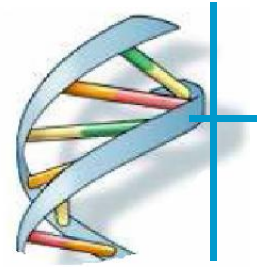
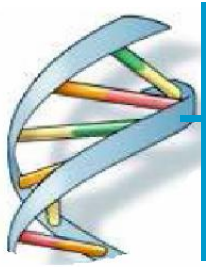


Паралелни изчисления за биоинформатика и изчислителна биология

гл. ас. Веска Ганчева

vgan@tu-sofia.bg



Паралелни изчисления за биоинформатика и изчислителна биология

Тема: Имплементиране на паралелен алгоритъм mpiBLAST за търсене на фрагмент от нуклеотидни последователности в човешкия геном

Алгоритъма за търсене на биологични секвенции (последователности) BLAST (Basic Local Alignment Search Tool) е разработен от NCBI (National Center for Biotechnology Information). mpiBLAST е паралелна имплементация на NCBI BLAST с отворен код и се използва за търсене на разпределени компютърни системи, например клъстери чрез MPI (Message Passing Interface) комуникации, като така оползотворява всички налични ресурси за разлика от стандартния BLAST, който може да се възползва най-много от мултипроцесор със споделена памет (SMPs).

Описание на BLAST

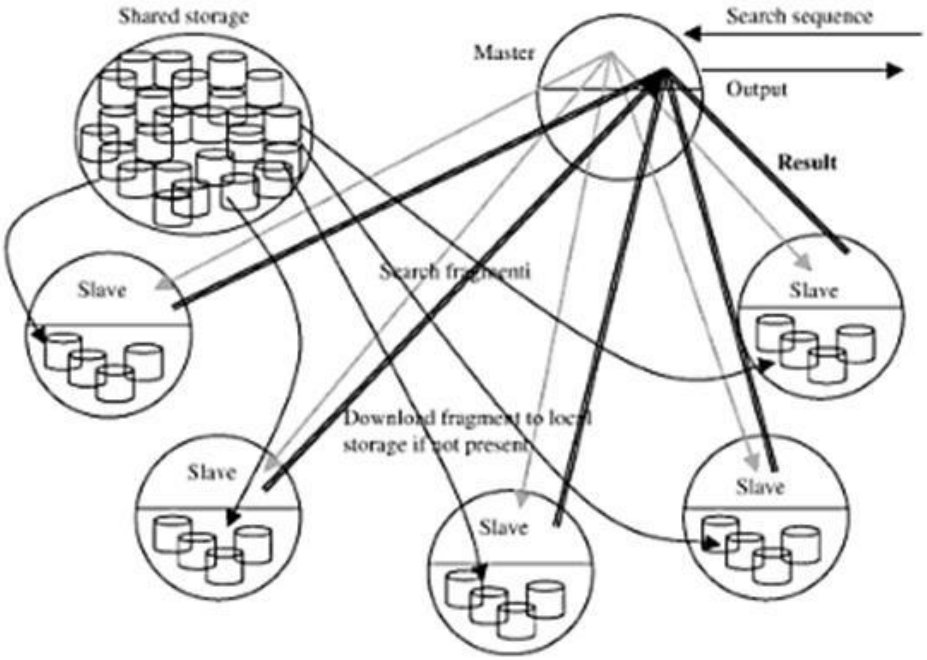
BLAST търси секвенция от нуклеотиди (ДНК) или пептиди (амино киселини) в база данни от нуклеотидни или пептидни секвенции. Могат да се правят сравнения между пептидни и нуклеотидни последователности. BLAST позволява да се сравняват всички възможни комбинации от типове на заявката и данните за търсене, като превежда съответните типове по време на изпълнение. Таблицата изброява имената на видовете търсения спрямо комбинациите от типа на търсената секвенция и типа на данните.

Search Name	Query Type	Database Type	Translation
blastn	Nucleotide	Nucleotide	None
tblastn	Peptide	Nucleotide	Database
blastx	Nucleotide	Peptide	Query
blastp	Peptide	Peptide	None
tblastx	Nucleotide	Nucleotide	Query and Database

Паралелни изчисления за биоинформатика и изчислителна биология

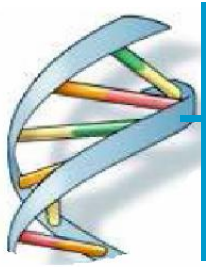
Тема: Имплементиране на паралелен алгоритъм mpiBLAST за търсене на фрагмент от нуклеотидни последователности в човешкия геном

mpiBLAST се състои от две програми mpiBLAST и mpiformatdb, които паралелно стартират BLAST задачи на клъстер от компютри с инсталиран MPI. mpiBLAST разделя данните на всички възли на клъстера. Декомпозицията на данните се прави офлайн (преди да е започнало търсенето). Фрагментите на базата данни се намират в споделена папка



Програмата може да се стартира на *n* възлов клъстер, търси секвенция в човешкия геном и при откриването на такива извежда резултатите във файл *blastresults.txt*, а статистиките в .clog2 файл, който се анализира с Jumpshot.

За експериментите се използват две реални бази данни с ДНК секвенции, с различни части от нея. Експериментите се извършват на 8, 16 и 32 ядра. Търсят се сходни участъци в двете бази данни с дадения ген. Сходен участък е участък, който съдържа максимално количество от секвенциите на гена. Използваните данни са от базата данни GenBank разработвана от NCBI (National Center for Biotechnology Information) (ftp://ftp.ncbi.nih.gov/blast/db/FASTA/est_human.gz) съвместно с EMBL (European Molecular Biology Laboratory).



Паралелни изчисления за биоинформатика и изчислителна биология

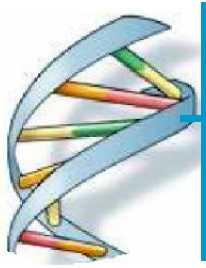
Тема: Имплементиране на паралелен алгоритъм mpiBLAST за търсене на фрагмент от нуклеотидни последователности в човешкия геном

Диаграма на Гант

Диаграмата на Гант е кръстена на откривателя и Хенри Гант. Тя е вид диаграма с дейност на върха и е най-предпочитаният начин за разглеждане на един проект, защото визуално представя и времетраенето на отделните задачи. Дейностите в диаграмата на Гант са наредени хоризонтално - за всяка дейност е заделен един ред, а всяка вертикална линия представлява момент във времето. Това става като всяка задача е правоъгълник, чиято дължина е пропорционална на времетраенето му. В последствие в диаграмите на Гант се налагат и специфични означения за дейности (своеобразни скоби), които са съставни, както и за ключови дати (малки ромбчета).

Времева диаграма(Хистограма)

Една опростена вариация на диаграмата на Гант е широко известна като *времева диаграма*. В нея няма стрелки, но е запазен табличният вид на представянето на задачите във времето. Така отново всеки ред е запазен за една задача, а времето напредва отляво-надясно. Редовете са запълнени само на тези места, където задачата е активна в определения момент.



Паралелни изчисления за биоинформатика и изчислителна биология

Тема: Имплементиране на паралелен алгоритъм mpiBLAST за търсене на фрагмент от нуклеотидни последователности в човешкия геном

За да се извършват търсения с mpiBLAST данните трябва да се форматират и сегментират използвайки командата *mpiformatdb.*

1 Форматиране на данните

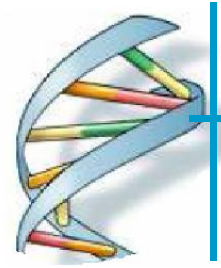
mpiformatdb –N 30 –i est_human

Горната команда ще форматира данните в 30 фрагмента, идеално за 32 работни възела. *N* е броят на фрагментите, а с флагът *i* се задават данните, които ще се форматират. Mpiformatdb създава форматираните фрагменти в споделената директория за данни (Shared Storage Directory). Базата данни използвана в този пример е реална човешка ДНК.

2 Извършване на заявка

mpiexec -n 30 mpiblast -d ./est_human -p blastn -i ./AlcoholGene.fa -o blastresults.txt

Горната команда ще стартира mpiBLAST търсене на 30 ядра, като ще търси секвенции от *AlcoholGene.fa* в *est_human.* Параметърът –p blastn показва, че ще се търсят само нуклеотидните секвенции в AlcoholGene. Резултатите ще се запишат във файл *blastresults.txt* в директорията, от която се стартира mpiBLAST. За да се постигне по-добра производителност се препоръчва да се стартират два процеса повече, тъй като един от процесите на mpiBLAST е за планиране и комуникация, а другият за входно/изходните операции.



Паралелни изчисления за биоинформатика и изчислителна биология

Тема: Имплементиране на паралелен алгоритъм ClustalW-MPI за множествено подравняване на протеинови секвенции на грипен вирус тип А

Описание на алгоритъма ClustalW

Алгоритъмът се състои от три основни стъпки:

1.Чрез подредба по двойки на последователностите, се изчисляват разликите между двойките. Пресмята се „разтоянието” между всеки две последователности. Методът е следния:

NKL-ON

-MLNON

разстояние = ¼ = 0,25

За всяка двойка се гледат двойките, където няма празно място. Разстоянието се пресмята като се раздели броя на двойките с различие, на общия брой двойки. В случая K и M се различават на 2-ра позиция, L=L , O=O и N=N, т.е. имаме една разлика от общо 4, което е ¼ или 0,25.

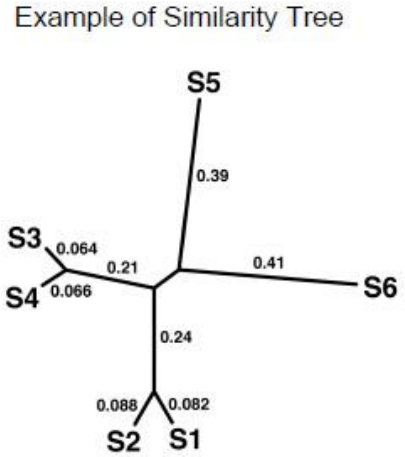
След изчисляване на „разстоянията” между всички двойки последователности, те се записват в матрица. Ако например има 6 последователности, матрицата ще изглежда така:

матрица на разстоянията

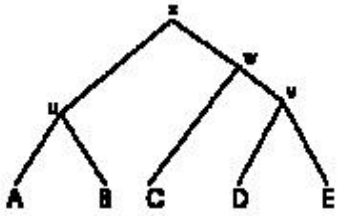
Seq.	S1	S2	S3	S4	S5	S6
S1	-					
S2	.17	-				
S3	.59	.60	-			
S4	.59	.59	.13	-		
S5	.77	.77	.75	.75	-	
S6	.81	.82	.73	.74	.80	-

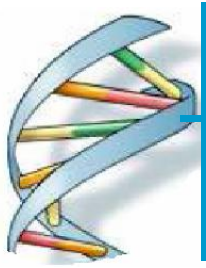
Тема: Имплементиране на паралелен алгоритъм ClustalW-MPI за множествено подравняване на протеинови секвенции на грипен вирус тип А

2.На базата та тази матрица се построява дърво на подобието. ClustalW използва матрицата и алгоритъма за свързването на съседни последователности, за да изгради дървото




```
A: RPCVCP___VLRQAAQ__QVLQRQIIQGPQQLRRLF_AA
B: RPCACP___VLRQVVQ__QALQRQIIQGPQQLRRLF_AA
C: KPCLCPKQAAVKQAAH__QQLYQGQLQGPKQVRRAFRLL
D: KPCVCPRQLVLRQAAHLAQQLYQGQ____RQVRRAF_VA
E: KPCVCPRQLVLRQAAH__QQLYQGQ____RQVRRLF_AA
```





Тема: Имплементиране на паралелен алгоритъм ClustalW-MPI за множествено подравняване на протеинови секвенции на грипен вирус тип А

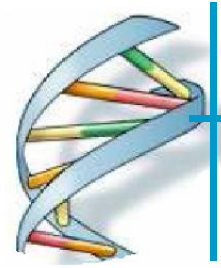
3.Комбинират се сравнените резултати, като се започне от най-близко свързаните групи и се завърши с най-отдалечените, като се върви от корена към дървото.

В синтезиран вид алгоритъма изглежда така:
Вход: Набор S от n на брой последователности (секвенции)
Изход: Множествено сравнен набор S

Сравнение на двойки: Изчисляват се по двойки всички последователности и резултатите се записват в матрица на подобието.

Справочно дърво: Съставя се дърво, което определя реда, по който двойките са подредени и комбинирани с предишни подредби на базата на матрицата на подобие и алгоритъма за свързването на съседни последователности

Множествено сравнение: Подреждат се последователностите на базата на справочното дърво. Започва се с двойките, които имат най-голямо сходство и след това всяка следваща последователност се сравнява с получената преди това подредба.

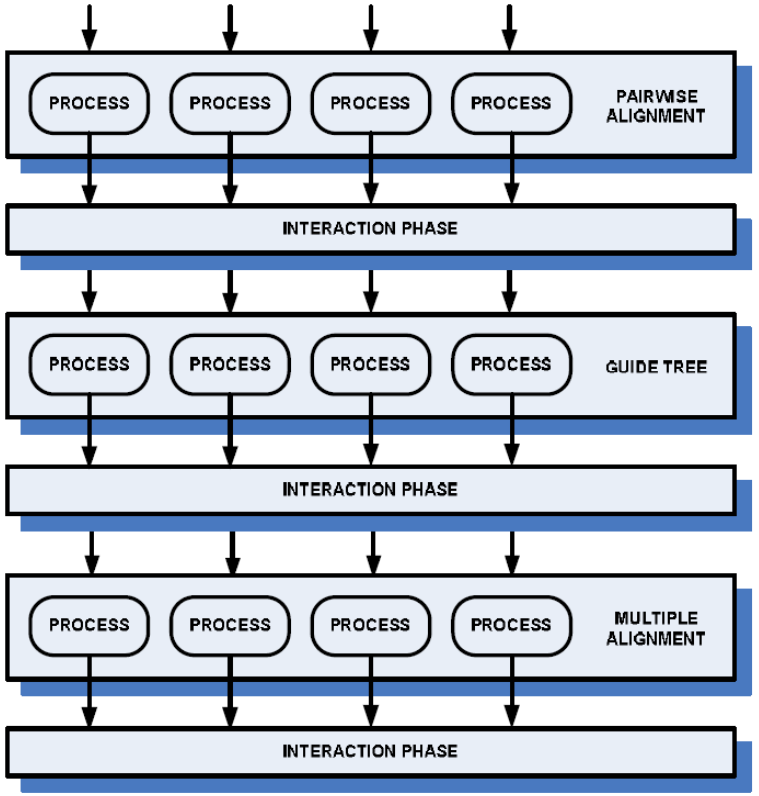


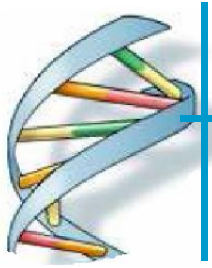
Паралелни изчисления за биоинформатика и изчислителна биология

Тема: Имплементиране на паралелен алгоритъм ClustalW-MPI за множествено подравняване на протеинови секвенции на грипен вирус тип А

Паралелизация на алгоритъма ClustalW

Паралелният изчислителен модел за множествено подравняване, базиран на алгоритъма ClustalW е базиран на фазово-паралелната алгоритмична парадигма и използва декомпозиция по данни. И трите стъпки в процеса на подредба са паралелизирани, за да се повиши бързодействието и да се намали времето за изпълнение. Софтуерът използва библиотека MPI (Message Passing Interface) и работи както на разпределени компютърни клъстери, така и на традиционни паралелни компютри.





Паралелни изчисления за биоинформатика и изчислителна биология

Тема: Имплементиране на паралелен алгоритъм ClustalW-MPI за множествено подравняване на протеинови секвенции на грипен вирус тип А

Фазата на взаимодействието включва разпространение на данни между паралелните процеси, събиране на резултатите и изпращане на нови дялове от данни на паралелни процеси.
Приема се, че разполагаме с p на брой процесора, като всеки от тях означаваме с P0, P1, P2,..., Pp-1
Вход: набор от последователности, (S) Изход: справочно дърво, Ts; Сравнен набор от последователност As.

1) Процесорът P0 прочита набора от последователности S. След това разпределя двойките последователности между процесорите, като се имат в предвид броя процесори p и броя последователности, n=|S|.

1.1) Процесорът P0 изпраща поднабора от последователности към останалите процесори, Pi

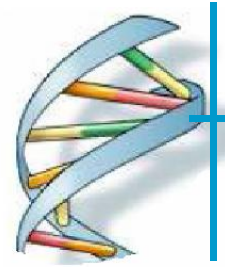
1.2) Всеки процесор Pi извършва сравнение по двойки на дадения му поднабор от последователности

1.3) Всеки процесор Pi изпраща полученото сравнение на P0.

2) Процесорът P0 построява справочното дърво, Ts, използвайки получените резултати от другите процесори.

3) Процесорът P0 анализира Ts, за да идентифицира двойките, които могат да бъдат изследвани самостоятелно в следващата стъпка. Тези последователности отново се изпращат на осталите процесори.

3.1) Процесорът P0 събира събира отново получените резултати и използвайки Ts, последователно завършва паралелното сравнение As.



Паралелни изчисления за биоинформатика и изчислителна биология

Тема: Имплементиране на паралелен алгоритъм ClustalW-MPI за множествено подравняване на протеинови секвенции на грипен вирус тип А

За експериментите е създадена папка AH1N1 за съхраняване на файлове с нуклеотидни последователности.
За да се стартира програмата, се използва следната команда:

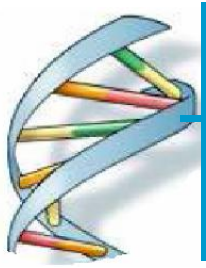
*mpiexec –n **N** ./clustalw-mpi –infile=**file.fa***

Тук **N** е броя на процесорите, които ще участват в изпълнението на подредбата, а **file.fa** е файла с нуклеотидните последователности. В конкретен случай командата ще изглежда така:

mpiexec -n 32 ./clustalw-mpi -infile=./AH1N1/H1N1_Protein_NA_1228.fa

Файловете, които обработва програмата, трябва да бъдат във FASTA формат (.fa), като самия файл има някои особености. Затова е хубаво да се ползват външни източници за автоматично генериране на подобни файлове.

Възможно е, въз основа на зададени критерии, да се генерира файл във FASTA формат, който след това да се подложи на обработка от ClustalW-MPI.
След успешното приключване на изпълнението на програмата, в папката, където се намира файла с последователностите, се записват два нови файла със същото име. Те са с други разширения - **.aln** и **.dnd** .ALN файла представлява подредените последователности.
Статистиките се записват в .clog2 файл, който може да се отвори с Jumpshot, за да се видят профилите.

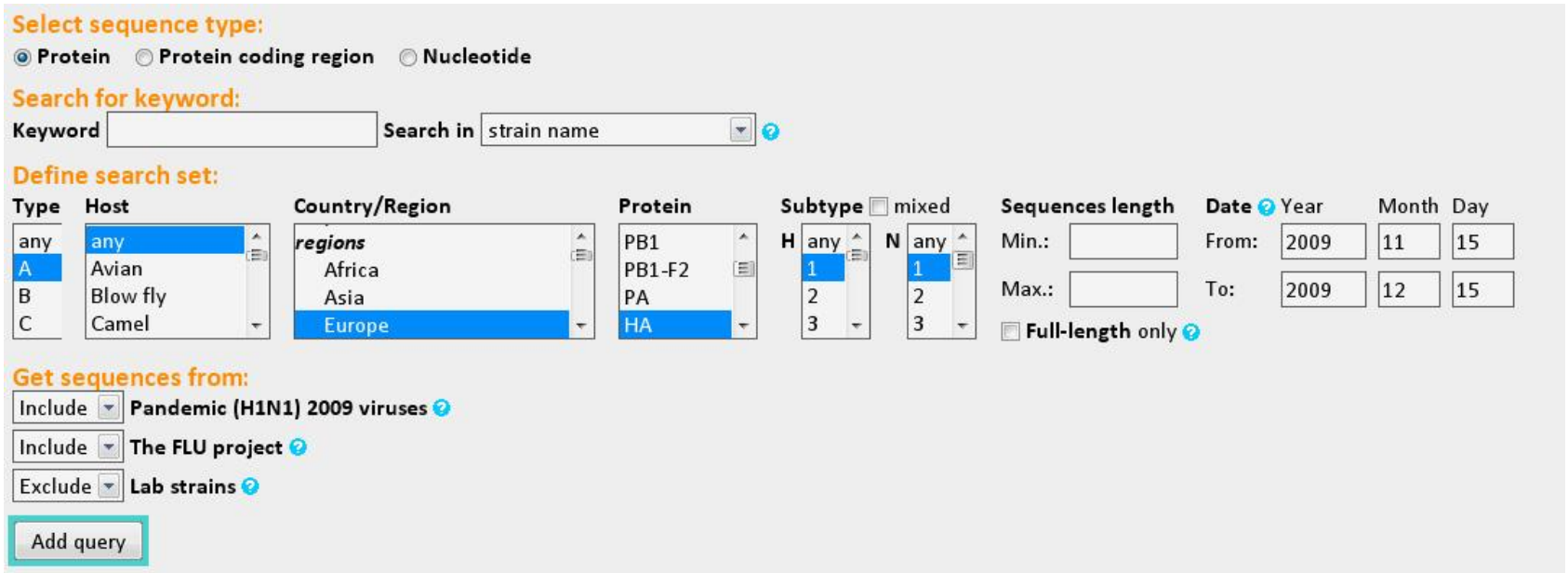


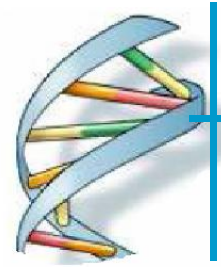
Паралелни изчисления за биоинформатика и изчислителна биология

Тема: Имплементиране на паралелен алгоритъм ClustalW-MPI за множествено подравняване на протеинови секвенции на грипен вирус тип А

Експерименти:

Използвано е множество от реални данни, снети от NCBI Influenza Virus Resource –> Database. На базата на специфицираните критерии като подтип, гостоприемник и др. са извлечени различни множества от данни за протеинови секвенции и са обработени паралелно.





Паралелни изчисления за биоинформатика и изчислителна биология

Тема: Имплементиране на паралелен алгоритъм ClustalW-MPI за множествено подравняване на протеинови секвенции на грипен вирус тип А

Number of sequences/ Time [s]	Serial execution	Parallel execution 8 cores	Parallel execution 16 cores	Parallel execution 32 cores
35	16	4.45	3.91	4.08
140	198	25.98	16.15	12.46
280	717	92	55	36
820	365	69	51	34

Number of sequences / Speedup	Parallel execution 8 cores	Parallel execution 16 cores	Parallel execution 32 cores
35	3.6	4.1	3.9
140	7.6	12.3	15.9
280	7.8	14.3	21.9
820	5.3	7.15	17.4

Number of sequences / Efficiency	Parallel execution 8 cores	Parallel execution 16 cores	Parallel execution 32 cores
35	0.45	0.26	0.12
140	0.95	0.77	0.5
280	0.97	0.89	0.68
820	0.66	0.45	0.54

Паралелни изчисления за биоинформатика и изчислителна биология

Тема: Имплементиране на паралелен алгоритъм ClustalW-MPI за множествено подравняване на протеинови секвенции на грипен вирус тип А

AN1H1 Segment/ Time [s]	Number of seq	Size of seq	PA	Guide Tree	MA	Total time in sec
PB1-F2	820	57	10	24	16	50
NS2	1955	121	195	343	123	661
NP	2107	498	3074	430	307	3811
NS1	2030	230	614	392	158	1164
PB1	1906	757	5676	311	536	6523
PB2	1894	759	5657	305	320	6282
NA	3404	470	7145	1854	929	9928
HA	3057	565	8351	1343	540	10234