РАЗМЕР ПРИРОДНЫХ ЛИНЕЙНЫХ ПЕПТИДНЫХ СТРУКТУР

Замятнин А.А., Белозерская Т.А.
Институт биохимии им. А.Н. Баха
ФИЦ "Фундаментальные основы биотехнологии» РАН
Ленинский просп., 33, г. Москва,119071, РФ; e-mail: aaz@inbi.ras.ru
Поступила в редакцию: 25.07.2019

Аннотация. Размер линейных молекул пептидной природы может рассматриваться как число содержащихся в них аминокислотных остатков (p). Целью работы был анализ области существования и встречаемости различных природных пептидных структур с различной величиной p. В работе использовались данные базы SwissProt о более чем $5{,}6\cdot10^5$ первичных структурах, определенных полностью (sequence complete). Из них нами были удалены структуры, содержащие нестандартные аминокислотные остатки, а также идентичные копии аминокислотных последовательностей. В результате для анализа использовано 463450 различных последовательностей длиной 2 до 35213 аминокислотных остатков. Анализ показал, что у разных биологических доменов и царств число встречающихся пептидных структур на шкале p характеризуются разными областями существования, а формы профилей полученных кривых близки к ряду известных распределений. В то же время у них могут наблюдаться острые высокие пики, свидетельствующие о наличии большого количества специфических белков с одинаковой величиной p. Таким образом продемонстрировано существование множества разных первичных структур белков одинаковой длины, обладающих одинаковыми функциями. Рассмотрены возможные причины особенностей полученных распределений.
Ключевые слова: *пептид, белок, аминокислотная последовательность, база данных UniProt.*

ВВЕДЕНИЕ

В общем случае пептидами называются природные вещества, состоящие из субмолекулярных блоков (двадцати стандартных аминокислотных остатков) соединенных пептидной связью [1]. Очевидно, что минимальным линейным пептидом является химическая структура, в которой два аминокислотных остатка соединены одной пептидной связью (дипептид). Природные пептидные структуры, содержащие много аминокислотных остатков (полипептиды), обычно называются белками, и длинные молекулы такого типа могут состоять из десятков тысяч аминокислотных остатков. Таким образом, размер (длина) молекулы пептидной природы p может варьировать в широком диапазоне числа аминокислотных остатков.

Также велико разнообразие аминокислотных последовательностей. Это основано на широком спектре возможных комбинаций 20 стандартных аминокислотных остатков. С ростом длины пептидной структуры p число таких комбинаций N_p стремительно растет в соответствии с известной формулой

$$N_p = 20^p,$$

(где p – число аминокислотных остатков в молекуле). В случае дипептидов $p = 2$, $N_2 = 400$, трипептидов $p = 3$, $N_2 = 8000$, тетрапептидов $p = 4$, $N_2 = 160000$ и т.д., а в конце интервала $2 \le p \le 50$, характеризующего олигопептиды [2, 3], т.е. при $p = 50$, это значение достигает величины $N_{50} = \sim10^{34}$. Конечно, не все комбинации аминокислотных остатков могут существовать в природе, но все же разнообразие возможных первичных структур следует признать гигантским. В данной работе изучено распределение числа различных (уникальных) природных пептидных структур по длине, т.е. по числу аминокислотных остатков p.

ОБЪЕКТ И МЕТОДЫ ИССЛЕДОВАНИЯ

Для получения информации о природных пептидных структурах наиболее часто используется база данных UniProt, в которой объединены данные о расшифрованных первичных структурах экспериментально изученных белков (SwissProt) [4] и базы TrEMBL [5] о первичных структурах, полученных в результате трансляции (Tr) нуклеотидных последовательностей на язык аминокислот. На момент исследования база UniProt содержала сведения о 159 022 877 аминокислотных последовательностях, полученных для представителей архей, прокариот и эукариот. Минимальное число аминокислотных остатков $p = 2$ в ней содержат три олигопептида разного происхождения, а максимальное ($p = 74\ 488$) – один транслированный бактериальный полипептид. Несмотря на такой большой диапазон p, большинство аминокислотных последовательностей сосредоточено в интервале $2 \le p \le 1000$ (154 685 385, т.е. более 97%). Распределение в этом интервале представлено на рисунке 1а. Характерной особенностью распределения является несколько острых пиков, свидетельствующих о существенно большем числе пептидных структур с данным числом аминокислотных остатков p по сравнению с соседними (близкими) величинами p. Среди них особенно выделяется пик при $p = 219$. Рассмотрение 594 841 структур, соответствующих этому пику, показало, что 227 570 из них являются белком cytochrome c oxidase subunit 1.

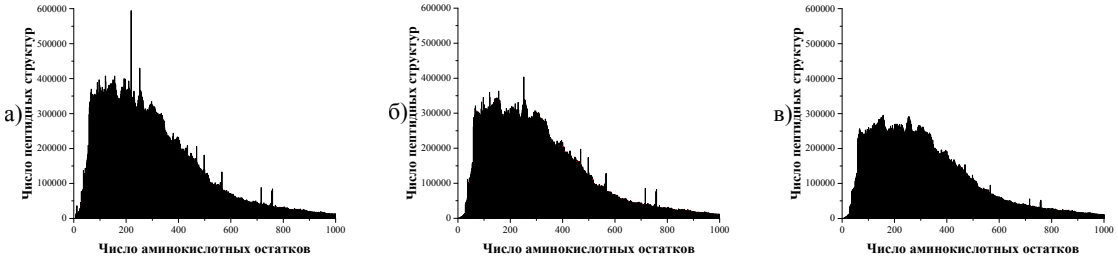

Рисунок 1. Зависимость числа аминокислотных последовательностей базы UniProt от числа аминокислотных остатков. (а) все последовательности; (б) последовательности, не являющиеся фрагментами; (в) последовательности, не являющиеся фрагментами, не содержащие нестандартные (O и U) или неидентифицированные (X) аминокислотные остатки, а также уникальные последовательности (без идентичных полных копий)

Данная величина примерно равна той части пика, которая возвышается над общей массой пептидных структур на рисунке 1а.

Однако данное распределение не дает представления об истинном числе встречаемости аминокислотных последовательностей в природе. Эти данные некорректны, поскольку в базе UniProt помимо полных последовательностей (complete) содержатся данные и о 15 495 873 неполных структурах (почти 10%), являющихся фрагментами (fragment). Распределение, представленное на рисунке 1б, характеризует массив 143 527 004 природных пептидных структур после исключения всех фрагментов из рассмотрения. В данном распределении также выявляется несколько пиков, но их величины и положение несколько отличаются от представленных на рисунке 1а.

Как выяснилось в процессе исследования, в базе UniProt имеются также последовательности, содержащие аминокислотные остатки, обозначенные буквами, не использующимися при описании стандартных остатков: O (оксипролина), U (α-аминомасляная кислоты), и X (неидентифицированного). Последовательности, в описании которых содержатся эти буквы, также были исключены из рассмотрения.

В базе UniProt имеется и большое число полностью идентичных аминокислотных последовательностей, полученных, как правило, для представителей разных, но таксономически близких живых организмов. Для того чтобы анализ был ограничен рассмотрением только разных структур, все дубликаты также были исключены из рассмотрения (рисунок 1в).

Описанные выше процедуры проводились на высокоскоростном сервере, позволяющем обрабатывать большие массивы информации. Возможность выделения полных последовательностей (complete) предусмотрена на сайте UniProt. В то же время экстракция «чистых» (содержащих только символы стандартных аминокислотных остатков) последовательностей и удаление дубликатов проводилось с помощью специально созданных программ. Специальная программа была сделана также для формирования распределений числа пептидных структур по числу аминокислотных остатков. Описанное ниже исследование проведено на базе SwissProt, представляющую часть базы UniProt и содержащую данные о выверенных (reviewed) аминокислотных последовательностях.

РЕЗУЛЬТАТЫ

На момент исследования база SwissProt содержала сведения о 560 118 аминокислотных последовательностях, полученных для представителей архей, прокариот и эукариот. Минимальное число аминокислотных остатков $p = 2$ в ней содержали два олигопептида разного происхождения, а максимальное ($p = 35\ 213$) – белок тайтин (titin) мыши [6]. Оказалось, что несмотря на такой большой диапазон p, большинство аминокислотных последовательностей сосредоточено в интервале $2 \le p \le 1000$ (542 302, т.е. около 97%). Распределение всех пептидных структур в этом интервале представлено на рисунке 2а. Характерной особенностью этого распределения является множество острых пиков, свидетельствующих о значительно большем числе пептидных структур с данным числом аминокислотных остатков p по сравнению с соседними величинами p. Среди них особенно выделяется два пика при $p = 156$ и 379. Рассмотрение структур, соответствующих этим пикам показало, что при $p = 156$, 2216 пептидных структур представляют собой большой набор самых разных белков. Однако в случае $p = 379$, из 1889 белков 1048 (две трети) являются митохондриальный цитохромом *b*. Данная величина примерно равна той части пика, которая возвышается над общей массой пептидных структур на рисунке 2а.

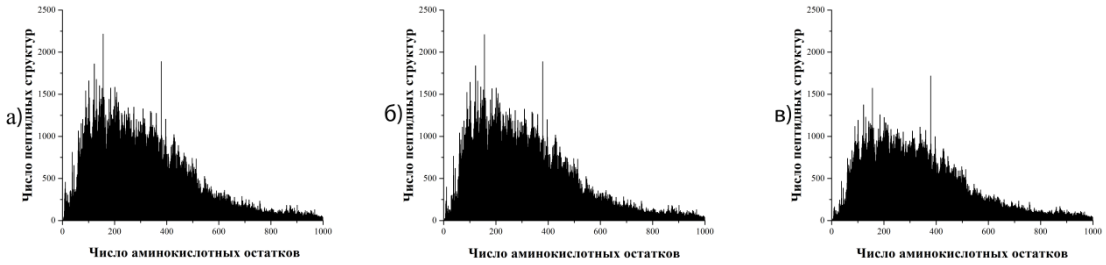

Рисунок 2. Зависимость числа аминокислотных последовательностей базы SwissProt от числа аминокислотных остатков. (а) все последовательности; (б) последовательности, не являющиеся фрагментами; (в) последовательности, не являющиеся фрагментами, не содержащие нестандартные (O и U) или неидентифицированные (X) аминокислотные остатки, а также уникальные последовательности (без идентичных полных копий)

Как и в случае UniProt, в данных базы SwissProt нами были удалены все неполные последовательности, являющиеся фрагментами (fragment). Распределение для полных последовательностей (complete) представлено на рисунке 2б. Этот рисунок мало отличается от рисунка 2а, поскольку фрагментарные последовательности базы SwissProt составляют лишь 1,6% (9167) от общего числа данных этой базы. Некоторое отличие заметно лишь в области олигопептидов, т.е. при малых значениях величины p.

Существенно измененное распределение получено нами после удаления последовательностей с нестандартными неидентифицированными аминокислотными остатками, а также после изъятия дубликатов (рисунок 2в). Отметим, что величина пика при p = 156 существенно уменьшилась, а при p = 379 осталась неизменной. Детальные данные о числе аминокислотных последовательностей, взятых для уже описанного и последующих анализов, собраны в таблице 1.

Из всей массы пептидных структур базы SwissProt нами были последовательно выделены и проанализированы уникальные последовательности таксономических групп различного уровня. На первом, самым высоким уровне были выбраны домены архей, прокариот и эукариот [7]. Несмотря на то, что число структур в этих доменах существенно различается (табл.1), общий характер распределения (рис. 3) у них одинаков. При этом подавляющая часть аминокислотных последовательностей, как и ранее, сосредоточена в интервале $2 \le p \le 1000$ ((99,1% у архей, 98,7% у прокариот и 92,7% у эукариот). У архей (рис. 3а) заметных пиков практически не наблюдается, у прокариот (рис. 3б) таких пиков довольно много, а у эукариот (рис. 3в) большинство пиков не слишком велики за исключением одного при p = 379.

На следующем таксономическом уровне раздельно проанализированы уникальные аминокислотные последовательности царств: грибов, растений и животных (рис. 4). Общий характер полученных распределений свидетельствует о том, что у грибов пики практически не наблюдаются (рис. 4а), у растений их довольно много (рис. 4б), а в случае животных также выделяется целый ряд пиков, среди которых пик при p = 379 выделяется еще более ярко, чем в случае распределения для эукариот (рис. 3в).

Таблица 1. Данные о числе аминокислотных последовательностей в различных таксономических группах и биологических видов в базе SwissProt

Таксономические группы	латинское название группы, вида	все в SwissProt	без фрагментов	уникальные	p_{min}	p_{max}
Все		560118	550951	463450	2	35213
Эукариоты	*Eukaryota*	189697	182592	174063	2	35213
Животные	*Metazoa*	106843	102155	98070	2	35213
человек	*Homo sapiens*	20421	20421	20358	2	34350
Растения	*Viridiplantae*	39930	38014	35093	5	5400
резуховидка	*Arabidopsis thaliana*	15856	15829	15768	5	5400
Грибы	*Fungi*	34084	33841	32431	3	11842
дрожжи	*Saccharomyces cerevisiae*	7919	7912	7290	16	4910
Прокариоты	*Bacteria*	334009	332477	255373	7	10746
бактерии грам-	*Escherichia coli*	23138	23121	10153	7	3289
бактерии грам+	*Staphylococcus aureus*	10175	10164	3171	9	10746
Археи	*Archaea*	19554	19482	18452	25	9159
Вирусы	*Viruses*	16858	16400	15607	11	7182

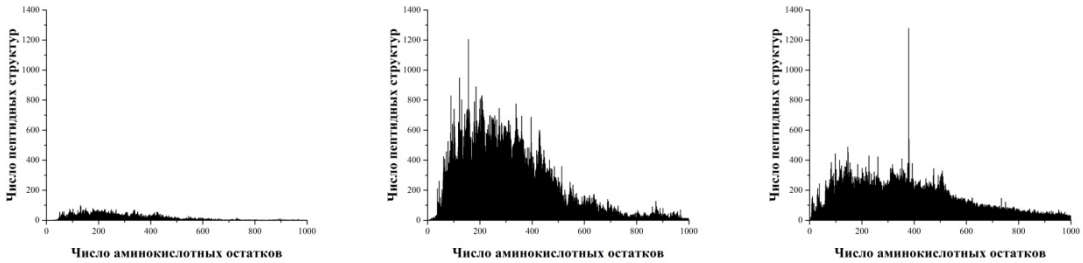

Рисунок 3. Распределение числа уникальных пептидных структур в базе SwissProt в различных биологических доменах. (а) археи, (б) прокариоты, (в) эукариоты

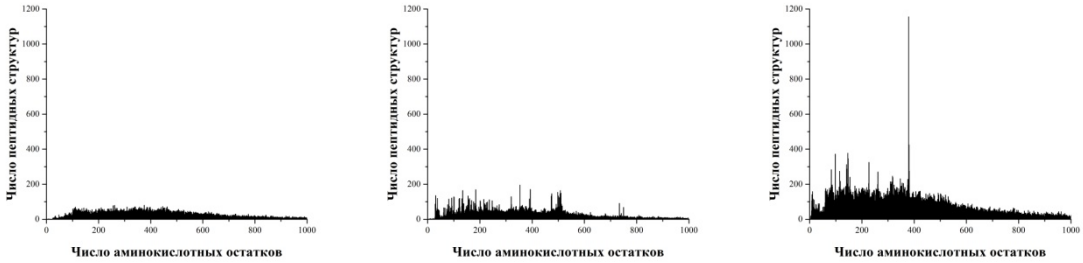

Рисунок 4. Распределение числа уникальных пептидных структур в базе SwissProt у различных эукариот. (а) грибы, (б) растения, (в) животные

Нами также были проанализированы уникальные аминокислотные последовательности отдельных представителей грибов, растений и животных (рис. 5). В случае дрожжей (*Saccharomyces cerevisiae*) получено распределение с заметным пиком при $p = 440$ (рис. 5а). Оказалось, что большинство структур с этим числом аминокислотных остатков представляют собой различные белки, называемые transposon polyprotein. Довольно много не очень больших пиков проявляется у растения резуховидки (рис. 5б). В то же время у человека ярко выделяются два пика (рис. 5в). Один из них при $p = 117$ более чем на 50% составлен из различных иммуноглобулинов, а другой – при $p = 312$, также более чем наполовину, характеризует наличие большого числа белков обонятельных рецепторов.

Нами получены также распределения для совокупности пептидных структур грам-отрицательных (рис. 6а) и грам-положительных (рис. 6б) бактерий и для всех вирусов (рис. 7). Все эти распределения характеризуются большим числом заметных пиков. При этом многие пики в распределении вирусных структур отличаются большими величинами. Особенно выделяется пик при $p = 252$, в котором значительную часть занимает белок matrix protein у Influenza A virus, играющий ключевую роль в репликации вирусов, а пик при $p = 498$ почти полностью соответствует белку nucleoprotein (также у Influenza A virus), защищающим вирусную РНК от нуклеаз.

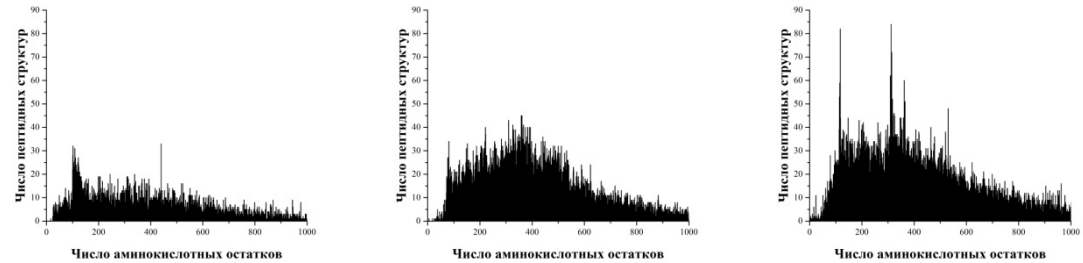

Рисунок 5. Распределение числа уникальных пептидных структур в базе SwissProt у отдельных представителей эукариот (грибов, растений, животных). (а) *Saccharomyces cerevisiae*, (б) *Arabidopsis thaliana*, (в) *Homo sapiens*

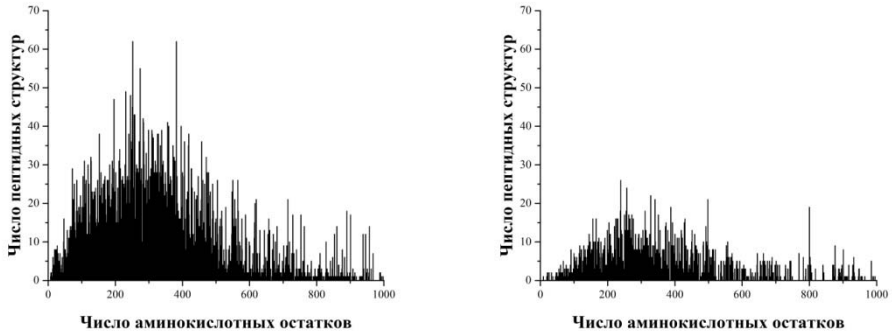

Рисунок 6. Распределение числа уникальных пептидных структур в базе SwissProt у представителей бактерий. (а) грам-отрицательные *Escherichia coli*, (б) грам-положительные *Staphylococcus aureus*

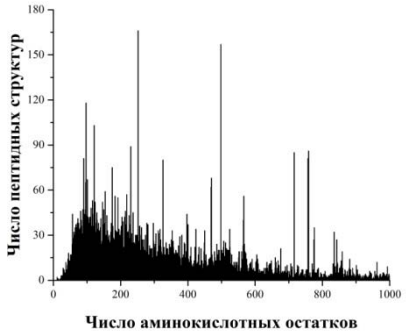

Рисунок 7. Распределение числа уникальных пептидных структур у вирусов в базе SwissProt

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Очевидно, что характер любого распределения нуждается в объяснении. Лучше всего объяснение могло бы быть сделано на основании аналитического выражения, выведенное из общих соображений о природе пептидных структур. Получение такого выражения пока что испытывает большие трудности, поскольку для этого необходим учет многих факторов: деталей генеза белков, особенности составляющих его элементов (физико-химическое разнообразие аминокислотных остатков), факторы эволюции и многое другое [8, 9].

Однако подбор математических выражений, позволяющий получать кривые, максимально близкие к реальным распределениям, проводился неоднократно. Так, для ряда отдельных организмов была использована функция логнормального распределения, которое вполне удовлетворительно описывало данные для 13 видов бактерий, 4 архей и 1 эукариота [10].

Этот же подход был осуществлен для описания распределений белков 1302 видов прокариот и 140 эукариот [11]. В указанном исследовании помимо логнормального распределения было использовано также распределение гамма типа. Совместное использование логнормального и гамма распределений [12] позволило авторам прийти к выводу о том, что средняя длина белков эукариот больше, чем у прокариот. Подобный вывод иллюстрируется и в данной нашей работе при рассмотрении рисунков 3б и 3в. Не менее успешно применялась также и лингвистическая модель Менсерат-Альтмана (Menzerath-Altmann) [13, 14]. С ее помощью было описаны данные 10 протеомов [15].

Однако во всех указанных работах основное внимание было уделено сглаживанию кривых распределений для подгонки к определенным математическим моделям. В то же время из приведенных нами результатов следует, что многочисленные пики практически у всех таксономических групп и отдельных видов живых организмов несут дополнительную информацию о пептидных структурах, формирующих эти распределения. Ярким примером является пик, прослеживаемый в распределении всех аминокислотных последовательностей базы SwissProt (рис. 2в), эукариот (рис. 3в) и животных (рис. 4в). Как уже было отмечено, этот пик представляет собой совокупность митохондриальных цитохромов *b*, имеющих одинаковую длину 379 аминокислотных остатков и обнаруженных только у животных. Дополнительный анализ распределения по длине цитохромов *b* животных (рис. 8) показал, что данные белки имеют преимущественную длину именно в 379 аминокислотных остатков (1028 белков). С несколько меньшим числом остатков насчитывается 99 белков, а с большим – 440.

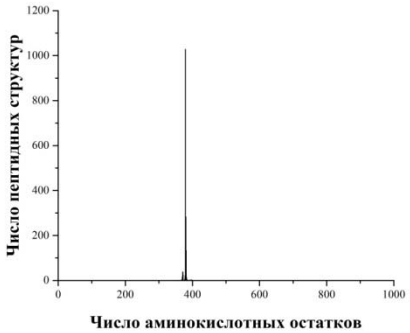

Рисунок 8. Распределение числа уникальных пептидных структур цитохрома *b* у эукариот в базе SwissProt

Столь большое число разных по аминокислотной последовательности, но с одинаковой величиной *p*, цитохромов *b* характерно именно для представителей животных. В SwissProt содержатся данные почти о 1700 разных аминокислотных последовательностей данного белка, полученных не только из животных. Они присутствуют и в бактериях, и растениях и в грибах. Однако все известные цитохромы *b* неживотного происхождения всегда содержат больше чем 379 аминокислотных остатков. Можно предположить, что минимальная величина этого трансмембранного белка в эволюционном процессе могла быть достигнута в результате отбора, приведшего к наиболее оптимальным размерам всех его 8 сайтов, пронизывающих мембрану [16].

Очевидно, что разнообразие функциональных свойств белков основано на разнообразии первичных структур их молекул. Можно найти много примеров тому, что малые одинаковые по длине пептидные структуры (олигопептиды) при одинаковом *p* обладают одинаковыми функциями [17]. В данной работе нами же, в частности, продемонстрировано, что существуют множества разных первичных структур белков одинаковой длины, обладающих одинаковыми функциями.

База UniProt постоянно пополняется. Однако уже сейчас большой объем информации, заключенный в ней, позволяет провести множество новых различных анализов связи размера, аминокислотных последовательностей и многих других физико-химических характеристик природных пептидных структур с многочисленными функциональными свойствами этих молекул.

Список литературы / References:

1. Sewald N., Jakubke H.-D. Peptides: *Chemistry and Biology*. Weinheim, WILEY-VCH Verlag GmbH, 2002, 562 p.

2. Замятнин А.А. Особенности совокупности природных олигопептидов. *Нейрохимия*, 2016, т. 33, № 4, с. 265-275. DOI: 10.1134/S1819712416040176. [Zamyatnin A.A. Features of the totality of natural oligopeptides. *Neurochemistry*, 2016, vol. 33, no. 4, pp. 265-275. DOI: 10.1134 / S1819712416040176. (In Russ.)]

3. Zamyatnin A.A. Structural–functional diversity of the natural oligopeptides. *Progr. Biophys. Mol. Biol*., 2018, vol. 133, pp. 1-8. DOI: 10.1016/j.pbiomolbio.2017.09.024.

4. Bairoch A., Boeckmann B. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res*., 1991, vol. 19, no 1, pp. 2247-2249. DOI: 10.1093/nar/19.suppl.2247.

5. Kneale G.G., Kennard O. The EMBL nucleotide sequence data library. *Biochem. Soc. Trans*., 1984, vol. 12, no. 6, pp. 1011-1014. DOI: 10.1042/bst0121011.

6. Church D.M., Goodstadt L., Hillier L.W., Zody M.C., Goldstein S., She X., Bult C.J., Agarwala R., Cherry J.L., DiCuccio M., Hlavina W., Kapustin Y., Meric P., Maglott D., Birtle Z., Marques A.C., Graves T., Zhou S., Teague B., Potamousis K., Churas C., Place M., Herschleb J., Runnheim R., Forrest D., Amos-Landgraf J., Schwartz D.C., Cheng Z., Lindblad-Toh K., Eichler E.E., Ponting C.P.; Mouse Genome Sequencing Consortium. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol*., 2009, vol. 7, no. 5, p. e1000112. DOI: 10.1371/journal.pbio.1000112.

7. Woese C.R., Kandler O., Wheelis M.L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA*, 1990, vol. 87, no. 12, pp. 4576-4579. DOI: 10.1073/pnas.87.12.4576.

8. Замятнин А.А. Биофизические проблемы олигопептидной регуляции. *Биофизика*, 2003, т. 48, № 6, с. 1030-1039. [Zamyatnin A.A. Biophysical problems of oligopeptide regulation. *Biophysics*, 2003, vol. 48, no. 6, pp. 1030-1039. (In Russ.)]

9. Замятнин А.А. Биохимические проблемы олигопептидной регуляции. *Биохимия*, 2004, т. 69, № 11, с. 1565-1573. DOI: 10.1007/s10541-005-0073-8. [Zamyatnin A.A. Biochemical problems of oligopeptide regulation. *Biochemistry*, 2004, vol. 69, no. 11, pp. 1565-1573. (In Russ.)]

10. Ramakumar S. Stochastic dynamics modeling of the protein sequence length distribution in genomes: implications for microbial evolution. *Physica A: Statistical Mechanics and its Applications*, 1999, vol. 273, no. 3, pp. 476-485. DOI: 10.1016/S0378-4371(99)00370-2.

11. Tiessen A., Pérez-Rodríguez P, Delaye-Arredondo L.J. Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. *BMC Research Notes*, 2012, vol. 5, no. 85. DOI: 10.1186/1756-0500-5-85.

12. Jhang G. Protein-length distributions for the three domains of life. *Trends in Genetics*, 2000, vol. 16, no. 3, pp. 107-109. DOI: 10.1016/S0168-9525(99)01922-8.

13. Menzerath P. *Architektonik des deutschen Wortschatzes*. Bonn, 1954, 131 p.

14. Altmann G. Prolegomena to Menzerath's law. *Glottometrika,* 1980, vol. 2, pp. 1-10.

15. Eroglu S. Language-like behavior of protein length distribution in proteomes. *Complexity*, 2014, vol. 20, pp. 12-21. DOI: 10.1002/cplx.21498.

16. Esposti M.D., De Vries S., Crimi M., Ghelli A., Patarnello T., Meyer A. Mitochondrial cytochrome b: evolution and structure of the protein. *Biochim. Biophys. Acta*, 1993, vol. 1143, no 3, pp. 243-271. DOI: 10.1016/0005-2728(93)90197-n.

17. Замятнин А.А. Физико-химические и функциональные характеристики полной системы природных олигопептидов. *Актуальные вопросы биологической физики и химии*, 2018, т. 3, № 1, с. 225-235. [Zamyatnin A.A. Physico-chemical and functional characteristics of a complete system of natural oligopeptides. *Modern trends in biological physics and chemistry*, 2018, vol. 3, no. 1, p. 225-235. (In Russ.)]

SIZE OF THE NATURAL LINEAR PEPTIDE STRUCTURES

Zamyatnin F.F., Belozerskaya T.A.
A.N. Bach Institute of Biochemistry,
Research Center of Biotechnology Russian Academy of Sciences
Leninsky prosp., 33, Moscow 119071, Russia; e-mail: aaz@inbi.ras.ru

Abstract. The size of linear peptide molecules is considered as a number of amino acid residues (*p*) contained in them. The aim of this work was to analyze the region of existence and occurrence of various natural peptide structures with different *p*-values. We used SwissProt database contained more than 560000 complete primary structures. We have removed structures containing non-standard amino acid residues, as well as identical amino acid sequences. As a result, 463450 different sequences with a length of 2 to 35,213 amino acid residues were obtained for analysis. The analysis showed that the number of peptide structures on *p*-scale is characterized by different areas of existence in different biological domains and kingdoms, and the shapes of the profiles of the curves obtained are close to classical distributions. However, they can have sharp high peaks, indicating the presence of a large number of specific proteins with the same *p*-value. Possible reasons for the existence and features of such distributions are considered.

Key words: *peptide, protein, amino acid sequence, UniProt database.*