

PepString: инструмент для поиска коротких пептидов

Мацуга Д.Г.^{1*}, Суваан Б.С.², Козин С.А.³, Анашкина А.А.^{2,3}

¹ ФГБОУ ВО Астраханский ГМУ Минздрава России, Астрахань, Россия

² ФГАОУ ВО Первый МГМУ им. И.М. Сеченова Минздрава России (Сеченовский Университет), Москва, Россия

³ ФГБУН Институт молекулярной биологии им. В.А. Энгельгардта РАН, Москва, Россия

* HitokiriDeepWeb@yandex.ru

Ключевые слова: короткая последовательность; PepString; короткий пептид; пептидный поиск

Мотивация и цель: В последнее время появляется все больше работ по изучению пептидов различных размеров и их роли в регуляции различных процессов в организме, а также по поиску потенциальных сайтов связывания в белках. Короткие пептиды могут появляться в клетке несколькими путями, например, в процессе созревания белков или в результате деградации. В настоящее время на сайте <https://www.uniprot.org> нет возможности выполнять поиск по пептидам размером менее 7 аминокислот. Кроме того, Uniprot не позволяет выбор произвольного таксона для поиска, например, «млекопитающие», а только конечный таксон в цепочке (вид/штамм). Целью данной работы было создание веб-интерфейса для поиска происхождения коротких пептидов с возможностью их сортировки по таксономии.

Методы и алгоритмы: Для создания базы данных последовательностей белков на основе UniprotKB (SwissProt+TrEMBL) и ее автоматического обновления использовался Python и PostgreSQL; веб-интерфейс написан с помощью фреймворка Django.

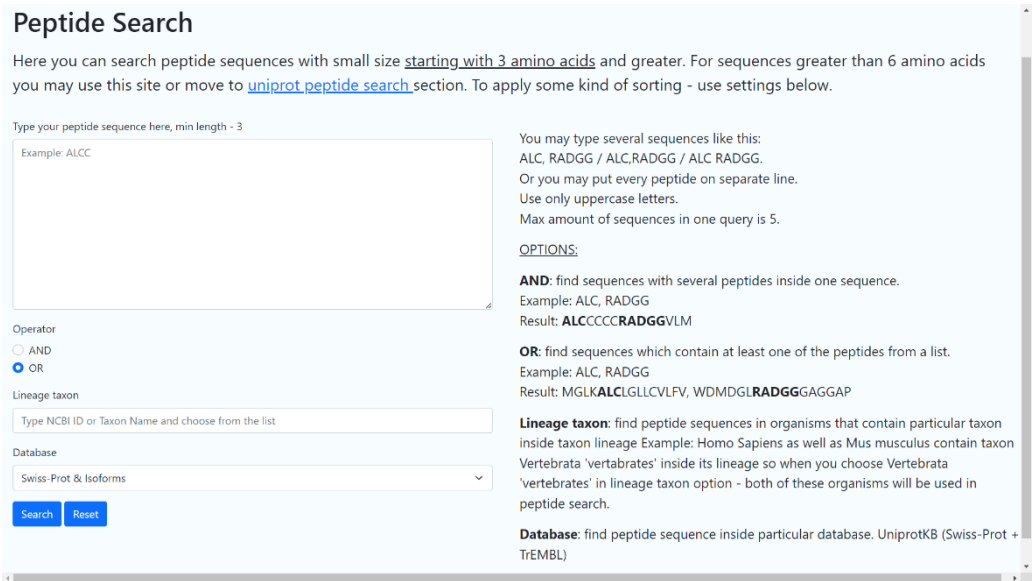


Рис. 2. Форма для запроса поиска коротких пептидов сервера PepString (<http://pepstring.eimb.ru>)

Main Peptide search									
Entry	Entry name	Peptide name	NCBI ID	Organism name	Uniprot name	Rank	Length	Sequence	
Q9N2Y8	RUSC2_HUMAN	AP-4 complex accessory subunit RUSC2	9606	Homo sapiens[9606]	Homo sapiens	species	1516	MDSPPKLTGETLVHHIPLVHCQVDRQCCGAGGGGGSTF	
Q80U22	RUSC2_MOUSE	AP-4 complex accessory subunit RUSC2	10090	Mus musculus[10090]	Mus musculus	species	1514	MDSPPKLTGETLVHHIPLVHCQVDRQCCGAGGGGGGTI	
Q6PDK2	KMT2D_MOUSE	Histone-lysine N-methyltransferase 2D	10090	Mus musculus[10090]	Mus musculus	species	5588	MDSQKPPAEDKSDPAADGLAAPEKPGATEPDLILCIGEV	

The amount of species is 2 of 9772 (Mammalia[40674] species);

Total results: 3

[Download CSV](#)
[Download FASTA](#)
[Download Tax info](#)

Page 1 of 1.

© 2024

Рис. 3. Форма выдачи результатов запроса поиска коротких пептидов. Перечислены все белковые последовательности организмов выбранного таксона, содержащие подпоследовательность, совпадающую с запросом

Результаты: Создан сервер PepString (<http://pepstring.eimb.ru>), позволяющий искать короткие пептиды как в нужном таксоне, так и во всей базе данных, причем запрос может быть построен на основе наличия нескольких точно совпадающих фрагментов в одной последовательности (до пяти таких фрагментов), а также наличия хотя бы одного из перечисленных фрагментов в последовательности белка. Кроме того, можно искать последовательности как в UniprotKB, так и отдельно в SwissProt. Скриншот страницы запроса представлен на рис. 1. На рис. 2 приведена форма выдачи результатов запроса. Результаты пользователь может сохранить в формате FASTA, CSV, а также может получить подробную информацию о таксономии найденных последовательностей.

Выводы: Создан удобный инструмент для поиска коротких последовательностей пептидов с возможностью сортировки по таксономии.

Финансирование: Исследование поддержано грантом Российского научного фонда № 19-74-30007 на тему «Разработка новых фармакологических мишеней и взаимодействующих с ними низкомолекулярных химических соединений для лечения болезни Альцгеймера».

PepString: a tool for short peptide search

Matsuga D.G.^{1*}, Suvaan B.S.², Kozin S.A.³, Anashkina A.A.^{2, 3}

¹FSBEI HE Astrakhan State Medical University of the Ministry of Health of Russia, Astrakhan, Russia

²FSAEI HE I.M. Sechenov First MSMU of the Ministry of Health of the Russian Federation (Sechenov University), Moscow, Russia

³Engelhardt Institute of Molecular Biology, RAS, Moscow, Russia

*HitokiriDeepWeb@yandex.ru

Key words: short sequence; PepString; short peptide; peptide search

Motivation and Aim: A lot of research works appear nowadays which explore various sized peptides and its role in different organism processes as well as research for finding potential binding sites in proteins. There are several ways for short peptides to appear in cell such as proteins maturation and degradation. Today there is no ability to search

sequences with size lesser than 7 amino acids on <https://www.uniprot.org>. Moreover, Uniprot taxonomy filtering doesn't provide ability to choose any taxon from lineage like mammalia. It only allows the last taxon in lineage (species/strain). The aim of this work was the creation of web interface for short protein origin search and ability for its taxonomy filtering.

Methods and Algorithms: Python and PostgreSQL were used to create protein sequence database on the base of UniprotKB (SwissProt+TrEmbl) and automatically update it; web interface was written using Django framework.

Peptide Search

Here you can search peptide sequences with small size [starting with 3 amino acids](#) and greater. For sequences greater than 6 amino acids you may use this site or move to [uniprot peptide search](#) section. To apply some kind of sorting - use settings below.

Type your peptide sequence here, min length - 3

Example: ALCC

Operator

☐ AND

☒ OR

Lineage taxon

Type NCBI ID or Taxon Name and choose from the list

Swiss-Prot & Isoforms

Search

Reset

You may type several sequences like this:
ALC, RADGG / ALC,RADGG / ALC RADGG.
Or you may put every peptide on separate line.
Use only uppercase letters.
Max amount of sequences in one query is 5.

OPTIONS:

AND: find sequences with several peptides inside one sequence.
Example: ALC, RADGG
Result: **ALCCCCRADGGVLM**

OR: find sequences which contain at least one of the peptides from a list.
Example: ALC, RADGG
Result: **MGLKALCLGLLCVLFV, WDMMDGLRADGGGAGGAP**

Lineage taxon: find peptide sequences in organisms that contain particular taxon inside taxon lineage Example: Homo Sapiens as well as Mus musculus contain taxon Vertebrata 'vertebrates' inside its lineage so when you choose Vertebrata 'vertebrates' in lineage taxon option - both of these organisms will be used in peptide search.

Database: find peptide sequence inside particular database. UniprotKB (Swiss-Prot + TrEMBL)

Fig. 4. Form that used to query for searching short peptides on the PepString server (<http://pepstring.eimb.ru>)

Main

Peptide search

Entry	Entry name	Peptide name	NCBI ID	Organism name	Uniprot name	Rank	Length	Sequence
Q8N2Y8	RUSC2_HUMAN	AP-4 complex accessory subunit RUSC2	9606	Homo sapiens[9606]	Homo sapiens	species	1516	MDSPPKLTGETLVHHIPLVHCQVPRQCCGAGGGGGSTF
Q8QU22	RUSC2_MOUSE	AP-4 complex accessory subunit RUSC2	10090	Mus musculus[10090]	Mus musculus	species	1514	MDSPPKLTGETLVHHIPLVHCQVPRQCCGAGGGGGGTI
Q6PDK2	KMT2D_MOUSE	Histone-lysine N-methyltransferase 2D	10090	Mus musculus[10090]	Mus musculus	species	5588	MDSQKPPAEDKDSDAADGLAAPEKPGATEPDLPLCIGVS

The amount of species is 2 of 9772 (Mammalia[40674] species);

Total results: 3

Download CSV

Download FASTA

Download Tax info

Page 1 of 1.

© 2024

Fig. 5. Result table of short peptide search. All protein sequences of the chosen taxon which contain the needed subsequence are listed

510

BGRS/SB-2024

Results: PepString server has been created (<http://pepstring.eimb.ru>) that allows user to search short peptides in a particular taxon or search it without filtering along the whole database. Also, query may be based on the presence of several fragments in the same sequence (five or less fragments) or on the presence of at least one fragment from a list in a sequence. In addition, there is a choice to search sequences in UniprotKB or in SwissProt. Page screenshot of the form is presented in the Fig. 1. In the Fig. 2 the result table is being shown. User also may save the result of query in FASTA and CSV formats and get thorough information about taxonomy of the resulting proteins.

Conclusion: A convenient tool for short peptide searching with taxonomy filtering was created.

Funding: The study is supported by grant from Russian science foundation No. 19-74-30007 on the topic “Development of new pharmacological targets and small molecular chemical compounds interacting with them for the treatment of Alzheimer’s disease”.