МЕЖДУНАРОДНЫЙ УНИВЕРСИТЕТ АСТАНА

Высшая школа естественных наук

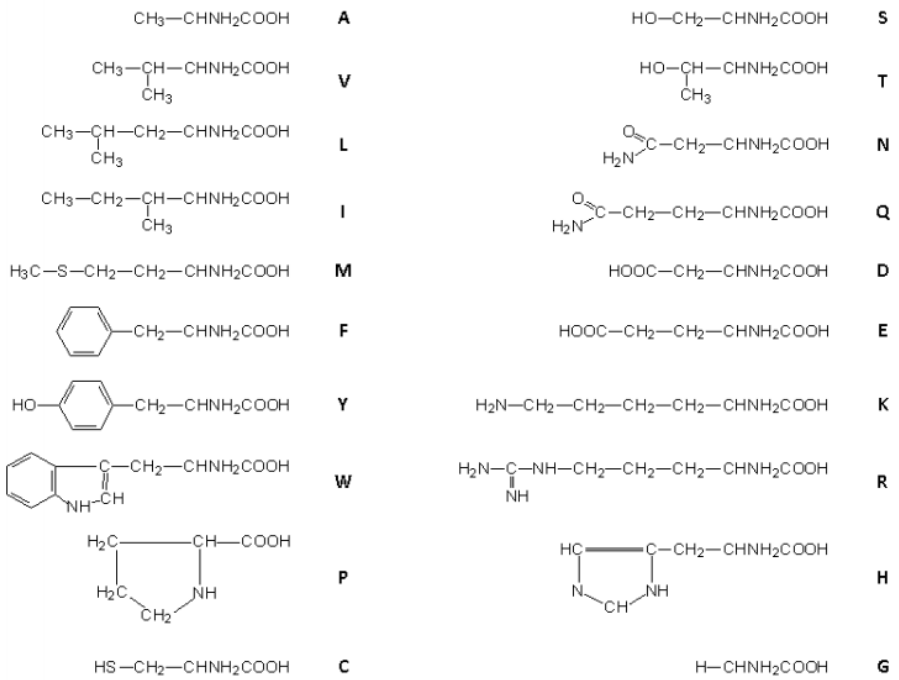
ОСНОВЫ БИОИНФОРМАТИКИ
(Практикум)

Поиск и сравнение последовательностей

ОРАЗОВ АЙДЫН ЕРҒАЛИҰЛЫ
PhD, и.о. ассоциированного профессора Высшей школы естественных наук международного университета Астана,
Заведующий научно-исследовательской лаборатории изучения окружающей среды «NatureLaB»

АСТАНА, 2025

Работа с первичной структурой биологических макромолекул – то есть с нуклеотидной или аминокислотной последовательностью – является, пожалуй, наиболее распространенной задачей биоинформатики. Этому способствует, прежде всего, относительная легкость получения информации о первичной структуре нуклеиновых кислот и белков на современном этапе развития биохимии и молекулярной биологии. На сравнении последовательностей нуклеиновых кислот основаны современные методы классификации живых организмов, а анализ множества гомологичных белковых последовательностей позволяет получить информацию о структурных и функциональных особенностях молекулы.



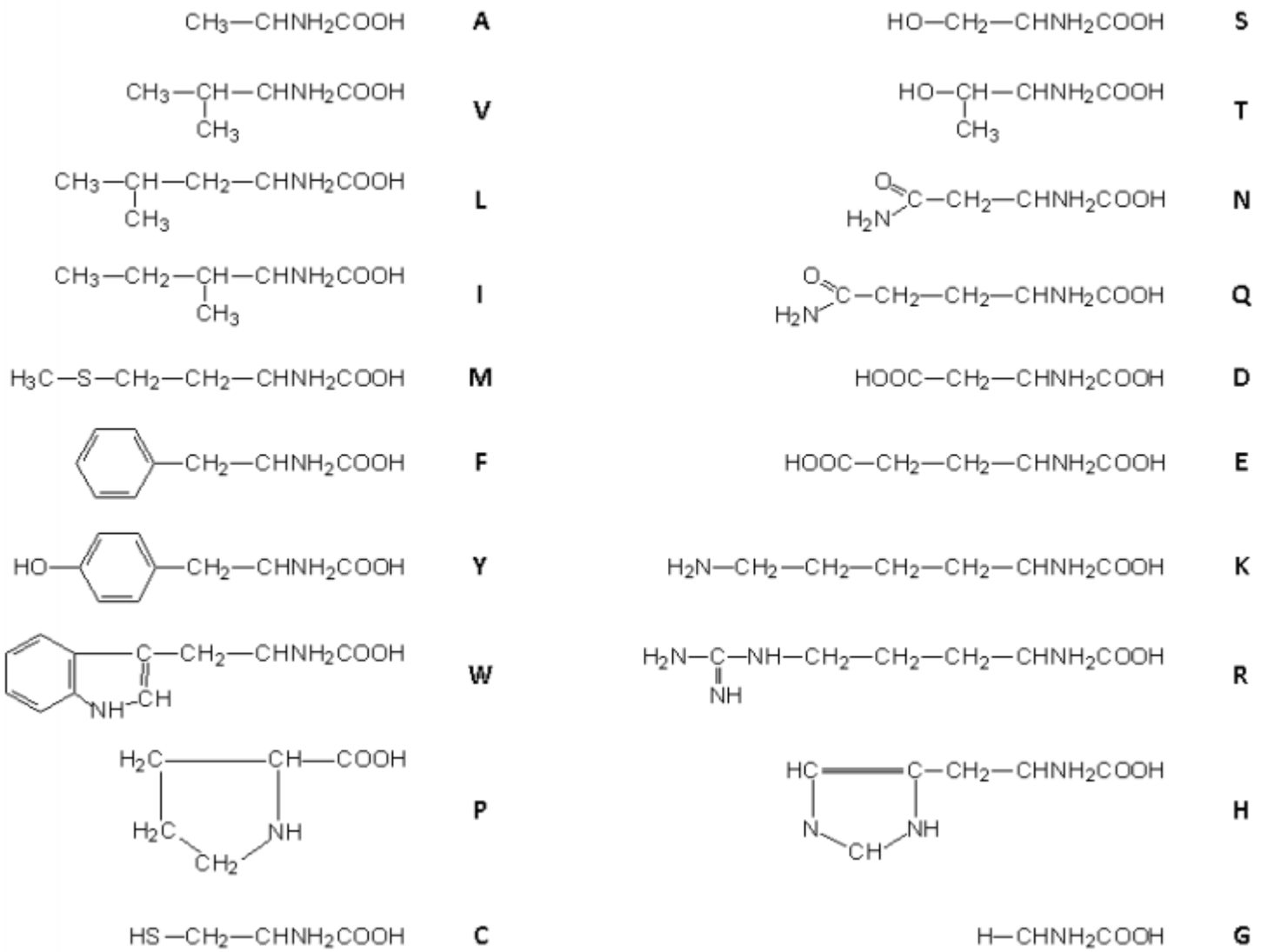
Однобуквенные обозначения аминокислотных остатков

Последовательности белков и нуклеиновых кислот очень легко могут быть записаны в виде текста, поскольку и те, и другие являются линейными, то есть не ветвятся. Первичная структура описывается как последовательность мономеров:

у белков - в направлении от N-конца к С-концу;

у нуклеиновых кислот – от 5'-конца к 3'-концу.

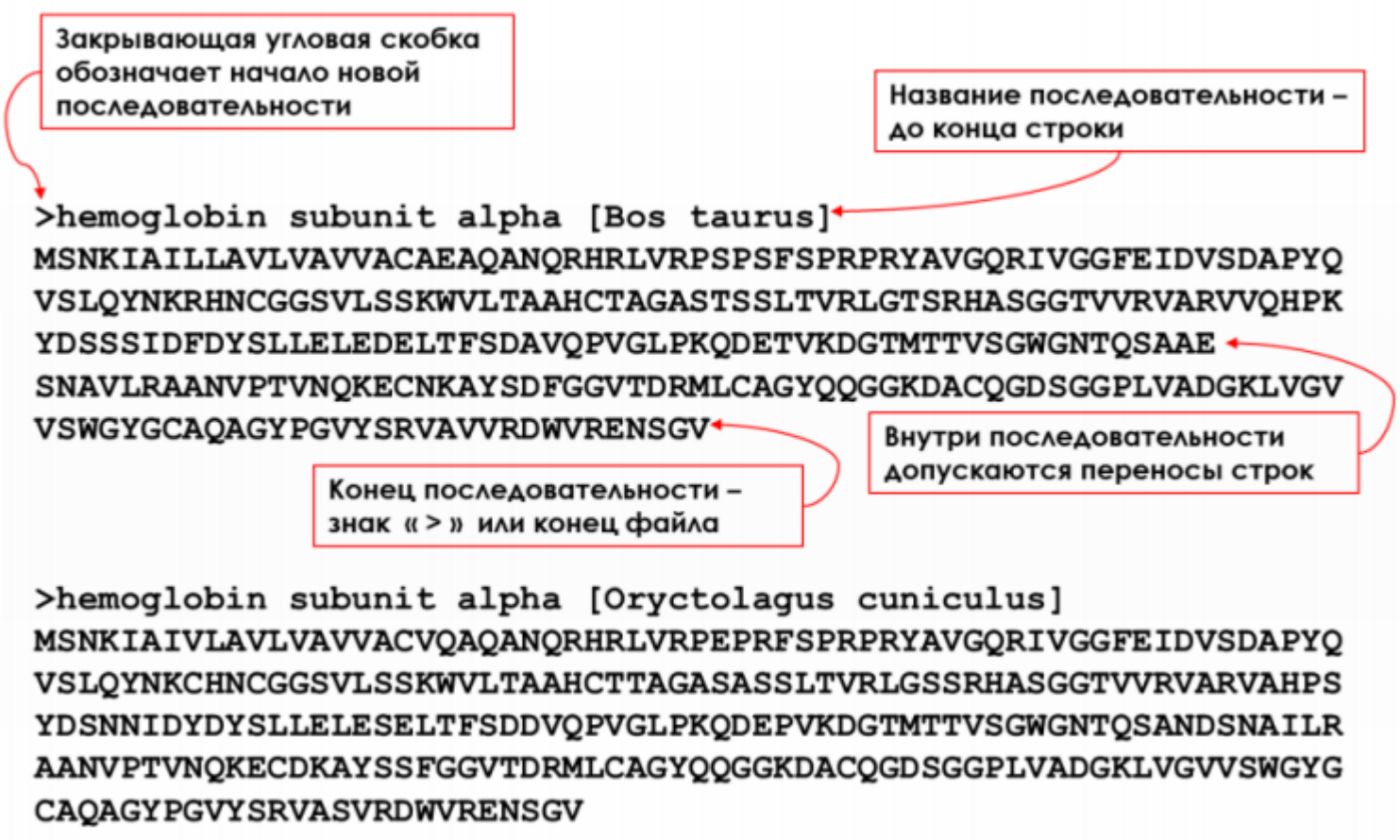
Последовательности записываются буквами английского алфавита в однобуквенной нотации.



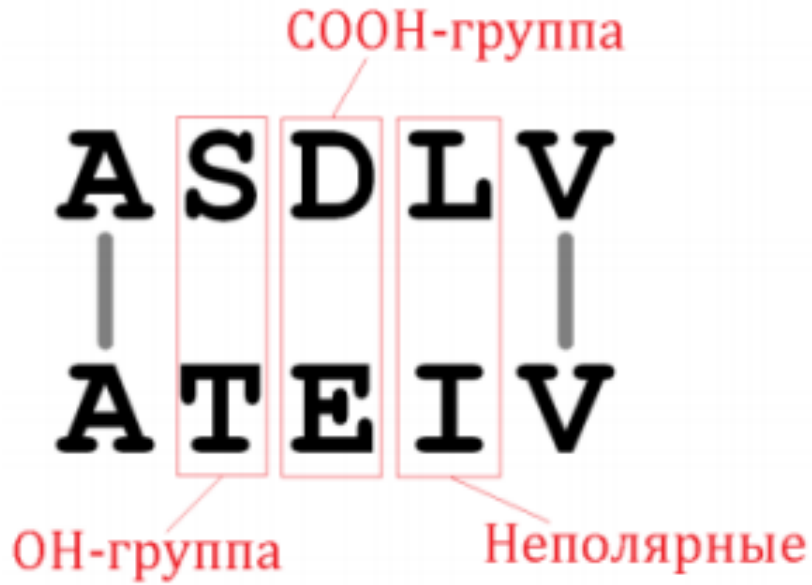
Однобуквенные обозначения нуклеотидных остатков

Символ	Обозначаемые остатки	Мнемоническое правило
A	A	**A**denine
C	C	**C**ytosine
G	G	**G**uanine
T	T	**T**hymine
U	U	**U**racil
R	A или G	pu**R**ine
Y	C, T или U	p**Y**rimidines
K	G, T или U	**К**етон
M	A или C	Содержит а**М**иногруппу
S	C или G	**S**trong interaction (три водородные связи)
W	A, T или U	**W**eak interaction (две водородные связи)
B	Любой, кроме A	после A
D	Любой, кроме C	после C
H	Любой, кроме G	после G
V	Любой, кроме T и U	после U
N	Любой	**N**ucleic acid
X	Любой	Неизвестная величина

Для записи биологических последовательностей используется формат Fasta – текст, размеченный особым образом. Каждая последовательность может быть предварена заголовком, содержащим знак ">" и название последовательности. Заголовок занимает отдельную строку; после переноса строки записывается последовательность. Текстовый файл (или поле ввода), содержащий несколько последовательностей (естественно, с разными заголовками), обозначается термином "multi-fasta". Если в файле или поле ввода содержится только одна последовательность, она может не иметь заголовка.



Замечательное свойство молекулярных последовательностей, способствующее их использованию в качестве источника данных для классификации организмов, заключается в возможности количественной оценки их сходства. Для количественной оценки сходства последовательностей используют выравнивание – расположение последовательностей одна под другой с сохранением порядка символов таким образом, чтобы достичь совпадения или сходства мономеров в наибольшем числе позиций. Выравнивание производится путем смещения частей последовательностей относительно друг друга и добавления разрывов в последовательности.

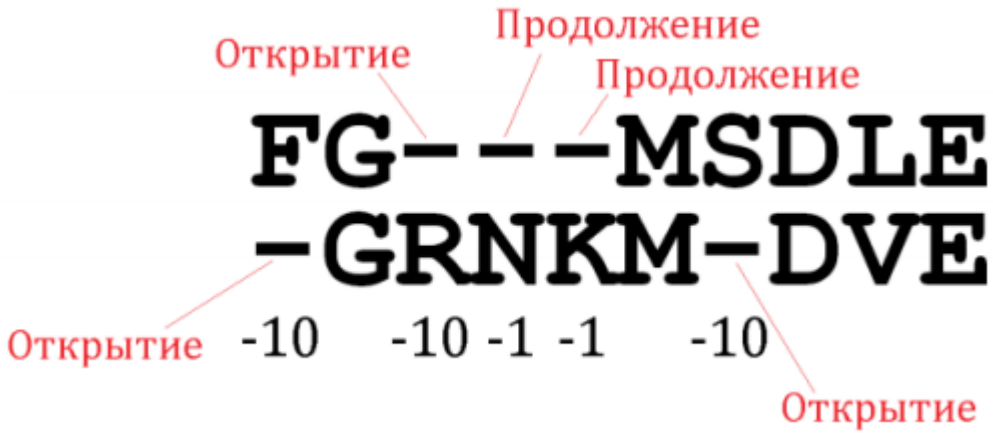


При количественном сравнении последовательностей биомолекул необходимо учитывать физико-химические свойства мономеров.

Для дифференцированной оценки совпадений или несовпадений различных мономеров используются матрицы замен, основанные на статистике аминокислотных или нуклеотидных замен в гомологичных молекулах с известной структурой. Числа на пересечениях символов показывают, насколько часто в молекулах происходит замена одного мономера на другой, а значит, насколько сильно соответствие друг другу этих мономеров в выравнивании свидетельствует о сходстве выравниваемых молекул.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala (A)	4																			
Arg (R)	−1	5																		
Asn (N)	−2	0	6																	
Asp (D)	−2	−2	1	6																
Cys (C)	0	−3	−3	−3	9															
Gln (Q)	−1	1	0	0	−3	5														
Glu (E)	−1	0	0	2	−4	2	5													
Gly (G)	0	−2	0	−1	−3	−2	−2	6												
His (H)	−2	0	1	−1	−3	0	0	−2	8											
Ile (I)	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu (L)	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys (K)	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met (M)	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe (F)	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro (P)	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser (S)	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr (T)	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp (W)	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr (Y)	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val (V)	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

Если выравнивание содержит разрывы, они также должны быть учтены. Очевидно, что разрывы – это признак различия в молекулах, поэтому наличие разрывов должно понижать счет выравнивания. Величина, отнимаемая от счета выравнивания для учета разрывов, называется штрафом за делецию (штрафом за разрыв) – gap penalty. Длинные разрывы в последовательностях, занимающие несколько позиций, оцениваются целиком – поскольку разрывы интерпретируются как мутации в одной из гомологичных молекул, длинный разрыв рассматривается как одна мутация, затронувшая сразу несколько мономеров. Для этого используются отдельные значения штрафов за «открытия» и за «продолжения» разрывов. Значения штрафов, как правило, выбираются исходя из характеристик выравниваемых последовательностей, особенностей конкретного алгоритма и сложившейся практики.



Расчет штрафов за делеции, если штраф за открытие разрыва равен 10, за продолжение – 1.

Ссылки:

https://www.ncbi.nlm.nih.gov/
http://www.uniprot.org/
https://lifemap-ncbi.univ-lyon1.fr/

https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D0%BB%D1%8B%D0%BD%D1%8C

https://ru.wikipedia.org/wiki/%D0%9E%D0%B4%D1%83%D0%B2%D0%B0%D0%BD%D1%87%D0%B8%D0%BA

https://xn--80abvyzg.xn--p1ai/%D0%B1%D0%B8%D0%BE%D0%B8%D0%BD%D1%84%D0%BE%D1%80%D0%BC%D0%B0%D1%82%D0%B8%D0%BA%D0%B0/#11