Алгоритъм ClustalW за множествено подравняване на биологични секвенции

проф. Пламенка Боровска

проф. Боровска

Алгоритъм ClustalW за множествено подравняване на биологични секвенции

- Един от най-често използваните алгоритми за множествено подравняване, който е базиран на прогресивното подравняване
- Основната идея - първоначално подравняване на най-тясно свързаните секвенции
- ClustalW е първият алгоритъм, при който се комбинира подравняването по двойки на секвенциите с глобалното подравняване с цел намаляване на изчислителното време.
- *Алгоритъмът ClustalW има сложност* $O(N^2)$ поради използването на метода на присъединяване на съседите (neighbor-joining method).

Алгоритъм ClustalW за множествено подравняване на биологични секвенции

- Методът *Clustal V* (предложен от Higgins и Sharp, 1989 г.) групира секвенциите в клъстери на основата на анализ на еволюционните дистанции на всяка двойка секвенции.
- Клъстерите се подреждат по двойки, след това колективно като групи секвенции с цел получаване на глобално подравняване.
- За полученото подравняване на секвенциите се прилага методът за присъединяване на съсед за реконструкция на филогенетичното дърво.
- Предназначението на метода *Clustal W* (Thompson, 1994 г.) е да генерира по-точни подреждания от метода Clustal V за секвенции с големи еволюционните дистанции.

ClustalW

- Той е евристичен алгоритъм и осигурява оптимално или най-често близко до оптималното подравняване.
- Въпреки недостатъците за качеството на решение, той е широко използван поради неговите предимства, а именно:

(1) приемливо време за подравняване на големи набори дълги секвенции, и

(2) направляващото дърво намалява чувствителността на подравняването към зашумеността на данните

ClustalW

- Clustalw е широко използван софтуерен продукт за подравняване на множество ДНК, РНК или протеинови секвенции.
- Резултатите показват общия произход в еволюцията на тези секвенции.
- ClustalW, в сравнение с други MSA алгоритми, се изпълнява като един от най-бързите, като все още поддържа приемливо ниво на точност.
- Точността за ClustalW е по-ниска при тестване срещу MAFFT, T-Coffee, Clustal Omega и други MSA софтуерни инструменти, но при него изискванията за RAM памет са най-малки.
- Има актуализации и подобрения на алгоритъма, напр., в ClustalW2, които водят до повишаване на точността, при запазване на висока скорост на изчисленията.

Алгоритъмът ClustalW се състои от три основни стъпки

1.Подравняване по двойки
(pairwise alignment)
2.Конструиране на направляващо дърво
(guide tree)
3.Множествено подравняване

ClustalW – стъпка 1: Подравняване по двойки (pairwise alignment)

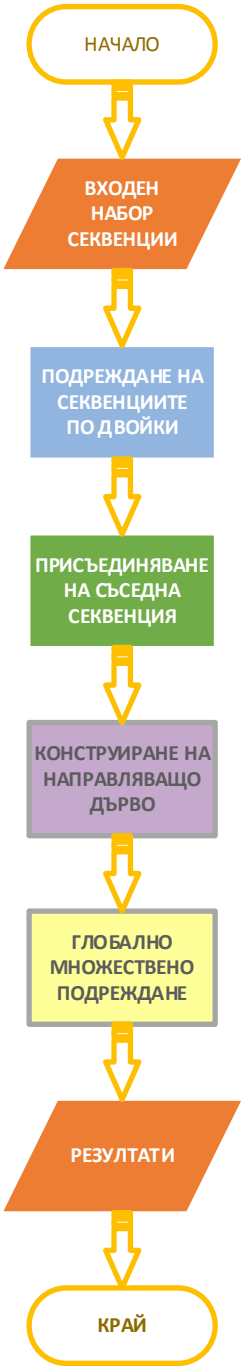
- Изчисляват се разликите между двойките секвенции чрез подравняване на секвенциите по двойки.
- Пресмята се дистанцията (разликите) между всеки две секвенции.
- За всяка двойка секвенции се извършва сравнение на двойките символи, при които няма празна позиция.
- Разстоянието се пресмята, като се раздели броят на двойките, при които съществува различие, на общия брой двойки.

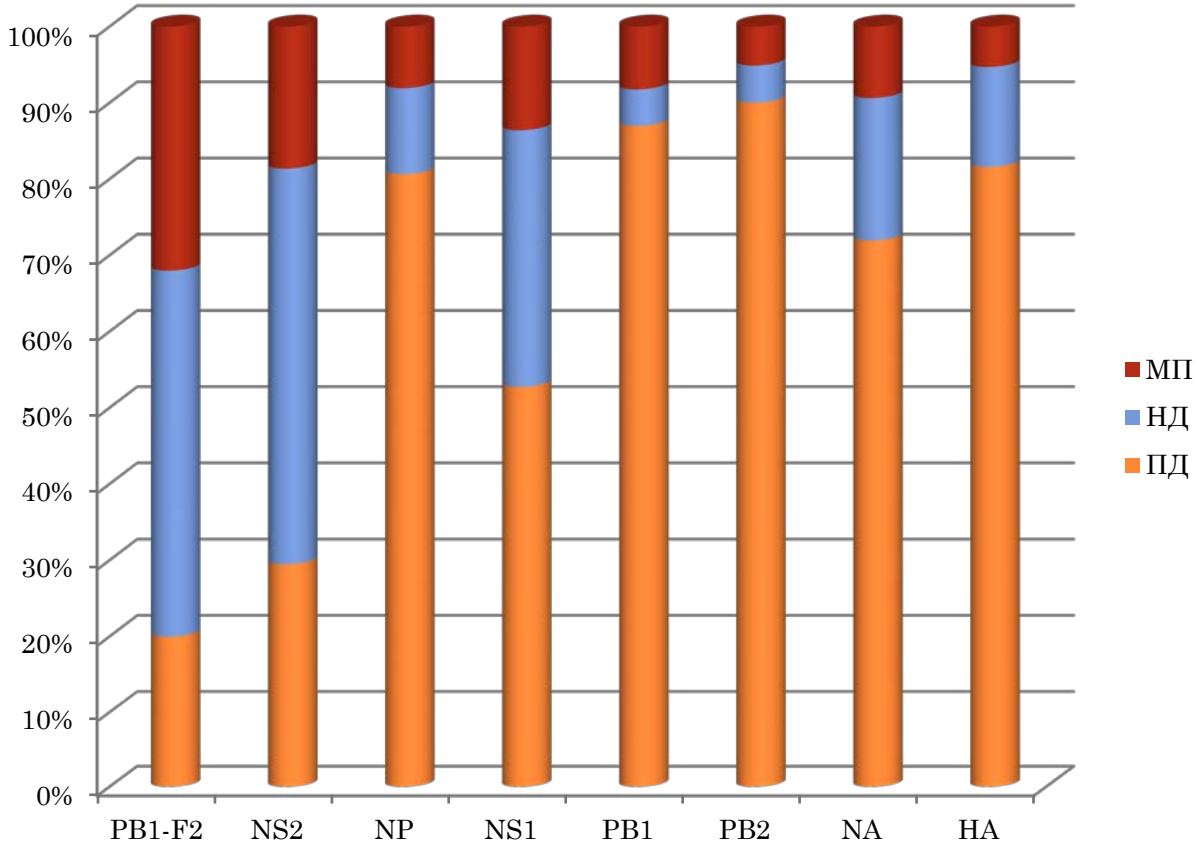
ClustalW – стъпка 2: Конструиране на направляващо дърво

- На базата на матрицата на разстоянията се построява направляващо дърво.
- При метода ClustalW се използва матрицата на дистанциите и методът на присъединяване на съседна секвенция (*Neighbor Joining)* за да се изгради направляващото дърво (guide tree).
- Комбинират се резултатите от подравняването по двойки, като се започне от най-близко свързаните секвенции и се завърши с най-отдалечените.
- Процесът започва от върховете до достигане на корена на дървото.

ClustalW – стъпка 3: Множествено подравняване

- Секвенциите се подреждат на базата на направляващото дърво.
- Започва се с подравняването на двойките, които имат най-голямо сходство.


НАЧАЛО
ВХОДЕН НАБОР СЕКВЕНЦИИ
ПОДРЕЖДАНЕ НА СЕКВЕНЦИИТЕ ПО ДВОЙКИ
ПРИСЪЕДИНЯВАНЕ НА СЪСЕДНА СЕКВЕНЦИЯ
КОНСТРУИРАНЕ НА НАПРАВЛЯВАЩО ДЪРВО
ГЛОБАЛНО МНОЖЕСТВЕНО ПОДРЕЖДАНЕ
РЕЗУЛТАТИ
КРАЙ



Относителни дялове на времената за паралелно изпълнение на различните фази при множествено подравняване с алгоритъма ClustalW на 8-те сегмента на РНК на вируса на свинския грип AH1/N1 на 32 ядра на компютърен клъстер: ПД - подравняване по двойки, НД – насочващо дърво, МП – множествено подравняване

Визуализацията на резултатите

За улесняване на интерпретацията на резултатите, получени от приложението на алгоритъма ClustalW е необходимо изходните данни да съдържат следното:

- Визуализация на подреденото множество секвенции;
- Конструираното филогенетично дърво;
- Таблица на дистанциите между секвенциите;
- Визуализация и хистограма на консенсусните региони в секвенциите;
- Възможност за редактиране на секвенциите.

Визуализацията на консенсусните региони в подравненото множество секвенции посредством многоцветното им маркиране по диапазони

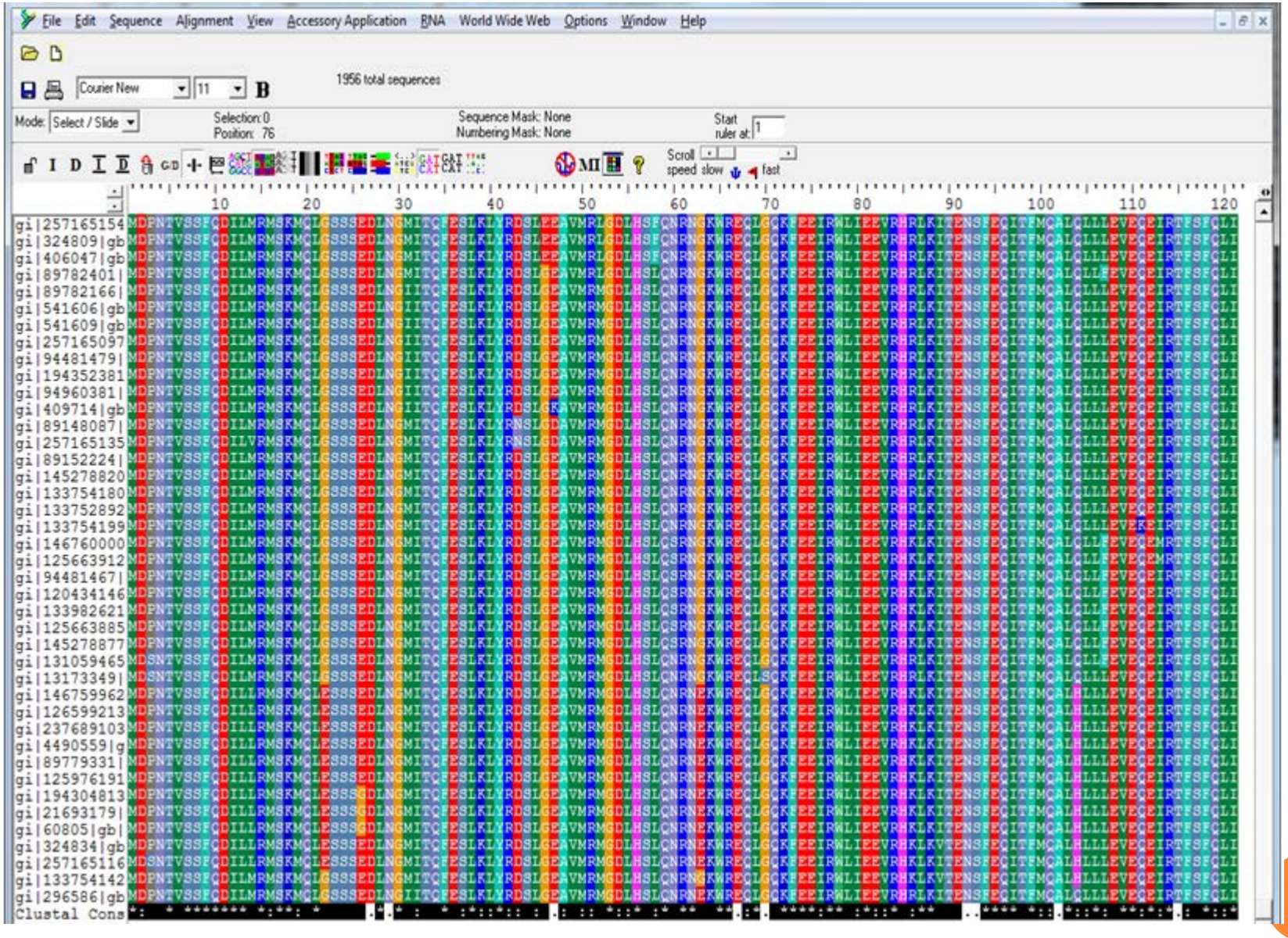
- Таблицата на дистанциите между секвенциите съдържа *процента на сходство за всяка двойка секвенции*, който се определя по следния начин:

Процент на сходство = (брой на съвпаденията x 100)/общ брой на символите в региона (вкл. вътрешните празни позиции, не се отчитат тези в края на секвенциите)

- За визуализацията на консенсусните региони в подреденото множество секвенции в общия случай се използва многоцветното им маркиране по диапазони

Визуализацията на консенсусните региони в подравненото множество секвенции посредством многоцветното им маркиране по диапазони

Степен на сходство	Цвят
<1%	
1-20%	
21-40%	
41-60%	
61-80%	
81-100%	


File Edit Sequence Alignment View Accessory Application BNA World Wide Web Options Window Help
Courier New
11
1956 total sequences
Mode: Select / Slide
Selection: 0
Position: 76
Sequence Mask: None
Numbering Mask: None
Start ruler at: 1
Scroll speed slow fast
10 20 30 40 50 60 70 80 90 100 110 120
gi|257165154
gi|324809|gb
gi|406047|gb
gi|89782401|
gi|89782166|
gi|541606|gb
gi|541609|gb
gi|257165097
gi|94481479|
gi|194352381
gi|94960381|
gi|409714|gb
gi|89148087|
gi|257165135
gi|89152224|
gi|145278820
gi|133754180
gi|133752892
gi|133754199
gi|146760000
gi|125663912
gi|94481467|
gi|120434146
gi|133982621
gi|125663885
gi|145278877
gi|131059465
gi|13173349|
gi|146759962
gi|126599213
gi|237689103
gi|4490559|g
gi|89779331|
gi|125976191
gi|194304813
gi|21693179|
gi|60805|gb|
gi|324834|gb
gi|257165116
gi|133754142
gi|296586|gb
Clustal Cons

Документиране на кода на Clustalw с Doxygen

- документиране на кода в HTML формат посредством стандартния софтуерен инструмент за документиране на софтуер от сорс файлове Doxygen
- Doxygen е стандартен инструмент за генериране на документация от анотирани източници на C ++, но също така поддържа други популярни езици за програмиране като C, C #, PHP, Java, Python, и др.
- Doxygen е разработен под операционните системи Mac OS X и Linux, но също така е преносим.

Функционалност на Doxygen

- Може да генерира онлайн браузър за документация (в HTML) и/или офлайн справочно ръководство от набор от документирани изходни файлове.
- Има и поддръжка за генериране на изходни данни в RTF (MS-Word), PostScript, хипервръзка PDF, компресиран HTML.
- Документацията се извлича директно от сорс кода, което много улеснява поддържането на актуална документацията в съответствие с изходния код.

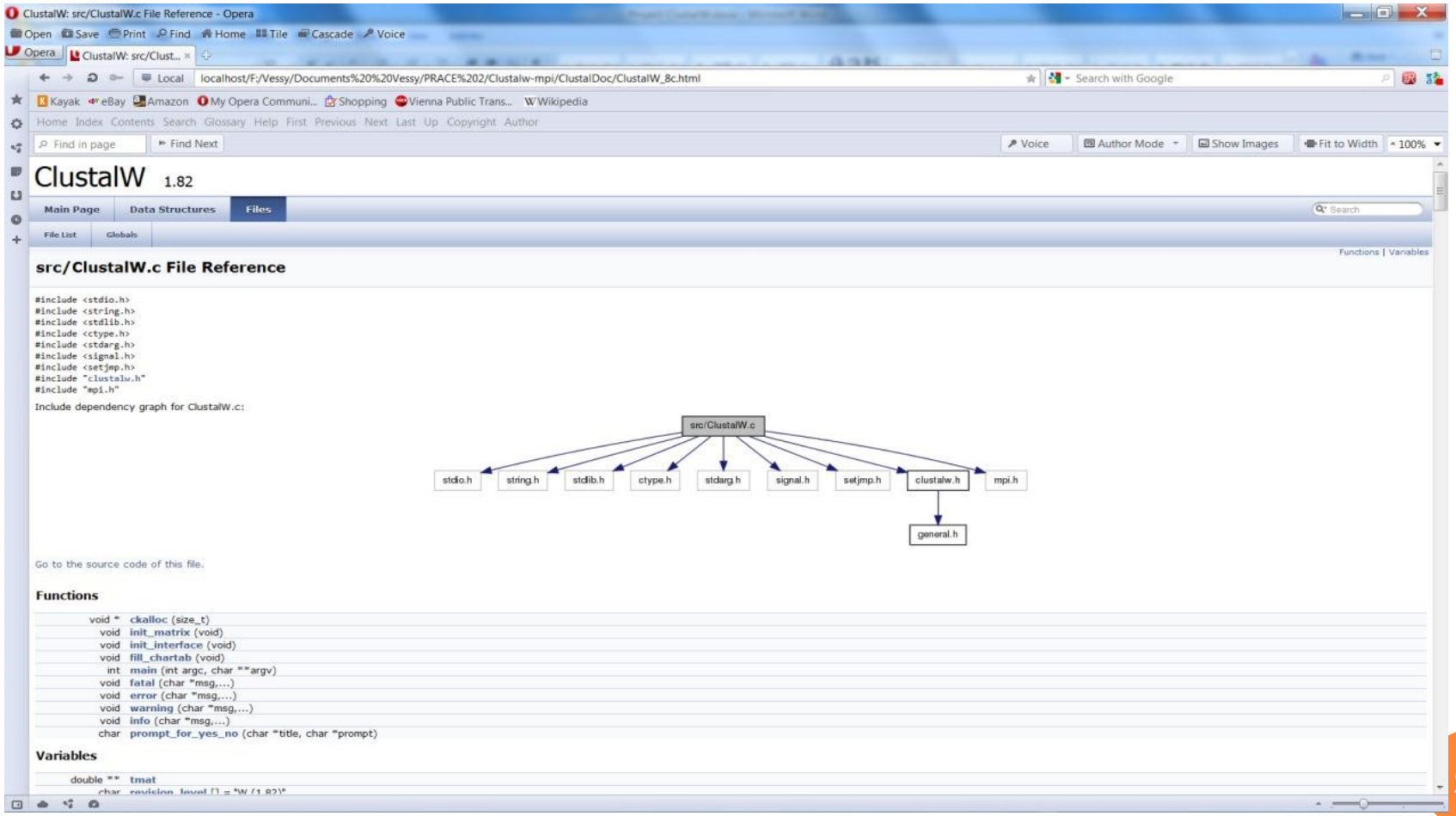
Функционалност на Doxygen

- Doxygen може да се конфигурира за извличане на структура на кода от недокументирани файлове на сорс кода.
- Doxygen може също така да визуализира релациите между различните компоненти на кода чрез генериране на графики на зависимост, диаграми за наследяване и колаборативни диаграми, които се генерират автоматично.
- Doxygen може също така да се използва за създаване на документация на кода.
- В документацията е включен и програмния код, като по този начин е възможно винаги да се поддържа актуална версия на кода по отношение на направени промени.

Функционалност на Doxygen

- Поддържат се връзки между отделните документирани сорс файлове и компонентите на софтуера – функции, структури от данни, променливи и др.
- Препратките позволяват да се проследяват връзките, извикванията на отделните функции, структурите данни, променливите, дефинициите и др. с цел улесняване на анализа на кода.
- Включени са метаданни за описание на кода.
- Също така е включена е възможност да се визуализират отношенията между различните елементи на програмната система чрез диаграми, представящи графи на зависимостите и връзките.

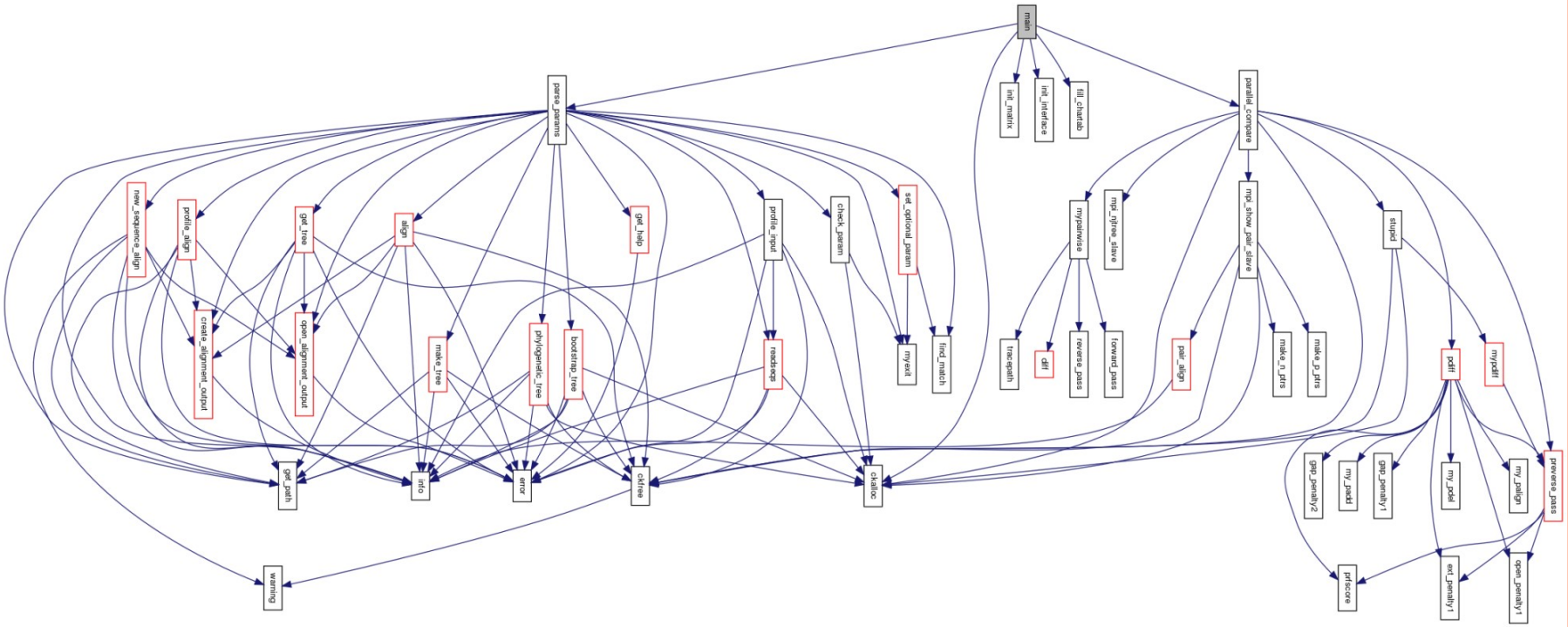
Структура, съдържание и референции на основния файл clustalw.c



Програмните файлове на софтуера ClustalW

clustalw.h - Main header file for ClustalW
general.h - General purpose header file
dayhoff.h - Table of estimated PAMS matrices
matrices.h – Scoring matrices
param.h – List of parameters
clustalw-mpi.c – function main()
parallel_compare.c – parallel pairwise alignment, calls function pair_align()
pairalign_new.c - parallel pairwise alignment
trees.c - phylogenetics tree calculating functions
malign_mpi_progressive.c -(function malign_mpi_progressive()) - full progressive multiple alignment
malign_mpi_pdiff.c (function malign_mpi_pdiff ()) - full progressive multiple alignment
malign.c (function malign()) - sequential full progressive alignment
alnscore.c - calculation overall score for the alignment
calcgapcoeff.c - calculation gap penalty
calctree.c - Calculation guide tree
interface.c - Command line interface
random.c – Random number generators
readmat.c – Matrix initialization
sequence.c – Read the sequences and structure data

Граф на извикванията на главната функция main()



Редактиране на региони

- След финализирането на множественото подравняване, има възможност за ръчно редактиране на региони от секвенциите посредством софтуерните инструменти на палетата „*Straighten Columns*“, „*Shuffle Right*“ и „*Shuffle Left*“.
- При „изправянето“ на колоните (*Straighten Columns)* на подравняването в ляво за маркирани региони от избрани секвенции се вмъкват празнини.
- При разместването (Shuffling) елементите на секвенциите се преместват от единия край на празнина до другия край.

Вирусна биоинформатика

- Вирусите са обект на еволюцията и на естествения отбор също като клетъчните форми на живот. Много от вирусите еволюират бързо.
- Когато два вируса инфектират една и съща клетка едновременно, те могат да си обменят генетичен материал и така да се получат нови, "смесени" вируси с уникални свойства. По този начин могат да се появят различни щамове грип.
- РНК вирусите имат високи нива на мутации, което им позволява да еволюират изключително бързо.
- Пример за това е ХИВ, който еволюира във форми, резистентни към лекарства

Рекомбинация

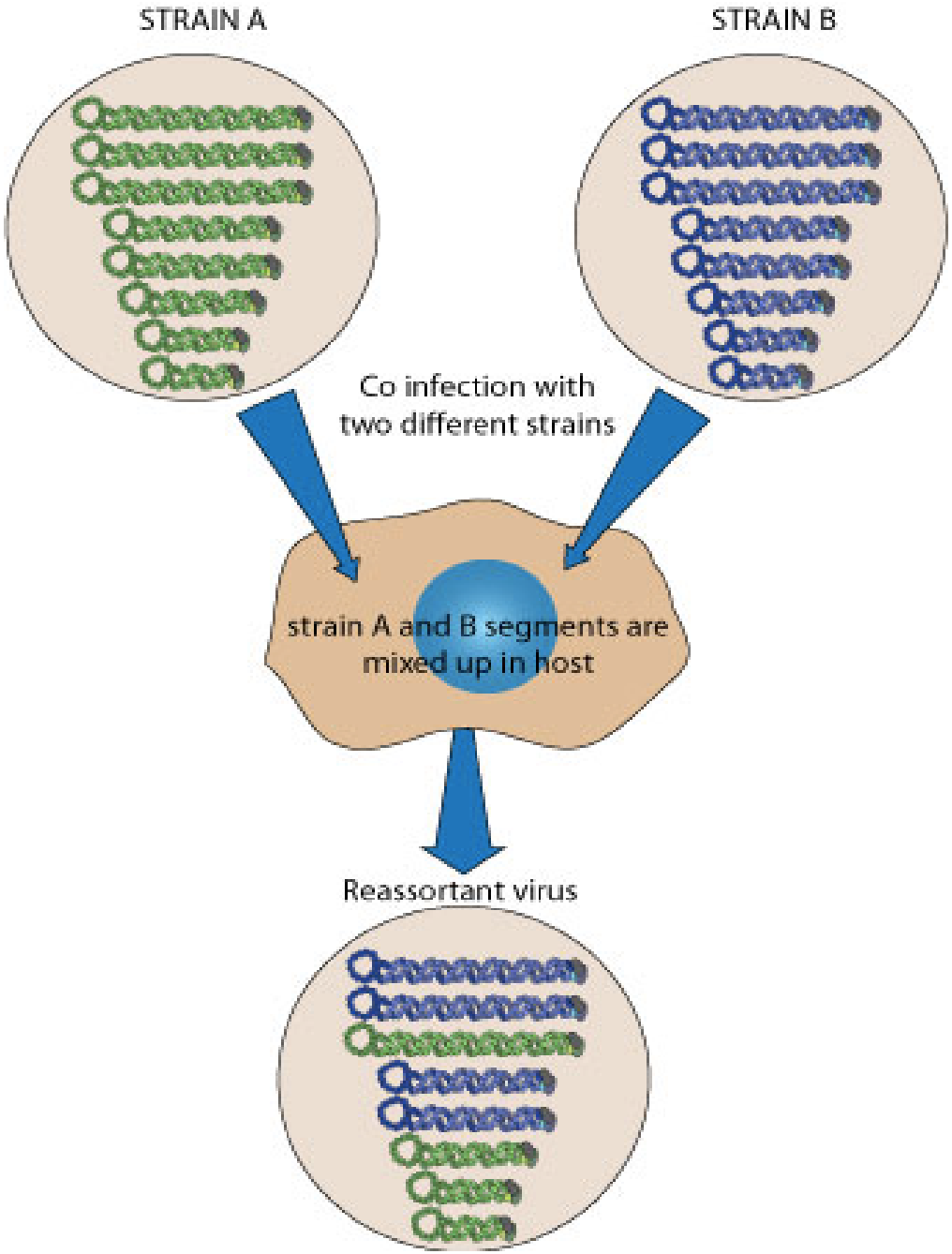
- Вирусите не просто еволюират, те еволюират по-бързо от гостоприемниците си, например човека.
- Това прави еволюцията на вирусите важна тема не само за биолозите, които изучават вируси, но също и за лекарите, сестрите и останалите хора, които работят в здравната система, както и за всеки, който може да бъде изложен на вируси.
- Вирусите си разменят ДНК или РНК чрез процес на име **рекомбинация**.
- Най-често рекомбинацията се случва, когато два вируса заразят едновременно една и съща клетка.
- Тъй като и двата вируса използват клетката, за да произведат нови вирусни частици, във вътрешността ѝ ще има много плаващи вирусни елементи по едно и също време.

Рекомбинация

- Рекомбинацията е процесът, при който две отделни молекули от ДНК или РНК обменят участъци от своя геном.
- Обменът често е в хомоложни области на генома и се извършва както в едноверижни, така и в двуверижни молекули на ДНК и РНК.
- По този начин рекомбинацията е основен механизъм, движещ еволюционната промяна.

Рекомбинация

- При тези обстоятелства рекомбинацията може да протече по два начина.
- Първият е подобни региони от вирусните геноми да се сдвоят и и да си обменят части от тях като физически разрушат и след това свържат отново ДНК или РНК.
- При втория начин вируси с различни сегменти, подобно на малки хромозоми, могат да си разменят някои от тези сегменти чрез процес, наречен ресортиране.


STRAIN A
STRAIN B
Co infection with
two different strains
strain A and B segments are
mixed up in host
Reassortant virus

Рекомбинация и инфлуенца

- Вирусите инфлуенца (вирусите на грипа) са “майстори” на ресортирането.
- Те имат осем сегмента РНК, всеки от които носи по един или по няколко гена.
- Когато два вируса инфлуенца заразят една и съща клетка по едно и също време, някои от новите вирусни частици, произведени в клетката могат да имат микс от сегменти (например сегменти 1-4 от щам А и сегменти 5-8 от щам В).

Как животинските вируси заразяват клетки?

- Животинските вируси, също като останалите вируси, зависят от клетка гостоприемник, за да изпълнят жизнения си цикъл.
- За да се възпроизведе, вирусът трябва да зарази клетка гостоприемник и да я програмира да произвежда нови вирусни частици.
- Първата ключова стъпка от инфектирането е разпознаване.
- Животинските вируси имат специални молекули на повърхността си, които им позволяват да се свържат с рецептори от мембраната на клетката гостоприемник.
- Щом се прикрепят за клетката гостоприемник, животинските вируси могат да навлязат в нея по различни начини

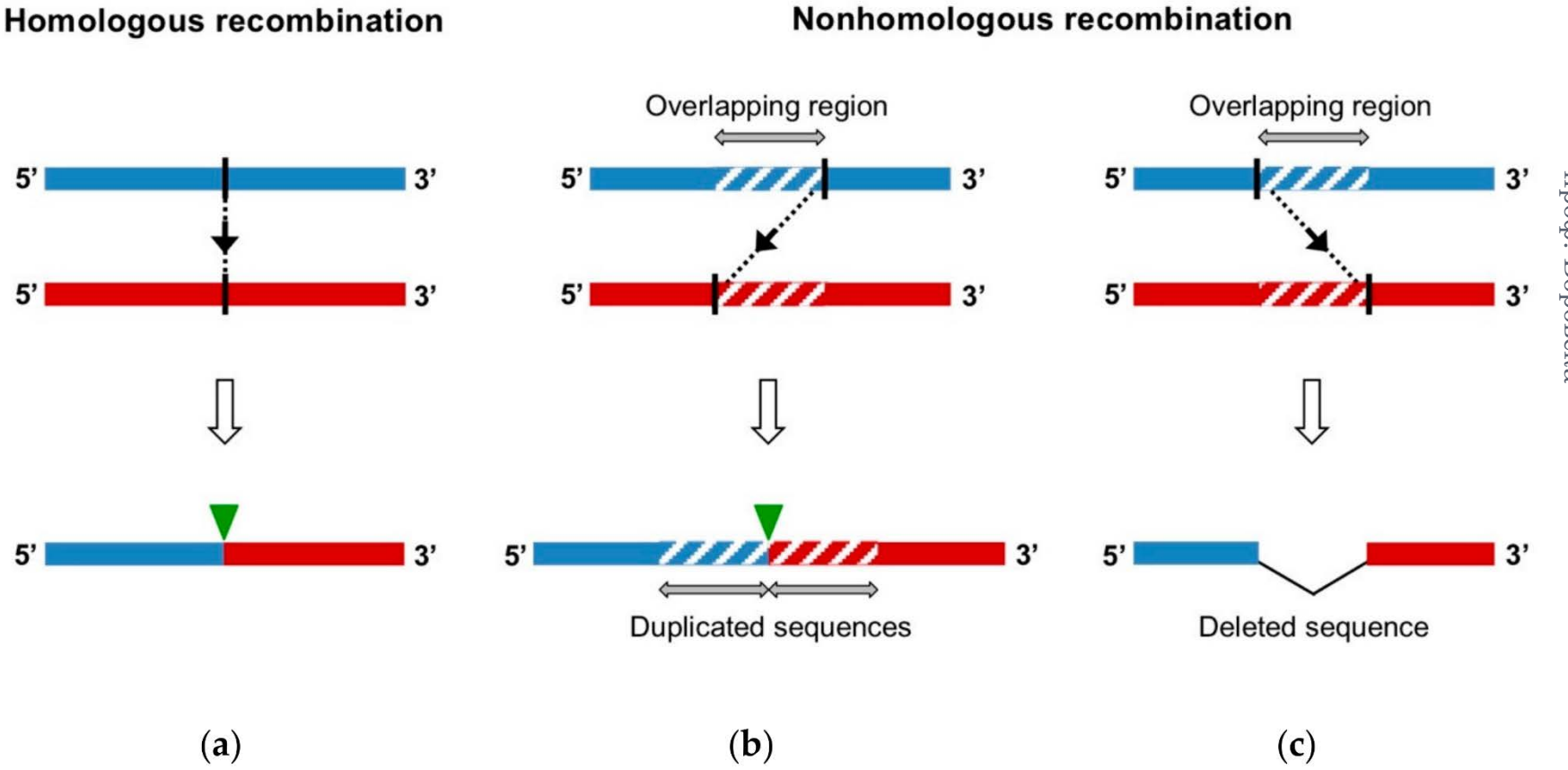
проф. Боровска

Рекомбинация

- Прасетата са добре известни "смесителни съдове" за вирусите инфлуенца
- Клетките на прасетата могат да бъдат разпознати и заразени от човешки и от птичи вируси на инфлуенца (както и от свински вируси, разбира се).
- Ако дадена клетка в прасето е инфектирана от два вида вирус по едно и също време, тя може да освободи нови вируси, които съдържат смес от генетичните материали на човешкия и на птичия вирус.

Рекомбинация

- Този вид обмяна е много често срещана в природата при вирусите инфлуенца.
- Например, щама H1N1 (свински грип), който предизвика пандемия през 2009 г.
- H1N1 съдържаше сегмент от човешки и птичи вируси, както и свински вируси от Северна Америка и Азия.
- Тази комбинация отразява поредица от ресортирания, които са се случили стъпка по стъпка в продължение на много години, за да се стигне до щама H1N1


Homologous recombination
Nonhomologous recombination
Overlapping region
Overlapping region
5'
3'
5'
3'
5'
3'
5'
3'
5'
3'
5'
3'
5'
3'
5'
3'
Duplicated sequences
Deleted sequence
(a)
(b)
(c)

Честота на мутации на вирусите

- Някои вируси имат много висока честота на мутации, което им помага да еволюират бързо, осигурявайки им по-голямо генетично разнообразие.
- Два други фактора, които допринасят за бързата еволюция са големият размер на популациите и краткият жизнен цикъл.
- Колкото по-голяма е популацията, толкова по-голям е шансът в нея да има вирус, който носи случайна мутация, която ще бъде селектирана от естествения отбор, например мутация, която прави вируса резистентен към лекарства или усилва вирулентността (инфекциозността) му.
- Освен това вирусите се възпроизвеждат бързо, така че популациите им еволюират по-бързо от тези на гостоприемниците им. Например вирусът ХИВ завършва жизнения си цикъл само за 52 часа, което е много по-кратко време в сравнение с жизнения цикъл на човека.

Рекомбинация на вирусни РНК-и

- Рекомбинация на РНК има по-голяма вероятност да възникне на места (сайтове), с висока степен на хомология между секвенциите на двата родителски вируса, а също и по-лесно се осъществява между тясно свързани вирусни РНК-и.
- Предполагаемите рекомбинационни горещи точки (hot spots) съответстват на хипервариантния регион, в който се появяват чести делеции (изтривания) след преминаване на вируса в тъканна култура или животни
- рекомбинация между суперинфектираща вирусна РНК и РНК

RAT (Recombination Analysis Tool)

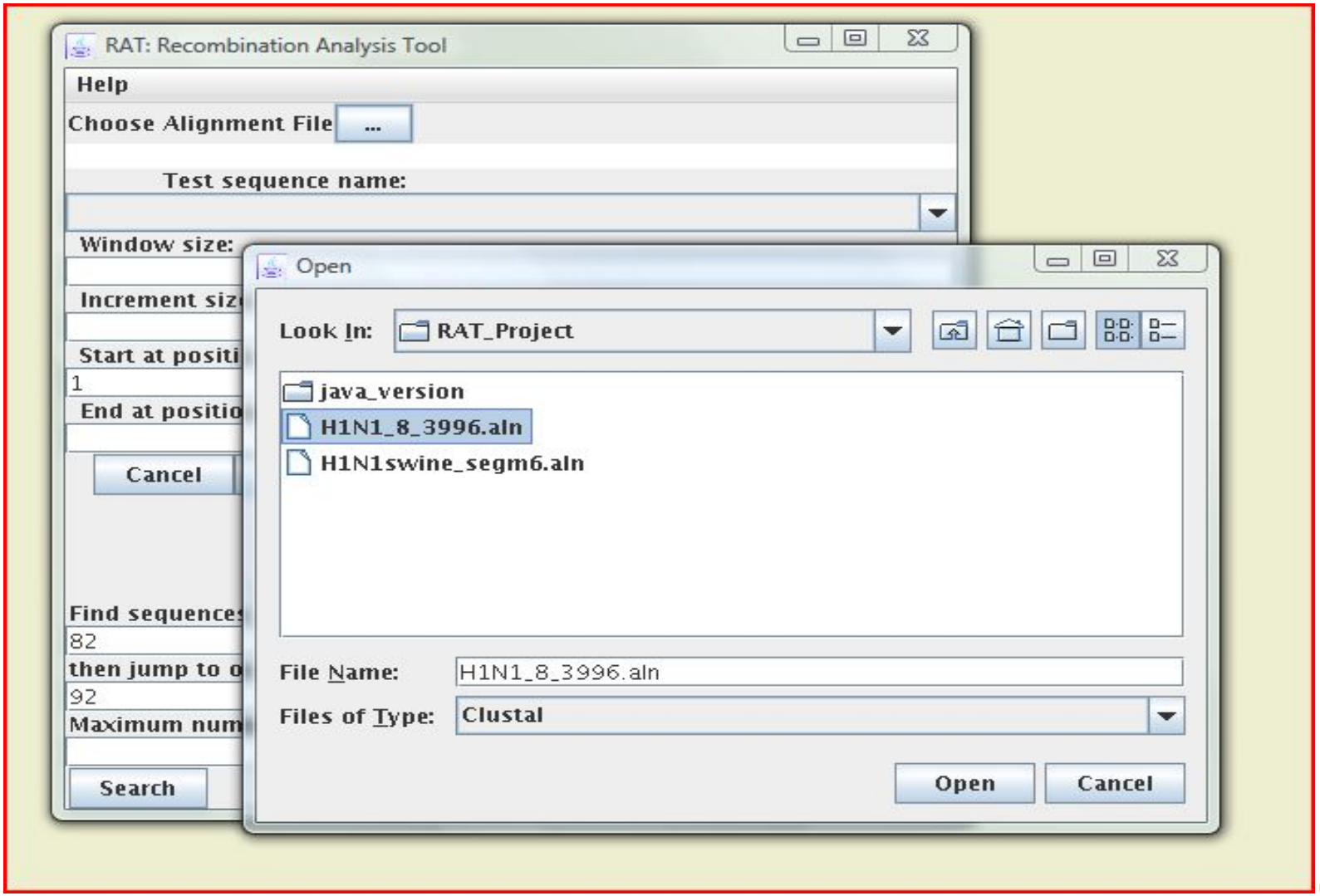
- RAT (Recombination Analysis Tool) е кросплатформено Java-базирано приложение за изследване на рекомбинацията в произволен брой подравнени секвенции (протеини или ДНК) с произволна дължина (къси секвенции на вируси до пълни геноми).
- Позволява анализа на данни от файлове с до 7 формата
- Използва метода на дистанциите за откриване на рекомбинация.
- Всички операции при RAT се изпълняват през графичния потребителски интерфейс GUI.

RAT (Recombination Analysis Tool)

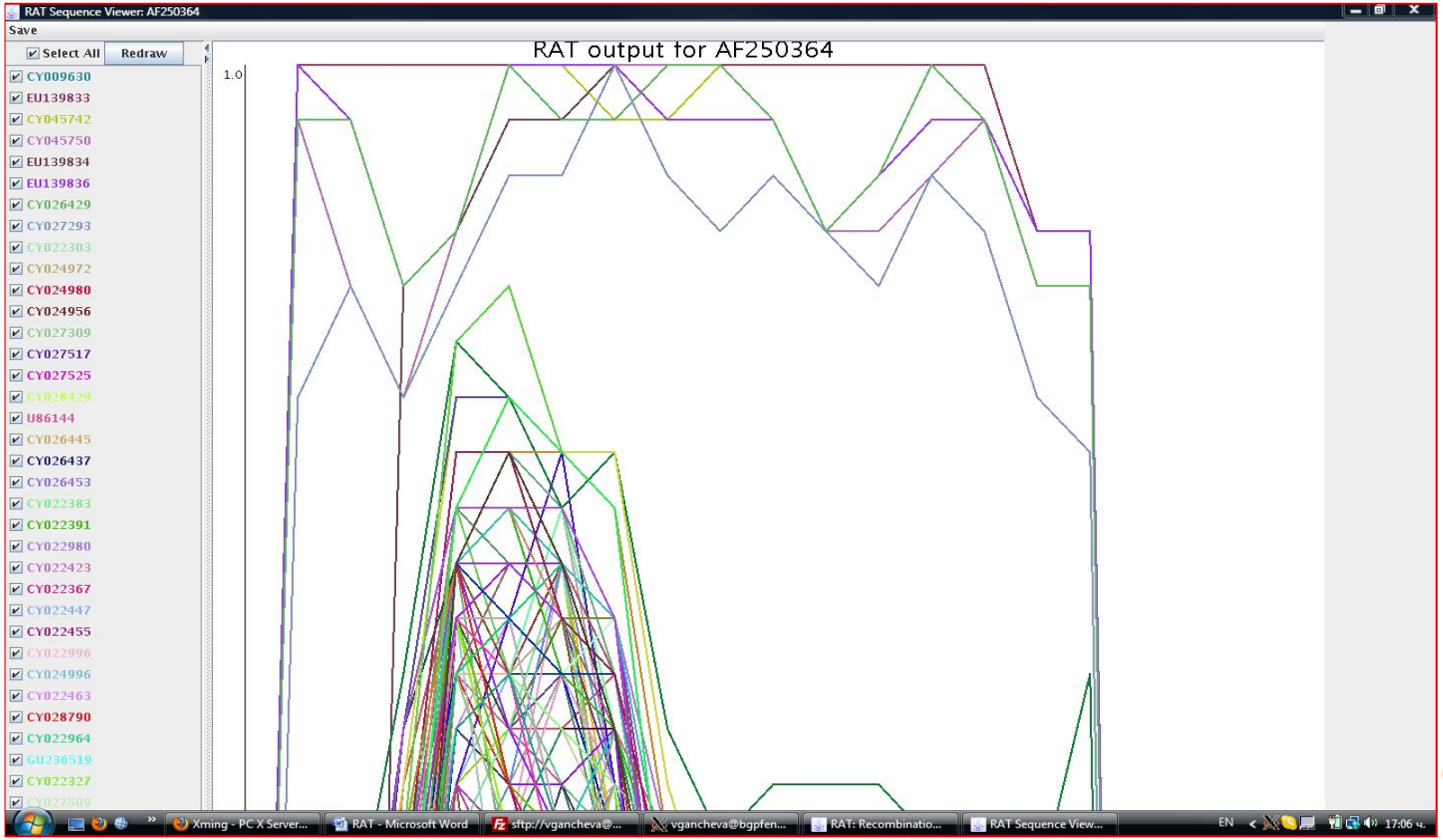
- Изходните данни се съхраняват във файлове с формат .txt, .xls, .csv или като .jpg файлове при графичен изход.
- Стартовата позиция за анализа по подразбиране е първата.
- При желание да се анализира дадена област в рамките на секвенцията, може да се промени стартовата позиция и това ще промени областта за анализ.
- Плъзгащият се прозорец на софтуера RAT се движи по сцените на кратки стъпки и се разглежда пространството с размера на зададения "размер на прозореца".

RAT (Recombination Analysis Tool)

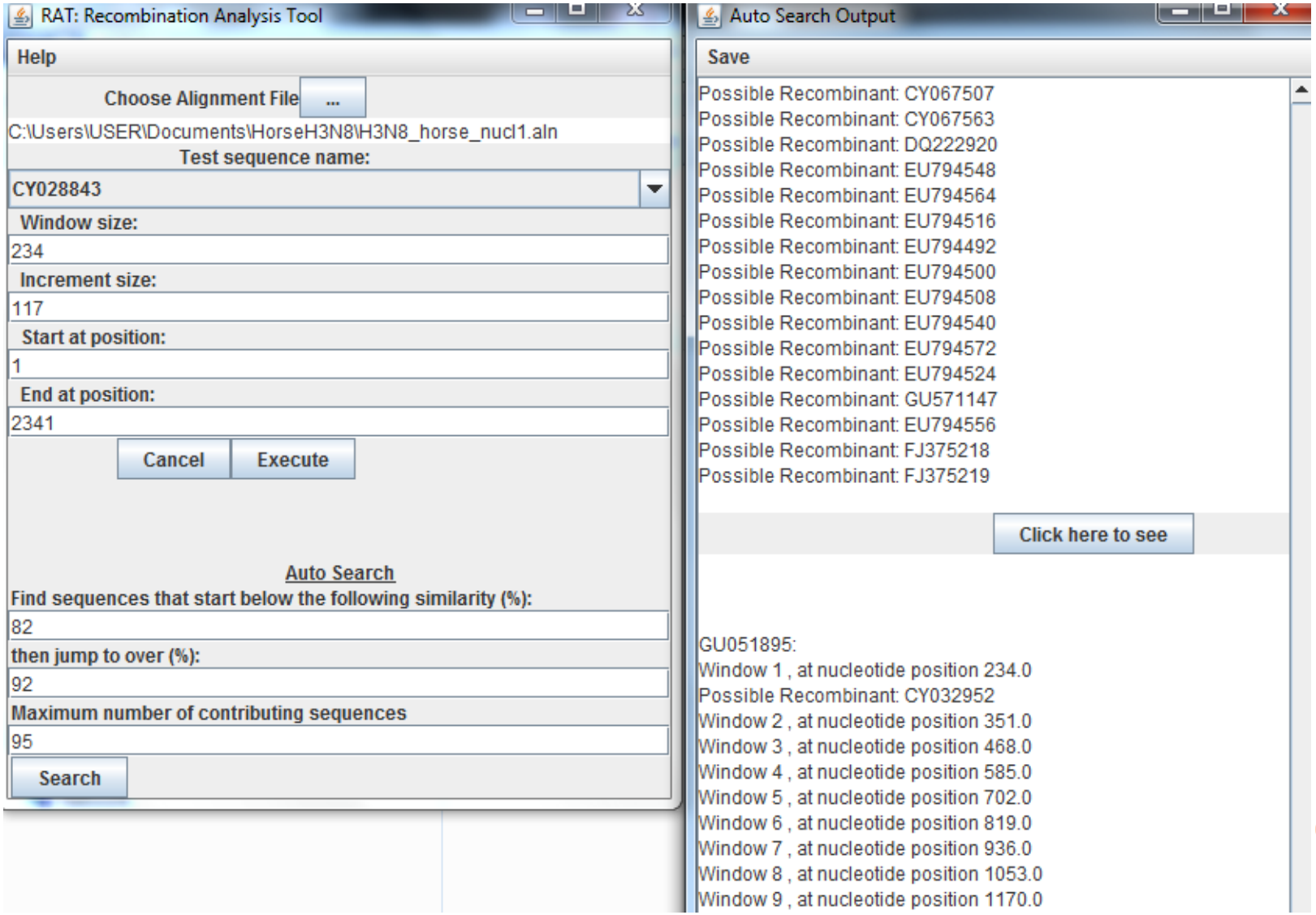
- Автоматичното търсене сканира секвенции, които отговарят на параметрите, представени в три полета. Разглеждат се всички секвенции в съответния ред и се сравняват с всички други секвенции.
- *Поле 1:* Търсят се участъци, които започват със сходство, зададени в %.
- *Поле 2:* Търсят се участъци с прилика над зададен %.
- *Поле 3:* Определя се максималният брой секвенции, които да се анализират за рекомбинация. Може да се ограничи броят на секвенциите, които да се включат като полезни в рекомбинацията. По подразбиране това е броят на всички секвенции (без ограничение).


RAT: Recombination Analysis Tool
Help
Choose Alignment File
...
Test sequence name:
Window size:
Increment siz
Start at positi
1
End at positio
Cancel
Find sequences
82
then jump to o
92
Maximum num
Search
Open
Look In:
RAT_Project
java_version
H1N1_8_3996.aln
H1N1swine_segm6.aln
File Name:
H1N1_8_3996.aln
Files of Type:
Clustal
Open
Cancel

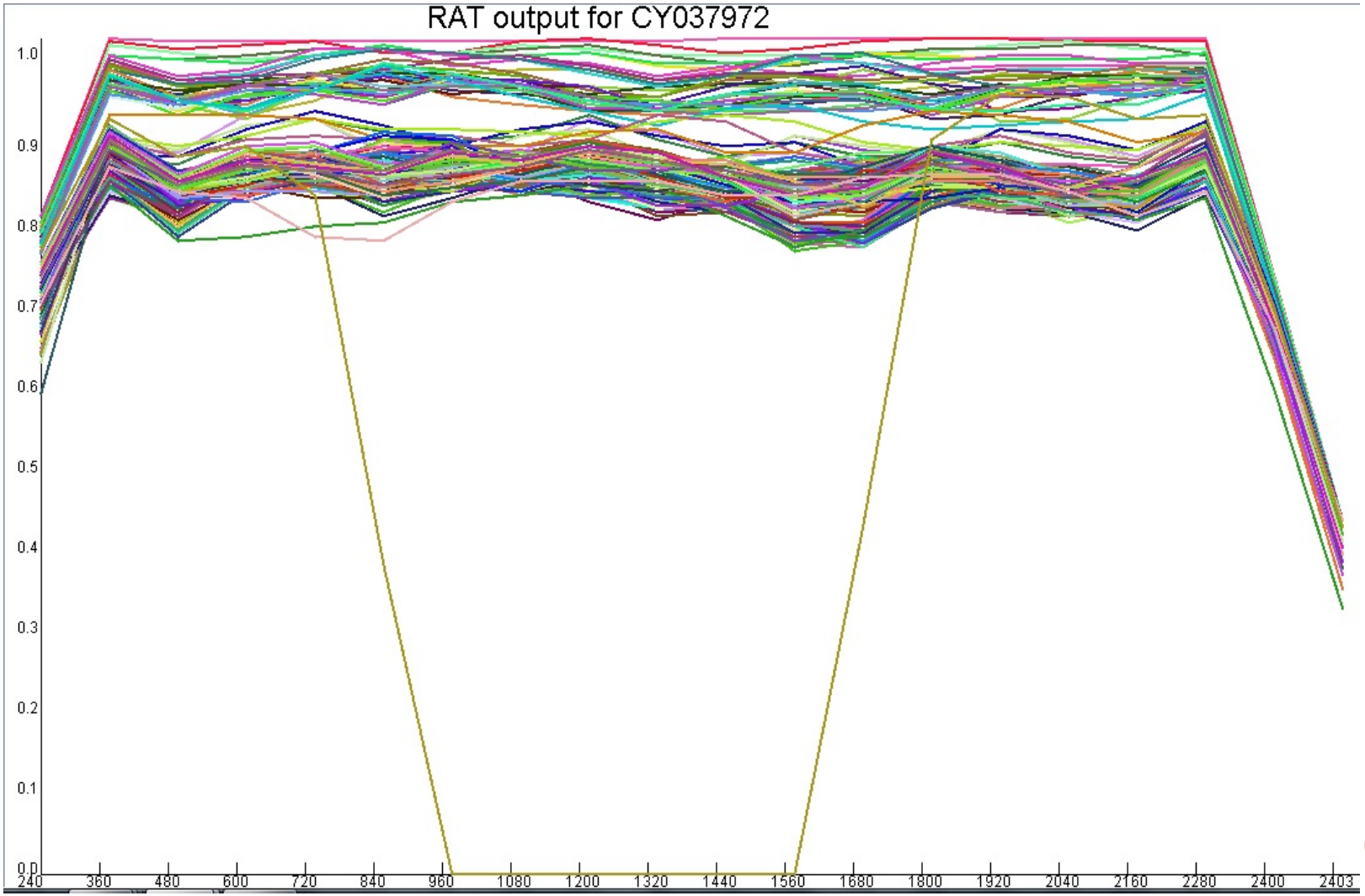
Експерименталните резултати за изследване на горещи точки на рекомбинация на грипен вирус A/H1N1, сегмент 1 (3780 секвенции от различни изолати)



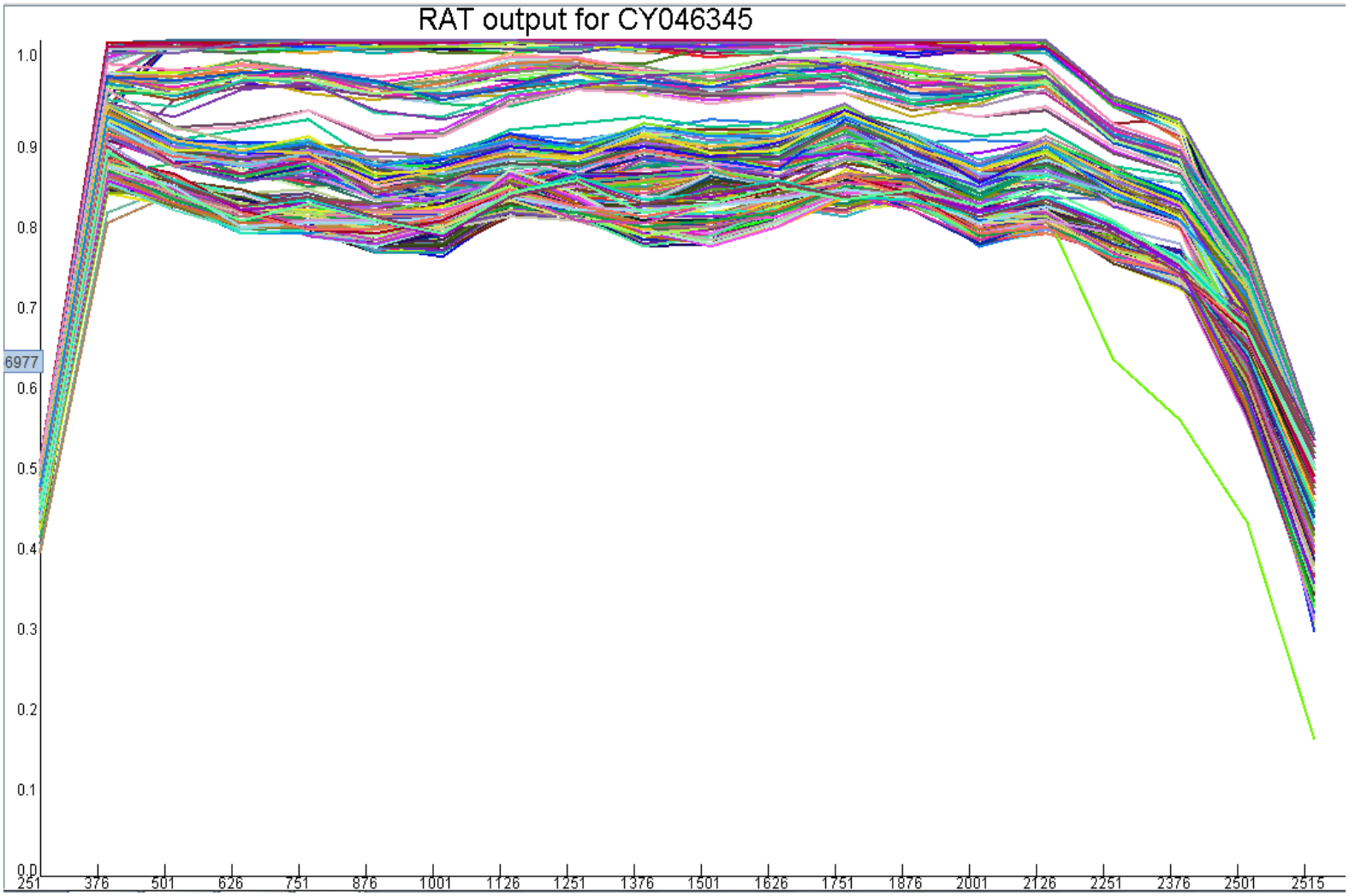
Изследване на горещи точки на рекомбинация грипен вирус тип A/H3N8



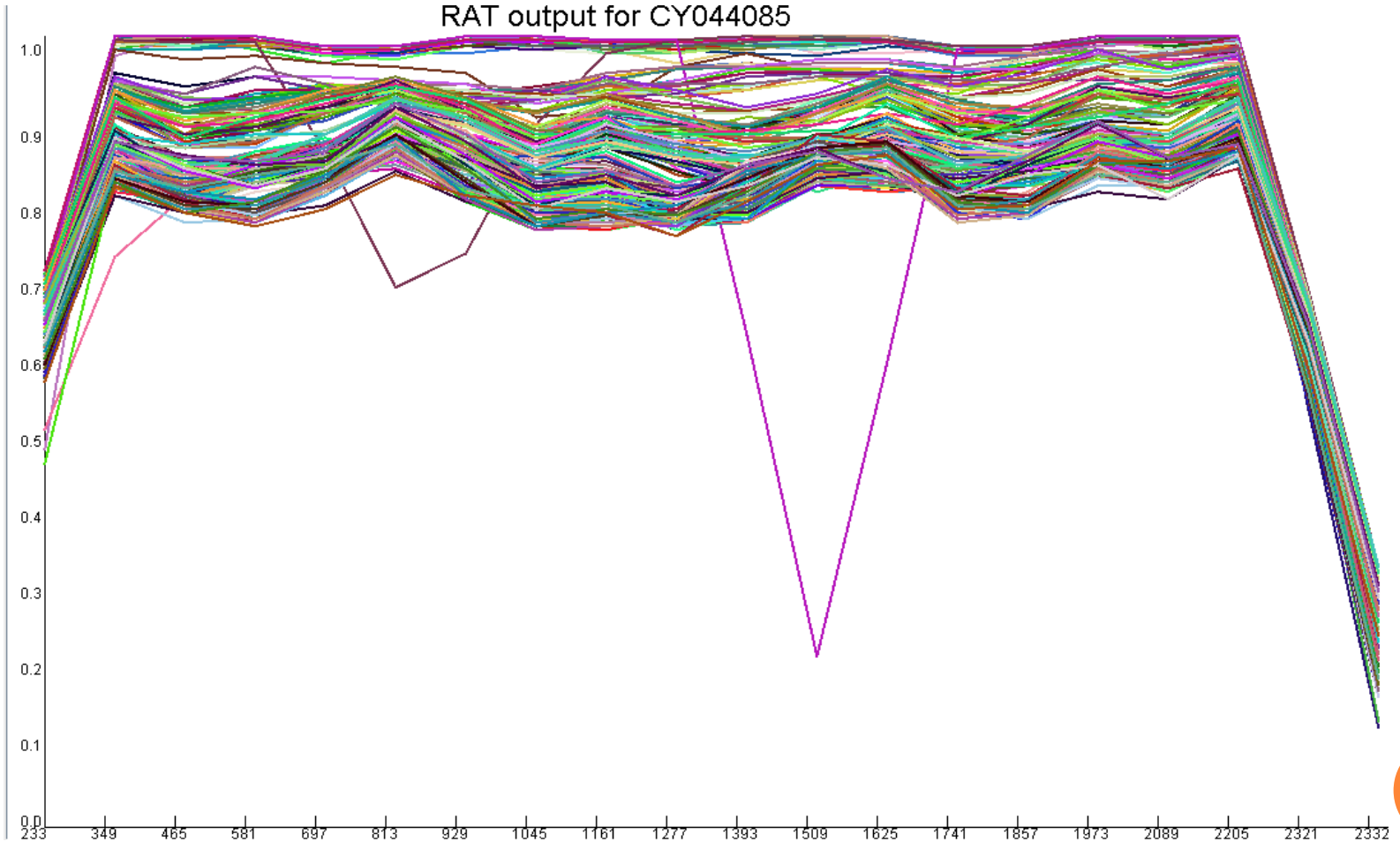
Точки на рекомбинация при грипен вирус A/H1N1, сегмент 1



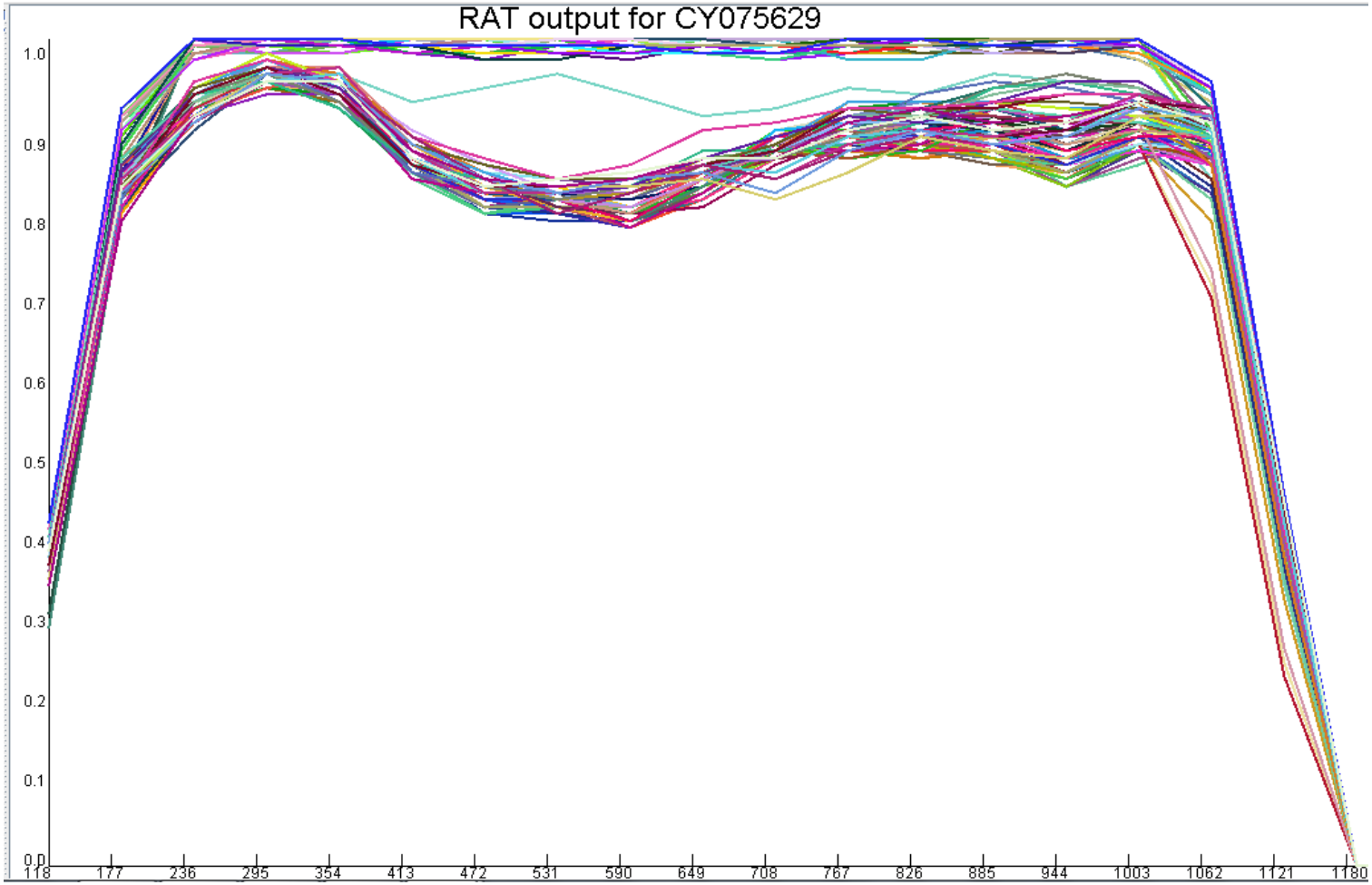
Точки на рекомбинация при грипен вирус A/H1N1, сегмент 2



Точки на рекомбинация при грипен вирус A/H1N1, сегмент 3



Точки на рекомбинация при грипен вирус A/H1N1, сегмент 7



Точки на рекомбинация при грипен вирус A/H1N1, сегмент 8