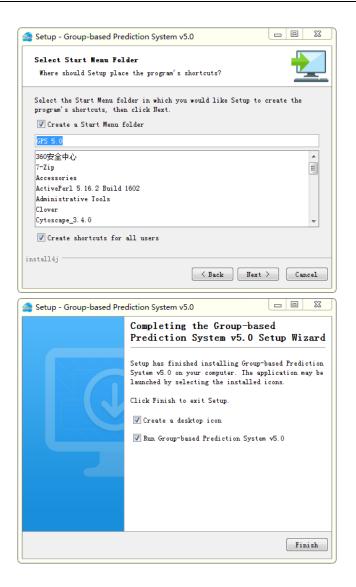
Group-based Prediction System v5.0 User Interface

Download & Installation

The software of GPS 5.0 was implemented in JAVA, and could be installed on Windows systems. GPS 5.0 distributions for Windows can be found at <u>http://gps.biocuckoo.cn/download.php</u>. We recommend that users could download the latest release.

After downloading, please double-click on the file *GPS_5.0_windows.exe* to begin installation. Follow the user prompts through the installation. And snapshots of the setup program are shown below:





Click on the **Finish** button to complete the setup program.

Prediction of Kinase-specific Phosphorylation Sites

Direct Prediction

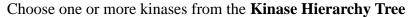
For convenience, the GPS 5.0 allows users to input their protein sequences into the "TEXT form" for prediction. One or more protein sequences should be prepared in **FATSA** format as below:

>protein1	
XXXXXXX	XXXXXXX
XXXXXXX	XX
>protein2	
XXXXXXX	XXXXXXXXXXX
>protein3	
XXXXXXX	XXXXXX

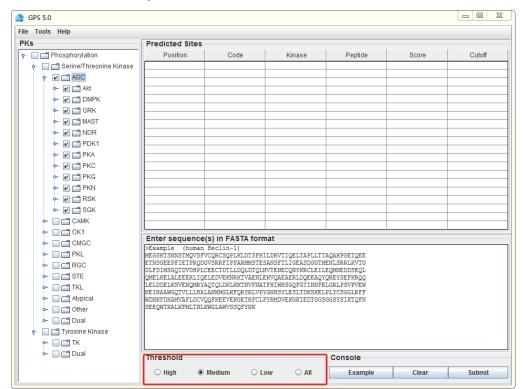
Please note: All irregular words, including non-amino acid word (eg, number) and blank, will be removed automatically. As an instance, we put **human Beclin-1** protein sequence as an example for GPS 5.0. Users could click on the "Example" button to access the example.

GPS 5.0						
File Tools Help						
PKs	Predicted Sites					
P− □ □ Phosphorylation	Position	Code	Kinase	Peptide	Score	Cutoff
👇 🔲 🗂 Serine/Threonine Kinase						
P→ □ □ □ AGC						
← 🖌 🚍 Akt						
🗠 🔲 🚍 DMPK						
🗠 🗌 🚍 GRK						
← 🔲 📑 MAST						
← 🗌 📑 NDR						
← 🗌 🚍 PDK1						
🗠 🖂 🚍 PKA						
← 🔲 📑 PKC						
🗠 🔲 🚍 PKG						
🗠 🗌 🚍 PKN						
🗠 🗌 📑 RSK						
🔶 🗌 🚍 SGK						
← 🔲 🚍 CAMK						
← 🗌 🚍 CK1	Enter sequence	(s) in FASTA for	nat			
	>Example (humar	Beclin-1)				
► C C PKL	MEGSKTSNNSTMQVSE	VCQRCSQPLKLDTSFK	ILDRVTIQELTAPLLTT SANSFTLIGEASDGGTM	AQAKPGETQEE		
► □ □ RGC	DLFDIMSGQTDVDHPI	CEECTDTLLDQLDTQL	NVTENECQNYKRCLEIL	EQMNEDDSEQL		
► C C STE			EKVQAEAERLDQEEAQY IFHIWHSGQFGTINNFR			
	NEINAAWGQTVLLLHA	LANKMGLKFQRYRLVP	YGNHSYLESLTDKSKEL	PLYCSGGLRFF		
Atypical	WDNKFDHAMVAFLDCV SEEQWTKALKFMLTNI		PYRMDVEKGKIEDTGGS	GGSYSIKTQFN		
- Other	SECONTRADICIDITAL	iniobanyoogi inin				
- Dual						
Tyrosine Kinase						
🔶 🔲 🚍 Dual	Threshold			Console		
	🔾 High 🖉	Medium 🔷 L	ow 🔾 All	Example	Clear	Submit

Serine/Threonine Kinase Image: Serine/Threonine Kinase	5	Predicted Sites					
♀- 🔲 🗂 Tyrosine Kinase			(5) in FASTA for Beclin-1) vcQRcSQCLAUTSEK GVSRRFIPPARMSTE GECTDTLLDQLDTQL LEEVDENIRKTVABL LANKRGLKFQQTALUW QFKEEVEKGETRFCL	mat ILLERVTIQELTAPLLTTA SANSTTLIGEASDGTME NVTENECONYRECLEILE EVQAEAZENDEGEAQYC THTIWISGGFOTINIFRI	AQAKPGETQEE INLSRRLKVTG CQMNEDDSEQL JKYSEFKRQQ GRLPSVFVEW	Score	Cutoff
Console	► 🗌 🚍 ТК						



Choose a Threshold what you need, default is Medium.



Click on the Submit button, then the predicted phosphorylation sites will be shown.

Ks	Predicted	Sites				
- Thosphorylation	Position	Code	Kinase	Peptide	Score	Cutoff
	90	S	AGC	IPPARMMSTESANSF		0.386
👇 🔲 📑 Serine/Threonine Kinase	295	S	AGC	FRLGRLPSVPVEWNE		0.386
P ■ AGC	10	S	AGC/Akt	GSKTSNNSTMQVSFV		8.272
🔶 🔽 🗂 Akt	90	S	AGC/Akt	IPPARMMSTESANSF		8.272
	234	S	AGC/Akt	AQYQREYSEFKRQQL		8.272
	295	S	AGC/Akt	FRLGRLPSVPVEWNE		8.272
🗠 🗹 🚍 GRK	7	S	AGC/DMPK	*MEGSKTSNNSTMQV	0.100	11.906
🗠 🗹 🚍 MAST	22	S	AGC/DMPK	SFVCQRCSQPLKLDT		11.906
	38	T	AGC/DMPK	FKILDRVTIQELTAP		11.906
	90	S	AGC/DMPK	IPPARMMSTESANSF		11,906
• 🗹 📑 PDK1	434	T	AGC/DMPK	KALKFMLTNLKWGLA	12.301	11.906
🗠 🗹 📑 PKA	10	S	AGC/GRK	GSKTSNNSTMQVSFV	61.024	59.564
🔶 🗹 🚍 PKC	93	S	AGC/GRK	ARMMSTESANSFTLI	61.212	59.564
🗠 🔽 📑 PKG	113	S	AGC/GRK	GGTMENLSRRLKVTG	64.169	59.564
	119	Т	AGC/GRK	LSRRLKVTGDLFDIM	60.887	59.564
🗠 🗹 🚍 PKN	155	Т	AGC/GRK	LDTQLNVTENECQNY	60.346	59.564
🔶 🖌 🚍 RSK	177	S	AGC/GRK	EQMNEDDSEQLQMEL	67.07	59.564
🗢 🔽 📑 SGK	295	S	AGC/GRK	FRLGRLPSVPVEWNE	63.699	59.564
	426	T	AGC/GRK	FNSEEQWTKALKFML	63.981	59.564
- CAMK	259	Т	AGC/MAST	NQMRYAQTQLDKLKK	2.109	1.801
🗠 🔲 🚍 СК1	Enter cog	uence(s) in FA	CTA format			
🔶 🔄 🔚 CMGC		human Beclin-				
🕶 🔲 🚍 PKL			LKLDTSFKILDRVTIQELTAP	LTTAOAKPGFTOFF		
			PARMMSTESANSFTLIGEASD			
			LDQLDTQLNVTENECQNYKRC			
► 🔲 🚍 STE			IRKIVAENLEKVQAEAERLDQE			
🗠 🔲 🚍 TKL			KKTNVFNATFHIWHSGQFGTI			
🖛 🔲 🚍 Atypical			(FQRYRLVPYGNHSYLESLTDK KGETRFCLPYRMDVEKGKIED			
⊷ 🔲 📑 Other		MLTNLKWGLAWVS		rooboobibinigin		
🔶 🔄 🚍 Dual						
👇 🔲 🗂 Tyrosine Kinase						
🔶 🔲 🚍 ТК						
🔶 🔲 🚍 Dual	Threshold			Console		

Batch Prediction

We also provide an alternative approach for processing multiple protein sequences. If the file is large, the **Batch Predictor** will be convenient for users.

The following steps show you how to use it: Put protein sequences into a file with **FATSA** format as below:

>protein1 XXXXXXXXXXX XXXXXXXX >protein2 XXXXXXXXXXXXXXXXXXX >protein3 XXXXXXXXXXXX

The names of proteins are necessary (the line with ">" and a protein name/accession number).

To run the Batch Predictor just select the **Batch Predictor** option in the **Tools** menu.

💩 Batch Predictor	and the second					X
Batch Predictor	Sequence File Result File Lis Result Expo	t.	move All	Remove	Add File	
	Threshold	roluer			Console	
	🔾 High 🏾 🖲	Medium	○ Low	O Ali	Clear	Submit

Click on the Add File button and add one or more protein sequence files in your hard disk.

Batch Predictor	X
PKs	Sequence File List
	eq Quence1.txt Quence2.txt Quence3.txt
← □ □ □ Dual 文件名(11):	"FASTA_sequence1.txt" "FASTA_sequence2.txt" "FASTA_sequence3.txt"
文件类型(I):	所有文件
	打开 取消
	Result Export Folder >>
	Threshold
	⊖ High

The name of added files will be shown in the Sequence File List

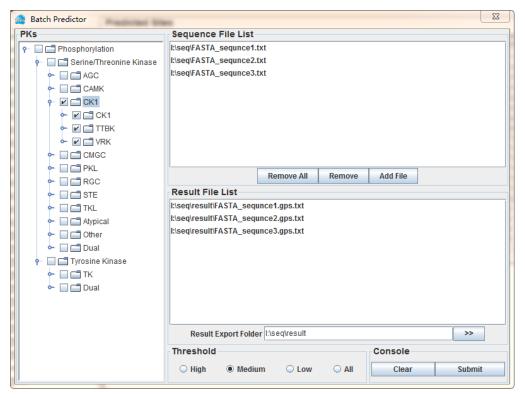
🙆 Batch Predictor					23	
PKs	Sequence File List					
🛉 🥐 🔲 🚍 Phosphorylation	I:\seq\FASTA_sequnce1.	txt				Ì
🛉 🔲 🚍 Serine/Threonine Kinase	I:\seq\FASTA_sequnce2.	txt				
- AGC	I:\seq\FASTA_sequnce3.	txt				
- CAMK						
🕶 🗌 🚍 СК1						
← □ 🗖 CMGC						
🕶 🗔 🚍 PKL						
🕶 🗔 🚍 RGC						
🕶 🔲 🚍 STE						
► 🗌 🚍 TKL			_			
🔶 🔲 🚍 Atypical		Remove All	Remove	Add File		
🔶 🔲 🚍 Other	Result File List					
🕶 🔲 🚍 Dual						
👇 🔲 🚍 Tyrosine Kinase						
🔶 🗌 🚍 ТК						
🕶 🔲 🚍 Dual						
	Result Export Folde	er			>>	1
	Threshold			Console		
	🔾 High 🛛 🖲 Mediu	m 🔾 Low	🔾 Ali	Clear	Submit	
<u></u>	-					⊒)

The output directory of prediction results should also be defined. Please click on the ">>" button to specify the export file fold.

👩 Batch Predictor				X
PKs	Sequence File List			
	I:\seq\FASTA_sequnce1.txt			
👇 🔲 🗂 Serine/Threonine Kinase	l:\seq\FASTA_sequnce2.txt			
- AGC	I:\seq\FASTA_sequnce3.txt			
← □ □ CAM @ 保存			23	
← □ □ СК1				
← □ □ CMG 保存: □ seq	1	- 6 6		
	-			
🗣 🗌 🚍 RGC 🔤 result				
► 🔲 🚍 STE				
► C C TKL				
🕶 🔄 🚍 Atypi				
← □ □ Othe				
- Dual				
🛉 🗌 🚍 Tyrosine				
 ← □ □ TK 	l:\seq			
文件类型(I):	Folder		-	
		保存	取消	
	-			
	Result Export Folder			>>
	Threshold		Console	
	🔾 High 💿 Medium	🔾 Low 🔷 All	Clear	Submit

5. Choose one or more kinases from **Kinase Hierarchy Tree**, and then pick a **Threshold**, click on the **Submit** button, then the **Batch Predictor** will begin to process all of the sequence files that have been added to the list. The results of predictions will be exported to the **Prediction Export Fold**, and the name of result files will be shown in the **Prediction**

File List.



Prediction of Species-specific Kinase-specific Phosphorylation Sites

Direct Prediction

An advance feature of GPS 5.0 is Prediction of Species-specific Kinase-specific Phosphorylation Sites. Users can predict phosphorylation sites in different species. Since there isn't a standard principle classification for all animals and plants, we applied the kinases classification for eukaryotes from a new published database - iEKPD, which is a eukaryotic protein kinases and protein phosphatases database. Kinases and phosphatases in 164 species are classified in a hierarchy structure, including group, family and single protein.

To predict phosphorylation sites in 161 species, users could open the species-specific prediction interface by clicking the item "GPS 5.0 – Species-specific" item in "Tools" menu.

@ GPS 5.0								
File Tools Help		redicted Si	<u> </u>					
PK GPS 5.0 - Species-speci	Ctrl-B	Position	Code	e	Kinase	Peptide	Score	Cutoff
Performance	Ctrl-P							
Domain Graph	Ctrl-D							
 CK1 CMGC PKL RGC STE TKL Atpical Other Dual Tyrosine Kinase TK 								
← 🗌 🗖 Dual		Enter seque	nce(s) in FAS	TA format				
		Throshold				Capacito		
		Threshold -	Modium	Olow	0.01	Console	Cloar	Submit
J		🔾 High	Medium	Low		Example	Clear	Submit

Similar with the classical version, this prediction interface also allows users to input their protein sequences into the "TEXT form" for prediction. One or more protein sequences should be prepared in **FATSA** format as below:

>protein1

XXXXXXXXXXXX XXXXXXXX >protein2 XXXXXXXXXXXXXXXXXXX >protein3 XXXXXXXXXXXXX

Please note: All irregular words, including non-amino acid word (eg, number) and blank, will be removed automatically. As an instance, we put **human Beclin-1** protein sequence as an example for GPS 5.0. Users could click on the "Example" button to access the example.

ile Tools Help						
PKs	Predicted Site	5				
👇 🔚 🔚 Phosphorylation	Position	Code	Kinase	Peptide	Score	Cutoff
👇 🔲 📑 Serine/Threonine Kinase						
P- □ □ AGC						
— 🖌 🗋 Akt						
— 🔲 🗋 DMPK						
- 🔲 🗋 GRK						
- 🔄 🗋 MAST						
- 🔲 🗋 NDR						
— 🛄 🎦 PDK1						
— 🛄 🎦 РКА						
— 🔛 🎦 РКС						
— 🔲 🗋 PKG						
- 🔲 🗋 PKN						
- 🔲 🗋 RSK						
- 🔄 🗋 SGK						
► 🔲 🚍 CAMK						
👇 🔲 🚍 CK1		- (-) :				1
🗠 🔄 🚍 CMGC	Enter sequence		format			
🗠 🔲 🚍 PKL			SFKILDRVTIQELT.	APLLTTAQAKPGETQE	E	
👇 🔲 🚍 RGC				SDGGTMENLSRRLKVT RCLEILEQMNEDDSEQ		
🗣 🔛 🚍 STE	QMELKELALEEERLI	QELEDVEKNRKIV	AENLEKVQAEAERLD	QEEAQYQREYSEFKRQ	Q	
🖛 🔲 🚍 TKL				TINNFRLGRLPSVPVE DKSKELPLYCSGGLRF		
🔶 🔄 🚍 Atypical	WDNKFDHAMVAFLDC	VQQFKEEVEKGETI	RECLEYRMDVEKGKI	EDIGGSGGSYSIKIQF		
🔶 🔛 🚍 Other	SEEQWIKALKFMLIN	LKWGLAWVSSQFYI	IK			
🖙 🔲 🚍 Dual						
👇 🔲 🚍 Tyrosine Kinase	Options			Console		
🔶 🔲 📑 TK	options			Console		

Choose a species from the Organism List.

Ks	Predicted Site	s				
Phosphorylation	Position	Code	Kinase	Peptide	Score	Cutoff
- Carlo Carl						
- R B Akt						
- GRK						
CK1	Enter sequence	e(s) in FASTA	format			
	>Example (huma					
PKL				APLLTTAQAKPGETQE SDGGTMENLSRRLKVT		
	DLFDIMSGQTDVDH	PLCEECTDTLLDQLI	TQLNVTENECQNYK	RCLEILEQMNEDDSEQ	L	
► □ 🗖 STE	QMELKELALIH. sap	Diens ARKIV		QEEAQYQREYSEFKRQ TINNFRLGRLPSVPVE		
	NEINAAWGQ R. noi	regrue FQRYI		DKSKELPLYCSGGLRF		
🔶 🔲 🚍 Atypical	WDNKFDHAM SEEOWTKALLD. me	lanogast SQFYI		EDTGGSGGSYSIKTQF	N	
🔶 🔄 🚍 Other	C. ele		in.			
🔶 🔛 🚍 Dual	S. por					
🖕 🔄 📑 Tyrosine Kinase	Options S. cer	evisiae		Console		

Choose one or more kinases from the Kinase Hierarchy Tree

File Tools Help						
PKs	Predicted Site	5				
👇 🔲 🚍 Phosphorylation	Position	Code	Kinase	Peptide	Score	Cutoff
👇 🔲 🚍 Serine/Threonine Kinase						
👇 🗹 🗂 AGC						
— 🗹 🗋 Akt						
- 🗹 🗋 DMPK						
- 🗹 🗋 GRK						
- 🗹 🗋 MAST						
- 🗹 🗋 NDR						
- 🗹 🗋 PDK1						
— 🔽 🗋 РКА						
- PKC						
- 🔽 🗋 PKG						
- 🗹 🗋 PKN						
- 🔽 🗋 RSK						
SGK						
⊷ □ 🗖 СК1						
	Enter sequend		format			
	>Example (huma		SEVILOPUTIOELT	APLLTTAQAKPGETQ	F	
- RGC	ETNSGEEPFIETPRO	DGVSRRFIPPARM	ISTESANSFTLIGEA	SDGGTMENLSRRLKVT	ſĠ	
- STE				RCLEILEQMNEDDSE(QEEAQYQREYSEFKR(
	LELDDELKSVENQMF	YAQTQLDKLKKTN	FNATFHIWHSGQFG	TINNFRLGRLPSVPVE	EW .	
- Atypical				DKSKELPLYCSGGLRH EDTGGSGGSYSIKTQI		
• Other	SEEQWIKALKFMLIN			EDIGGGGGGGGGGGGGG	1.14	
Culei						
- Tyrosine Kinase	ļ L					
Tyrosine kinase	Options			Consol	e	
	Organism H. sap	iens 🔻	Threshold Mediu	Im 🔻 Exam	ple Clear	Submit

Choose a Threshold what you need, default is Medium

e Tools Help						
Ks	Predicted Sites	5				
- 🔄 🚍 Phosphorylation	Position	Code	Kinase	Peptide	Score	Cutoff
👇 🔲 🚍 Serine/Threonine Kinase						
👇 🗹 🗂 AGC						
— 🗹 🗋 Akt						
- 🗹 🗋 DMPK						
– 🗹 🗋 GRK						
- 🗹 🗋 MAST						
- 🗹 🗋 NDR						
- 🗹 🗋 PDK1						
- 🔽 🗋 PKA						
— 🔽 🗋 РКС						
– 🗹 🗋 PKG						
- 🔽 🗋 PKN						
- 🔽 🗋 RSK						
SGK						
~ □ □ CK1						
	Enter sequenc		format			
🔶 🔲 📑 PKL	>Example (huma MEGSKTSNNSTMOVS		SEKTLORVITOELT	APLLTTAQAKPGETQE	2	
- 🗖 🗖 RGC	ETNSGEEPFIETPRQ	DGVSRRFIPPARM	ISTESANSFILIGEA	SDGGTMENLSRRLKVT	3	
🗣 🔲 📑 STE				RCLEILEQMNEDDSEQ QEEAQYQREYSEFKRQ		
	LELDDELKSVENQMR	YAQTQLDKLKKTN	/FNATFHIWHSGQFG	TINNFRLGRLPSVPVEV	v.	
Atypical				DKSKELPLYCSGGLRFI EDTGGSGGSYSIKTQF1		
Other	SEEQWIKALKFMLIN					
- Dual						
Tyrosine Kinase						
← □ □ □ TK	Options			Console		

Click on the **Submit** button, then the predicted phosphorylation sites will be shown.

ile Tools Help						
PKs	Predicted	l Sites				
P- 🔲 🗂 Phosphorylation	Position	Code	Kinase	Peptide	Score	Cutoff
🔶 🥅 🗂 Serine/Threonine Kinase	90	S	AGC	IPPARMMSTESANSF	1.127	0.386
	295	S	AGC	FRLGRLPSVPVEWNE	2.955	0.386
	10	S	AGC/Akt	GSKTSNNSTMQVSFV	8.377	8.272
— 🖌 🗋 Akt	90	S	AGC/Akt	IPPARMMSTESANSF	9.777	8.272
- 🗹 🗋 DMPK	234	S	AGC/Akt	AQYQREYSEFKRQQL	10.066	8.272
- 🔽 🗋 GRK	295	S	AGC/Akt	FRLGRLPSVPVEWNE	9.483	8.272
	7	S	AGC/DMPK	*MEGSKTSNNSTMQV	14.012	11.906
- 🗹 🗋 MAST	22	S	AGC/DMPK	SFVCQRCSQPLKLDT	12.001	11.906
- 🗹 🗋 NDR	38	Т	AGC/DMPK	FKILDRVTIQELTAP	15.24	11.906
- 🔽 🗋 PDK1	90	S	AGC/DMPK	IPPARMMSTESANSF	13.036	11.906
	434	Т	AGC/DMPK	KALKFMLTNLKWGLA	12.301	11.906
- 🗹 🗋 PKA	10	S	AGC/GRK	GSKISNNSIMQVSFV	61.024	59.564
— 🗹 🗋 PKC	93	S	AGC/GRK	ARMMSTESANSFILI	61.212	59.564
- 🔽 🗋 PKG	113	S	AGC/GRK	GGTMENLSRRLKVTG	64.169	59.564
	119	Т	AGC/GRK	LSRRLKVTGDLFDIM	60.887	59.564
	155	Т	AGC/GRK	LDTQLNVTENECQNY	60.346	59.564
– 🗹 🗋 RSK	177	S	AGC/GRK	EQMNEDDSEQLQMEL	67.07	59.564
SGK	295	S	AGC/GRK	FRLGRLPSVPVEWNE	63.699	59.564
	426	Т	AGC/GRK	FNSEEQWTKALKFML	63.981	59.564
	259	T	AGC/MAST	NQMRYAQTQLDKLKK	2.109	1.801
• 🔲 🚍 CK1	Enter sec	uence(s)	in FASTA format			
🗠 🔲 🚍 CMGC	>Example	(human Bec				
🔶 🔲 🚍 PKL			CSQPLKLDTSFKILDRVTIQ	ELTAPLLTTAQAKPGETQEE		
🖙 🔲 🗂 RGC			RFIPPARMMSTESANSFILIG			
			TDTLLDQLDTQLNVTENECQN			
			VEKNRKIVAENLEKVQAEAEF DLDKLKKTNVFNATFHIWHSG			
🗠 🔲 🚍 TKL			MGLKFQRYRLVPYGNHSYLES			
🖛 🔄 🚍 Atypical			EEVEKGETRFCLPYRMDVEK			
🔶 🔲 🗖 Other	SEEQWIKAL	KFMLTNLKWGI	AWVSSQFYNK	-		
- Dual						
🛉 🔲 🚍 Tyrosine Kinase	Options			Console		
🔶 🔲 🗂 TK						

Batch Prediction

In GPS 5.0 - Species-specific prediction interface, we also provide an alternative approach for processing multiple protein sequences. If the file is large, the **Batch Predictor** will be convenient for users. The usage is similar to the **Batch Prediction** of **Direct Prediction**.

Algorithms and Prediction Performance

Algorithm Design

To predict kinase-specific phosphorylation sites, a series of GPS (Group-based Phosphorylation Scoring) algorithm were maintained and improved in the past few years 2,3 . The basic hypothesis of GPS algorithm is that highly similar short peptides bear similar biochemical properties for the modification. Therefore, we defined a *phosphorylation site peptide* PSP(*m*, *n*) as a pS, pT or pY amino acid flanked by *m* residues upstream and *n* residues downstream. Then we used the amino acid substitution matrix BLOSUM62 to calculate the similarity between two PSP (m, n) peptides.

In GPS 5.0, a new hypothesis is that long flanking regions around p-sites might be important for the recognition of PKs, which are bulky molecules to interact with phosphorylatable residues. Thus the GPS 5.0 is upgraded in basis of GPS 2.1 and contained two parts: the scoring strategy and performance improvement.

The scoring strategy defined the average similarity score (S) between a PSP(30,30) peptide P and peptides around all known p-sites in the training data set:

$$S = \frac{1}{N} \sum_{j=-30}^{L-31} (\sum_{i=1}^{N} M_{train}[P_j, T_{ij}]) \times W_j$$

Where *L* is the length of the PSP(30, 30) peptide and equal to 61 to represent a relatively long flanking region. *N* is the number of known p-sites in the positive data set. T_{ij} is the amino acid at position *j* around a known p-site T_i (*i* = 1, 2, 3, ..., *N*). W_j is the weight value of position *j*, and M_{train} represents the optimized amino acid substitution matrix.

The performance improvement comprises two parts, the MLS and MaM in GPS 2.1 were upgraded into PWD and SMO, respectively.

(i) PWD: The amino acid substitution matrix BLOSUM62 ($M_{BLOSUM62}$) was used to calculate an average similarity score at the position *j* of a PSP(30, 30) peptide *P* as S_i :

$$S'_{j} = W_{j} \frac{1}{N} \sum_{i=1}^{N} M_{BLOSUM62}[P_{j}, T_{ij}])$$

Initially, the weight value of each position W_j in the PSP(30, 30) peptide was set to 1. Then the one-vs-rest (OVR) classifier with the ridge (L2) penalty of the LR algorithm was used to optimize W_j values, by applying the "newton-cg" solver in the class LogisticRegressionCV of scikit-learn v0.21.0 (https://scikit-learn.org/). To avoid overfitting, such a procedure was repeated for 100 times and 10-fold cross-validation was conducted to determine the inverse of regularization strength of each time. Receiver operating characteristic (ROC) curves were illustrated, and area under curve (AUC) values were calculated. The optimal W_j vectors were determined based on the highest AUC value:

$$W_j = W_{-30}, \dots, W_{-1}, W_0, W_1, \dots, W_{30}$$

In order to evaluate position-specific contributions of flanking regions around p-sites for different PK clusters, the W_j vectors were normalized into -1 to 1 based on the maximum absolute value.

(ii) SMO: The average similarity score of an amino acid a in the given PSP(30, 30) peptide P and a residue b in peptides around all known p-sites was defined as S_{ab} :

$$S_{ab} = \frac{1}{N} \sum_{j=-30}^{L-31} C_j \times M_{BLOSUM62}[a, b] \times W_j$$

Where C_j is the number of *ab* amino acid pairs at position *j*. In BLOSUM62, there were 24 types of characters including 20 types of amino acids and 4 non-canonical characters (B, Z X and *). Thus, a number of [24*(24+1)]/2=300 unique S_{ab} scores ($S_{ab} = S_{ba}$) were generated. Then, the same LR algorithm was used to optimize all of S_{ab} scores to produce a new matrix M_{train} :

$$M_{train} = (S_{AA}, S_{AC}, S_{AD}, \cdots, S_{**})_{300}$$

Evaluation of Prediction Performances

Performance measurements

To evaluate the prediction performances, four standard measurements were used, including accuracy (Ac), sensitivity (Sn), specificity (Sp) and Mathew correlation coefficient (MCC). Accuracy (Ac) represents the correct ratio between both positive (+) and negative (-) data sets, while sensitivity (Sn) and specificity (Sp) illustrate the correct prediction ratios of positive (+) and negative data (-) sets respectively. Since the number of positive data and negative data differed too much from each other, the Mathew correlation coefficient (MCC) was also included. The value of MCC ranges from -1 to 1, and a larger MCC value stands for better prediction performance.

Among the data with positive hits by GPS 5.0, the real positives were defined as *true* positives (*TP*), while the others were defined as *false positives* (*FP*). Among the data with negative predictions, the real positives were defined as *false negatives* (*FN*), while the others were defined as *true negatives* (*TN*). The four measurements of sensitivity (*Sn*), specificity (*Sp*), accuracy (*Ac*), and Mathew correlation coefficient (*MCC*) were defined as below:

$$Sn = \frac{TP}{TP + FN}, \quad Sp = \frac{TN}{TN + FP},$$

$$Ac = \frac{TP + TN}{TP + FP + TN + FN}, \text{ and}$$

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}.$$

The self-consistency, leave-one-out validation and n-fold cross-validation

The self-consistency used the training positive data and negative data directly to evaluate the prediction performance, and represented the computational power of the prediction system. However, the robustness and stability of the software should also be evaluated by leave-one-out validation and *n*-fold cross-validation. In the leave-one-out validation, which is also called as Jack-Knife validation, each sites in the dataset was picked out in turn as an independent test sample, and all the remaining sites were regarded as training data. This process was repeated until each site was used as test data one time. In *n*-fold cross-validation all the (+) sites and (-) sites were combined and then divided equally into *n* parts, keeping the same distribution of (+) and (-) sites in each part. Then *n*-1 parts were merged into a training data set while the remanent part was taken as a testing data set. This process was repeated twenty times and the average performance of *n*-fold cross- validation was used to estimate the performance. In this work, the performances of validation were calculated for all PK groups. And the 10-fold cross-validations were performed for PK groups with the number of positive data ≥ 30 and leave-one-out validation were calculated for PK groups with the number of positive data < 30. All validation performances of GPS 5.0 have been included in **Performance** option in the **Tools** menu.

Accessing of prediction performances of GPS 5.0 with related information

To check the prediction performance of the **479** human PKs just select the **Performance** option in the **Tools** menu.

PKs	Kinase Info	ormation					
P → □ P hosphorylation	Name		Group	Family	U	niProt	
👇 🗐 Serine/Threonine Kinase	AKT1		AGC	Akt	AKT1	HUMAN	7.
	AKT2		AGC	Akt	AKT2	HUMAN	
- C) Akt	AKT3		AGC	Akt	AKT3	HUMAN	
	CDC42BF		AGC	DMPK	MRCK	A HUMAN	
	CDC42BF	-	AGC	DMPK	MRCKE	<u>B HUMAN</u>	
– 🗋 GRK	CDC42BF	°G	AGC	DMPK	MRCK	<u>G HUMAN</u>	=
- 🗋 MAST	CIT		AGC	DMPK		HUMAN	
- 🗋 NDR	DMPK		AGC	DMPK		HUMAN	
- 🗋 PDK1	GRK1		AGC	GRK		HUMAN	_
	GRK2		AGC	GRK		I HUMAN	
	GRK3		AGC	GRK		2 HUMAN	-1
- D PKC	GRK4 GRK5		AGC	GRK GRK		HUMAN	-1
– 🗋 PKG	GRK5		AGC	GRK		HUMAN	-1
- 🗋 PKN	GRK0 GRK7		AGC	GRK		HUMAN	-1
- 🗋 RSK	LATS1		AGC	NDR		HUMAN HUMAN	-1
- D SGK	LATS2		AGC	NDR		HUMAN	-1
	MAST1		AGC	MAST		HUMAN	-1
	MAST2		AGC	MAST		2 HUMAN	-1
•- 📑 CK1	MAST3		AGC	MAST		HUMAN	
← 🛄 CMGC	MAST4		AGC	MAST		HUMAN	
🗠 🛄 PKL	MASTL		AGC	MAST		HUMAN	
🗠 🛄 RGC	PDPK1		AGC	PDK1		1 HUMAN	
⊷ 🗂 STE	PKN1		AGC	PKN	PKN1 HUMAN		
🔶 🚍 TKL	PKN2		AGC	PKN	PKN2	HUMAN	1
🕶 🚞 Atypical	Prediction	Performand	e				
🗠 🗂 Other		Cutoff	Ac	Sn	Sp	Mcc	
🔶 🚍 Dual	High	1.9250	96.42%	39.48%	97.65%	0.3070	-
🛉 📑 Tyrosine Kinase							-11
•- 🛄 TK	Medium	0.3860	92.72%	57.38%	93.49%	0.2768	
🔶 🛄 Dual	Low	-0.3970	88.88%	67.86%	89.33%	0.2548	

Choose a kinase what you want check from the Kinase Hierarchy Tree.

Then **Kinase Information** and **Prediction Performance** of the kinase are shown in the tables.

S	Kinase Info	ormation					
Phosphorylation	Name		Group		Family	U	niProt
🛉 📑 Serine/Threonine Kinase	AKT1		AGC		Akt	AKT1	HUMAN
- CARC	AKT2		AGC		Akt	AKT2	HUMAN
Akt	AKT3		AGC		Akt	AKT3	HUMAN
	CDC42BF		AGC		DMPK	MRCK/	A HUMAN
	CDC42BF	-	AGC		DMPK	MRCK	<u>B HUMAN</u>
- GRK	CDC42BF	PG	AGC		DMPK		<u>g human</u>
- 🗋 MAST	CIT		AGC		DMPK		HUMAN
- 🗋 NDR	DMPK		AGC		DMPK		HUMAN
- 🗋 PDK1	GRK1		AGC		GRK		HUMAN
- 🗋 PKA	GRK2 GRK3		AGC		GRK GRK		<u>1 HUMAN</u>
	GRK3 GRK4		AGC AGC		GRK		2 HUMAN
- D PKC	GRK4 GRK5		AGC		GRK		HUMAN
- 🗋 PKG	GRK5 GRK6		AGC		GRK		HUMAN
– 🗋 PKN	GRK0		AGC		GRK		HUMAN HUMAN
- 🗋 RSK	LATS1		AGC		NDR		HUMAN
- n sgk	LATS2		AGC		NDR		2 HUMAN
	MAST1		AGC		MAST		I HUMAN
	MAST2		AGC		MAST		2 HUMAN
• 📑 CK1	MAST3		AGC		MAST		3 HUMAN
	MAST4		AGC		MAST		4 HUMAN
► 🚍 PKL	MASTL		AGC		MAST	GWL	HUMAN
← 📑 RGC	PDPK1		AGC		PDK1	PDPK1	1 HUMAN
← 🗂 STE	PKN1		AGC		PKN	PKN1	HUMAN
🗠 🚍 TKL	PKN2		AGC		PKN	PKN2	HUMAN
🕶 🚍 Atypical	Prediction	Perform	ance				
🕶 🚍 Other		Cutoff	A	с	Sn	Sp	Mcc
← 🗂 Dual ← 🗂 Tyrosine Kinase	High	1.9250	96.4	2%	39.48%	97.65%	0.3070
► C TK	Medium	0.3860	92.7	2%	57.38%	93.49%	0.2768
🕶 🗂 Dual	Low	-0.3970	88.8		67.86%	89.33%	0.2548

If you want get more kinase information, you can click on the hyperlinks in the table.

Ks	Kinase Info	ormation					
- 🗂 Phosphorylation	Name	(Group	Family	Ur	niProt	
👇 🗂 Serine/Threonine Kinase	AKT1		AGC	Akt	AKT1	HUMAN	_
	AKT2		AGC	Akt	AK12	HUMAN	-
Akt	AKT3		AGC	Akt	AKT3	HUMAN	_
	CDC42B		AGC	DMPK	MRCKA	HUMAN	_
	CDC42BF		AGC	DMPK	MRCKE	B HUMAN	
– 🗋 GRK	CDC42BF	-	AGC	DMPK	MRCKG	G HUMAN	_
- 🗋 MAST	CIT		AGC	DMPK	CTRO	HUMAN	
	DMPK		AGC	DMPK	DMPK	HUMAN	
- DPDK1	GRK1		AGC	GRK	RK I	HUMAN	
	GRK2		AGC	GRK	ARBK1	HUMAN	
- 🗋 PKA	GRK3		AGC	GRK	ARBK2	<u>HUMAN</u>	
- 🗋 PKC	GRK4		AGC	GRK	GRK4	HUMAN	
- 🗋 PKG	GRK5		AGC	GRK	GRK5	HUMAN	
- D PKN	GRK6		AGC	GRK	GRK6	HUMAN	
	GRK7		AGC	GRK	GRK7	HUMAN	
	LATS1		AGC	NDR		HUMAN	
- 🗋 SGK	LATS2		AGC	NDR		HUMAN	
► □ CAMK	MAST1		AGC	MAST		HUMAN	
← 🚍 CK1	MAST2		AGC	MAST		HUMAN	
	MAST3		AGC	MAST		HUMAN	
	MAST4		AGC	MAST		HUMAN	
← C RGC	MASTL		AGC	MAST		HUMAN	_
	PDPK1		AGC	PDK1		HUMAN	_
	PKN1 PKN2		AGC AGC	PKN PKN		HUMAN	
• 🚍 TKL				PKN	PKN2	HUMAN	
- C Atypical	Prediction	Performanc	e	1			
🕶 🔚 Other		Cutoff	Ac	Sn	Sp	Mcc	
🖕 🚍 Dual	High	1.9250	96.42%	39.48%	97.65%	0.3070	
Tyrosine Kinase	Medium	0.3860	92.72%	57.38%	93.49%	0.2768	
	weaturn	0.3000	32.12.10	57.5070	33.4370	0.2100	_

The hyperlinks will access the **UniProt** database and show you the detailed information of the kinases.

1 1 1 1		
UniProt	UniProlKB +	Advanced - Q Search
BLAST Align Retrieve/ID map	ping Peptide search	Help Contact
UniProtKB - P3	31749 (AKT1_HUMAN)	📾 Basket 🗸
Display	SLAST ≡ Align Gromat market @ History	Other tutorials and videos Help videos Feedback
Entry	Protein RAC-alpha serine/threonine-protein kinase	
Publications	Gene AKT1	
Feature viewer	Organism Homo sapiens (Human)	
Feature table	Status 🛐 Reviewed - Annotation score: 🗰 🗰 - Experimental evidence at protein level ⁱ	
Function	Function ⁱ	
 Names & Taxonomy Subcellular location 	AKT1 is one of 3 closely related serine/threonine-protein kinases (AKT1, AKT2 and AKT3) called the AKT kinase, and which n survival, growth and angiogenesis. This is mediated through serine and/or threonine phosphorylation of a range of downstre	

References

- 1. G. Manning, D. B. Whyte, R. Martinez et al., Science 298 (5600), 1912 (2002).
- 2. Y. Xue, F. Zhou, M. Zhu et al., Nucleic Acids Res 33 (Web Server issue), W184 (2005).
- 3. F. F. Zhou, Y. Xue, G. L. Chen et al., Biochem Biophys Res Commun 325 (4), 1443 (2004).
- 4. Y. Xue, A. Li, L. Wang et al., BMC Bioinformatics 7, 163 (2006).
- 5. N. Blom, T. Sicheritz-Ponten, R. Gupta et al., Proteomics 4 (6), 1633 (2004).
- 6. J. C. Obenauer, L. C. Cantley, and M. B. Yaffe, Nucleic Acids Res 31 (13), 3635 (2003).
- 7. H. D. Huang, T. Y. Lee, S. W. Tzeng et al., Nucleic Acids Res 33 (Web Server issue), W226 (2005).
- 8. J. H. Kim, J. Lee, B. Oh et al., *Bioinformatics* **20** (17), 3179 (2004).
- 9. R. I. Brinkworth, R. A. Breinl, and B. Kobe, Proc Natl Acad Sci U S A 100 (1), 74 (2003).
- 10. T. Li, F. Li, and X. Zhang, Proteins (2007).
- 11. G. Neuberger, G. Schneider, and F. Eisenhaber, *Biology direct* 2, 1 (2007).
- 12. Y. H. Wong, T. Y. Lee, H. K. Liang et al., Nucleic Acids Res 35 (Web Server issue), W588 (2007).
- 13. F. Diella, S. Cameron, C. Gemund et al., BMC Bioinformatics 5 (1), 79 (2004).
- 14. P. V. Hornbeck, I. Chabra, J. M. Kornhauser et al., Proteomics 4 (6), 1551 (2004).
- 15. L. M. Iakoucheva, P. Radivojac, C. J. Brown et al., Nucleic Acids Res 32 (3), 1037 (2004).
- 16. N. Blom, S. Gammeltoft, and S. Brunak, J Mol Biol 294 (5), 1351 (1999).
- 17. C. R. Ingrell, M. L. Miller, O. N. Jensen et al., Bioinformatics 23 (7), 895 (2007).
- 18. Y. R. Tang, Y. Z. Chen, C. A. Canchaya et al., Protein Eng Des Sel 20 (8), 405 (2007).
- 19. R. Linding, L. J. Jensen, G. J. Ostheimer et al., Cell 129 (7), 1415 (2007).

Release Note

- 1. Jan. 1st, 2008, the online service and the local stand-alone packages of GPS 2.0 were released. The stand-alone software of GPS 2.0 could support Windows Operating Systems.
- Jan. 29th, 2008, a bug was found that the version 2.0 couldn't be used under non-English Operating Systems. We fixed the bug and released the version 2.0.1 beta version. We thank Dr. Miguel Angel Sanchez (Malaga, Spain) and Dr. Gilles Vachon (Universite J. Fourier, France) to send us feedbacks.
- 3. Apr. 13th, 2008, The GPS 2.0.1 was released, with online service and local packages, to support three major Operating Systems, including Windows, Linux/Unix and Mac. Also, the GPS 2.0.1 manual was updated and included in the packages.
- 4. Mar. 1st, 2009, The GPS 2.1 was released, with online service and local packages and support three major Operating Systems including Windows, Linux/Unix and Mac. In this version, the newly developed motif length selection (MLS) method was introduced and the robustness of the prediction system was greatly improved.
- 5. Jul. 21st, 2009, The GPS 2.1.1 was released, with online service and local packages and support three major Operating Systems including Windows, Linux/Unix and Mac.
- 6. Sep. 13th, 2012, The GPS 2.1.2 was released, with online service and local packages and support three major Operating Systems including Windows, Linux/Unix and Mac.
- 7. Dec. 14th, 2014, The GPS 3.0 was released, with online service and local packages and support three major Operating Systems including Windows, Linux/Unix and Mac. In this version, *k-means* clustering, peptide selection (PS), and weight training (WT) procedures were added to enhance the prediction performance.
- 8. Jul. 20th, 2019, The GPS 5.0 was released, with online service and local packages and support three major Operating Systems including Windows, Linux/Unix and Mac. In this version, two novel methods of position weight determination (PWD) and scoring matrix optimization (SMO) were developed to improve the performance for kinase-specific phosphorylation sites prediction. Meanwhile, the GPS 5.0 manual was updated and included in the packages.