

Основы биоинформатики в исследовании вирусов гриппа

{ Краткий курс

Цели и задачи курса

- Научиться применять методы биоинформатики к данным эпидемического надзора за гриппом.
- Научиться работать с основными базами данных генетических последовательностей.
- Научиться строить множественные выравнивания генетических последовательностей и анализировать их результаты.
- Научиться строить, анализировать и интерпретировать филогенетические деревья.

Программное обеспечение

- MEGA6 (<http://www.megasoftware.net/>) – программа для филогенетического анализа
- Ugene (<http://ugene.unipro.ru/>) – набор биоинформатических инструментов
- Notepad++ (<http://notepad-plus-plus.org/>) – текстовый редактор
- Paint.NET (<http://www.getpaint.net/>) – графический редактор

Понятие биоинформатики

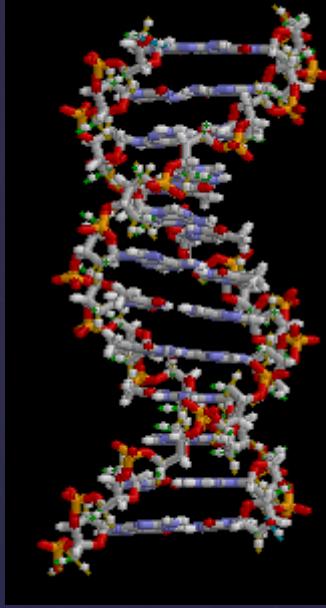
- ¶ Термин «биоинформатика» был впервые введен голландским биологом Полиной Хогевег в 1970 для описания процессов передачи информации в биологических системах.
- ¶ Современная биоинформатика возникла в конце 70-х годов XX века с появлением эффективных методов расшифровки последовательностей ДНК.
- ¶ На сегодняшний день биоинформатика представляет собой широкий спектр научных направлений, лежащих на стыке биологии, медицины, статистики, математики и информатики и занимающихся обработкой и анализом различных биологических данных.
- ¶ В приложении к молекулярной генетики биоинформатика занимается сборкой и анализом генетических последовательностей.

Генетические последовательности

- ❖ Генетические последовательности – это последовательности мономеров, определяющих первичную структуру основных биополимеров – нуклеиновых кислот и белков.
- ❖ Генетические последовательности можно разделить на 2 типа: нуклеотидные и аминокислотные.

Нуклеотидные последовательности

- ❖ Нуклеотидные последовательности – это последовательности мономеров нуклеиновых кислот – биополимеров, основной функцией которых является хранение и передача наследственной информации.
- ❖ Алфавит нуклеотидных последовательностей содержит 4 «буквы», соответствующие 4 азотистым основаниям, однако также используются дополнительные символы для обозначения вырожденных позиций.
- ❖ Нуклеиновые кислоты способны образовывать двуцепочечную спираль из двух антипараллельных цепей, между азотистыми основаниями которых возникают водородные связи. В таких случаях принято выделять кодирующую цепь положительной полярности и комплементарную ей цепь отрицательной полярности. Последовательности одной цепи можно получить, взяв в обратном порядке последовательность, комплементарную другой.



Символ	Обозначает	Объяснение
G	G	Guanine - гуанин
A	A	Adenine - аденин
T(U)	T(U в PHK)	Thymine(Uracil) – тимин(урацил в PHK)
C	C	Cytosine - цитозин
R	A или G	puRine - пурин
Y	C или T	pYrimidine - пиридин
M	A или C	aMino - амино
K	G или T	Keto -кето
S	C или G	сильное /Strong/ взаимодействие – три водородные связи
W	A или T	слабое /Weak/ взаимодействие – две водородные связи
H	(A, C, T) но не G	H следует за G в алфавите
B	(C, G, T) но не A	B следует за A в алфавите
V	(A, C, G) но не T(U)	V следует за T(U) в алфавите
D	(A, G, T) но не C	D следует за C в алфавите
N	(A, G, C, T)	любое основание / Nucleotide

Небольшое упражнение:

- ¤ имеется фрагмент (первые 30 нуклеотидов с 5'-конца) последовательности сегмента NS вируса A/Puerto Rico/8/1934 из базы данных GenBank:

5'-agcaaaagcagggtgacaaagacataATGg-3'

- ¤ Необходимо определить полярность данной последовательности и установить соответствующую данной последовательность РНК, упакованную в вирион.

Правильный ответ:

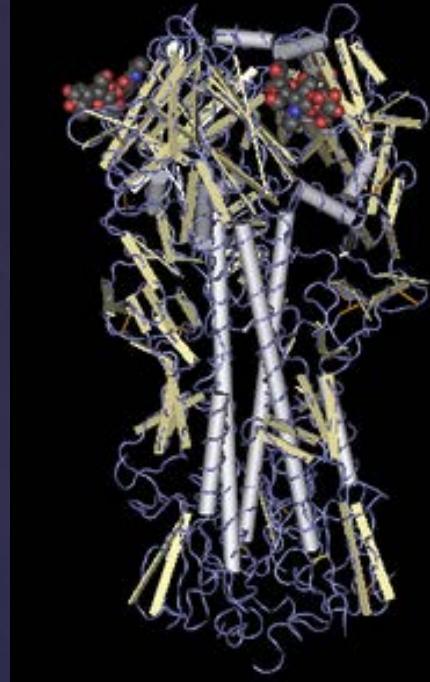
- ¶ В базе данных находится последовательность РНК, комплементарная вирусной и представленная в так называемом «ДНК-алфавите», то есть вместо урацила U в последовательности находится тимин T.
- ¶ Таким образом, реальная последовательность РНК, упакованная в вирион будет иметь вид:

5'-ссааиаугисиииугисасссиугсиииугси-3'

Аминокислотные последовательности

❖ Аминокислотные последовательности – это последовательности мономеров белков – биополимеров, выполняющих огромнейший набор функций в живом организме.

❖ Алфавит аминокислотных последовательностей содержит 20 «букв», соответствующих 20 аминокислотным остаткам.



Название	Трехбуквенный код	Однобуквенный код
Глицин	Gly	G
Аланин	Ala	A
Валин	Val	V
Изолейцин	Ile	I
Лейцин	Leu	L
Пролин	Pro	P
Серин	Ser	S
Треонин	Thr	T
Цистеин	Cys	C
Метионин	Met	M
Аспарагиновая кислота	Asp	D
Аспарагин	Asn	N
Глутаминовая кислота	Glu	E
Глутамин	Gln	Q
Лизин	Lys	K
Аргинин	Arg	R
Гистидин	His	H
Фенилаланин	Phe	F
Тирозин	Tyr	Y
Триптофан	Trp	W

Информация о первичной структуре белковых молекул хранится в последовательностях нуклеиновых кислот. Правило, определяющее направление передачи генетической информации между различными типами биомолекул в живом организме называют Центральной Догмой молекулярной биологии.

DNA



Replication

Reverse transcription

Transcription

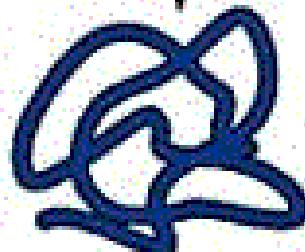
RNA



Replication

Translation

Protein



- ¶ При матричном синтезе молекул нуклеиновых кислот в процессе репликации и транскрипции передача информации происходит по правилу комплементарности.
- ¶ При матричном синтезе белковых молекул в процессе трансляции для передачи информации используется генетический код - система соответствий между последовательностью нуклеотидов в кодирующем фрагменте мРНК (ДНК) и последовательностью аминокислот в закодированном белке.

		Second Letter					
		T	C	A	G		
First Letter	T	TTT TTC TTA TTG } Phe	TCT TCC TCA TCG } Ser	TAT TAC TAA TAG } Tyr Stop Stop	TGT TGC TGA TGG } Cys Stop Trp	T C A G	
	C	CTT CTC CTA CTG } Leu	CCT CCC CCA CCG } Pro	CAT CAC CAA CAG } His Gln	CGT CGC CGA CGG } Arg	T C A G	
	A	ATT ATC ATA ATG } Ile Met	ACT ACC ACA ACG } Thr	AAT AAC AAA AAG } Asn Lys	AGT AGC AGA AGG } Ser Arg	T C A G	
	G	GTT GTC GTA GTG } Val	GCT GCC GCA GCG } Ala	GAT GAC GAA GAG } Asp Glu	GGT GGC GGA GGG } Gly	T C A G	
		Third Letter					

Свойства генетического кода:

- ¶ Триплетность – каждая аминокислота кодирована последовательностью из 3 нуклеотидов – триплетом (кодоном). Число возможных кодонов $4^3=64$, 61 кодон кодирует аминокислоты, 3 – сигнал окончания трансляции (стоп-кодоны).
- ¶ Вырожденность – аминокислоты и сигнал окончания транскрипции могут кодироваться более чем одним кодоном. Кодоны, кодирующие одну и ту же аминокислоту – синонимичные.
- ¶ Неперекрываемость – один и тот же нуклеотид не может входить одновременно в состав двух или более триплетов.
- ¶ Непрерывность – между соседними кодонами нет незначащих нуклеотидов.
- ¶ Однозначность (специфичность) – определённый кодон соответствует только одной аминокислоте.
- ¶ Универсальность – генетический код работает практически одинаково в организмах разного уровня сложности – от вирусов до человека.

Вновь небольшое упражнение:

¶ имеется следующая нуклеотидная последовательность:

AGCAAAAGCAGGGGAAAAATAAAA
ACAAACCAAAATGAAGGCCAACCTA
CTGGTCCTGTTAAGTGCACTGCA
GCTGCA

¶ Необходимо найти в ней кодирующую часть и получить соответствующую ей аминокислотную последовательность

Правильный ответ:

В данной последовательности есть одна открытая рамка считывания, которая кодирует следующий пептид:

МКАНLLVLLSALAAA

Получение биологических последовательностей

Генетические последовательности для биоинформационического анализа можно получить двумя способами:

- Секвенированием последовательностей имеющихся изолятов
- Поиском последовательностей в международных базах данных

Секвенирование генов вирусов гриппа

¶ Так как вирус гриппа является РНК-содержащим, то путь «от образца до последовательности» принимает следующий вид:

- Выделение РНК
- Обратная транскрипция
- ПЦР
- Очистка результатов ПЦР
- Секвенирование

- ¶ Не существует универсального протокола для полногеномного секвенирования любых вирусов гриппа.
- ¶ Основной сложностью является подбор праймеров для ПЦР и капиллярного секвенирования.
- ¶ Существуют различные подходы к решению этой проблемы, такие как:
 - ☒ поиск, комбинация и оптимизация праймеров и протоколов, доступных по литературным источникам
 - ☒ подбор собственных праймеров либо под конкретную последовательность, либо с учетом консервативности различных участков последовательности по результатам анализа множественных выравниваний.

**Universal primer set for the full-length amplification
of all influenza A viruses**

E. Hoffmann¹, J. Stech², Y. Guan³, R. G. Webster^{1,4}, and D. R. Perez¹

¹Department of Virology and Molecular Biology, St. Jude Children's Research Hospital,
Memphis, Tennessee, U.S.A.

²Institute for Virology, Marburg, Germany

³Department of Microbiology, The University of Hong Kong,
Queen Mary Hospital, Hong Kong, China

⁴Department of Pathology, University of Tennessee, Memphis, Tennessee, U.S.A.

Accepted August 29, 2001

JOURNAL OF CLINICAL MICROBIOLOGY, Sept. 2008, p. 3048–3055
0095-1137/08/\$08.00+0 doi:10.1128/JCM.02386-07
Copyright © 2008, American Society for Microbiology. All Rights Reserved.

Vol. 46, No. 9

**Subtyping of Avian Influenza Viruses H1 to H15 on the Basis of
Hemagglutinin Genes by PCR Assay and Molecular
Determination of Pathogenic Potential^V**

Kenji Tsukamoto,^{1,*} Hisayoshi Ashizawa,² Koji Nakanishi,³ Noriyuki Kaji,⁴ Kotaro Suzuki,¹
Masatoshi Okamatsu,¹ Shigeo Yamaguchi,¹ and Masaji Mase¹

*Research Team for Zoonotic Diseases, National Institute of Animal Health, 3-1-5 Kannondai, Tsukuba, Ibaraki 305-0856, Japan¹;
Tyouu Livestock Hygiene Service Center of Chiba Prefecture, Iwatomi, Sakura, Chiba 285-0072, Japan²;
Livestock Hygiene Service Center of Shiga Prefecture, Nishihongo, Omihachiman, Shiga 523-0813,
Japan³; and Livestock Hygiene Service Center of Shimane Prefecture, Kaminiishioki,
Izumo, Shimane 699-0822, Japan⁴*

Received 12 December 2007/Returned for modification 17 April 2008/Accepted 17 June 2008

Rescue of influenza B virus from eight plasmids

Erich Hoffmann*, Kutubuddin Mahmood*, Chin-Fen Yang*, Robert G. Webster[§], Harry B. Greenberg*,
and George Kemble*

*MedImmune Vaccines, 297 North Bernardo Avenue, Mountain View, CA 94043; [§]Division of Virology, Department of Infectious Diseases, St. Jude Children's Research Hospital, Memphis, TN 38105; and [§]Department of Pathology, University of Tennessee, Memphis, TN 38163

Contributed by Robert G. Webster, July 3, 2002

- ❖ Для обработки секвенированных последовательностей существует широкий набор платных и бесплатных программ, осуществляющих сборку единой последовательности из коротких фрагментов - ридов, полученных при секвенировании.
- ❖ Большинство таких программ реализуют алгоритм CAP3 [Huang and Madan, 1999], принцип действия которого заключается в поиске перекрытий между ридами с помощью алгоритмов локального выравнивания последовательностей.

Получение
биологических
последовательностей из
международных баз
данных

База данных GenBank

¶ Крупнейшая международная база данных GenBank содержит особый раздел посвященный вирусам гриппа: Influenza Virus Resourse, расположенный по адресу:

<http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database>

¶ Этот ресурс является наиболее полным открытым банком генетических последовательностей.

Форма поиска последовательностей Influenza Virus Resource

NCBI Resources How To

Influenza Virus Resource Information, Search and Analysis

Influenza Virus Database

Contact us Help

Flu home Database Genome Set Alignment Tree BLAST Annotation Submission FTP Virus resources ▾

Protein or nucleotide sequences can be retrieved from the database using GenBank accession numbers or search terms. Multiple queries can be built by clicking the "Add Query" button every time a new query is made, and queries in any combination from the Query Builder can be selected to get sequences in the database. Sequences can be downloaded, and it is possible to analyze them using the multiple sequence alignment or tree building tool integrated to the database.

Get sequences by accession

Enter a comma or space separated list of sequence accessions or upload text file with this list.

Upload Choose File No file chosen Accessions

Add query Show results

Select sequence type:

Protein Protein coding region Nucleotide

Search for keyword:

Keyword A/California/04/2009 Search in strain name

Define search set:

Type	Host	Country/Region	Segment	Subtype	Sequence length	Collection date	Release date
any	any	any	any	H any	Min.: <input type="text"/> Max.: <input type="text"/>	From: <input type="text"/> Year	<input type="text"/> Year
A	Avian	regions	1 (PB2)	1	<input type="checkbox"/> Full-length only	<input type="text"/> Month	<input type="text"/> Month
B	Bat	Northern temperate	2 (PB1)	2	<input type="checkbox"/> Full-length plus	<input type="text"/> Day	<input type="text"/> Day
C	Blow fly	Southern temperate	3 (PA/PA)	3			

Additional filters: show

Add query Show results Collapse identical sequences Clear form

Форма выдачи результатов запроса Influenza Virus Resource

NCBI Resources How To

Influenza Virus Resource Information, Search and Analysis

Influenza Virus Database

Contact us Help

Flu home Database Genome Set Alignment Tree BLAST Annotation Submission FTP Virus resources ▾

Add your own sequences Do multiple alignment Download Protein (FASTA) ▾ Customize FASTA defline ⓘ Permanent link

Show query

! Warning: Different virus species and/or segments selected.
Alignment and clustering cannot be performed.

Hold Ctrl or Shift key while clicking on column headers
to select/deselect multiple columns for sequential sorting.

Accession	Length	Host	Segment	Subtype	Country	Region	Date	Virus name	Mutations	Age	Gender	Lineage	Vac	Str	Complete
GQ280797	1701	Human	4 (HA)	H1N1	USA	N	2009/06/05	Influenza A virus (A/California/04/2009(H1N1))							c
FJ966079	2280	Human	1 (PB2)	H1N1	USA	N	2009/04/01	Influenza A virus (A/California/04/2009(H1N1))							c
FJ966080	2274	Human	2 (PB1)	H1N1	USA	N	2009/04/01	Influenza A virus (A/California/04/2009(H1N1))							c
FJ966081	2151	Human	3 (PA)	H1N1	USA	N	2009/04/01	Influenza A virus (A/California/04/2009(H1N1))							c
FJ966082	1701	Human	4 (HA)	H1N1	USA	N	2009/04/01	Influenza A virus (A/California/04/2009(H1N1))							c
FJ966083	1497	Human	5 (NP)	H1N1	USA	N	2009/04/01	Influenza A virus (A/California/04/2009(H1N1))							c
FJ966084	1410	Human	6 (NA)	H1N1	USA	N	2009/04/01	Influenza A virus (A/California/04/2009(H1N1))							c
FJ966085	972	Human	7 (MP)	H1N1	USA	N	2009/04/01	Influenza A virus (A/California/04/2009(H1N1))	S31N						p
FJ966086	838	Human	8 (NS)	H1N1	USA	N	2009/04/01	Influenza A virus (A/California/04/2009(H1N1))							c
FJ969512	1497	Human	5 (NP)	H1N1	USA	N	2009/04/01	Influenza A virus (A/California/04/2009(H1N1))							c
FJ969513	982	Human	7 (MP)	H1N1	USA	N	2009/04/01	Influenza A virus (A/California/04/2009(H1N1))	S31N						c
FJ969514	863	Human	8 (NS)	H1N1	USA	N	2009/04/01	Influenza A virus (A/California/04/2009(H1N1))							c
FJ969515	2151	Human	3 (PA)	H1N1	USA	N	2009/04/01	Influenza A virus (A/California/04/2009(H1N1))							c
FJ969516	2280	Human	1 (PB2)	H1N1	USA	N	2009/04/01	Influenza A virus (A/California/04/2009(H1N1))							c
FJ969517	1410	Human	6 (NA)	H1N1	USA	N	2009/04/01	Influenza A virus (A/California/04/2009(H1N1))							c
GQ117044	1701	Human	4 (HA)	H1N1	USA	N	2009/04/01	Influenza A virus (A/California/04/2009(H1N1))							c
GQ377049	2274	Human	2 (PB1)	H1N1	USA	N	2009/04/01	Influenza A virus (A/California/04/2009(H1N1))							c

Форма просмотра и загрузки последовательностей Influenza Virus Resource

Display Settings: GenBank

Influenza A virus (A/California/04/2009(H1N1)) segment 4 hemagglutinin gene, complete cds

GenBank: GQ280797.1
FASTA Graphics

Go to: □

LOCUS GQ280797 1701 bp cRNA linear VRL 15-JUN-2009

DEFINITION Influenza A virus (A/California/04/2009(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds.

ACCESSION GQ280797

VERSION GQ280797.1 GI:240130134

DBLINK BioProject: PRJNA37513

KEYWORDS .

SOURCE Influenza A virus (A/California/04/2009(H1N1))

ORGANISM Influenza A virus (A/California/04/2009(H1N1))
Viruses; ssRNA negative-strand viruses; Orthomyxoviridae;
Influenavirinae A.

REFERENCE 1 (bases 1 to 1701)

AUTHORS Ye, Z. and Iouwe, O.

TITLE Direct Submission

JOURNAL Submitted (17-JUN-2009) CBER, FDA, 5600 Rockville Pike, Bethesda, MD 20892, USA

CODING_SW Swine influenza A (H1N1) virus isolated during human swine flu outbreak of 2009.

FEATURES Location/Qualifiers

source 1..1701
/organism="Influenza A virus (A/California/04/2009(H1N1))"
/mol_type="viral cRNA"
/strain="A/California/04/2009"
/serotype="H1N1"
/host="Homo sapiens"
/db_xref="taxon:961501"
/segment="4"
/country="USA: California"
/collection_date="05-Jun-2009"
/note="lineage: xw1"

gene 1..1701
/gene="HA"
1..1701
/gene="HA"
/codon_start=1
/product="hemagglutinin"
/protein_id="AC545035.1"
/db_xref="GI:240130135"
/translation="MIAKIVVLLYTFATANADTLCLGVMANNSTDIVDVTILENNVTVI
HSTVLLDEDKMKNGHCKLRLGVAPPLHLGNCIAGHILQPECESLSTASSWNVIVETPSS
DNNTCCYPDQFIDQVEELRQLLSSTVSESTERPELTPTKTSWPNMDNSNGVIAACPAGAKS
FVQHLLVNUVKGNSMSPKLEKSTIMDNGKEVVLVLMGZKMPKTSADQJSSVYVQADITVTV
GSSRISKNTTPRILAPRNTRDQEGRGTTVWVNEPGDKITTEATGNLVVUPVRKATAMER
MAGSQQIIISSTPVWHDGTTICPTPGAAINTSLPQYQNPNTTIGCPRWTKSTLRLPAM
LRNRPISQSGHLYGAIAAGTGEGQHNTQVQDGTGTYQHNGQJQSGTAAADLKSQTQMADEI
TNNVWSVLEEDQGQFTAVWGNFHNLKSERLUNHNNVBDGFLDWVTHAEVLVLLNEER
TLOVHDSDWVONLYENVVA5QJUNRAKINGCCTFEYKNCNDTCMES5VNSGTYTDDVWYSE
EANLNRKEDIGVVKLESTRIZYQJILAIYSTVASSLLVVLVSLGAISFVNCSN2SLQCRICCI
"

misc_Features 55..1695
/gene="HA"
/note="Haemagglutinin; Region: Hemagglutinin; pCam0505"
/db_xref="CDD:249917"

ORIGIN

1 atggccgtt tactgtatcat acattttggaa ccgcacatgtc agacacatca
61 tptatgtttt atatcgaaa caatccaaa gacccatgtc acaccatgtt agaaaaatgt
121 gtccaaatgtt cacatctttt tttccatgtt gatggccatgg atatcgaaa
181 ctatggggat tagccccatgtt gatgtttttt acatgtttttt ttgttgtttt gatccatgg
241 acatccatgtt gttggccatgtt atccatgtt gttttttttt acatgtttttt gatccatgg
301 apttggatgtt atggccatgtt ttgtttttt gatccatgtt gttggccatgtt
361 ccattttgtt cttttttttt ttgttgtttt tttccatgtt gatggccatgtt
421 ccatccatgtt acatgtttttt gttggccatgtt ttgttgtttt gatccatgg
481 ttatccatgtt atttttttt ttgttgtttt tttccatgtt gatggccatgtt

Send: □

• Complete Record
• Coding Sequences
• Gene Features

Choose Destination

• File • Clipboard
• Collections • Analysis
Tool

Download 1 items.

Format: GenBank ▾

Create File

Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features
Find in this Sequence

Influenza Viral Resource
Flu-related NCBI resources in sequences, alignments, phylogeny, literature.

LinkOut to external resources:
Influenza Virus-IRD Flu Database [Influenza Resear...

Related Information
Related Sequences
BioProject
Full text in PMC
Protein
Taxonomy

Recent activity

Influenza A Virus (A/California/04/2009(H1N1))

База данных GISAID EPIFLU

¤ База данных GISAID EPIFLU, расположенная по адресу:

<http://platform.gisaid.org/epi3/frontend>

содержит большее число последовательностей вирусов гриппа, чем GenBank, однако для пользования ею необходима регистрация.

Форма поиска последовательностей GISAID EPIFLU

© 2008 - 2014 | The GISAID Initiative | Terms of Use | Contact | System Requirements  

You are logged in as Andrey Komissarov - [logout](#)

Welcome News Registered Users EpiFlu™ FAQ My profile About GISAID

Browse Back to results Workssets Upload Batch Upload Settings Analysis

Count 4 isolates GISAID published 34,859 isolates (90,045 sequences) Total isolate count 113,631 isolates (3)

Basic filters

Predefined search

Search in Released files My released files My unreleased files Workssets

Search patterns A/California/04/2009

Type	H	N	Lineage	Host	Location
A				-all- Human Animal Avian Chicken	-all- Africa Antarctica Asia Europe
B					
C					

Additional filters

Collection date (YYYY-MM-DD) From To

Submission date (YYYY-MM-DD) From To

Originating Laboratory [Albania, Tirana] Institute of Public Health
[Algeria, Algiers] Institut Pasteur d'Algérie
[Argentina, Buenos Aires] CEMIC University Hospital
[Argentina, Buenos Aires] Instituto Nacional de Enfermedades Infecciosas
[Argentina, CABA Pcia. de Buenos Aires] Servicio de Virosis Respiratorias INEI ANLIS Carlos ...

Submitting Laboratory [Argentina, Buenos Aires] Malbran
[Australia, Casuarina] Royal Darwin Hospital
[Australia, Geelong] CSIRO Australian Animal Health Laboratory
[Australia, North Melbourne] WHO Collaborating Centre for Reference and Research on Influe ...
[Austria, Vienna] Medical University Vienna

Required Segments PB2 PB1 PA HA NP NA MP NS HE P3
 full genome only complete Min Length
 only GISAID uploaded isolates only INSDC imported isolates

New features Help Reset Search

Форма выдачи результатов запроса GISAID EPIFLU

© 2008 - 2014 | The GISAID Initiative | Terms of Use | Contact | System Requirements  

You are logged in as **Andrey Komissarov** - [logout](#)

Welcome News Registered Users **EpiFlu™** FAQ My profile About GISAID

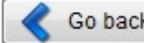
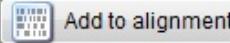
 [Browse](#)  [Back to results](#)  [Worksets](#)  [Upload](#)  [Batch Upload](#)  [Settings](#)  [Analysis](#)

Released files

	edit	Name	Isolate ID	Subtype	Passage	PB2	PB1	PA	HA	NP	NA	MP	NS	HE	P3
	edit	A/California/04/2009	EPI_ISL_98817	H1N1	P1 passage	2280	2274	2174	1701	1497	1410	982	838	---	---
	edit	A/California/04/2009	EPI_ISL_31103	H1N1		---	---	---	1701	---	---	---	---	---	---
	edit	A/California/04/2009	EPI_ISL_29618	H1N1	C2	2280	2274	2151	1701	1497	1410	982	863	---	---
	edit	A/California/04/2009	EPI_ISL_29573	H1N1	CX	2280	2274	2151	1701	1497	1410	972	838	---	---

Total: 4 isolates [<< first](#) [< prev](#) **1** [next >](#) [last >>](#)

Search in results

 [Go back](#)  [Help](#)  [Copy to...](#)  [Add to alignment](#)  [Download](#)

Форма загрузки последовательностей GISAID EPIFLU

Download

1 isolates selected.

Format

Isolates as XLS
 Sequences (DNA) as FASTA
 Sequences (proteins) as FASTA
 Acknowledgment table

DNA

all PB2 PB1 PA HA NP NA MP NS HE P3

FASTA Header

Isolate name _ Collection date

Isolate

Isolate name
Isolate ID
Type
Passage details/history
Lineage

Date format

Year fraction (2009.162)

Replace spaces with underscores in FASTA header

Example for copied segment

```
>A/DARWIN/36/2010_ | 2010.512
atgaaggcaataactagtagtccctgcttatatacattacaaccgcaaatgccgacacattatgtataggttatcatgcaaa
caattcaactgcacccgtagacacaataactagaaaagaatgtAACAGTAACACACTCTGTCAACCTTCTAGAAACCAGGC
ataatggaaaactatgtaaactaagagggtagctccattgcattggtaatgtAACATTGCTGGCTCCCTGGGA
```

Example for uploaded segment

```
>_ |
atgaaggcaataactagtagtccctgcttatatacattacaaccgcaaatgccgacacattatgtataggttatcatgcaaa
caattcaactgcacccgtagacacaataactagaaaagaatgtAACAGTAACACACTCTGTCAACCTTCTAGAAACCAGGC
```

[Go back](#) [Help](#) [Download](#)

Выравнивание генетических последовательностей

- ❖ Основным методом сравнительного анализа полученных генетических последовательностей является построение выравнивания.
- ❖ Задачу построения выравнивания генетических последовательностей можно условно разделить на две части: построение выравнивания двух последовательностей и множественное выравнивание.

Попарное выравнивание генетических последовательностей

¶ Небольшое лирическое отступление о массивах и матрицах

- ¶ Массив — это упорядоченный набор данных, идентифицируемых с помощью одного или нескольких индексов.
- ¶ Количество используемых индексов массива может быть различным. Массивы с одним индексом называют одномерными или векторами, с двумя — двумерными или матрицами.

$(1 \ 2)$

Вектор

$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$

матрица

Глобальное выравнивание: алгоритм Нидмана-Вунша

- ¶ Данный алгоритм был предложен в 1970 году Нидлманом и Вуншем [Needleman and Wunsch, 1970] для выравнивания аминокислотных последовательностей, а впоследствии использован и для выравнивания нуклеотидных последовательностей.
- ¶ Алгоритм основывается на принципе динамического программирования и состоит в поиске оптимального пути в матрице выравнивания, построенной с учетом коэффициентов пенализации за замены в последовательности, определяемых матрицой штрафов, а также коэффициентов пенализации за открытие и продолжение пробелов.
- ¶ Метод ориентирован на получение глобального выравнивания – то есть выравнивания двух последовательностей по всей их длине.

- ❖ Матрица штрафов вводит коэффициенты стоимости различных типов мутаций (например транзиций и трансверсий в ДНК).
- ❖ Даные коэффициенты могут быть заданы эмпирически, а также получены путем статистического анализа уже имеющихся данных.
- ❖ Для нуклеотидных выравниваний наиболее часто используется матрица IUB, в которой всем совпадениям присваивается коэффициент 1.9, а несовпадениям 0.
- ❖ На сегодняшний день наиболее используемыми для построения аминокислотных выравниваний являются семейства матриц замен PAM [Dayhoff, 1972] и BLOSUM [Henikoff and Henikoff, 1992].

Матрица аминокислотных замен BLOSUM62 – наиболее часто используемая по умолчанию для аминокислотных выравниваний

	C	S	T	P	A	G	N	D	E	O	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-3	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	5													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	3										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		F	
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

- ¶ Рассмотрим принцип работы алгоритма Нидлмана-Вунша на простом примере.
- ¶ Выровняем две нуклеотидные последовательности:
ACTGATTCA и **ACGCATCA**
- ¶ Примем коэффициент пенализации за любое несовпадение равным -3, за открытие и продолжение пробела равным -2, а за совпадение равным 2.

¶ Процесс работы алгоритма Нидмана-Вунша можно разделить на 3 этапа:

1. Начальное заполнение матрицы выравнивания.
2. Прямой проход метода – модификация матрицы выравнивания.
3. Обратный проход метода – поиск оптимального пути в модифицированной матрице.

Начальное заполнение матрицы

Расставим начальные штрафы за пробел

Расставим баллы за совпадения нуклеотидов

		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2				2				2
C	-4		2						2	
G	-6				2					
C	-8		2						2	
A	-10	2				2				2
T	-12			2			2	2		
C	-14		2						2	
A	-16	2				2				2

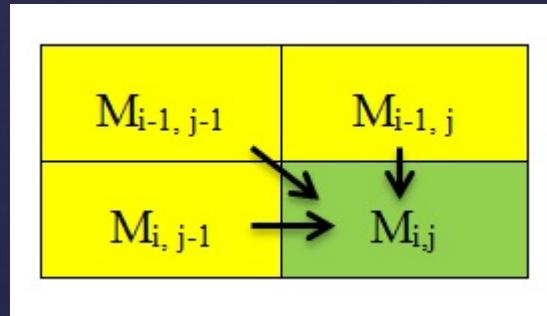
Расставим баллы за несовпадения нуклеотидов

		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	-3	-3	-3	2	-3	-3	-3	2
C	-4	-3	2	-3	-3	-3	-3	-3	2	-3
G	-6	-3	-3	-3	2	-3	-3	-3	-3	-3
C	-8	-3	2	-3	-3	-3	-3	-3	2	-3
A	-10	2	-3	-3	-3	2	-3	-3	-3	2
T	-12	-3	-3	2	-3	-3	2	2	-3	-3
C	-14	-3	2	-3	-3	-3	-3	-3	2	-3
A	-16	2	-3	-3	-3	2	-3	-3	-3	2

Прямой ход метода Нидлмана-Вунша

¤ Прямой ход метода Нидлмана-Вунша состоит в модификации элементов матрицы по следующему правилу:

$$M_{i,j} = \max \begin{pmatrix} M_{i-1,j-1} + S_{i,j} \\ M_{i-1,j} + gap \\ M_{i,j-1} + gap \end{pmatrix}$$



Выделенная область подвергнется модификации

		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	-3	-3	-3	2	-3	-3	-3	2
C	-4	-3	2	-3	-3	-3	-3	-3	2	-3
G	-6	-3	-3	-3	2	-3	-3	-3	-3	-3
C	-8	-3	2	-3	-3	-3	-3	-3	2	-3
A	-10	2	-3	-3	-3	2	-3	-3	-3	2
T	-12	-3	-3	2	-3	-3	2	2	-3	-3
C	-14	-3	2	-3	-3	-3	-3	-3	2	-3
A	-16	2	-3	-3	-3	2	-3	-3	-3	2

Прямой ход метода Нидлмана-Вунша: пример

		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	-3	-3	-3	2	-3	-3	-3	2
C	-4	-3	2	-3	-3	-3	-3	-3	2	-3
G	-6	-3	-3	-3	2	-3	-3	-3	-3	-3
C	-8	-3	2	-3	-3	-3	-3	-3	2	-3
A	-10	2	-3	-3	-3	2	-3	-3	-3	2
T	-12	-3	-3	2	-3	-3	2	2	-3	-3
C	-14	-3	2	-3	-3	-3	-3	-3	2	-3
A	-16	2	-3	-3	-3	2	-3	-3	-3	2

Прямой ход метода Нидлмана-Вунша: пример

		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	-3	-3	-3	2	-3	-3	-3	2
C	-4	-3	2	-3	-3	-3	-3	-3	2	-3
G	-6	-3	-3	-3	2	-3	-3	-3	-3	-3
C	-8	-3	2	-3	-3	-3	-3	-3	2	-3
A	-10	2	-3	-3	-3	2	-3	-3	-3	2
T	-12	-3	-3	2	-3	-3	2	2	-3	-3
C	-14	-3	2	-3	-3	-3	-3	-3	2	-3
A	-16	2	-3	-3	-3	2	-3	-3	-3	2

Прямой ход метода Нидлмана-Вунша: пример

		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	0	-3	-3	2	-3	-3	-3	2
C	-4	-3	2	-3	-3	-3	-3	-3	2	-3
G	-6	-3	-3	-3	2	-3	-3	-3	-3	-3
C	-8	-3	2	-3	-3	-3	-3	-3	2	-3
A	-10	2	-3	-3	-3	2	-3	-3	-3	2
T	-12	-3	-3	2	-3	-3	2	2	-3	-3
C	-14	-3	2	-3	-3	-3	-3	-3	2	-3
A	-16	2	-3	-3	-3	2	-3	-3	-3	2

Прямой ход метода Нидлмана-Вунша: пример

		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	0	-2	-4	-6	-8	-10	-12	-14
C	-4	-3	2	-3	-3	-3	-3	-3	2	-3
G	-6	-3	-3	-3	2	-3	-3	-3	-3	-3
C	-8	-3	2	-3	-3	-3	-3	-3	2	-3
A	-10	2	-3	-3	-3	2	-3	-3	-3	2
T	-12	-3	-3	2	-3	-3	2	2	-3	-3
C	-14	-3	2	-3	-3	-3	-3	-3	2	-3
A	-16	2	-3	-3	-3	2	-3	-3	-3	2

Прямой ход метода Нидлмана-Вунша: пример

		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	0	-2	-4	-6	-8	-10	-12	-14
C	-4	0	2	-3	-3	-3	-3	-3	2	-3
G	-6	-3	-3	-3	2	-3	-3	-3	-3	-3
C	-8	-3	2	-3	-3	-3	-3	-3	2	-3
A	-10	2	-3	-3	-3	2	-3	-3	-3	2
T	-12	-3	-3	2	-3	-3	2	2	-3	-3
C	-14	-3	2	-3	-3	-3	-3	-3	2	-3
A	-16	2	-3	-3	-3	2	-3	-3	-3	2

Прямой ход метода Нидлмана-Вунша: результат

		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	0	-2	-4	-6	-8	-10	-12	-14
C	-4	0	4	2	0	-2	-4	-6	-8	-10
G	-6	-2	2	1	4	2	0	-2	-4	-6
C	-8	-4	0	-1	2	1	-1	-3	0	-2
A	-10	-6	-2	-3	0	4	2	0	-2	2
T	-12	-8	-4	0	-2	2	6	4	2	0
C	-14	-10	-6	-2	-4	0	4	2	6	4
A	-16	-12	-8	-4	-5	-2	2	1	4	8

Обратный ход метода Нидлмана-Вунша

¶ Обратный ход метода Нидлмана-Вунша
состоит в поиске пути из нижнего правого угла
полученной матрицы в левый верхний с
наибольшей суммой баллов.

Обратный ход метода Нидлмана-Вунша

		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	0	-2	-4	-6	-8	-10	-12	-14
C	-4	0	4	2	0	-2	-4	-6	-8	-10
G	-6	-2	2	1	4	2	0	-2	-4	-6
C	-8	-4	0	-1	2	1	-1	-3	0	-2
A	-10	-6	-2	-3	0	4	2	0	-2	2
T	-12	-8	-4	0	-2	2	6	4	2	0
C	-14	-10	-6	-2	-4	0	4	2	6	4
A	-16	-12	-8	-4	-5	-2	2	1	4	8

Обратный ход метода Нидлмана-Вунша: результат

		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	0	-2	-4	-6	-8	-10	-12	-14
C	-4	0	4	2	0	-2	-4	-6	-8	-10
G	-6	-2	2	1	4	2	0	-2	-4	-6
C	-8	-4	0	-1	2	1	-1	-3	0	-2
A	-10	-6	-2	-3	0	4	2	0	-2	2
T	-12	-8	-4	0	-2	2	6	4	2	0
C	-14	-10	-6	-2	-4	0	4	2	6	4
A	-16	-12	-8	-4	-5	-2	2	1	4	8

Обратный ход метода Нидлмана-Вунша: результат

¶ Таким образом получаем выравнивание:

The diagram shows two sequences aligned vertically. The top sequence is "ACTG-ATTCA" and the bottom sequence is "AC-GCAT-CA". Vertical arrows connect the 'A' in ACTG to the 'A' in AC, the 'C' in ACTG to the 'G' in GCAT, the 'T' in ATTCA to the 'C' in CAT, and the 'A' in ATTCA to the 'A' in CA. The 'T' in ATTCA has no arrow pointing to it from the bottom sequence.

ACTG-ATTCA

AC-GCAT-CA

¶ Метод Нидлмана-Вунша направлен на выравнивание двух последовательностей по всей длине и позволяет найти лучшее для данной системы оценки глобальное выравнивание.

Вновь упражнение:

- ❖ Необходимо выровнять две аминокислотные последовательности:

NRLVLATGLRN и **NTLLATGMN**

- ❖ Примем коэффициент пенализации за любое несовпадение равным -3, за открытие и продолжение пробела равным -2, а за совпадение равным 2.

Решение задачи: инициализация матрицы

Решение задачи: прямой ход метода Нидлмана-Вунша

		N	R	L	V	L	A	T	G	L	R	N
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18	-20	-22
N	-2	2	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
T	-4	0	-1	-3	-5	-7	-9	-6	-8	-10	-12	-14
L	-6	-2	-3	1	-1	-3	-5	-7	-9	-6	-8	-10
L	-8	-4	-5	-1	-2	1	-1	-3	-5	-7	-9	-11
A	-10	-6	-7	-3	-4	-1	3	1	-1	-3	-5	-7
T	-12	-8	-9	-5	-6	-3	1	5	3	1	-1	-3
G	-14	-10	-11	-7	-8	-5	-1	3	7	5	3	1
M	-16	-12	-13	-9	-10	-7	-3	1	5	4	2	0
N	-18	-14	-15	-11	-12	-9	-5	-1	3	2	1	4

Решение задачи: обратный ход метода Нидлмана-Вунша

		N	R	L	V	L	A	T	G	L	R	N
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18	-20	-22
N	-2	2	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
T	-4	0	-1	-3	-5	-7	-9	-6	-8	-10	-12	-14
L	-6	-2	-3	1	-1	-3	-5	-7	-9	-6	-8	-10
L	-8	-4	-5	-1	-2	1	-1	-3	-5	-7	-9	-11
A	-10	-6	-7	-3	-4	-1	3	1	-1	-3	-5	-7
T	-12	-8	-9	-5	-6	-3	1	5	3	1	-1	-3
G	-14	-10	-11	-7	-8	-5	-1	3	7	5	3	1
M	-16	-12	-13	-9	-10	-7	-3	1	5	4	2	0
N	-18	-14	-15	-11	-12	-9	-5	-1	3	2	1	4

Правильный ответ:

¤ В результате получаем выравнивание:

NRLVLATGLRN

| * | ||| | | * |

NTL-LATG-MN

Локальное выравнивание: алгоритм Смита-Ватермана

- ❖ Данный алгоритм, развивающий принцип динамического программирования, заложенный в методе Нидлмана-Вунша, был предложен [Smith and Waterman, 1981] для получения локального выравнивания последовательностей, то есть для выявления сходных участков двух нуклеотидных или белковых последовательностей.
- ❖ Нововведения метода Смита-Ватермана заключаются в появлении возможности на стадии модификации присваивать 0 ячейкам с отрицательными значениями, а также возможности начинать поиск пути обратного хода не в крайней ячейке, а в ячейке, имеющей наибольшую оценку и заканчивать его по достижении ячейки с оценкой, равной 0.

¶ Рассмотрим принцип работы алгоритма Смита-Ватермана на простом примере.

¶ Выровняем две нуклеотидные последовательности:

ATGCATCC и ATCCGT

¶ Примем коэффициент penaлизации за любое несовпадение равным -3, за открытие и продолжение пробела равным -2, а за совпадение равным 2.

¤ Процесс работы алгоритма Смита-Ватермана также можно разделить на 3 Этапа:

1. Начальное заполнение матрицы выравнивания.
2. Прямой проход метода – модификация матрицы выравнивания.
3. Обратный проход метода – поиск оптимального пути в модифицированной матрице.

Начальное заполнение матрицы

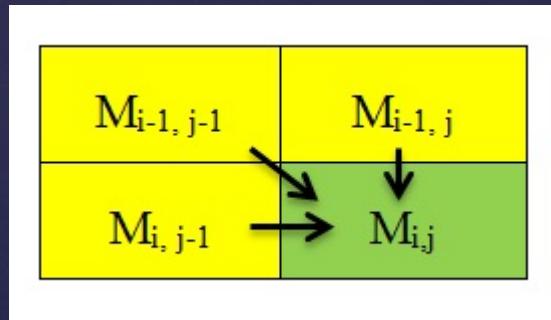
Начальное заполнение матрицы: результат

		A	T	G	C	A	T	C	C
	0	-2	-4	-6	-8	-10	-12	-14	-16
A	-2	2	-3	-3	-3	2	-3	-3	-3
T	-4	-3	2	-3	-3	-3	2	-3	-3
G	-6	-3	-3	-3	2	-3	-3	2	2
C	-8	-3	-3	-3	2	-3	-3	2	2
	-10	-3	-3	2	-3	-3	-3	-3	-3
T	-12	-3	2	-3	-3	-3	2	-3	-3

Прямой ход метода Смита-Ватермана

- ¶ Прямой ход метода Смита-Ватермана состоит в модификации элементов матрицы по следующему правилу:

$$M_{i,j} = \max \begin{pmatrix} M_{i-1,j-1} + S_{i,j} \\ M_{i-1,j} + gap \\ M_{i,j-1} + gap \\ 0 \end{pmatrix}$$



- ¶ В данном методе модификации подвергнутся все ячейки матрицы

Прямой ход метода Смита-Ватермана: пример

		A	T	G	C	A	T	C	C
	0	0	0	0	0	0	0	0	0
A	0	2	-3	-3	-3	2	-3	-3	-3
T	0	-3	2	-3	-3	-3	2	-3	-3
C	0	-3	-3	-3	2	-3	-3	2	2
C	0	-3	-3	-3	2	-3	-3	2	2
G	0	-3	-3	2	-3	-3	-3	-3	-3
T	0	-3	2	-3	-3	-3	2	-3	-3

Прямой ход метода Смита-Ватермана: пример

		A	T	G	C	A	T	C	C
	0	0	0	0	0	0	0	0	0
A	0	2	-3	-3	-3	2	-3	-3	-3
T	0	-3	2	-3	-3	-3	2	-3	-3
C	0	-3	-3	-3	2	-3	-3	2	2
C	0	-3	-3	-3	2	-3	-3	2	2
G	0	-3	-3	2	-3	-3	-3	-3	-3
T	0	-3	2	-3	-3	-3	2	-3	-3

Прямой ход метода Смита-Ватермана: пример

		A	T	G	C	A	T	C	C
	0	0	0	0	0	0	0	0	0
A	0	2	0	-3	-3	2	-3	-3	-3
T	0	-3	2	-3	-3	-3	2	-3	-3
C	0	-3	-3	-3	2	-3	-3	2	2
C	0	-3	-3	-3	2	-3	-3	2	2
G	0	-3	-3	2	-3	-3	-3	-3	-3
T	0	-3	2	-3	-3	-3	2	-3	-3

Прямой ход метода Смита-Ватермана: пример

		A	T	G	C	A	T	C	C
	0	0	0	0	0	0	0	0	0
A	0	2	0	0	0	2	0	0	0
T	0	0	2	-3	-3	-3	2	-3	-3
C	0	-3	-3	-3	2	-3	-3	2	2
C	0	-3	-3	-3	2	-3	-3	2	2
G	0	-3	-3	2	-3	-3	-3	-3	-3
T	0	-3	2	-3	-3	-3	2	-3	-3

Прямой ход метода Смита-Ватермана: пример

		A	T	G	C	A	T	C	C
	0	0	0	0	0	0	0	0	0
A	0	2	0	0	0	2	0	0	0
T	0	0	4	-3	-3	-3	2	-3	-3
C	0	-3	-3	-3	2	-3	-3	2	2
C	0	-3	-3	-3	2	-3	-3	2	2
G	0	-3	-3	2	-3	-3	-3	-3	-3
T	0	-3	2	-3	-3	-3	2	-3	-3

Прямой ход метода Смита-Ватермана: результат

		A	T	G	C	A	T	C	C
	0	0	0	0	0	0	0	0	0
A	0	2	0	0	0	2	0	0	0
T	0	0	4	2	0	0	4	2	0
C	0	0	2	1	4	2	2	6	4
C	0	0	0	0	3	1	0	4	8
G	0	0	0	2	1	0	0	2	6
T	0	0	2	0	0	0	2	0	4

Обратный ход метода Смита-Ватермана

¶ Обратный ход метода Смита-Ватермана
состоит в поиске пути с наибольшей суммой
баллов из ячейки с наивысшей оценкой.

Обратный ход метода Смита-Ватермана

		A	T	G	C	A	T	C	C
	0	0	0	0	0	0	0	0	0
A	0	2	0	0	0	2	0	0	0
T	0	0	4	2	0	0	4	2	0
C	0	0	2	1	4	2	2	6	4
C	0	0	0	0	3	1	0	4	8
G	0	0	0	2	1	0	0	2	6
T	0	0	2	0	0	0	2	0	4

Обратный ход метода Смита-Ватермана: результат

		A	T	G	C	A	T	C	C
	0	0	0	0	0	0	0	0	0
A	0	2	0	0	0	2	0	0	0
T	0	0	4	2	0	0	4	2	0
C	0	0	2	1	4	2	2	6	4
C	0	0	0	0	3	1	0	4	8
G	0	0	0	2	1	0	0	2	6
T	0	0	2	0	0	0	2	0	4

Обратный ход метода Смита-Ватермана: результат

¶ Таким образом получаем выравнивание:

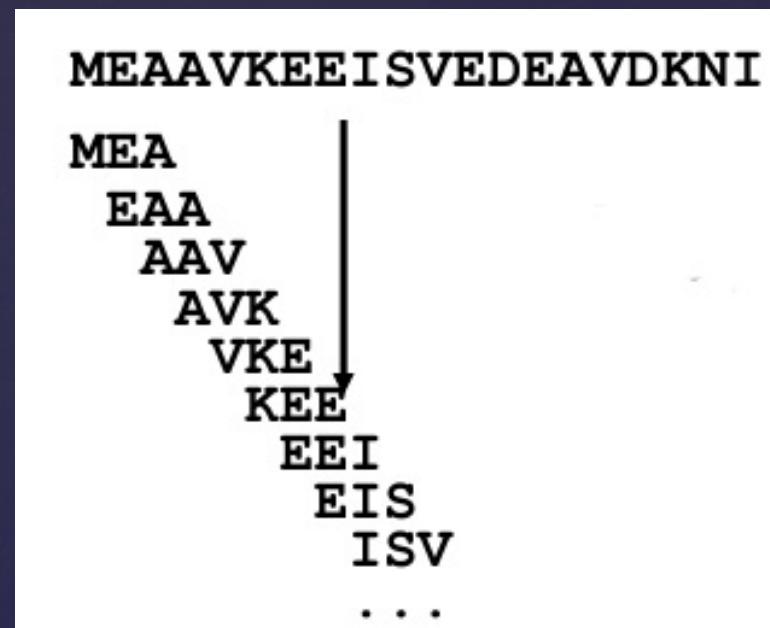
ATGCATCC--
| | | |
----ATCCGT

¶ Метод Смита-Ватермана направлен на выравнивание только гомологичных участков двух последовательностей и позволяет найти лучшее для данной системы оценки локальное выравнивание.

Методы слов: алгоритм BLAST

¶ Принцип работы методов слов заключается в получении из исследуемых последовательностей всех возможных подпоследовательностей определенной длины (слов) с последующим сравнением встречаемости каждого из полученных слов в обеих последовательностях.

¶ Таким способом можно легко устанавливать участки последовательности, которые могут быть выровнены, а затем применить более точный алгоритм только для этих участков.



- ¶ Описанный принцип лежит в основе методов семейства BLAST (Basic Local Alignment Search Tool), предложенных в 1990 году группой исследователей из NCBI [Altschul et al., 1990].
- ¶ Программы семейства BLAST предназначены для поиска локальных выравниваний интересующей последовательности в больших базах данных.
- ¶ В отличие от методов динамического программирования, дающих гарантированно лучшее для данной системы оценки локальное выравнивание, методы BLAST позволяют находить выравнивание только с определенной долей статистического приближения.
- ¶ В то же время метод BLAST требует в 50 раз меньше компьютерного времени чем метод Смита-Ватермана для выполнения одной и той же задачи

Семейство программ BLAST можно разделить на несколько групп:

- ¶ Нуклеотидные - предназначены для сравнения изучаемой нуклеотидной последовательности с базой данных секвенированных нуклеиновых кислот и их участков (*megablast* , *dc-megablast*, *blastn*).
- ¶ Белковые - предназначены для сравнения изучаемой аминокислотной последовательности с имеющейся базой данных белков и их участков (*blastp*, *cdart*, *rpsblast* , *psi-blast* , *phi-blast*).
- ¶ Транслирующие - способны транслировать нуклеотидные последовательности в аминокислотные (*blastx* , *tblastn* , *tblastx*).

Общий принцип работы методов методов BLAST состоит из следующих этапов:

1. Исследуемая последовательность разбивается на набор всех возможных подпоследовательностей определенной длины (по умолчанию 28 для нуклеотидных и 3 для аминокислотных последовательностей)
2. Производится поиск полных и близких (1-2 замены) совпадений по базе последовательностей, также разбитых на наборы слов
3. При нахождении совпадения программа пытается удлинить его в обе стороны до первого разрыва либо до набора определенного числа штрафов за несовпадение
4. На последнем этапе проводится выравнивание методом Смита-Ватермана исследуемой последовательности с теми последовательностями из базы, для которых на предыдущем шаге обнаружены статистически достоверные совпадения

Web-server BLAST: поиск гомологичных последовательностей по базе данных GenBank

¶ Программы семейства BLAST доступны через веб-интерфейс на сайте NCBI по адресу

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

¶ Данный набор программ является наиболее удобным средством поиска гомологичных последовательностей.

Главная страница портала NCBI BLAST

 **BLAST®** Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

► NCBI BLAST Home

BLAST finds regions of similarity between biological sequences. [more...](#)

New DELTA-BLAST, a more sensitive protein-protein search [Go](#)

BLAST Assembled RefSeq Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

<input type="checkbox"/> Human	<input type="checkbox"/> Dog	<input type="checkbox"/> Fruit fly	<input type="checkbox"/> Arabidopsis
<input type="checkbox"/> Mouse	<input type="checkbox"/> Rabbit	<input type="checkbox"/> Honey bee	<input type="checkbox"/> Rice
<input type="checkbox"/> Rat	<input type="checkbox"/> Chimp	<input type="checkbox"/> Chicken	<input type="checkbox"/> Yeast
<input type="checkbox"/> Cow	<input type="checkbox"/> Guinea pig	<input type="checkbox"/> Zebrafish	<input type="checkbox"/> Neurospora crassa
<input type="checkbox"/> Pig	<input type="checkbox"/> Sheep	<input type="checkbox"/> Clawed frog	<input type="checkbox"/> Microbes

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms:</i> blastn , megablast , discontiguous megablast
protein blast	Search protein database using a protein query <i>Algorithms:</i> blastp , psi-blast , phi-blast , delta-blast
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

Форма поискового запроса NCBI BLAST

BLAST® Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST/ blastn suite Standard Nucleotide BLAST

blastn **blastp** **blastx** **tblastn** **tblastx**

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#)
gggggggggg gggggggggg gggggggggg gggggggggg tttttttttt tttttttttt

[Clear](#) [Query subrange](#) [?](#)
From _____
To _____

Or, upload file Выберите файл | Файл не выбран [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

Choose Search Set

Database Human genomic + transcript Mouse genomic + transcript Others (nr etc.):
Nucleotide collection (nr/nt) [?](#)

Organism Optional
Enter organism name or id—completions will be sug Exclude [+](#) [?](#)
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude Optional
 Models (XM/XP) Uncultured/environmental sample sequences

Limit to Optional
 Sequences from type material

Entrez Query Optional
 [YouTube](#) [Create custom database](#)
Enter an Entrez query to limit search [?](#)

Program Selection

Optimize for Highly similar sequences (megablast)
 More dissimilar sequences (discontiguous megablast)
 Somewhat similar sequences (blastn)
Choose a BLAST algorithm [?](#)

BLAST Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)
 Show results in a new window

[Algorithm parameters](#)

Форма выдачи результатов поиска NCBI BLAST

Descriptions

Sequences producing significant alignments:

Select: All None Selected: 0

Alignments Download GenBank Graphics Distance tree of results

	Description	Max score	Total score	Query cover	E value	Ident	Accession
1	Influenza A virus (A/Puerto Rico/8-CIP045_RG89697/1934(H1N1)) segment 8 sequence	111	111	100%	4e-22	100%	CY077098.1
2	Influenza A virus (A/Puerto Rico/8-CIP045_RG83841/1934(H1N1)) segment 8 sequence	111	111	100%	4e-22	100%	CY077097.1
3	Influenza A virus (A/swine/Fujian/43/2007(H3N2)) segment 8 nuclear export protein (NEP) and nonstructural protein 1 (NS1) genes, complete cds	111	111	100%	4e-22	100%	GU088101.1
4	Influenza A virus (A/reassortant/NYMC X-179A(California/07/2009 x NYMC X-157(H1N1)) segment 8 sequence	111	111	100%	4e-22	100%	CY058523.1
5	Influenza A virus (A/reassortant/NYMC X-175C(Uruguay/716/2007 x Puerto Rico/8/1934)(H3N2)) segment 8 sequence	111	111	100%	4e-22	100%	CY058507.1
6	Influenza A virus (A/lvPR8/34(H1N1)) segment 8, complete sequence	111	111	100%	4e-22	100%	EF190986.1
7	Influenza A virus (A/hvPR8/34(H1N1)) segment 8, complete sequence	111	111	100%	4e-22	100%	EF190978.1
8	Influenza A virus (A/Puerto Rico/8/34/Mount Sinai(H1N1)) segment 8, complete sequence	111	111	100%	4e-22	100%	AF389122.1
9	Influenza A virus (X-31(H3N2)) RNA segment 8 for NS1, NS2, complete sequence	111	111	100%	4e-22	100%	AB036777.1
10	Influenza A virus (A/WS/1933(H1N1)) non-structural protein 2 (NS2) and non-structural protein 1 (NS1) genes, complete cds	111	111	100%	4e-22	100%	U13883.1

Alignments

Download GenBank Graphics Next Previous Descriptions

Influenza A virus (A/Puerto Rico/8-CIP045_RG89697/1934(H1N1)) segment 8 sequence
Sequence ID: qb|CY077098.1 Length: 890 Number of Matches: 1

Range 1: 1 to 60 GenBank Graphics Next Match Previous Match

Score	Expect	Identites	Gaps	Strand
111 bits(60)	4e-22	60/60(100%)	0/60(0%)	Plus/Plus

Query 1 AGCAAAAGCAGGGTGACAAAGACATAATGGATCCAAAACACTGTGTCAGCTTTCAGGTAG 60
Sbjct 1 AGCAAAAGCAGGGTGACAAAGACATAATGGATCCAAAACACTGTGTCAGCTTTCAGGTAG 60

Related Information

Множественное
выравнивание
биологических
последовательностей

❖ Для выравнивания более чем двух последовательностей классические алгоритмы попарного выравнивания оказываются сложно применимыми, поскольку количество необходимых математических операций для процедуры динамического программирования определяется как m^n , где m – суммарная длина последовательностей, а n – их число.

❖ Для решения этой задачи был предложен метод прогрессивного множественного выравнивания.

Принцип метода прогрессивного выравнивания

Процесс прогрессивного множественного выравнивания включает в себя 3 этапа.

1. Все последовательности попарно выравниваются между собой с использованием какого-либо алгоритма выравнивания, и среди них выявляются группы схожих между собой последовательностей.
2. Выравнивание последовательностей в каждой такой группе.
3. Выравнивание групп между собой.

Алгоритм CLUSTAL

‐ CLUSTAL, предложенный Хиггинсом и Шарпом в 1988 году [Higgins and Sharp, 1988] стал первым алгоритмом множественного прогрессивного выравнивания.

‐ Несмотря на появление более совершенных программ для прогрессивного выравнивания CLUSTAL по сей день одним из наиболее используемых алгоритмов.

❖ Процедура работы CLUSTAL включает в себя 3 этапа:

1. На первом этапе производится попарное выравнивание всех последовательностей. Для каждой пары последовательностей определяется генетическая дистанция.
2. На втором этапе происходит построение руководящего дерева путем кластеризации наиболее схожих последовательностей.
3. На третьем этапе происходит построение выравнивания по руководящему дереву путем выравнивания соответствующих последовательностей и групп последовательностей. Окончательное выравнивание получается при достижении корня руководящего дерева

- ¶ Рассмотрим принцип работы CLUSTAL на простом примере.
- ¶ Выровняем 4 коротких нуклеотидных последовательности:

AC, ATG, TCG и TCC

- ❖ На первом этапе необходимо выровнять все последовательности попарно друг на друга.
- ❖ По результатам попарных выравниваний видно, что последовательности TCG и TCC имеют наибольший непрерывный общий участок.

A-C	A-C	A-C
ATG	-TCG	-TCC
AT-G	ATG	
-TCG	-T-CC	
TC-G		
TCC-		

- ❖ На втором этапе необходимо сгруппировать попарно наиболее близкие последовательности.
- ❖ Соответственно, TCG и TCC составят первую пару, а AC и ATG – вторую.

?

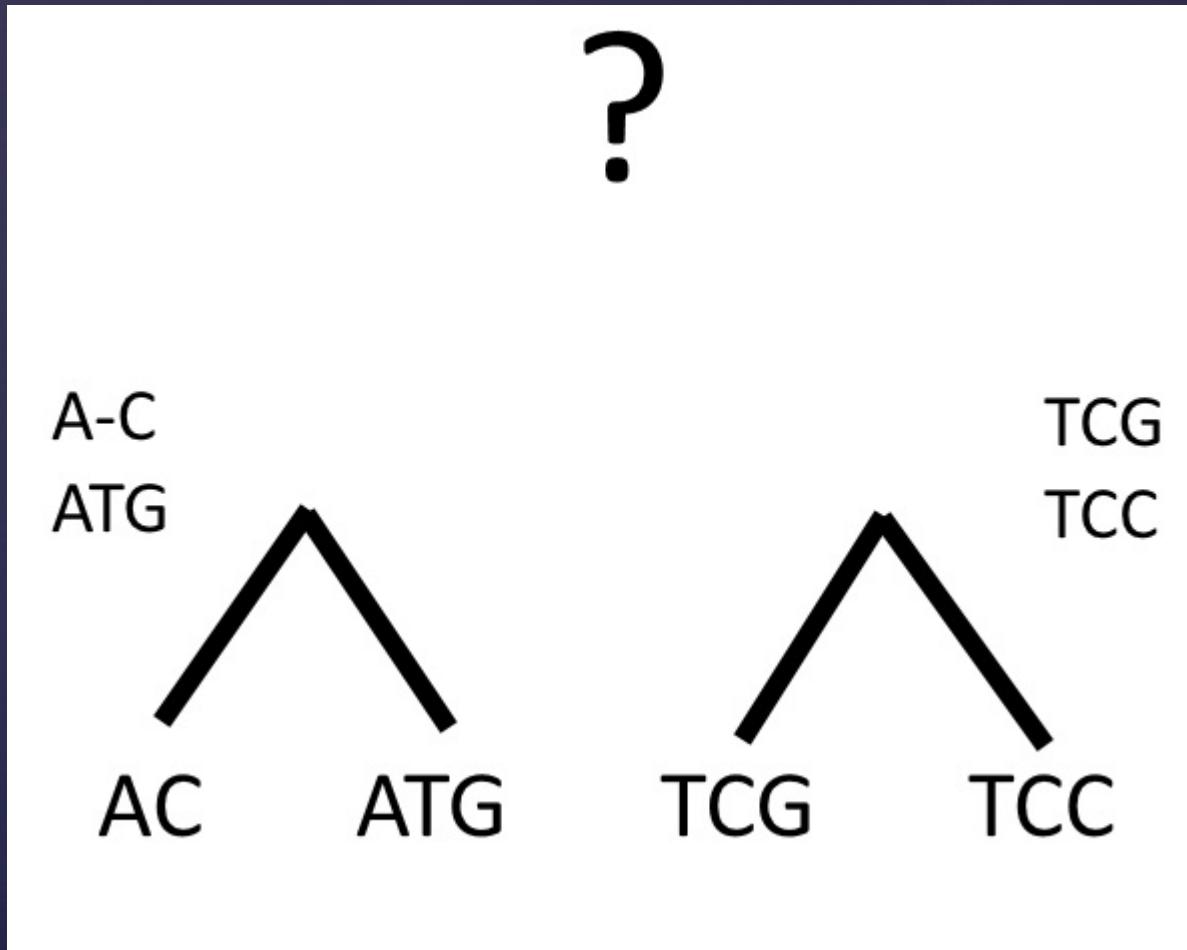
AC

ATG

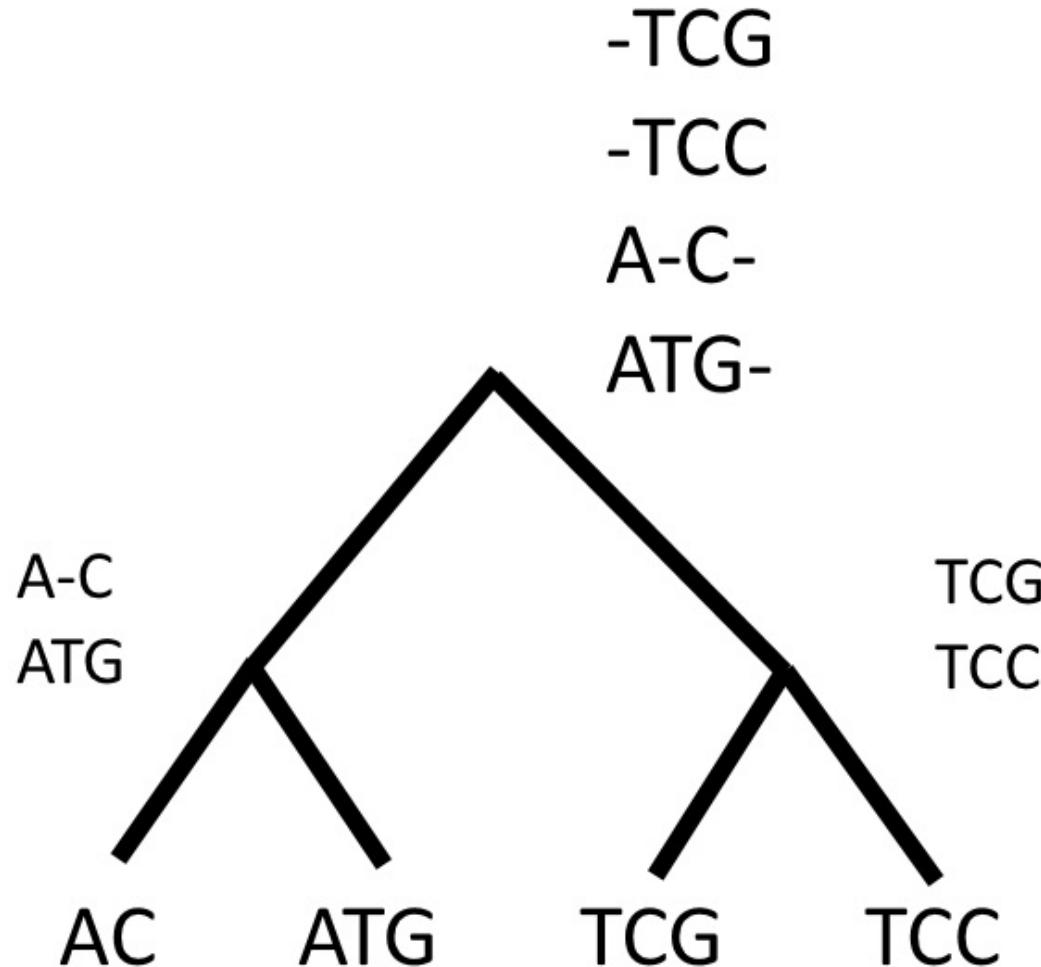
TCG

TCC

¶ По результатам группировки начинаем строить руководящее дерево и выравнивать последовательности в группах между собой.



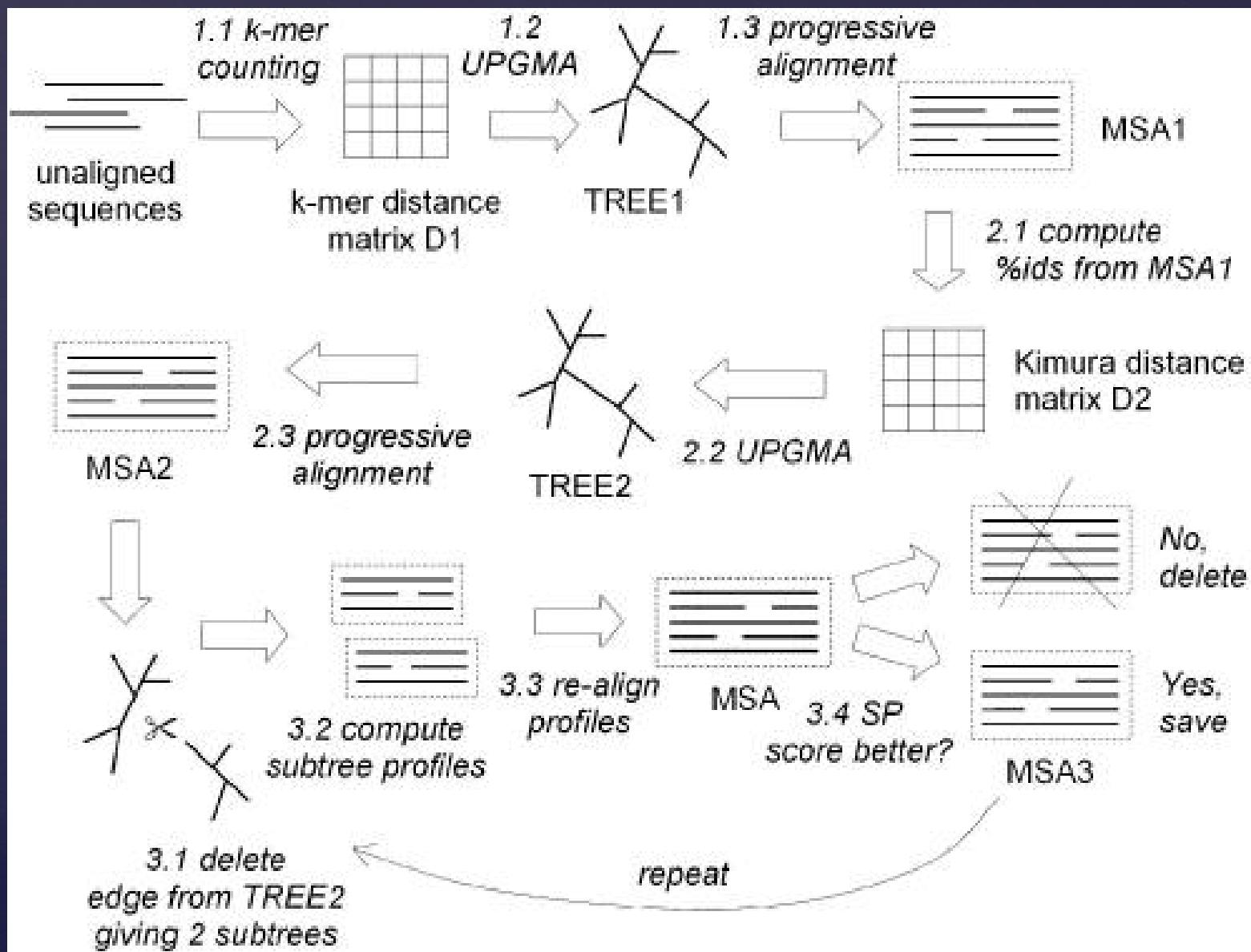
¶ Построение руководящего дерева заканчивается с выравниванием групп относительно друг друга и получением итогового результата.



Алгоритм MUSCLE

- ❖ Алгоритм MUSCLE [Edgar et al., 2004] является развитием идеи, положенной в основу программы CLUSTAL, реализуя принцип прогрессивно-итеративного множественного выравнивания.
- ❖ Принципиальным новшеством этого алгоритма является комбинация методов динамического программирования и методов слов, а также минимум двукратное перестроение руководящего дерева в процессе выравнивания, что позволяет значительно снизить вероятность возникновения случайных ошибок в выравнивании.

Схема работы алгоритма MUSCLE (по Edgar, 2004)



Анализ множественных выравниваний генетических последовательностей

Поиск консервативных
участков нуклеотидных
последовательностей и
подбор ПЦР-праймеров

¶ Построение нуклеотидных выравниваний позволяет выявлять консервативные участки последовательностей, пригодные для подбора относительно универсальных ПЦР-праймеров.



- ¶ Основные параметры праймеров:
 - ¤ Длина 18-24 нуклеотида.
 - ¤ Четыре и более 3'-концевых нуклеотида не должны быть комплементарны самому праймеру, праймеру в паре или иным добавленным в реакцию олигонуклеотидам.
 - ¤ Температура отжига должна лежать в диапазоне 60 – 70 °C.
 - ¤ Температура плавления праймеров, работающих в паре должна быть максимально близкой.
 - ¤ Желательно, чтобы температура плавления 5'-концевой части была выше чем у 3'-концевой части.
 - ¤ С 5'-конца праймера может быть добавлена не комплементарная матрице последовательность практически любой длины.
- ¶ Также при подборе праймеров для секвенирования необходимо помнить, что длина прочтения капиллярного секвенатора обычно составляет 600-700 нуклеотидов.

¶ Также полезным инструментом в подборе и оценке специфичности праймеров является веб-сервер NCBI Primer-BLAST, расположенный по адресу:

<http://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi>

Веб-интерфейс NCBI Primer-BLAST

Primer-BLAST A tool for finding specific primers

NCBI Primer-BLAST: Finding primers specific to your PCR template (using Primer3 and BLAST). More... Tips for finding specific primers

Reset page Save search parameters Retrieve recent results

PCR Template

Enter accession, gI, or FASTA sequence (A refseq record is preferred)

Range

Forward primer From To Clear
Reverse primer

Or, upload FASTA file Выберите файл Файл не выбран

Primer Parameters

Use my own forward primer (5'->3' on plus strand) Forward primer
Use my own reverse primer (5'->3' on minus strand) Reverse primer

PCR product size Min Max
500 1800

of primers to return
10

Primer melting temperature (Tm) Min Opt Max Max Tm difference
57.0 60.0 63.0 3

Exon/Intron selection

A refseq mRNA sequence as PCR template input is required for options in this section

Exon Junction span No preference
Exon Junction match
Intron Inclusion
Intron length range

Note: Parameter values that differ from the default are highlighted in yellow

Primer Pair Specificity Checking Parameters

Specificity check
Beezorth mode Automatic
Database nr
Organism Homo sapiens
Canis lupus (taxid 9612)
Enter an organism name, taxonomy id or select from the suggestion list as you type.
Add more organisms

Exclusion (optional)
Entrez query (optional)
Primer specificity stringency
Misprimed product size deviation
Splice variant handling

Get Primers Show results in a new window Use new graphic view

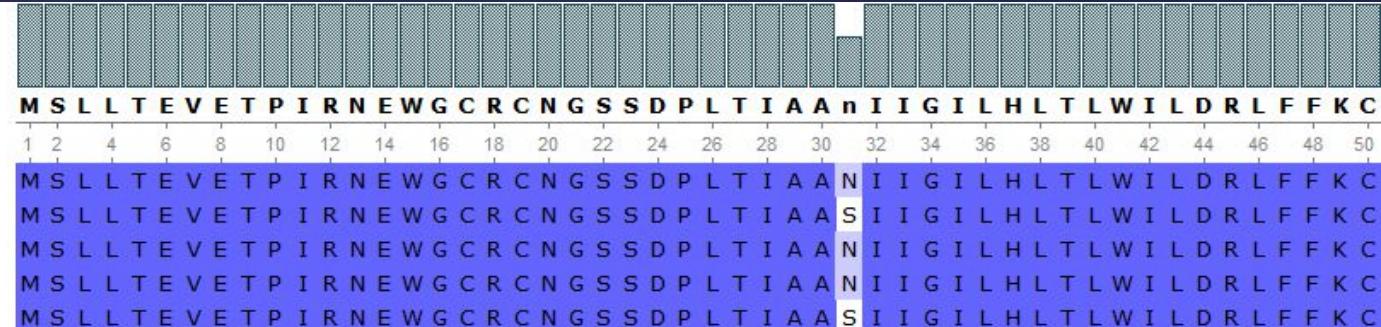
Advanced parameters Note: Parameter values that differ from the default are highlighted in yellow

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback on new interface

Результаты поиска NCBI Primer-BLAST

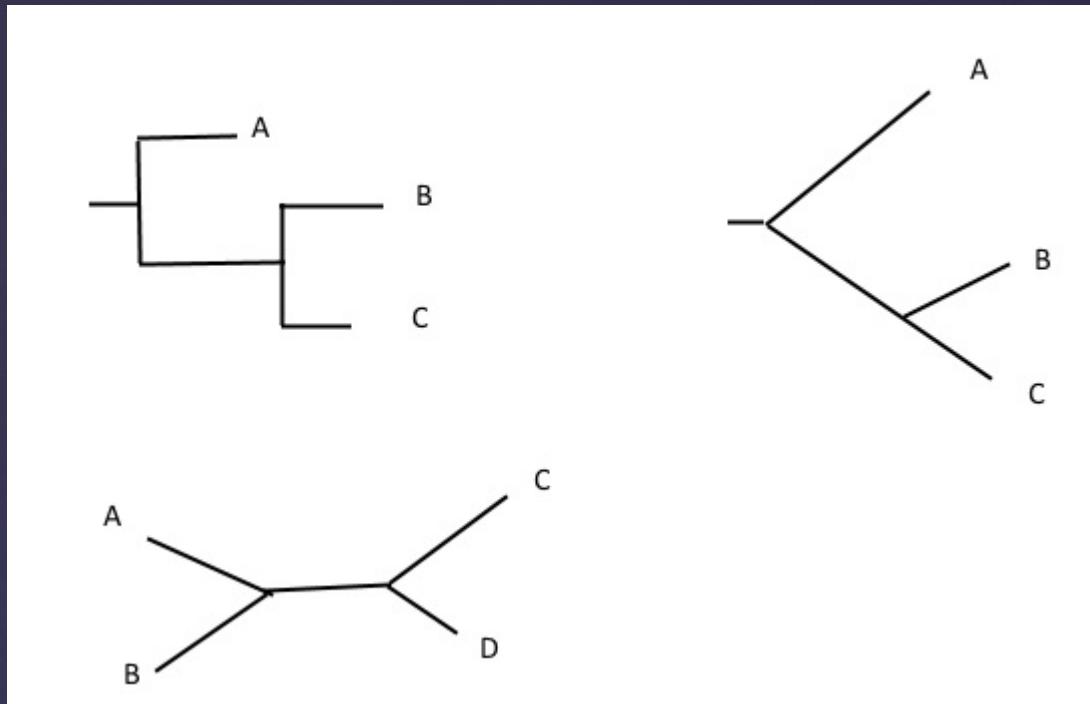
Анализ аминокислотных
последовательностей,
поиск мутаций
устойчивости и замен в
антигенных сайтах

¶ Построение аминокислотных выравниваний позволяет анализировать наличие замен в антигенных сайтах первой субъединицы гемагглютинина, активном сайте нейраминидазы, а также позволяет отслеживать появление мутаций устойчивости к основным противовирусным препаратам – адамантанам (замена S31N в белке M2) и ингибиторам нейраминидазы (замена H275Y в последовательности нейраминидазы N1).



Основные понятия филогенетики

- ❖ Филогенетика — область биологической систематики, которая занимается идентификацией и прояснением эволюционных взаимоотношений среди разных видов жизни на Земле, как современных, так и вымерших.
- ❖ Молекулярная филогенетика решает задачи поиска эволюционных связей путем анализа генетических последовательностей.
- ❖ Графическим представлением эволюционных отношений служит филогенетическое дерево.



- ❖ Филогенетическое дерево состоит из внутренних и внешних ветвей, узлов и, если при построении выбрана соответствующая опция, - корня
- ❖ Принято разделять узлы в дереве на внешние и внутренние.
- ❖ Внешние узлы также иногда называют листьями дерева.

Генетические дистанции и эволюционные модели

- ❖ Для количественной оценки различий между генетическими последовательностями используется понятие генетических или эволюционных дистанций.
- ❖ В случае, если имеется более двух последовательностей, для записи дистанций принято использовать матрицы дистанций.

p-дистанция

¶ Простейшим способом выражения генетических дистанций является наблюдаемая дистанция или р-дистанция – отношение числа различающихся нуклеотидов или аминокислот у двух выровненных последовательностей к их длине, выраженное волях от единицы либо в процентах

$$p = \frac{m}{n}$$

где m – число несовпадений, а n – длина последовательностей

¶ В общем случае, наблюдаемая р-дистанция не равна истинной эволюционной дистанции.

¶ Поэтому для определения эволюционной дистанции используются различные эволюционные модели

Модель Джукса-Кантора

- ¶ Модель Джукса-Кантора JC69 [Jukes and Cantor, 1969] стала первой предложенной эволюционной моделью.
- ¶ Данная модель исходит из того, что частоты встречаемости всех нуклеотидов в последовательности равны 0.25, а вероятность любой нуклеотидной замены равна μ .
- ¶ Таким образом, данная модель является однопараметрической.
- ¶ Согласно модели Джукса-Кантора эволюционная дистанция принимает вид:

$$D_{JC69} = -\frac{3}{4} \ln \left(1 - \frac{4}{3} p \right)$$

где p – наблюдаемая p -дистанция

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

Модель Кимуры

- Модель Кимуры K80 [Kimura, 1980], также как и модель Джукса-Кантора исходит из предположения, о том, что частоты встречаемости всех нуклеотидов в последовательности равны 0.25, но учитывает различную частоту для транзиций и трансверсий.
- Таким образом, данная модель является двухпараметрической.
- Согласно модели Кимуры эволюционная дистанция примет вид:

$$D_{K80} = -\frac{1}{2} \ln(1 - 2P - Q) - \frac{1}{4} \ln(1 - 2Q)$$

где P и Q – наблюдаемые пропорции транзиций и трансверсий , то есть отношение наблюдаемых транзиций и трансверсий к числу наблюдаемых замен.

$$Q = \begin{pmatrix} * & \kappa & 1 & 1 \\ \kappa & * & 1 & 1 \\ 1 & 1 & * & \kappa \\ 1 & 1 & \kappa & * \end{pmatrix}$$

Модель Хасегавы- Кишино-Яно

¶ Модель Хасегавы-Кишино-Яно HKY85 [Hasegawa et al., 1985) расширяет модель Кимуры, допуская различные частоты встречаемости нуклеотидов в последовательности при сохранении различной вероятности транзиций и трансверсий.

$$Q = \begin{pmatrix} * & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & * \end{pmatrix}$$

Модель GTR

¶ Модель GTR [Tavare, 1986] является наиболее сложной эволюционной моделью, допуская индивидуальные частоты встречаемости и вероятности замен для каждого нуклеотида.

$$Q = \begin{pmatrix} -(x_1 + x_2 + x_3) & \frac{\pi_1 x_1}{\pi_2} & \frac{\pi_1 x_2}{\pi_3} & \frac{\pi_1 x_3}{\pi_4} \\ x_1 & -\left(\frac{\pi_1 x_1}{\pi_2} + x_4 + x_5\right) & \frac{\pi_2 x_4}{\pi_3} & \frac{\pi_2 x_5}{\pi_4} \\ x_2 & x_4 & -\left(\frac{\pi_1 x_2}{\pi_3} + \frac{\pi_2 x_4}{\pi_3} + x_6\right) & \frac{\pi_3 x_6}{\pi_4} \\ x_3 & x_5 & x_6 & -\left(\frac{\pi_1 x_3}{\pi_4} + \frac{\pi_2 x_5}{\pi_4} + \frac{\pi_3 x_6}{\pi_4}\right) \end{pmatrix}$$

Расширения эволюционных моделей

- ❖ Различные позиции в последовательности в силу каких-либо причин могут эволюционировать с различной скоростью.
- ❖ Для учета неоднородности в скорости эволюции применяются два подхода:
 - ❖ принцип инвариантных сайтов: в случае значительных различий в скорости эволюции сайты делятся на 2 категории – сайтам с низкой скоростью эволюции присваивается нулевая, скорость остальных сайтов определяется эволюционной моделью.
 - ❖ аппроксимация неоднородностей с помощью гамма-распределения – используется в случае незначительных различий в скорости эволюции.
- ❖ Также возможна комбинация этих подходов.

Выбор оптимальной эволюционной модели

- ¶ Выбор эволюционной модели для расчета эволюционных дистанций проводится с помощью статистических критериев.
- ¶ Для этого вводится функция правдоподобия
$$L(Tree, Model) = P(Data|Tree, Model)$$
- ¶ Далее применяется один из двух информационных критериев:
 - ¤ информационный критерий Акаике
$$AIC = -2 \ln L + 2N,$$
где N – число свободных параметров в модели
 - ¤ байесовский информационный критерий
$$BIC = -2 \ln L + 2N \ln n,$$
где n – длина последовательности
- ¶ Наиболее подходящая модель будет обладать наименьшим значением критерия.

Небольшое упражнение

‐ Имеется 5 последовательностей:

AGCCAAAAGCAGGGGAAAA
AGCAATAGCAGGGGAAAA
AGCAAAAGCAGGGGTAA
AGCAAAAAGCAGGGGAAAA
AGCAATAGCAGGGGTAA

‐ Необходимо построить матрицу наблюдаемых дистанций между ними и рассчитать эволюционные дистанции по модели JC69

Правильный ответ

¶ Пронумеруем последовательности и подсчитаем количество замен для каждой пары

	1	2	3	4	5
1	0	2	2	1	3
2	2	0	2	1	1
3	2	2	0	1	1
4	1	1	1	0	2
5	3	1	1	2	0

¶ Разделим полученные значения на длину последовательностей и получим матрицу р-дистанций, после чего по формуле Джукса-Кантора рассчитаем эволюционные дистанции

	1	2	3	4	5
1	0	0.117	0.117	0.058	0.176
2	0.117	0	0.117	0.058	0.058
3	0.117	0.117	0	0.058	0.058
4	0.058	0.058	0.058	0	0.117
5	0.176	0.058	0.058	0.117	0

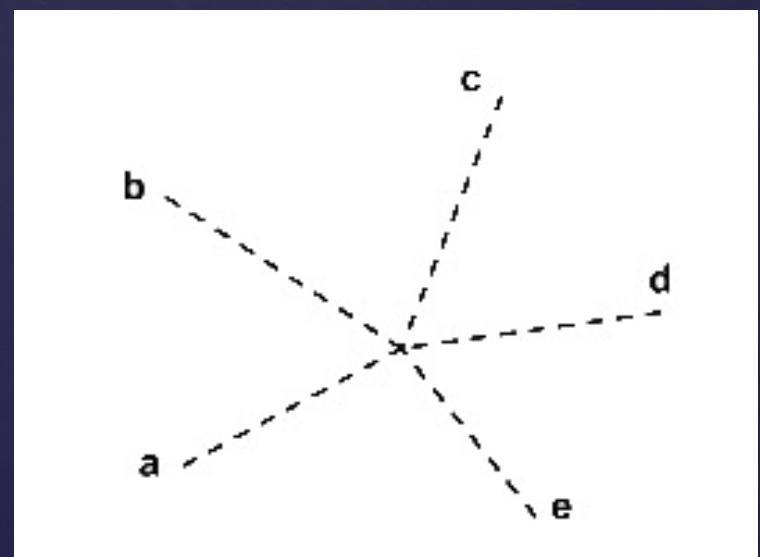
	1	2	3	4	5
1	0	0.128	0.128	0.061	0.201
2	0.128	0	0.128	0.061	0.061
3	0.128	0.128	0	0.061	0.061
4	0.061	0.061	0.061	0	0.128
5	0.201	0.061	0.061	0.128	0

Дистанционные методы
построения
филогенетических
деревьев: метод
присоединения
ближайших соседей
(Neighbor Joining)

- ¶ Метод присоединения ближайших соседей является одним из наиболее простых филогенетических методов.
- ¶ При построении дерева данным методом используются результаты расчета эволюционных дистанций.
- ¶ Метод рассматривает последовательности целиком, без разделения на отдельные позиции.
- ¶ Принцип метода состоит в группировке последовательностей таким образом, чтобы минимизировать суммарную длину ветвей в дереве.

- ¶ Рассмотрим принцип работы метода на небольшом примере.
- ¶ Представим, что у нас есть 5 последовательностей, эволюционные дистанции между которыми описываются следующей матрицей.

	a	b	c	d	e
a	0	5	9	9	8
b	5	0	10	10	9
c	9	10	0	8	7
d	9	10	8	0	3
e	8	9	7	3	0



¶ В начале расчитаем параметр Q, который пропорционален суммарной длине ветвей возможного дерева.

$$Q(i, j) = (n - 2)d(i, j) - \sum_{k=1}^n d(i, k) - \sum_{k=1}^n d(j, k)$$

¶ Получим следующую матрицу результатов:

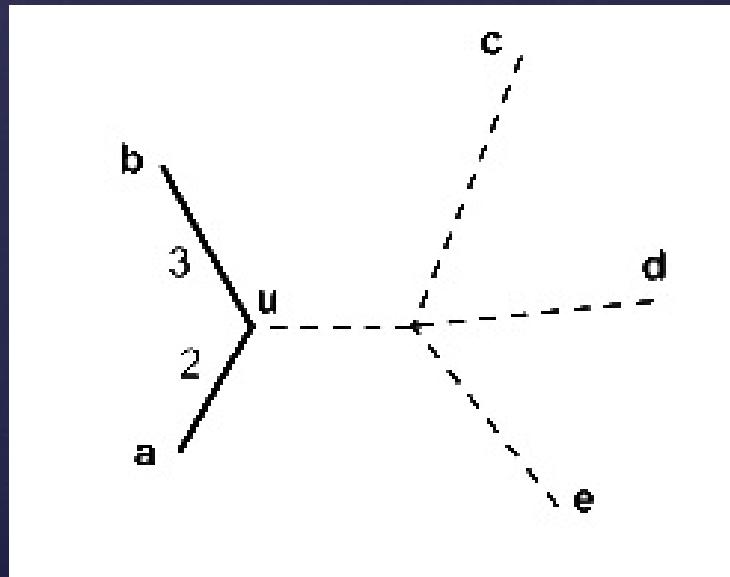
	a	b	c	d	e
a		-50	-38	-34	-34
b	-50		-38	-34	-34
c	-38	-38		-40	-40
d	-34	-34	-40		-48
e	-34	-34	-40	-48	

- ❖ Наименьшее значение Q дает объединение последовательностей a и b .
- ❖ Поэтому мы можем сгруппировать их в узел u .
- ❖ Рассчитаем длины полученных ветвей

$$\delta(a, u) = \frac{1}{2} d(a, b) + \frac{1}{2(n-2)} \left[\sum_{k=1}^n d(a, k) - \sum_{k=1}^n d(b, k) \right]$$

$$\delta(b, u) = d(a, b) - \delta(a, u)$$

- ❖ Наше дерево примет вид:



¶ Теперь мы можем пересчитать матрицу дистанций между полученным узлом и оставшимися последовательностями.

$$d(u, k) = \frac{1}{2} [d(a, k) + d(b, k) - d(a, b)]$$

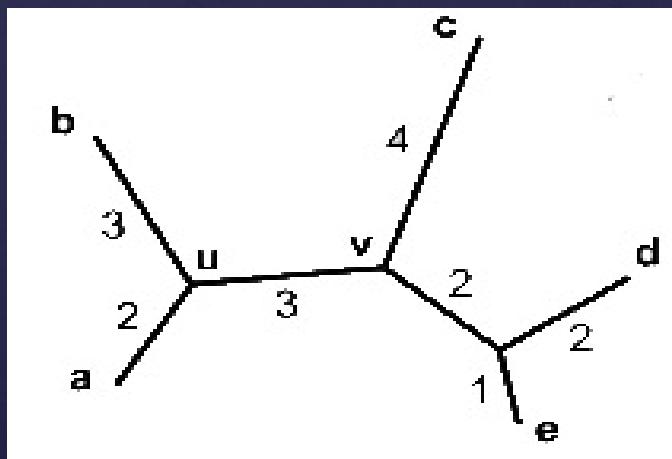
¶ Получим матрицу:

	u	c	d	e
u	0	7	7	6
c	7	0	8	7
d	7	8	0	3
e	6	7	3	0

¶ Для полученной матрицы пересчитаем Q:

	u	c	d	e
u		-28	-24	-24
c	-28		-24	-24
d	-24	-24		-28
e	-24	-24	-28	

- ¶ Здесь возможно 2 варианта группировки – u – c и d – e.
- ¶ Сгруппируем u и c, в результате чего образуется новый узел v. Дерево построено, остается только пересчитать длины оставшихся ветвей.



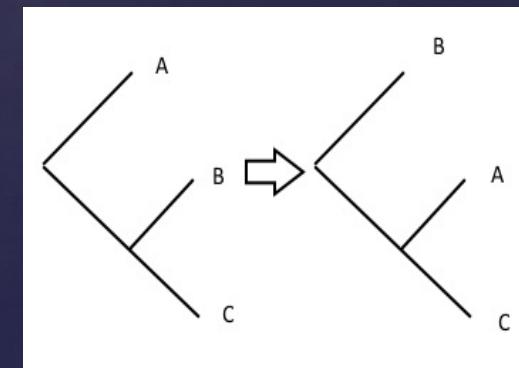
Вероятностные методы
построения
филогенетических
деревьев: метод
максимального
правдоподобия
(Maximum likelihood)

- ❖ Другим подходом к построению филогенетических деревьев является применение статистических оценок.
- ❖ Этот подход реализуется в методе максимального правдоподобия.
- ❖ Для математической оценки дерева вводится функция правдоподобия

$$L = P(Data|Parameters)$$

- ❖ Функция правдоподобия определяется как условная вероятность получения имеющихся данных при заданной модели.

- ❖ Идея метода – поиск максимума функции правдоподобия.
- ❖ Для этого производится перестройка начального дерева, которое может быть получено различными способами – как генерировано случайно, так и построено более простым методом (например NJ).
- ❖ Существуют два пути перестройки дерева:
 - ❖ перестановка ближайших соседей
 - ❖ выделение и перенос поддерева
- ❖ В случае если перестроенное дерево имеет большее значение функции правдоподобия, они сохраняются.
- ❖ Процесс оптимизации дерева продолжается до тех пор, пока изменения функции правдоподобия не становятся меньше установленного ограничения.



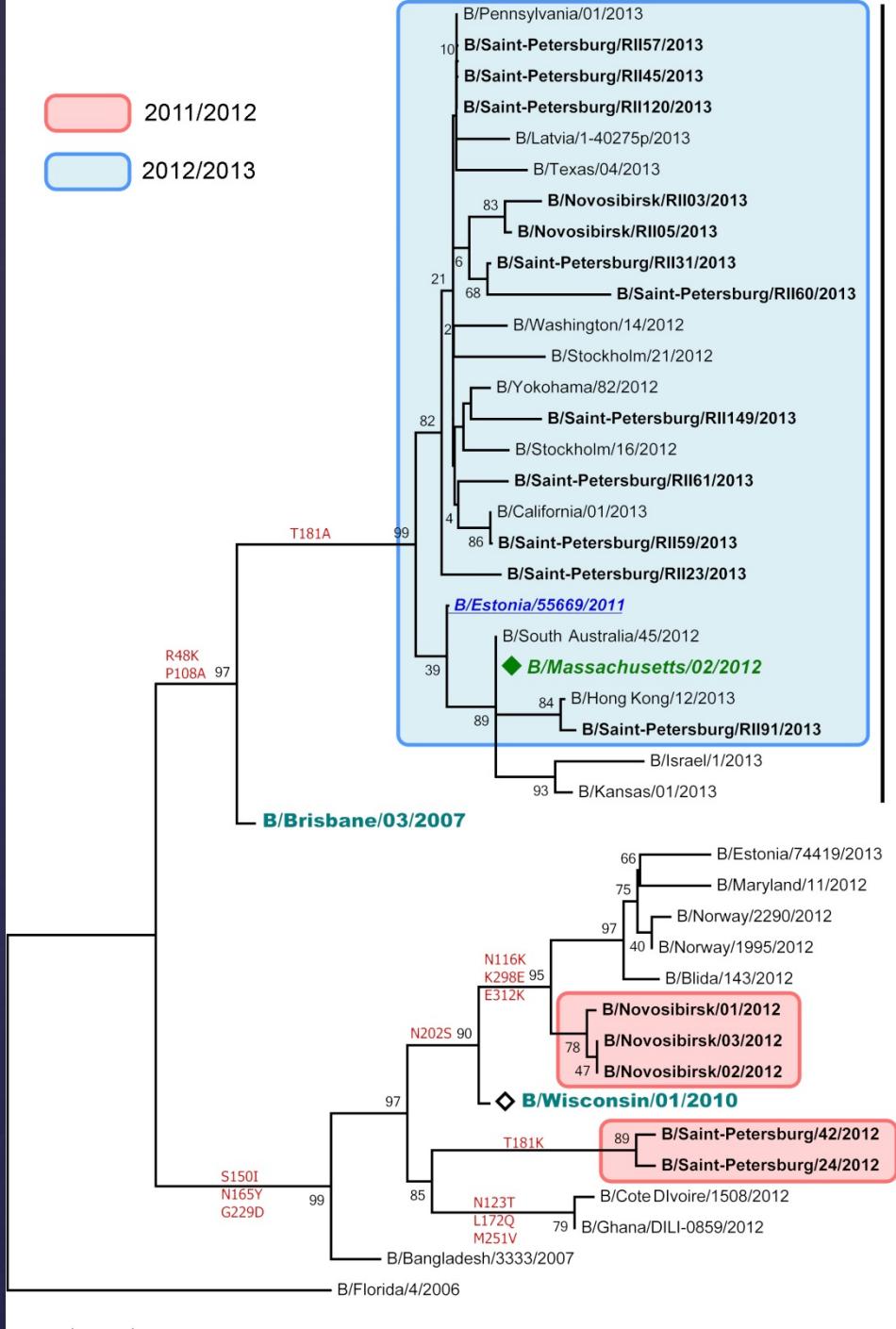
Статистическая оценка филогении: бутстреп

- ❖ Для проверки статистической достоверности полученных деревьев обычно используется бутстреп.
- ❖ В этом случае программа произвольным образом из имеющегося набора данных генерирует новые наборы того же размера и повторяет поиск дерева по ним.
- ❖ Далее производится статистический анализ встречаемости узлов в полученных деревьях.
- ❖ Обычно достаточной выборкой считают 500-1000 бутстреп репликаций.



Интерпретация филогенетических деревьев

Филогенетическое дерево вирусов гриппа В ямагатской линии, построенное методом ML, модель HKY+G, бутстреп 1000 репликаций



Генетическое разнообразие вирусов гриппа

- ❖ Ценным источником информации о текущей структуре популяции вирусов гриппа в мире являются регулярные отчеты британского Национального Института Медицинских Исследований.
- ❖ Отчеты NIMR регулярно размещаются по адресу:
<http://www.nimr.mrc.ac.uk/who-influenza-centre/annual-and-interim-reports/>
- ❖ Рассмотрим разнообразие циркулирующих вирусов гриппа по данным отчета NIMR за февраль 2014 года.

Вирусы гриппа А подтип H1N1pdm09

Figure 4. Phylogenetic comparison of influenza A(H1N1)pdm09 HA genes

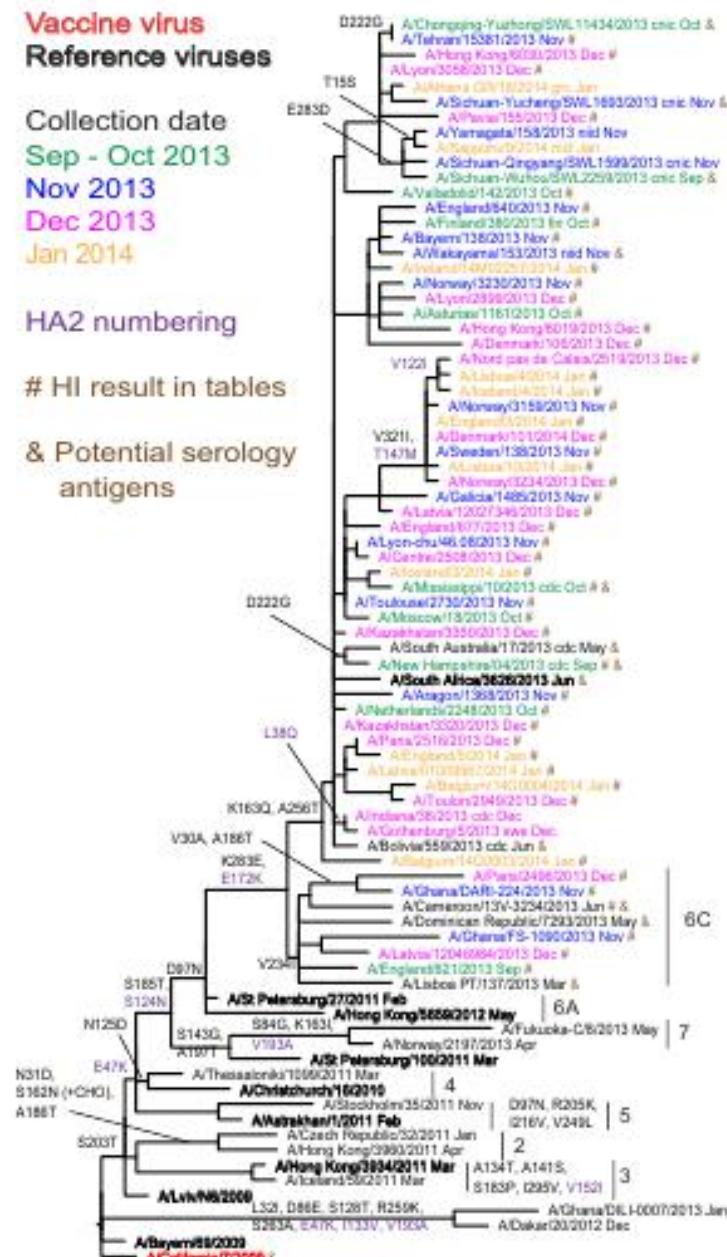
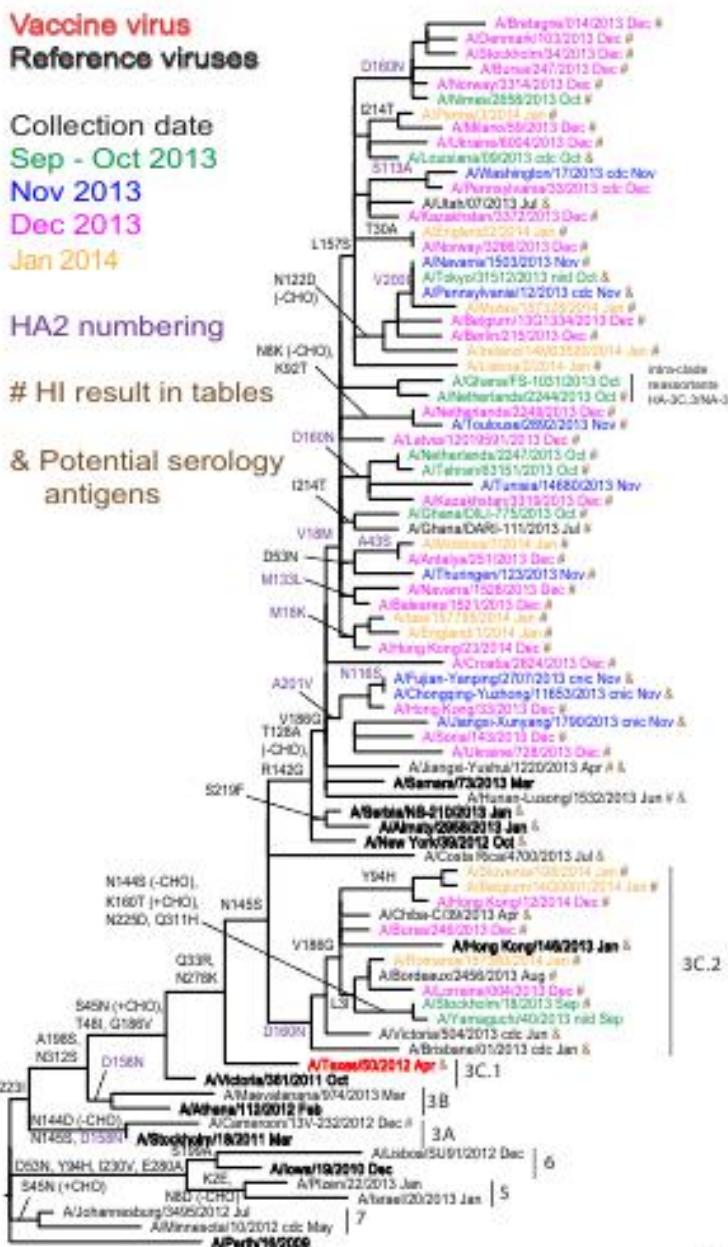


Figure 9. Phylogenetic comparison of influenza A(H3N2) HA genes

Вирусы гриппа А подтипа H3N2



Вирусы гриппа В викторианской линии

Figure 15. Phylogenetic comparison of influenza B (Victoria-lineage) HA genes

Vaccine virus

Reference viruses

Collection date

Oct 2013

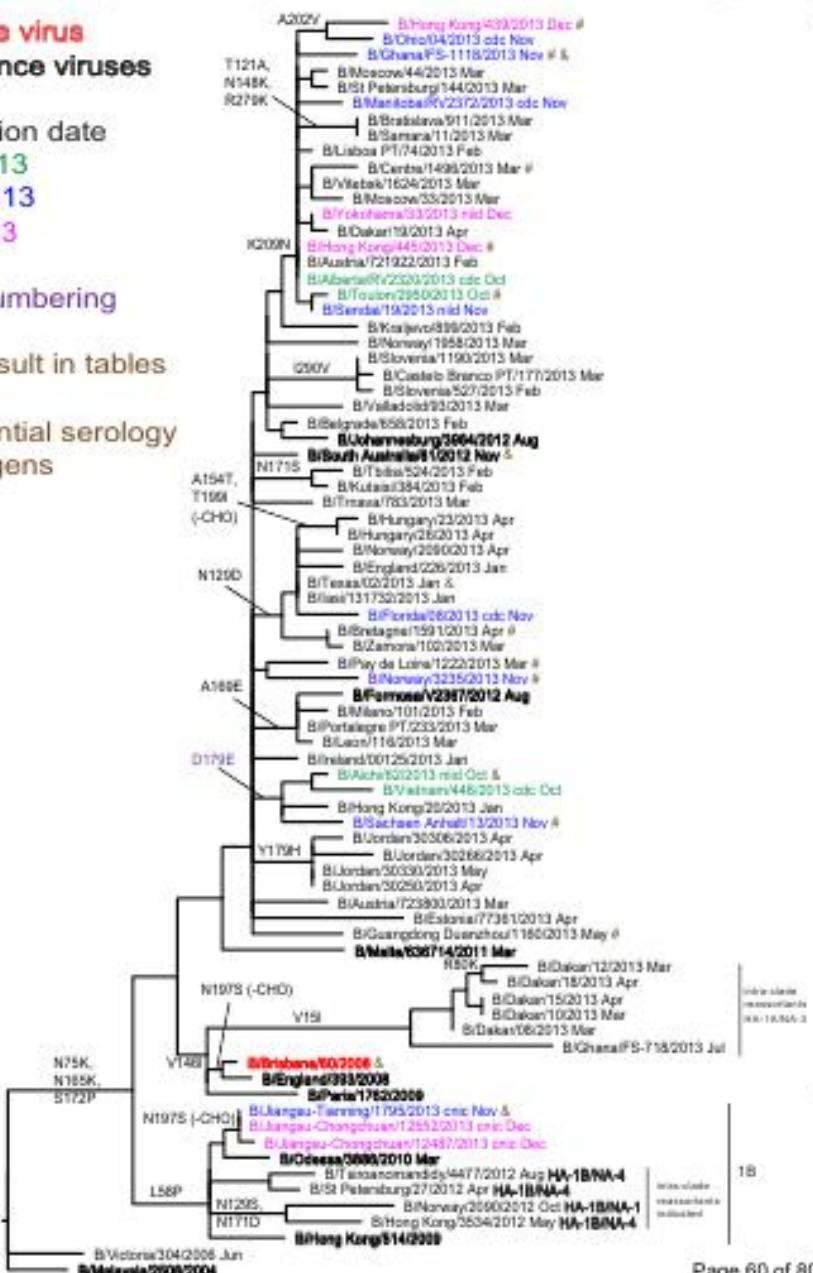
Nov 2013

Dec 2013

HA2 numbering

HI result in tables

& Potential serology
antigens



Вирусы гриппа В ямагатской линии

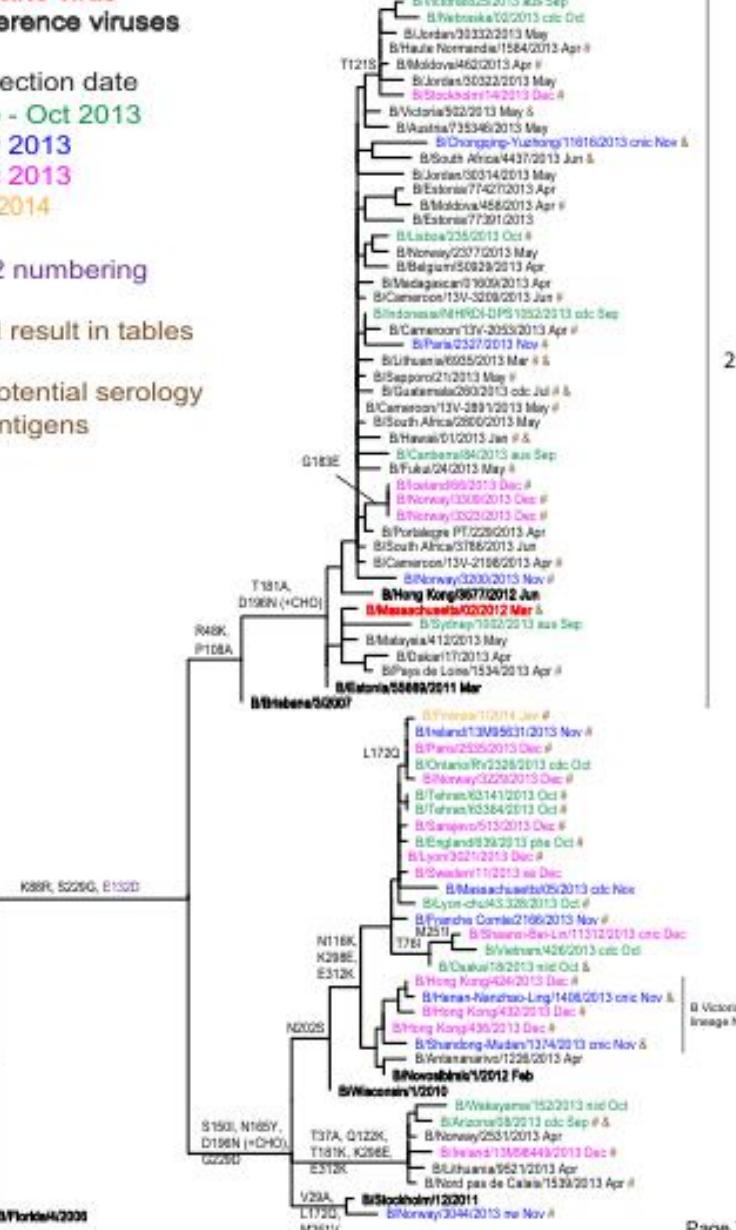
Figure 19. Phylogenetic comparison of influenza B (Yamagata-lineage) HA genes

Vaccine virus
Reference viruses

Collection date
Sep - Oct 2013
Nov 2013
Dec 2013
Jan 2014

HA2 numbering

HI result in tables
& Potential serology
antigens



0.002

Практическая часть