ВЕСЕЛИН БАЕВ | ЕЛЕНА АПОСТОЛОВА
ЕВЕЛИНА ДАСКАЛОВА | ГЕОРГИ МИНКОВ

РЪКОВОДСТВО ПО БИОИНФОРМАТИКА

Университетско издателство „Паисий Хилендарски”

Ръководство по биоинформатика
Първо електронно издание (v.1.0), 2013

ISBN: 978-954-423-835-3

университетско издателство
Паисий Хилендарски

2013

ISBN 978-954-423-835-3
90000 >
9 789544 238353

СЪДЪРЖАНИЕ

СРАВНЯВАНЕ НА ДВОЙКА СЕКВЕНЦИИ (PAIRWISE ALIGNMENT) 64

СРАВНЯВАНЕ НА МНОЖЕСТВО СЕКВЕНЦИИ (MULTIPLE ALIGNMENT) 98

Използване на бази данни за нуклеотидни секвенции

В тази глава:

- Работа с GenBank;
- Търсене на информация за даден ген;
- Работа с идентификатори на секвенцията;
- Работа със бази данни.

“Тайната на успеха е да знаеш нещо, което никой друг не знае”.

Аристотел Онасис (1902-1975)

Базите данни (БД) за нуклеотидни секвенции са един от най-важните инструменти в биоинформатиката, понеже ни предлагат уникален поглед върху наличните молекулярно-биологични данни. Те дават възможност да отговорим на съвременните биологични въпроси чрез анализиране на секвенции, които в някои случаи са били открити още преди 30 години. Тези БД са единствения начин за лесно и интуитивно индексиране на всички данни въпреки бързо-растящия им брой през последните няколко години.

Първите БД са били създадени като един вид “музей” за секвенции, където са били запазвани в първична форма, точно както са открити, интерпретирани и публикувани от техния истински автор. Тази историческа перспектива присъства и в GeneBank – водещото хранилище за нуклеотидни секвенции – поддържано от консорциума между U.S.National Center for Biotechnology Information (NCBI), European Molecular Biology Laboratory (EMBL) и DNA Data Bank of Japan (DDBJ). В тази глава ще ви покажем как да използвате GeneBank и да разчитате форматите, в които са записани всички видове секвенции.

Първоначално, базите данни са били създадени просто като хранилища за информация без никакви взаимовръзки, но след време е възникнала необходимостта за свързване на информация от различен тип, като например генна информация, библиография, таксономия и т.н. По тази причина, второто

поколение БД за нуклеотидни секвенции били създадени като генно-центрирани БД, където цялата информация за секвенцията е свързана с даден ген и е достъпна от едно място. За да ви помогнем да разгадаете мистериите на по-старомодния GeneBank, ние ще ви покажем как да използвате Entrez Gene на NCBI, като добър пример за генно-центрирана БД.

В момента, ние имаме информация не само за индивидуални генни последователности, но също така и за относителното им разположение, ориентацията им, както и за наличието или отсъствието на биохимични функции. Използвайки тази по-глобална информация, учените трябвало да създадат единна система за управление на информацията, към която може да свържат специализирани колекции от секвенции. В тази глава, вие ще се запознаете с някои от най-добрите геномни, ресурси отнасящи се за вируси,бактерии и еукариоти, както и за човека.

Преди да започнете да ползвате тази система за управление на информация по-задълбочено, добре е да прочетете и следващата секция. Там ще резюмираме основните характеристики на гените и геномите, и ще ви въведем в лексиката, която ви трябва, за да разчитате записите в GeneBank.

Четене на генни и геномни записи

Наистина, нуклеотидните последователности са универсални, но структурата на гените при прокариотните и еукариотните организми е доста различна. Трябва да знаете основните принципи на изграждане на двата типа геномни и генни записи, да си създадете “усещане” за GeneBank форматите. Запомнете, че е нужно да дефинираме гена като геномен сегмент, съдържащ цялата необходима информация като нуклеотидна секвенция, водеща до успешната експресия на РНК и протеин. Това важи както за кодиращата, така и за не-кодиращата област на гена.

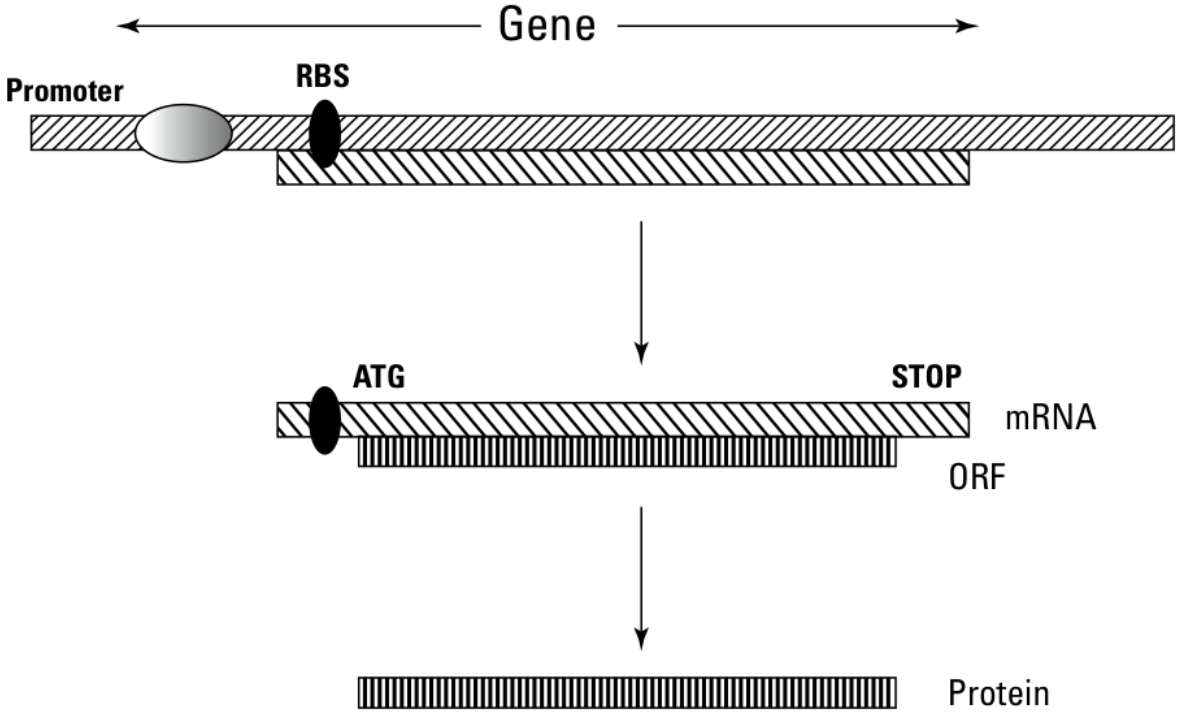
Прокариоти: малки микроорганизми, прости гени

Трите най-основни класа живи организми са прокариоти (най-често бактерии), Археа (archeae, бактерио-подобни организми, живеещи в екстремални условия) и еукариоти.

От биоинформатична гледна точка, прокариотите и Археа (archeae) са в голяма степен еднакви. С малки изключения – далеч извън сферата на тази книга – те имат следните характеристики както следва:

- Те са микроскопични организми;
- Геномът им е изграден от единична, циклична ДНК молекула;

- Размерът на геномът им е в порядъка на няколко милиона базови двойки (0.6-8Mbp);
- Генетичната гъстота (плътност) на броя гени към базовите двойки в генома е приблизително един ген на 1000 бд;
- Геномът им е беден на регулаторни участъци (70% от секвенциите кодират протеини);
- Гените не се застъпват;
- Гените се транскрибират в иРНК точно след промоторния участък;
- Един ген представлява една ОРЧ и кодира един протеин;
- Протеиновите секвенции произхождат от транслиране на най-дългите рамки на четене (от ATG до стоп кодона), включващи секвенцията на генния транскрипт.



Фигура 1-1. Колинеарни взаимоотношения между бактериалната геномна (генна) секвенция, транскриптът (иРНК), отворената рамка на четене (ORF) и крайния протеин.

Съответно, записите в БД, които описват кодиращата прокариотна секвенция трябва да съдържат 3 важни особености:

- Координатите на няколко промоторни елементи;

- Координатите на сайтовете за свързване на рибозомата (RBS);
- Координатите на ORF.

Не всички гени кодират протеини; за някои от тях, тази функцията се изпълнява от транскрибираната РНК молекула. Това включва тРНК, рРНК и още няколко класа РНК, като малките РНК, сиРНК и миРНК и др.

Еукариоти: големи геноми, комплексни гени

Еукариотите обхващат широко разнообразие от организми от микроскопичните дрожди и гъби до растения и животни. Те имат някои характерни особености, които са общи за всички:

- Техният геном се състои от множество линейни молекули ДНК наречени хромозоми;
- Размера на генома им (10- 670 000 милиона бази) – особено за животни и растения – е много по-голям от този на прокариотите;
- Генната им гъстота е много по-ниска от тази при прокариотите (1 човешки ген на 100 000 бази);
- Геномът им съдържа много регулаторни части, известни преди време като “junk DNA” (по- малко от 5 процента от човешкия геном кодира протеини);
- Гените на срещуположните вериги може да се препокриват, както частично, така и изцяло;
- Гените им се транскрибират точно след контролният регион, наречен промотор, но елементи, локализирани надалече могат да имат силно въздействие върху този процес;
- Генните им секвенции не са колинеарни с крайната иРНК и протеиновата секвенция. Само малки фрагменти (екзони) се запазват в зрялата иРНК, която кодира крайния белтък;
- Гените им често проявяват повече от една иРНК (или протеинова) форма – алтернативен сплайсинг.

Гените на висшите еукариоти (животните) могат да обхванат милиони базови двойки – човешкият ген за dystrophin например е 2.2 милиона бази на дължина, поради което разбирането на биологичните БД може да отнеме доста време. Освен това, в много от БД има хипервръзки към други БД. В остатъка от тази глава ще ви покажем няколко примера, така че след малко разучаване, вие сами ще можете да разчитате и най-сложните GeneBank записи.

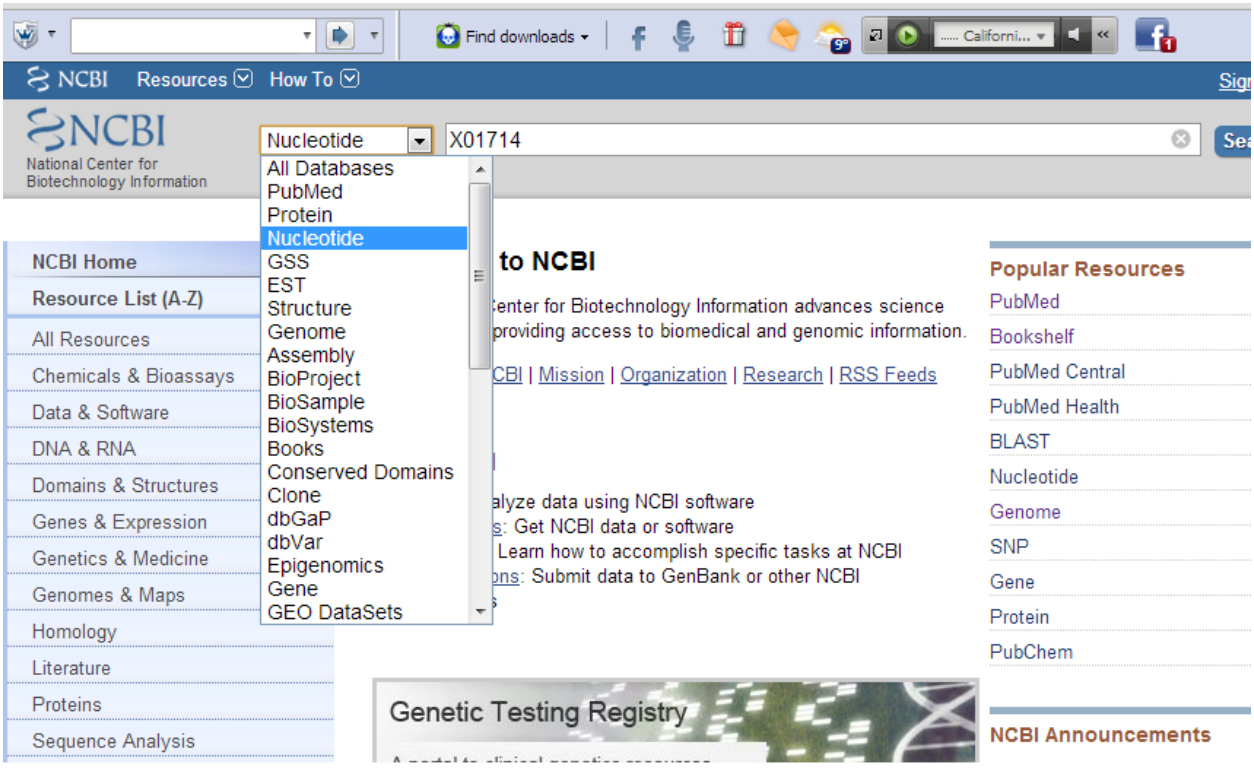
Използване на GeneBank

Записите в БД, които съответстват на бактериалните гени, са относително лесни за четене и разбиране. В тази секция ще ви преведем през записа на dUTPase ген на *E. coli* стъпка по стъпка (GeneBank ID: X01714).

Тълкуване на прокариотни GeneBank записи

Първо, трябва да извлечем нашият GeneBank запис както следва:

- Въведете във вашия браузър www.ncbi.nih.gov/
  Показва се началната страница на NCBI PubMed.
- От падащото меню за търсене изберете Nucleotide (както е показано на Фигура 1-2).
- Въведете GeneBank ID „X01714" в полето за търсене и натиснете „Search".
- X01714 се появява в подразбиращия се GeneBank формат както е показано на Фигура 1-3.
- Изберете опцията „File" от падащото меню „Send" за да генерирате файл за ID X01714.
- Браузърът ви ще предложи да запаметите файл на вашия диск. Направете го.

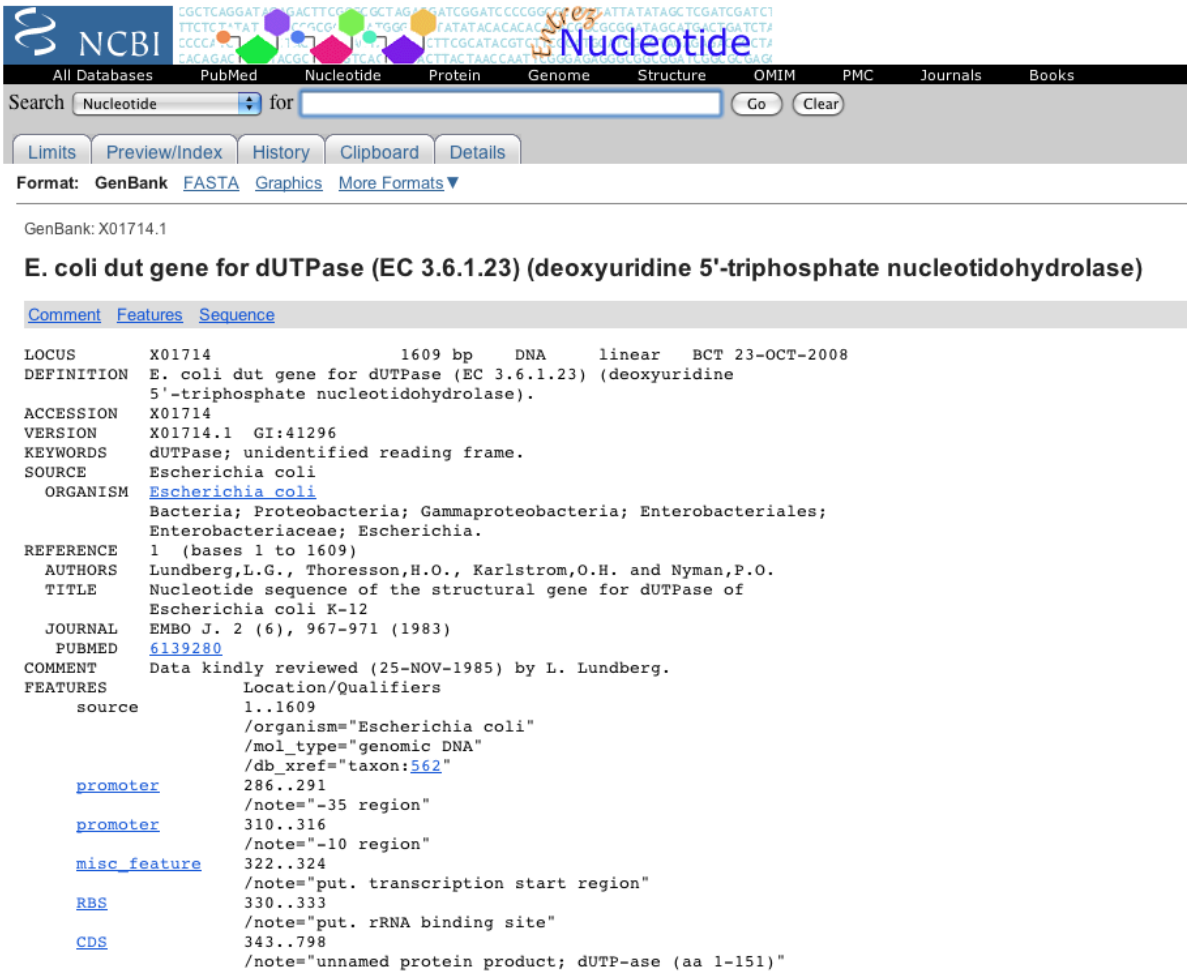


Фигура 1-2. Главна страница на NCBI.

GeneBank запис – идентификатори на секвенцията

Записаният текст от предишните стъпки е разделен на секции с ключови думи, които ни въвеждат във всяка секция. Типични ключови думи (идентификатори на секвенцията) са LOCUS, DEFINITION, ACCESSION, VERSION, KEYWORDS и SOURCE. Нека обясним тези ключови думи по-подробно:

- **LOCUS** дава информация за locus името, размера на нуклеотидната секвенция в базови двойки, вид на молекулата (тук тя е ДНК) и топологията и: линейна или циклична (в този случай-линейна).
- **DEFINITION** - осигурява ни кратко описание на гена, който съответства на зададената секвенция. Тук той е dUTPase ген от *E. coli.*
- **ACCESSION** дава номера за достъп – уникален идентификатор за дадената секвенция. Единствено този идентификатор е уникален и по него може да намерите винаги вашия запис.
- **VERSION** дава версия на секвенцията, както и синоними или стари ID номера.
- **KEYWORDS** - въвежда ни в списък с термини, които характеризират записа. Може да използвате тези ключови думи за търсене в БД.
- **SOURCE** разкрива името на съответния организъм, на който принадлежи секвенцията.
- **ORGANISM** - дава ни по-пълна идентификация на организма, допълнена с неговата таксономична класификация.
  Обърнете внимание как думата „ORGANISM" е изместена надясно в колоната си. Това не е грешка; компютърните специалисти наричат това подреждане „назъбване." Изместването на дадена ключова дума показва нейната зависимост спрямо друга ключова дума. Ключови думи с едно и също изместване са на едно и също ниво. Най-лявата позиция на SOURCE ни показва, че записът е независим от предишните пет, докато изместената позиция на ORGANISM ни показва, че той е част от SOURCE.
- **REFERENCE** съдържа следните ключови думи: AUTOR, TITLE, JOURNAL, PUBMED. Тук обикновено се споменава литературната библиография на конкретната секвенция/ген.
- **COMMENT** съдържа свободно-форматиран текст с информация, която не присъства в предишните секции.



Фигура 1-3. Основни данни за секвенцията.

Следващата секция в записа е **FEATURES** table (виж Фигура 1-4). Тя описва точните генни региони и асоциираните биологични характеристики, които са били идентифицирани в нуклеотидната секвенция. Тази секция включва голямо разнообразие от ключови думи, като например:

- **Source** ни дава произхода на конкретни региони в секвенцията. Това е полезно, когато искате да различите клониращи вектори от множество секвенции. Обикновено показва дължината на секвенцията.
- **Promoter** показва точните координати на промоторните елементи. При X01714 35-ти регион е отбелязан от позиция 286 до 291 (286..291) в нуклеотидната секвенция. Promoter съдържа и втори ред, с координатите на друг промоторен елемент, този път 10ти регион на позиции 310 до 316.
- **Misc. feature** ни дава предполагаемата локализация на транскрипционния старт (иРНК синтеза). За X01714 е от 322 до 324.

- **RBS** (Ribosome Binding site) показва позицията на последния upstream елемент. За X01714 е от 330 до 333.
- **CDS** (CoDing Segment) представя комплексна част, описваща отворената рамка на четене на гена, тази част е кодиращата секвенция.
    - Първият ред посочва координатите на ORF от инициращият ATГ до последния нуклеотид на първия стоп кодон TAA (за X01714 е от 343 до 798).
    - Всеки от следващите редове дава името на протеиновия продукт, рамката на четене (тук 343 е първата база от първият кодон), генетичния код и ID номерата за протеиновата секвенция.
    - /translation, последната ключова дума в CDS секцията, ни показва аминокиселинна секвенция, свързаната с кодиращия сегмент.

FEATURES секцията на записа X01714 е типичен пример за добре анотирана бактериална секвенция на добре идентифициран ген. Избрахме този запис, понеже се интересуваме от един ген: dUTPase. Освен това, записът също ни показва допълнителен предполагаем ген, показан от допълнителен RBS елемент и втора CDS секция.

```
FEATURES             Location/Qualifiers
     source          1..1609
                     /organism="Escherichia coli"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:562"
     promoter        286..291
                     /note="-35 region"
     promoter        310..316
                     /note="-10 region"
     misc_feature    322..324
                     /note="put. transcription start region"
     RBS             330..333
                     /note="put. rRNA binding site"
     CDS             343..798
                     /note="unnamed protein product; dUTP-ase (aa 1-151)"
                     /codon_start=1
                     /transl_table=11
                     /protein_id="CAA25859.1"
                     /db_xref="GI:41297"
                     /db_xref="GOA:P06968"
                     /db_xref="InterPro:IPR008180"
                     /db_xref="InterPro:IPR008181"
                     /db_xref="PDB:1DUD"
                     /db_xref="PDB:1DUP"
                     /db_xref="PDB:1EU5"
                     /db_xref="PDB:1EUW"
                     /db_xref="PDB:1RN8"
                     /db_xref="PDB:1RNJ"
                     /db_xref="PDB:1SEH"
                     /db_xref="PDB:1SYL"
                     /db_xref="PDB:2HR6"
                     /db_xref="PDB:2HRM"
                     /db_xref="UniProtKB/Swiss-Prot:P06968"
                     /translation="MKKIDVKILDPRVGKEFPLPTYATSGSAGLDLRACLNDAVELAP
                     GDTTLVPTGLAIHIADPSLAAMMLPRSGLGHKHGIVLGNLVGLIDSDYQGQLMISVWN
                     RGQDSFTIQPGERIAQMIFVPVVQAEFNLVEDFDATDRGEGGFGHSGRQ"
     misc_feature    831..851
                     /note="put.stem-loop structure"
     repeat_region   831..838
                     /note="inverted repeat A"
     repeat_region   844..851
                     /note="inverted repeat A'"
     misc_feature    866..893
                     /note="put. stem-loop structure"
     repeat_region   866..872
                     /note="imp. inverted repeat B"
     repeat_region   888..893
                     /note="imp. inverted repeat B'"
     RBS             889..895
                     /note="pot. rRNA binding site"
     CDS             905..1540
                     /note="unnamed protein product; unidentified reading
```

*Фигура 1-4. Характеристика на секвенцията (**FEATURES**).*

Използване на секция от прокариотните записи

Последната секция на записа X01714 в GeneBank ни дава самата нуклеотидна секвенция. Започва с ключовата дума ORIGIN и завършва със знак за край, представена от две наклонени черти. Всеки ред от нуклеиновата секвенция започва с номера на позицията на първия нуклеотид на този ред. Всеки ред съдържа 60 нуклеотида.

Понеже секцията със секвенцията на GeneBank записа включва числа и нуклеотиди, вие не можете да я използвате директно като входни данни за повечето биоинформатични програми за анализ на секвенции. Първо е необходимо да генерирате секвенция във FASTA формат, както следва:

> Изберете опцията „File" от падащото меню „Send" и от падащото меню „Format" изберете „FASTA".
> Сега нуклеотидната секвенция се появява в подходяща форма за повечето програми за анализиране на секвенции, а именно – във FASTA формат.

Еукариотни GeneBank **записи**

Сега ще продължим с dUTPase гена, понеже се среща и при прокариоти и при еукариоти, както и при човека. Поглеждайки в GeneBank записите описващи човешката версия на dUTPase ген, ясно се вижда повишената сложност на гените при висшите еукариоти, в сравнение с прокариотните гени. Сега ще започнем с GeneBank запис U90223:

- Въведете във вашия браузър www.ncbi.nih.gov/
  Показва се началната страница на NCBI PubMed.
- От падащото меню за търсене изберете Nucleotide (както е показано на Фигура 1-2).
- Въведете GeneBank ID „U90223" в полето за търсене и натиснете „Search".
- Появява се flat- file формат за ID U90223.
- Запаметете своя резултат като изберете „File" от падащото меню „Send", след което изберете „GeneBank" от падащото меню „Format".

Human deoxyuridine triphosphate nucleotidohydrolase precursor mRNA,

Features Sequence

```
LOCUS       HSU90223                960 bp    mRNA    linear   PRI 03-JAN-1998
DEFINITION  Human deoxyuridine triphosphate nucleotidohydrolase precursor mRNA,
            nuclear gene encoding mitochondrial protein, complete cds.
ACCESSION   U90223
VERSION     U90223.1  GI:2735291
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 960)
  AUTHORS   Ladner,R.D. and Caradonna,S.J.
  TITLE     The Human dUTPase Gene Encodes Both Nuclear and Mitochondrial
            Isoforms: Differential Expression of the Isoforms and
            Characterization of a cDNA Encoding the Mitochondrial Species
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 960)
  AUTHORS   Ladner,R.D. and Caradonna,S.J.
  TITLE     Direct Submission
  JOURNAL   Submitted (19-FEB-1997) Dept. of Molecular Biology, Univ. of Med.
            and Dent. of NJ-School of Osteopathic Medicine, 2 Medical Center
            Drive, Stratford, NJ 08084, USA
```

Фигура 1-5. Секвенция U90223.

Въпреки, че е свързан с човешки ген, GeneBank записът U90223 не изглежда по-различно от записа X01714, който описва бактериален хомолог. Главната част на записа е последвана от обща информация с ключови думи: LOCUS, ACCESSION, DEFINITION и VERSION.

KEYWORDS редът, където са търсените термини (като dUTPase), е празен за U90223 записа. За съжаление това не е случайно. Основен проблем на базите данни за секвенции е, че записът може да е непълен. В резултат, търсенето в БД чрез ключови думи обикновено не намира всички съответни секвенции в GeneBank. Това важи и за полетата SOURCE и REFERENCE, които също могат да липсват или да са непълни. Например, U90223 записът показва, че секвенцията идва от неизвестен източник.

Фигура 1-6 описва секцията FEATURES на записа U90223. Ключовата дума CDS посочва кодиращия регион (63 - 821), който съответства на митохондриалната форма на човешката dUTPase. Следва информация, отнасяща се за транслацията на ORF. Ключовата дума sig peptide посочва локуса на митохондриалната таргетна секвенция, а ключовата дума mat peptide се грижи за точните граници на зрелия пептид.

Човешкият запис в GeneBank наистина не е по-сложен от този на бактериалния хомолог понеже наблюдавате записа описващ иРНК, а не геномна информация. Както е показано в следващата секция, оригиналната геномна информация може да бъде доста по-сложна.

Значение на еукариотните геномни записи в GeneBank

Предишните секции описаха GeneBank записа, отговарящ на иРНК секвенцията на човешкия dUTPase ген. Ако искаме да разгледаме GeneBank записа AF018430, който е свързан с генната секвенция (на хромозомата), от която тази иРНК произлиза, трябва да направим следните стъпки:

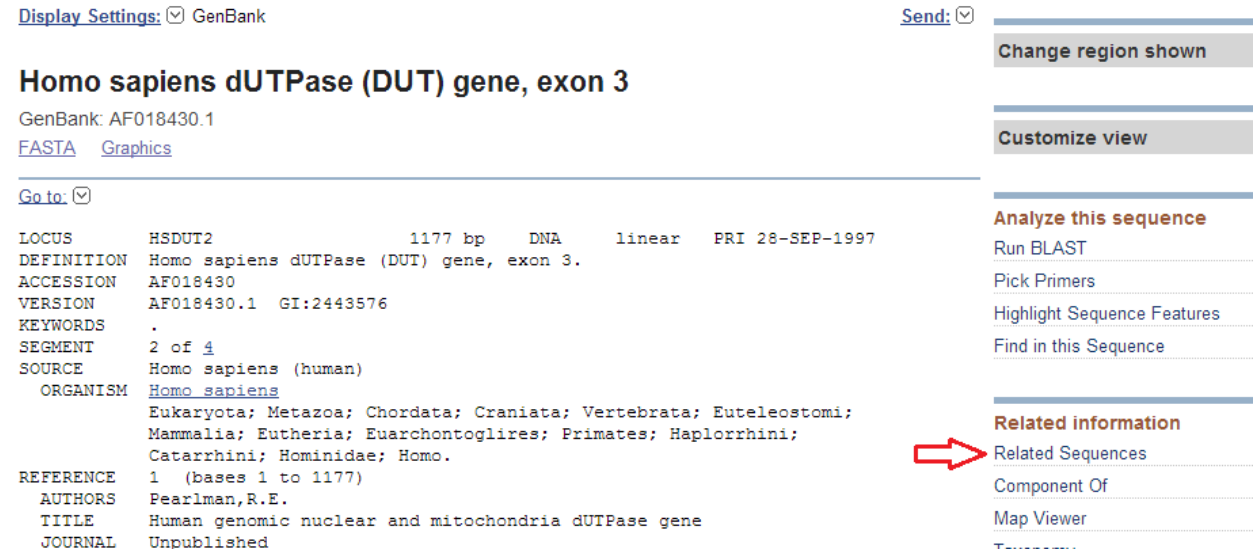
```
FEATURES             Location/Qualifiers
     source          1..960
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
     CDS             63..821
                     /note="mitochondrial dUTPase isoform; DUT-M"
                     /codon_start=1
                     /product="deoxyuridine triphosphate nucleotidohydrolase
                     precursor"
                     /protein_id="AAB94642.1"
                     /db_xref="GI:2735292"
                     /translation="MTPLCPRPALCYHFLTSLLRSAMQNARGTAEGRSRGTLRARPAP
                     RPPAAQHGIPRPLSSAGRLSQGCRGASTVGAAGWKGELPKAGGSPAPGPETPAISPSK
                     RARPAEVGGMQLRFARLSEHATAPTRGSARAAGYDLYSAYDYTIPPMEKAVVKTDIQI
                     ALPSGCYGRVAPRSGLAAKHFIDVGAGVIDEDYRGNVGVVLFNFGKEKFEVKKGDRIA
                     QLICERIFYPEIEEVQALDDTERGSGGFGSTGKN"
     sig_peptide     63..269
                     /note="mitochondrial targeting presequence"
     mat_peptide     270..818
                     /product="deoxyuridine triphosphate nucleotidohydrolase"
```

Фигура 1-6. Характеристика на U90223.

Първо трябва да вземем записа както преди:

- Въведете във вашия браузър www.ncbi.nih.gov/
  Показва се началната страница на NCBI PubMed.
- От падащото меню за търсене изберете Nucleotide.
- Въведете GeneBank ID „AF018430" в полето за търсене и натиснете „Search".
  Появява се AF018430 записа, както е показано на Фигура 1-7.

Вече сте готови да прочетете първия си човешки геномен GeneBank запис. Избрахме нещо просто, с което да започнете, но което все пак съдържа някои специфични особености, срещащи се само при еукариотите.



Фигура 1-7. Секвенция AF018430.

Горната част съдържа обща информация, която използва обичайните ключови думи:

- LOCUS: локус името е HSDUT2. Останалата част от реда ни показва , че секвенцията е линейна ДНК молекула с дължина 1177 бази.
- DEFINITION: този ред ни показва, че името на гена е DUT и уточнява, че записа включва екзон 3 на гена. Това ни напомня, че човешките гени са изградени от екзони и интрони, които имат различна функция.
- ACCESSION, VERSION и KEYWORDS са стандартни.
- SEGMENT: това поле се отнася за мозаичната структура на еукариотните гени. То уточнява, че текущият запис е втори сегмент от общо 4. Трябват ви всички 4 записа, за да реконструирате пълната иРНК секвенция, която можете да използвате като шаблон за синтез на протеин.
- ORGANISM, SOURCE и REFERENCE са стандартни.

Списъкът FEATURES е това, което всъщност прави еукариотния геномен запис специален. Този списък е много по-дълъг от този за прокариотния организъм. Той съдържа следните елементи (виж Фигура 1-8):

- source съдържа /map секция. За AF018430, която показва, че секвенцията е разположена в хромозома 15, по-точно в нейното дълго рамо (q), в хромозомен локус q21.1.
- Ключовата дума gene е комплексна. Нейната цел е да опише реконструкцията на редица иРНК, обхващащи няколко отделни записа.

Концентрирайте се над първата gene order формула, за да разберете как работи тя. Това не е нищо повече от екзон-сплайсингов шаблон (exon splicing recipe), който показва как да реконструираме иРНК секвенцията. Всичко, което ни казва е:

- Вземете нуклеотидите от позиции 1 до 1735 от запис AF018429.
- Прибавете нуклеотидите от позиции 1 до 1177 от текущия запис (AF018430).
- Добавете нуклеотидите от 1 до 45 от запис AF018431.
- Добавете нуклеотидите от 658 до 732 от запис AF018432.

```
FEATURES             Location/Qualifiers
     source          1..1177
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /map="15q15-q21.1"
     gene            order(AF018429.1:<1..1735,1..1177,AF018431.1:1..45,
                     AF018432.1:658..732,AF018432.1:884..954,
                     AF018432.1:1391..>1447)
                     /gene="DUT"
     mRNA            join(AF018429.1:<282..561,AF018429.1:1034..1172,560..651,
                     AF018431.1:1..45,AF018432.1:658..732,AF018432.1:884..954,
                     AF018432.1:1391..>1447)
                     /gene="DUT"
                     /product="dUTPase"
                     /note="alternatively spliced; encodes mitochondrial form
                     of the protein"
     CDS             join(AF018429.1:282..561,AF018429.1:1034..1172,560..651,
                     AF018431.1:1..45,AF018432.1:658..732,AF018432.1:884..954,
                     AF018432.1:1391..1447)
                     /gene="DUT"
                     /note="DUT-M; alternatively spliced; mitochondrial form of
                     the protein; similar to H. sapiens dUTPase encoded by
                     GenBank Accession Number U90224"
                     /codon_start=1
                     /product="dUTPase"
                     /protein_id="AAB71393.1"
                     /db_xref="GI:2443580"
                     /translation="MTPLCPRPALCYHFLTSLLRSAMQNARGTAEGRSRGTLRARPAP
                     RPPAAQHGIPRPLSSAGRLSQGCRGASTVGAAGWKGELPKAGGSPAPGPETPAISPSK
                     RARPAEVGGMQLRFARLSEHATAPTRGSARAAGYDLYSAYDYTIPPMEKAVVKTDIQI
                     ALPSGCYGRVAPRSGLAAKHFIDVGAGVIDEDYRGNVGVVLFNFGKEKFEVKKGDRIA
                     QLICERIFYPEIEEVQALDDTERGSGGFGSTGKN"
     mRNA            join(AF018429.1:<1018..1172,560..651,AF018431.1:1..45,
                     AF018432.1:658..732,AF018432.1:884..954,
                     AF018432.1:1391..>1447)
                     /gene="DUT"
                     /product="dUTPase"
                     /note="alternatively spliced; encodes nuclear form of the
                     protein"
     CDS             join(AF018429.1:1018..1172,560..651,AF018431.1:1..45,
                     AF018432.1:658..732,AF018432.1:884..954,
                     AF018432.1:1391..1447)
```

Фигура 1-8. GenBank запис за номер AF018432.

- Добавете нуклеотидите от 884 до 954 от същия запис (AF018432).
- Добавете нуклеотидите от 1391 до 1447 от същия запис отново.

“<” (знакът по-малко) в началото на формулата показва, че генът всъщност може да започне преди посочената позиция. > (знакът по-голямо) в края на формулата означава че генът всъщност може да продължи извън посочената позиция.

mRNA: записът AF018430 има 2 полета за иРНК, които са интересни, понеже илюстрират пътя, по който GeneBank представя шаблони за алтернативния сплайсинг. Това е общо свойство на генната експресия при еукариотите.

Таблица 3-1. Видове иРНК в AF0118430.

Mrna	***AF018429***	***AF018430***	***AF018431***	***AF018432***
Type 1 (Mitochondria)	282-561 1034-1172	560-651	1-45	658-732 884-954 1391-1447 ->
Type 2 (Nuclear)	<1018- 1172	560-651	1-45	658-732 884-954 1391-1447 ->

Таблица 3-1 ни помага да разберем този запис:

- Има 2 алтернативни иРНК: тип 1 (за митохондрии) и тип 2 (за ядро).
- Ядрената иРНК не съдържа първия екзон (запазен в запис AF018429), въпреки че митохондриалната иРНК го използва.
- Ядрената иРНК използва част от втория екзон (запазен в AF018429) в различна рамка на четене от този в митохондриалното копие/дубликат.
- И двете протеинови секвенции стават идентични при третия екзон. Можете да го видите в двете /translation полета (ETPAISPSK и т.н.). Вижте Фигура 1-9 за секвенцията на ядрената форма на протеина.

Ключовата дума exon показва позицията на единствения екзон, представен в тази секвенция. Това е част от записа AF018430 до края на FEATURES (отнася се за Фигура 1-9, позиции 560- 651). Разгледайте запис AF018432, ако искате да видите множество екзонни полета в единичен запис.

Частта със секвенцията в долната част на записа е форматирана както обикновено (погледнете отново Фигура 1-9).

```
                     /gene="DUT"
                     /note="DUT-N; alternatively spliced; nuclear form of the
                     protein; similar to H. sapiens dUTPase encoded by GenBank
                     Accession Number U90224"
                     /codon_start=1
                     /product="dUTPase"
                     /protein_id="AAB71394.1"
                     /db_xref="GI:2443581"
                     /translation="MPCSEETPAISPSKRARPAEVGGMQLRFARLSEHATAPTRGSAR
                     AAGYDLYSAYDYTIPPMEKAVVKTDIQIALPSGCYGRVAPRSGLAAKHFIDVGAGVID
                     EDYRGNVGVVLFNFGKEKFEVKKGDRIAQLICERIFYPEIEEVQALDDTERGSGGFGS
                     TGKN"
     exon            560..651
                     /gene="DUT"
                     /number=3
ORIGIN
        1 tccctaaatc aacacagatc atgtggagga ataaaatggg gttaatatat gtaaaaccaa
       61 ttaggaaact gtttctgggg caacacagta aagggcttat tcaatggata ggctagtatt
      121 attagttagt aattgggccc ttttttttctt tgtttctttt cttcattttt ttccttttca
      181 aactatgggt tgtaaagcat ccacctttg aaagtttgcc tttctgccct ttcacgctga
      241 taagtacctc agtttccaat aaacttttgt tcaggggcaa acatttacaa tgttgacatc
      301 tcttcacacc accaaaaata ttcatggaga attattttat ctaaagctgt ctttttaata
      361 ataaaatagc cacctctacc ttcttcataa acttttaaga tgaattggta attcatcata
      421 gcaaggttga ttttagaaac taaagttgca ttaattcatt aaatacactg aaagtaattt
      481 tgtatgcttg gtcacaaaga aaatataaaa acaattttat aaatagattt gcagttattt
      541 tctttcaata ttttcttagt gcctatgatt acacaatacc acctatggag aaagctgttg
      601 tgaaaacgga cattcagata gcgctccctt ctgggtgtta tggaagagtg ggtaagtcat
      661 ttaagaaaca ggtaactatt tgtcaagttc tcctttgtga tagattcttc atgtttcatt
      721 tggggtaata agcaggcaat attgcttggg ctgtgtccta aaagaagcac catttgtgat
      781 agcaaatgca ctctttgaaa ggctttattt acatctctgc tttgcctctt tttgaccctt
      841 ttatttttct ccttcctcac tggagctttt aggctcacac tggcctagaa ggctgttctc
      901 agaacatggc attttatatt atgagagtaa aacttctgac ctgttggtcc cagaatgtgt
      961 aagcctactt aacctttttct tgtttggcca tggggtttag ggtaagggat actcttcagt
     1021 gtttgtagag gcactgggag gaagctagga caaaatggag ttacacgtca acaggtttga
     1081 tttttcctgg aagcgaattc agtgtttacc agacagttcc tttgcagagc gttagttcct
     1141 ttttgactac ttccaagtta acttaaggag gcatgga
//
```

Фигура 1-9. *GenBank запис за секвенцията на AF018432.*

Разбирането на различни фрагменти от нуклеотидни секвенции от иРНК, както и на съответните кодиращи региони от отделени записи в GeneBank беше основната трудност в тази глава. Ако сте го разбрали, поздравления! Можете да поседнете и починете, докато продължим нашата разходка из GeneBank.

Работа с близки/сходни записи в GeneBank

В примерите до момента използвахме номерата за достъп на записите, които искахме да видим – X01714, U90223 и AF018430 например. По принцип, вие може да намерите тези номера за достъп в статии, които ясно ги цитират за дадените секвенции.

Това е добра стратегия като за начало, но след това може да ви потрябват и други сходни гени. Тях можете да намерите през опцията „Related Sequences” от

меню „Related information”, което се намира от дясната страна на екрана за всеки запис в GeneBank (погледнете Фигура 1-7).

Намиране на записи без номер за достъп. Търсене с абревиатура на гена.

Въпреки, че GeneBank не е най-добрата БД за търсене по ключови думи, това все пак е възможно. Допускаме, че искате да намерите нуклеотидната секвенция на човешката dUTPase, но нямате номера за достъп под ръка. Следващите стъпки ще демонстрират как да постъпите в тази ситуация:

- Въведете във вашия браузър www.ncbi.nih.gov/entrez/
  Показва се началната страница на NCBI PubMed.
- От падащото меню за търсене изберете Nucleotide.
- Въведете human [organism] AND dUTPase [Protein name] в прозореца за търсене и кликнете върху „Search”.
  Ще ви се покажат 5 записа: AH005568, AF018429, AF018430, AF018431, AF018432. Те включват екзон 1 и 2 (AH005568), екзон 3 (AF018430) и екзони 5, 6 и 7 (AF018432). Последните 4 записа посочват аминокиселинните секвенции на 2 от формите (ядрена и митохондриална) на протеина dUTPase. До тук добре, не е лошо като начало!
- Кликнете върху „Related Sequences” под записа AF018432.
  Това ви връща общо 26 записа. Някои от тях съдържат иРНК секвенции също като U90223.

Пробвайте да използвате други ключови думи, за да опитате да намерите и други ограничени търсения в GeneBank.

Използване на генно-центрирани (Gene-Centric) БД

В допълнение на традиционния GeneBank, NCBI разработиха друга БД (Genes), която е по-адаптирана към генна информация или както наричаме такива БД – генно-центрирани.

Заявката към такива БД включва задаването на ключови думи, който са свързани директно с конкретен ген. Основното преимущество на тази БД, е че получените резултати са по-конкретни за дадения ген от тези в дългия списък от GeneBank. Накратко, вземате цялата информация наведнъж.

В тази секция ви превеждаме през Entrez/Gene ресурсите, които са достъпни на сървъра на NCBI. С тях можете да съберете важна информация, свързана с генния локус, както и със специфични места по хромозомата, където е идентифициран даден ген. Благодарение на тази услуга, вие бързо можете да откриете всичко известно за вашия любим ген или неговото геномно обкръжение.

- Въведете във вашия браузър www.ncbi.nih.gov/
  Показва се началната страница на NCBI PubMed.
- От падащото меню за търсене изберете Gene.
- Въведете DUT [gene] human [organism] в полето за търсене и натиснете бутона „Search".
  Ще се появи екран, приличащ на Фигура 1-11.

Различни заявки, като например DUT [gene name] human [organism] дават същия резултат. Това е много добър начин да намерите всичко необходимо, свързано със секвенцията и информацията за съответния ген!

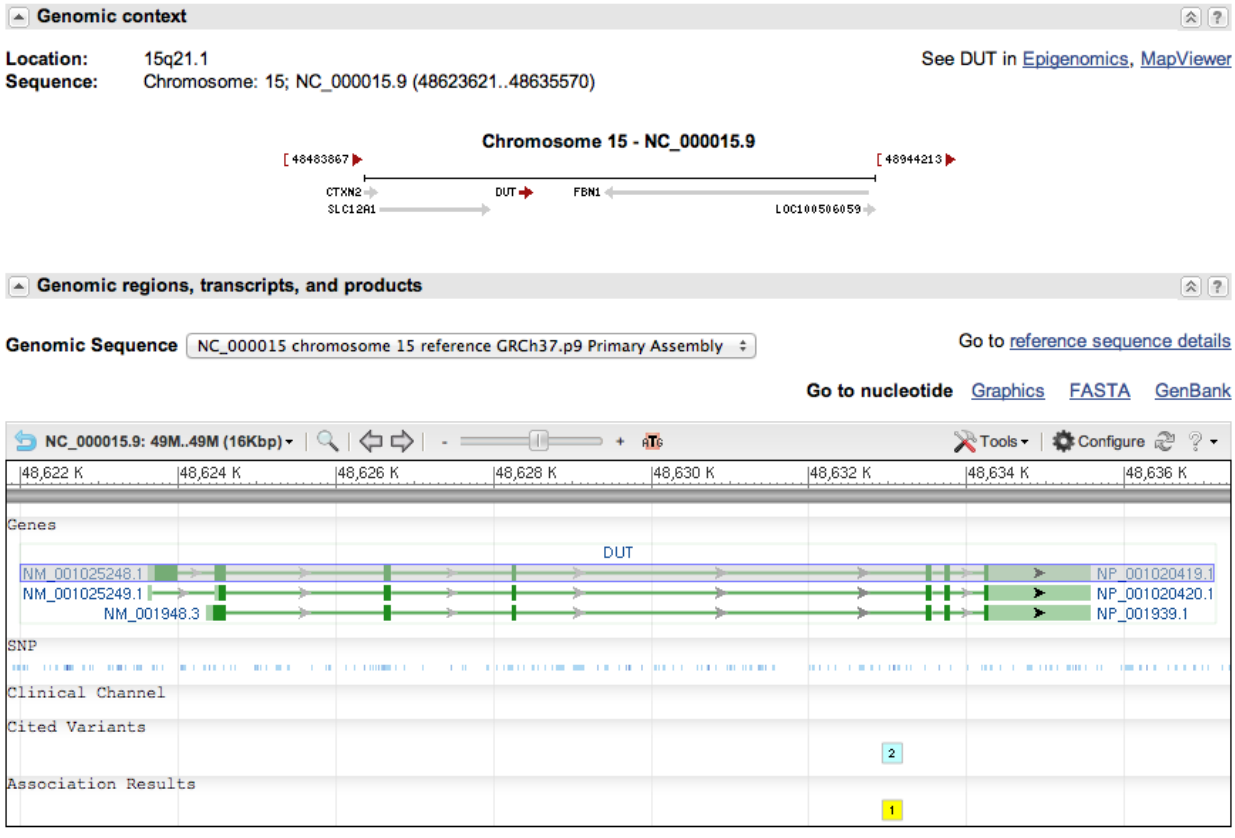
Най- горната част на записа DUT осигурява общото описание на това, какъв е този ген и какво се знае за неговите функции.

Фигура 1-12 показва следващата част от записа – схематично изображение на структурата на човешкия DUT ген с неговите 7 екзона, които обхващат 11 909 базови двойки от хромозома 15.

Дългият запис продължава с информация за потенциално взаимодействие с други генни продукти, хомоложни секвенции в други организми, протеинови функции и евентуално участие в метаболитни пътища. Всяка секция осигурява множество линкове, за да ви даде пълна представа какво е известно за вашия ген.

Тук са всички типове иРНК секвенции, които са наблюдавани и записани в GeneBank за този ген. Тези варианти (алтернативни транскрипти) включват митохондриална и ядрена форми на dUTPase (Таблица 3-1).

Фигура 1-11. Информация за човешкият ген DUT.



Фигура 1-12. Геномна карта за гена DUT.

По-долу на страницата можете да видите подробната генната структура.

Работа с цели геноми (геномно-центрични, Whole-Genome) БД

По-съвременните геномно-центрични БД са възникнали в отговор на големия брой проекти за секвениране на геноми. Те предоставят цялата информация, която ви трябва за даден организъм. Така вие можете по-лесно да насочите вашите анализи към гените от определен геном. Тези нови ресурси също така позволяват сравняването на цели геноми или цели геномни региони от различни организми (компаративна или сравнителна геномика).

Тези нови БД също така осигуряват и информация за определени секвенции и ги представят в зависимост един от друг на фона на глобалната рамка на пълния геном за даден организъм.

Най-добрият начин за изучаването на тези БД е да се придвижваме отвън навътре, докато достигнем до информацията, от която се интересуваме. Така, геномните БД осигуряват най-лесния начин за събиране на коя да е генна или протеинова секвенция на даден организъм наведнъж, само с няколко кликвания с мишката.

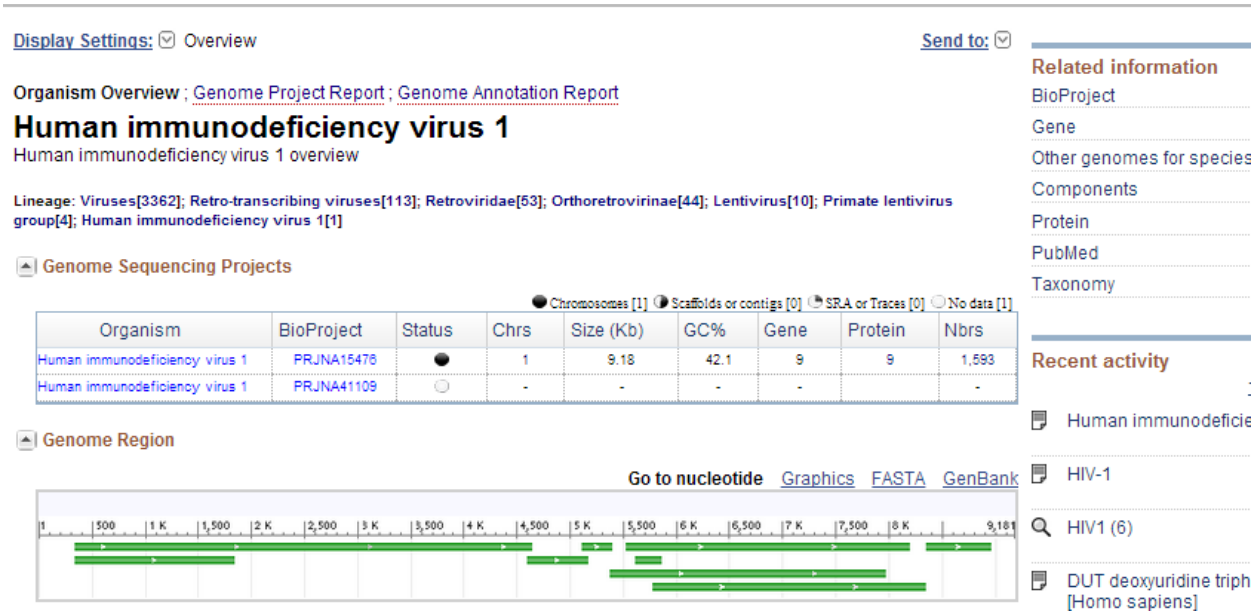
Нека да започнем нашето проучване на геномните БД чрез секция на вирусен геном, достъпна на NCBI сървъра.

Работа с вирусен геном

Вирусите са проектирани да осигуряват мултиплицирането на молекулите на нуклеиновите си киселини (вирусния геном) за сметка на клетъчния гостоприемник (еукариоти, бактерии и архебактерии). Вирусните геноми се проявяват в разнообразие от биохимични форми (РНК/ДНК, циклична/линейна, едноверижна/двойноверижна), и размери (от няколко кб до милион бази).
Ресурсите на NCBI за вирусни геноми са добър начин да намерите повече информация за тези молекули. За да демонстрираме това, ще ви покажем какво можете да намерите за AIDS-причинителя, тип-1 човешки имунодефицитен вирус или HIV-1.

- Въведете във вашия браузър www.ncbi.nih.gov/
  Вероятно сте запаметили този адрес. Показва се началната страница на NCBI.
- От падащото меню, изберете „All Databases”.

- Въведете „HIV1” в прозореца за търсене във всички бази данни и натиснете бутона „Search”.
- От списъка с резултатите, изберете този от тип „BioProject” (база данни за геномни проекти) и от следващият списък (6 обекта) изберете вирусния геном.
  Резултатът е глобална извадка на HIV-1 генома, показана на Фигура 1-14. В долната част на страницата има интерактивна карта на генома, на която можете да видите позицията на всички гени.



Фигура 1-14. Информация за генома на HIV- 1.

Подобни страници са налични за всички вируси, независимо от техния размер.

Над геномната карта можете да намерите различни формати – GenBank format, FASTA format – за пълната вирусна секвенция.

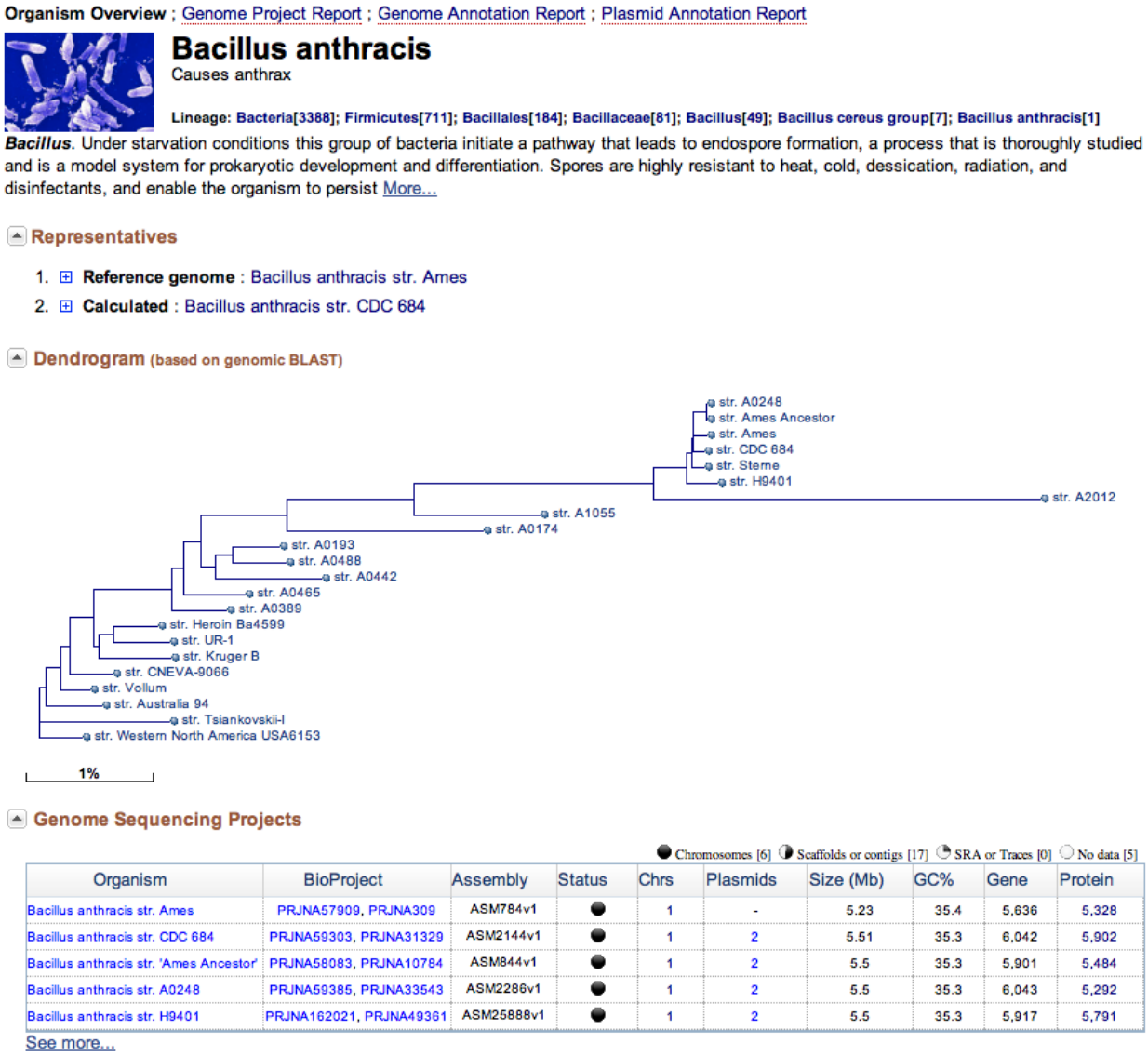
Работа с бактериални геноми

NCBI Entrez предлага приятен интерфейс на БД с цели бактериални геномни секвенции. За бърз преглед, направете следното:

- Въведете във вашия браузър www.ncbi.nih.gov/entrez/
  Показва се началната страница на NCBI PubMed.
- От падащото меню изберете „Genome”.
- В полето за търсене въведете „NC_007530” съответстващ на биотерористичния агент *Bacillus anthracis.*

Браузърът ви показва Таблица, обобщаваща бактериалния геном. Подобно на тази за вирусите (Фигура 1-18).

Така получавате бързо описание на бактерията, заедно с картинка, детайли за ареала, болестите, които причинява и т.н.



Фигура 1-18. Информация за генома на Bacillus anthracis.

Изследване на човешкия геном

Секвенирането на човешкия геном е едно от най-великите научни постижения. С три милиарда нуклеотида, простиращи се в 23 хромозоми, нашият геном определено е комплексен обект за работа. Една от главните задачи на биоинформатиката е да се интегрира минала, настояща и бъдеща информация, за човешките гени в тези геномни ресурси.

Ако искате да разберете човешкия геном, трябва да имате ясна идея за сегашното състояние на наличните данни.

- Пълна нуклеотидна секвенция на човешкия геном вече е налична, с изключение на няколко хиляди „дупки", които продължават да бъдат запълвани. Това е така, защото периодично се пускат нови версии (builds) на генома. Това динамично състояние е валидно за всички животински геноми.
- Първоначално, секвенциите са в raw формат; следващото предизвикателство било самото анотиране на преките данни. Това представлявало създаване на детайлна и точна FEATURES Таблица на човешкия геном.
- Постоянно се генерира нова информация за човешките генни, както и за техните характерни особености и функции. Някой трябва да събере цялата тази информация, да я пакетира добре, да я предложи на цялата изследователска общност безплатно!

Проектът Ensembl

Проектът Ensembl (www.ensembl.org) е част от проект между European Bioinformatics Institute (EBI) и Sanger Institute, които се намират в Кеймбридж. Те са разработили интегрирана БД и софтуер като система за извеждане,обслужване и поддържане на автоматични анотации за животинските геноми, като е обърнато специално внимание към нашите най-близко родствени организми – гръбначните. Всичко започнало като част от интернационалния проект „Човешки Геном," последван от мишия геномен проект, сега продължава за много други животни. Данните и софтуерът се разпространяват свободно между многобройни БД и биоинформатични центрове навсякъде по света, което позволява комплексната взаимовръзка и публичният достъп на информацията. Ще опишем тази ДБ подобно на другите в тази глава, наблягайки на характерните особености свързани с информацията за секвенциите.

Основният геномен браузър за момента осигурява набор от генни транскрипти и протеини за геномите на 9 организма.

Информацията в ENSEMBL се представя чрез т.нар изгледи (Views), като всеки изглед представя различно ниво на организация и детайли. Въпреки че анализите, описани в NCBI могат да бъдат осъществени и с ENSEMBL, при тази БД се подхожда към човешкия геном по по-различен начин.

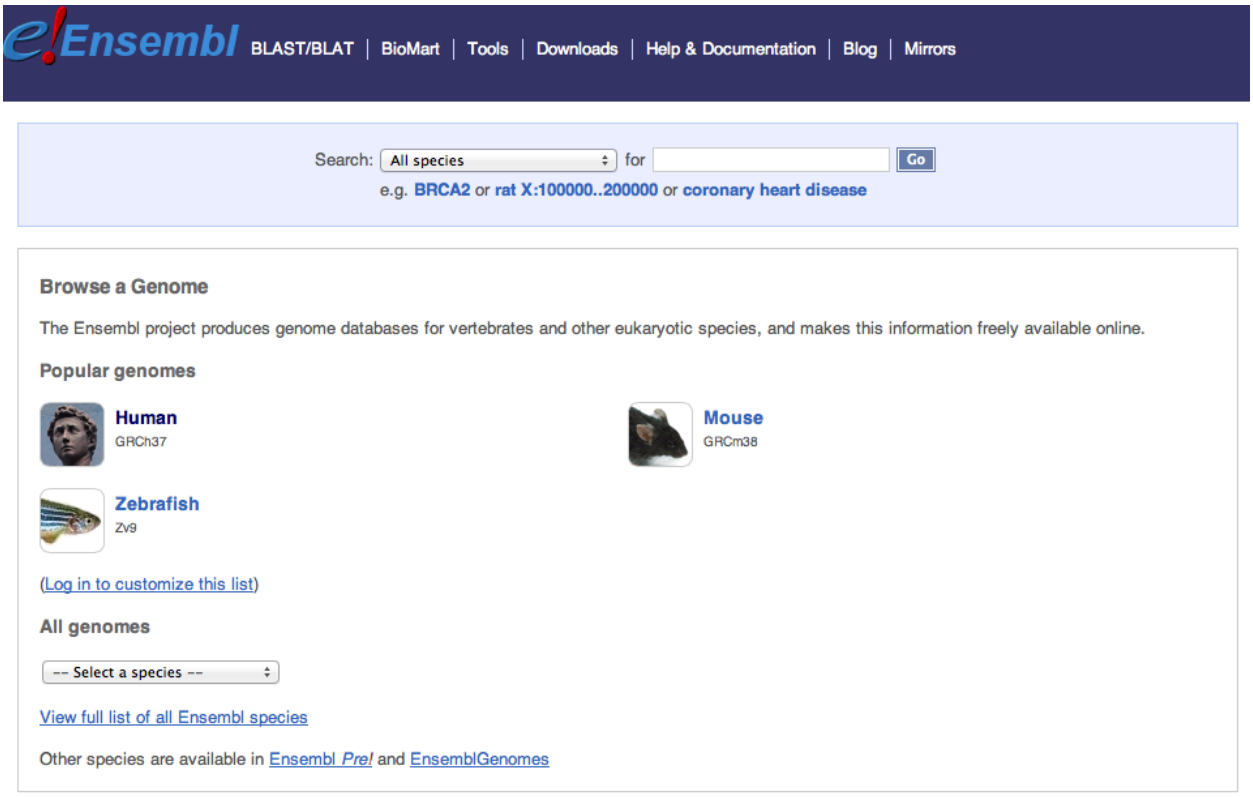
Началната страница за всеки организъм осигурява информация за геномната секвенция, както и датата, в която анотацията е обновена последно. В момента,

ENSEMBL е с версия GRCh37 за човешкия геном. Вторият номер – 37 – отговаря на версията (build) на геномната секвенция в случая 37 от NCBI.

Въведение в страницата Ensembl

Ресурсите на Ensembl могат да бъдат разглеждани по много начини. За да разберете бързо всички налични опции, най-добре да започнете от раздела „guided tour," който Ensembl ви предлага на своята страница.

1. Въведете във вашия браузър www.ensembl.org.
   Появява се страницата на Ensembl, както е показано на Фигура 1-23.
   За сега нека да се фокусираме на лявата половина от страницата, където са линковете свързани с животинската геномна сфера. Тук сме за да разгледаме човешкия геном.



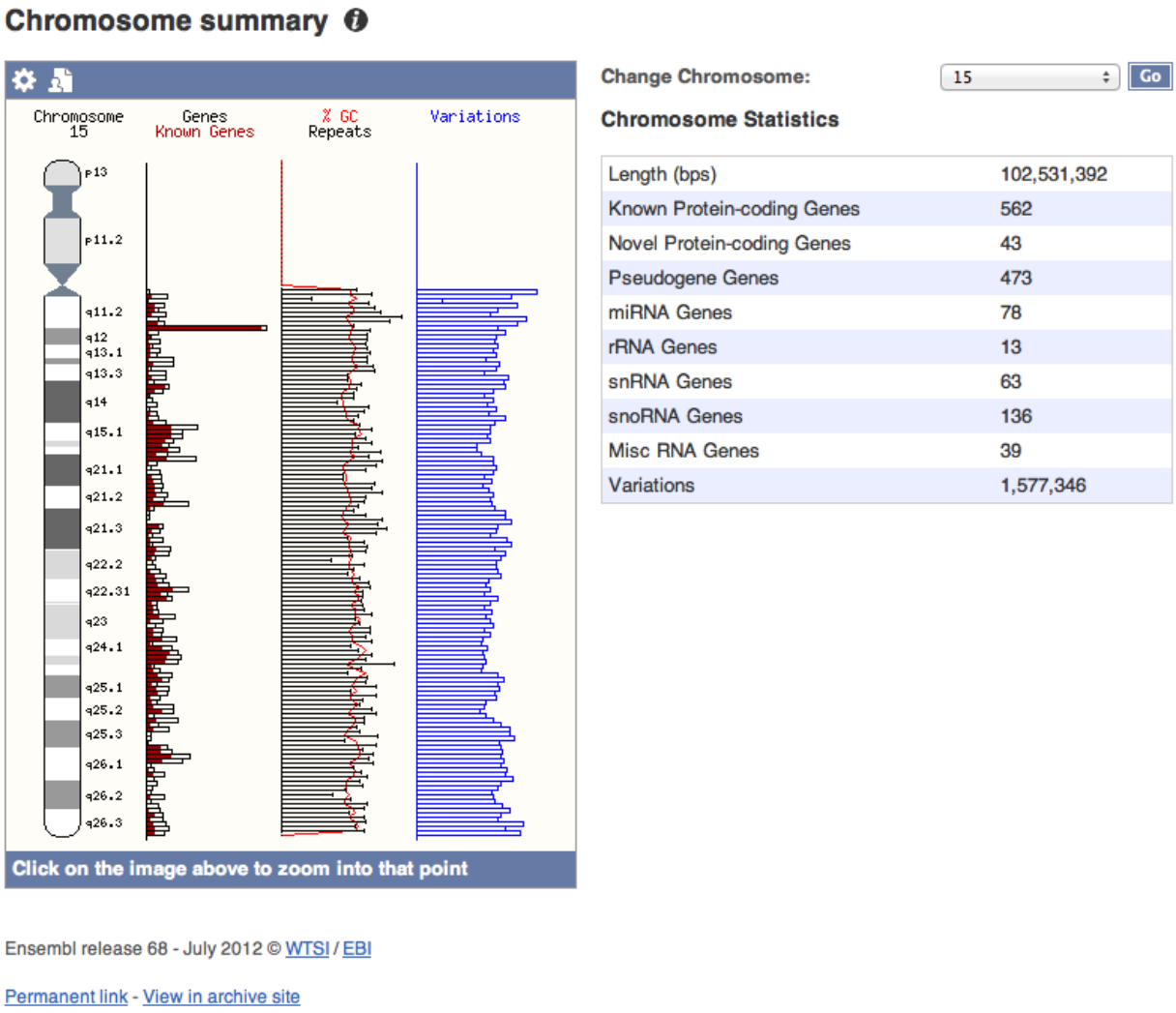
Фигура 1-23. База данни Ensembl.

2. Кликнете на иконата на Homo sapiens.
   На страницата може да видите кратък преглед на информацията за човешкия геном.
3. Кликнете на менюто Karyotype и след това на хромозома 15 – Chromosome summary.

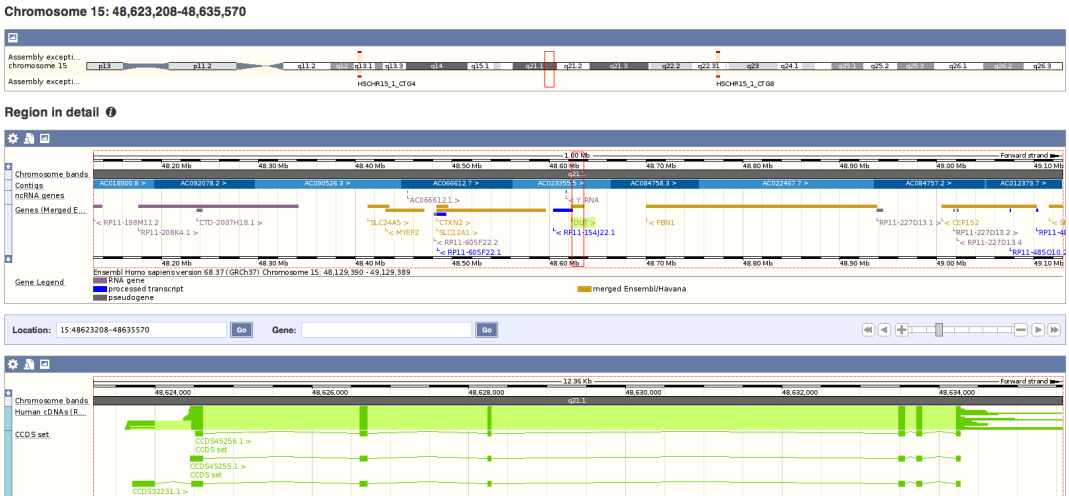
Това ви води до данни за самата хромозома 15 (Фигура 1-25), където можем да зададем конкретни въпроси като например: къде е гена dUTPase? Можем да направим това по-трудния начин, чрез кликване върху региона (локуса) на q21.1 бенда и тогава ръчно да сканираме региона. (Видяхме тази информация по-рано в тази глава – припомнете си Фигура 1-8) По-лесният начин е да използваме полето за търсене горе на страницата:

- Цъкнете на регион q21.1 и изберете местонахождение 15:44800001-49500000.
- Въведете DUT в прозореца за търсене „Gene”.
- Намерен е един запис за този ген. Сега хвърляте поглед на комплексния резултат, който ви показва структурата, геномния контекст на DUT гена, като и от-горе на-долу: къде се намира генът в хромозомата, генна карта и карта на транскриптите (Фигура 1-26).

  Тук може да проучите конкретен хромозомен регион, и да изучавате вътрешната структура и алтернативните форми на експресия на конкретни гени, както и да определяте единични нуклеотидни полиморфизми в този ген (SNP).



Фигура 1-25. Човешкият геном визуализиран по хромозоми.



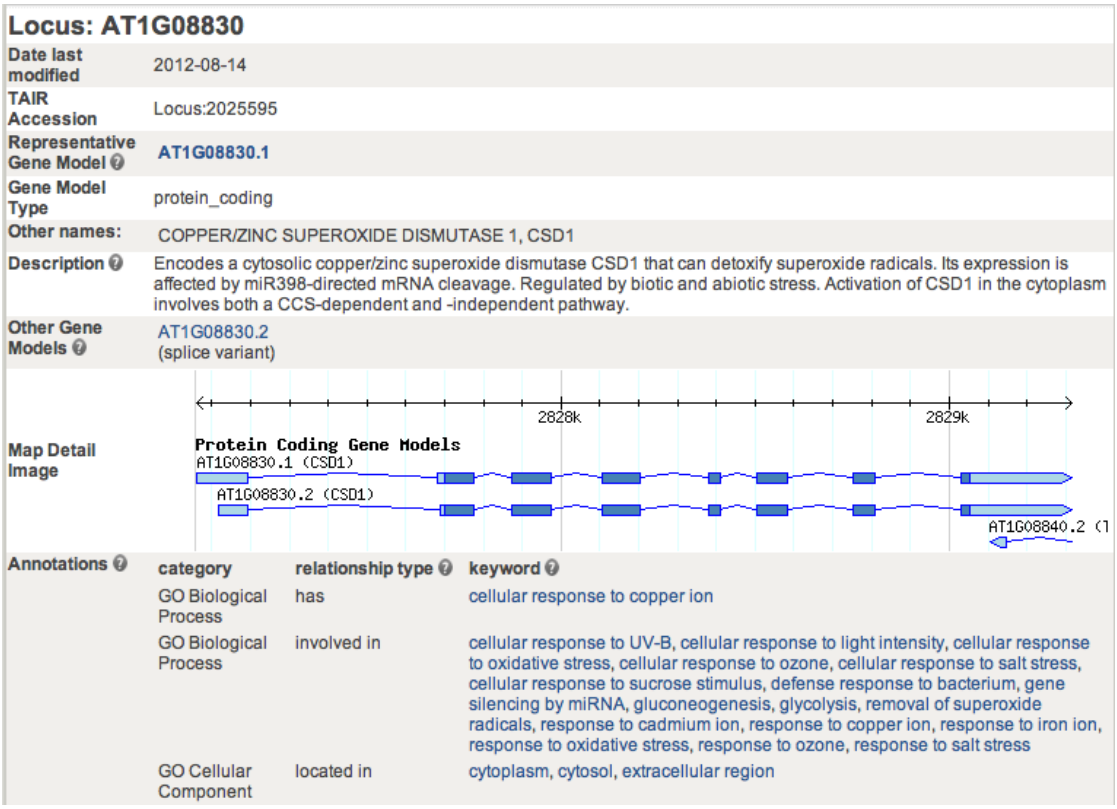
Фигура 1-26. Информация за гена DUT в Ensemble.

Работа с растителни геноми

Ако сте растителен молекулярен биолог, вие несъмнено ще имате нужда от база данни с пълната анотация на моделното растение *Arabidopsis*. Такава е базата данни наречена TAIR – The Arabidopsis Information Resource (http://www.arabidopsis.org). Тук можете да разглеждате структурата на различните гени и геномната им организация, както и да записвате локално секвенциите на отделните молекули в различни формати.

Можете директно да започнете разглеждането на даден ген от полето за търсене най-горе на страницата като посочите името на гена или неговият номер за достъп.

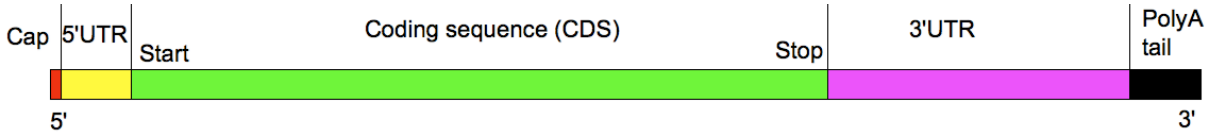
- Потърсете гена на медната супероксид дисмутаза CSD1.
- От намерените резултати, номер за достъп/локус „AT1G08830".
- Тук може да видите пълната анотация на гена с описание на неговата функция, генна карта както и линкове към пълната геномна секвенция (Фигура 1-27)



Фигура1-27. Информация за гена CSD1 с номер на достъп AT1G08830.1.

Упражнения

Био-бази данни 1 GenBank/NCBI



Задача 1. Извличане на секвенции от NCBI – GENEBANK.

Влезте в **NCBI – http://www.ncbi.nlm.nih.gov/ .**

Намерете секвенция с номер за достъп: **NM_001005862.**

Името на гена с дадената секвенция е: _____

Дължината на секвенцията е: _____

Типа на молекулата е: _____

Организъм: _____

Въведете координатите на кодиращата област: _____ - _____

Въведете координатите на 3’ не-кодиращата област (UTR): _____ - _____

Въведете координатите на полиА-опашката: _____ - _____

Извлечете секвенцията на нуклеотидната последователност във FASTA формат, както и на протеина и ги съхранете в отделни файлове чрез текстов редактор (Notepad) с разширение *.fa, *.fasta или *.seq.

Задача 2. Търсене на гени по техните имена от NCBI – GENEBANK.

Влезте в **NCBI – GENE.**

Потърсете гена от първа задача в базата данни от **гени** и намерете „database entry" за **човешкия** ген по име на гена.

Съществуват ли ортолози в базата данни на този ген? _____

В коя хромозома се намира генът? _____

Напишете хромозомния локус: _____

Избройте някои от синонимите на гена: _____

Посочете позицията на нуклеотидите в хромозомата, където се намира гена – (вижте генната карта) _____

Колко транскрипционни варианта има генът? _____

Защо даден ген може да има транскрипционни варианти? Напишете номерата за достъп на транскрипта/те _____

Ако генът има повече от един транскрипт, сравнете кодиращите области, коя е по-голяма? _____, _____

Извлечете секвенциите на всички транскрипти и техните протеини във ФАСТА формат. Съхранете ги в отделни файлове чрез текстов редактор (Notepad) с разширение *.fa , *.fasta или *.seq.

Задача 3. Извличане на геномният участък на изследваният ген.

Влезте в геномният браузър и в хромозомата, в която се намира гена (от Resources менюто изберете Genomes and Maps → Genomes → Human Genome) и влезте в хромозомата, в която се намира генът от предишната задача. Чрез полетата за детайлно разглеждане (в менюто от ляво) на хромозомата въведете геномният участък (намерен от вас в задача 2).

Използвайте функцията за Download/View Sequence/Evidence, за да изтеглите секвенцията.

Био-бази данни 2

EMBL

Задача 4. Извличане на секвенции от EMBL.

Влезте в **EMBL-SRS – http://www.ebi.ac.uk/embl/**

Намерете секвенция с номер за достъп: **BC104789**.

Името на гена с дадената секвенция е: _____

Дължината на секвенцията е: _____

Типа на молекулата е: _____

Организъм: _____

Въведете координатите на кодиращата област: _____ – _____

Извлечете секвенцията на нуклеотидната последователност във ФАСТА формат, както и на протеина и ги съхранете в отделни файлове чрез текстов редактор (Notepad) с разширение *.fa, *.fasta или *.seq.

Био-бази данни 3

Ensembl

Задача 5. Извличане на геномния участък на изследвания ген от Ensembl.

Влезте в **Ensembl http://www.ensembl.org .**

Влезте в човешкия геномен браузър и чрез полето за търсене (горе в дясно) намерете гена, изследван от вас в задача 1. Изберете от резултатите най-точния база данни запис.

Разгледайте записа за гена. Това същият ген ли е? Проверете хромозомата и локализацията му!

Опитайте се да свалите секвенцията на геномния участък, като разгледате полето – Genomic Location и използвате опцията Export data.

Био-бази данни 4

www.arabidopsis.org

Задача 6. Търсене на геномен участък в генома на арабидопсис и работа със геномната карта.

Влезте в БД за арабидопсис **www.arabidopsis.org.**

Влезте е **Seqviewer (геномната карта)** и изберете хромозома 1. Въведете геномни координати: 13719388-13719488. Задайте резолюция 10кб.

Въведеният регион се фланкира от гените: _____ и _____ (запишете номерата за достъп на гените).

Влезте в ретротранспозонния ген – с каква геномна локализация е? _____ – _____ нт.

Влезте в Sequence Ruler-а и съхранете секвенцията на ретротранспозонния екзон.

Използване на бази данни за протеинови секвенции, специализирани бази данни

- Ще изследваме формирането на протеините (protein maturation);
- Ще се ориентираме в UniProtKB/Swiss-Prot запис;
- Ще научим повече подробности за биохимията на белтъците;
- Ще научим каква е функцията на изследван от нас белтък;
- Ще се справяме с полезни информационни ресурси, касаещи белтъци protein-related information resources.

> *Човечеството е ензим, катализиращ прехода от въглерод-базирана към силикон-базирана интелигентност.*
>
> Gerard Bricogne

Белтъчните бази данни не се ограничават само с белтъчните секвенции. В тази глава ще се научите да използвате (понякога доста заплетената) мрежа от Интернет ресурси и връзки, за да свързвате секвенциите на протеините с техните функции и дори с техните 3-D структури.

Както и в огромния свят на нуклеотидните секвенции, където всичко се върти около един централен ресурс (например GenBank), повечето връзки между гените, протеините и техните функции са организирани около **UniProtKB/Swiss-Prot** – централният ресурс за анотирани белтъчни секвенции. Той се поддържа от Швейцарския институт по биоинформатика (Swiss Institute of Bioinformatics, SIB) и Европейският институт по биоинформатика (European Bioinformatics Institute, EBI). UniProt страницата е част от биоинформатичния портал ExPASy.

Има много начини за достигане до базата данни UniProt knowledge base (доскоро - Swiss-Prot). Повечето геномни бази данни, геномни браузъри или сайтове поддържащи система за достъп до секвенции (sequence retrieval system, SRS), ви свързват директно към съответния UniProt запис.

В тази глава ще научите да разбирате съдържанието на Swiss-Prot записи и да се ориентирате в изключително подробните й анотации в, а също и как да използвате една анотация като стартова точка към още по-специализирани бази данни. Те ще ви дадат информация например за структурата или функцията на изследвания от вас белтък.

Ще започнем с някои идеи как клетките превръщат един новосинтезиран (нативен) протеин в зрял, готов да функционира протеин чрез комбинация от различни пост-транслационни модификации. Информацията за тези модификации е една от основните причини да искате да използвате протеинова база данни.

От транслацията на ORF до зрелия протеин

Секвенирането на ДНК молекули е много по-бързо и евтино от секвенирането на протеини, затова повечето от аминокиселинните секвенции, които познаваме, на практика представляват компютърно транслирани **отворени рамки на четене** (ОРЧ, open reading frames, ORFs), установени при анализа на геномните данни. Повечето протеини, съответстващи на тези ORFs всъщност никога не са били изолирани от никого. Ето защо специалистите по белтъци мразят, когато молекулярните биолози и биоинформатиците говорят за транслираните ОРЧ, като че ли те са истински белтъци. За кратко време, в началото на геномната ера имаше тенденция тези специалисти да бъдат третирани като банда сърдити старчета и да бъдат игнорирани. След секвенирането на множество еукариотни геноми обаче, интересът към протеините се възвърна на друго ниво, в много по-мащабен формат, известен като **протеомика (proteomics**). Това е научна област, занимаваща се с визуализация и характеризиране на целия набор от протеини, срещащи се в дадена тъкан или организъм. Протеомните анализи доведоха до бързо натрупване на данни.

ORFs: това което виждаш, НЕ е което получаваш

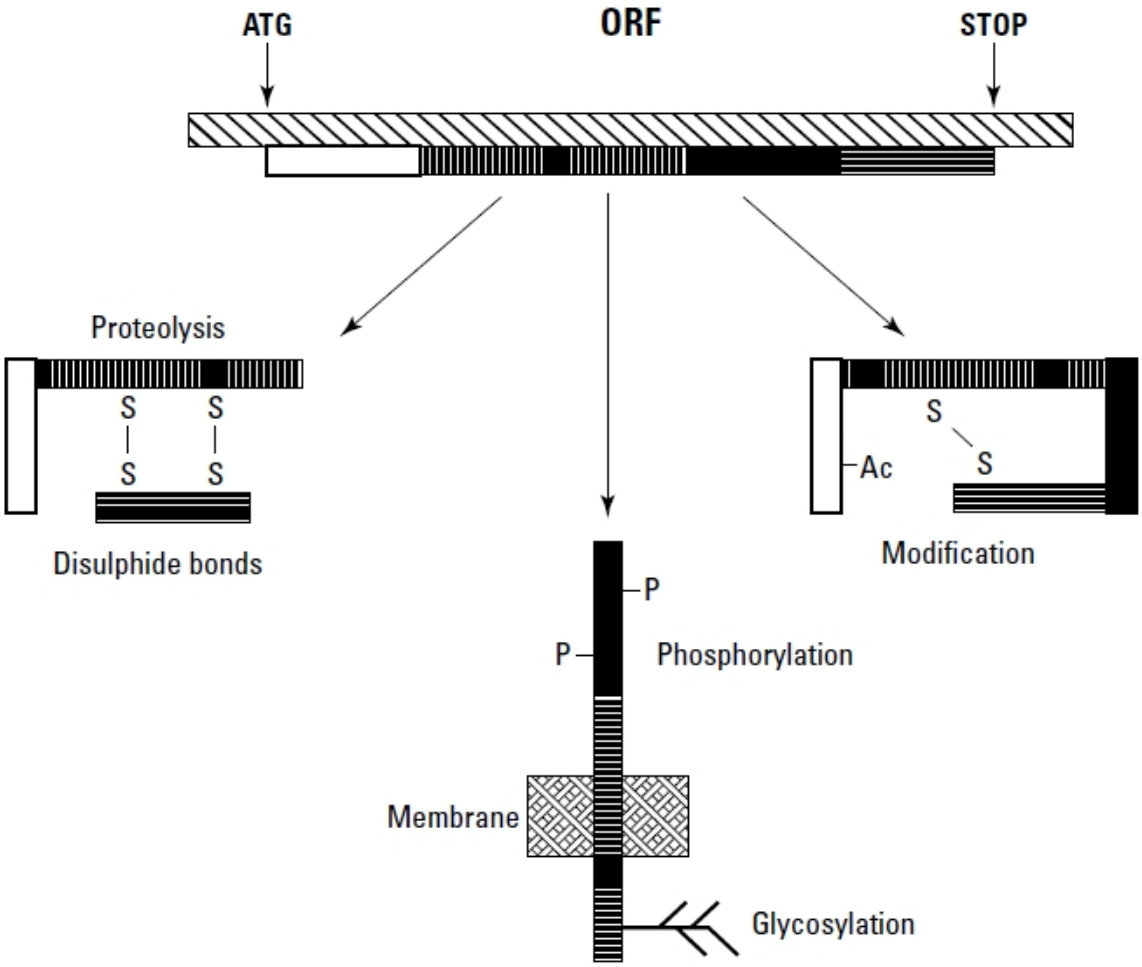
Когато биолозите провеждат 2D гел електрофореза (техника, при която протеините се разделят първо по маса, а след това по заряд), реалните протеини почти никога не се държат, както се очаква на базата на компютърна транслация на техните ORFs. В света на ORFs и протеините, това което виждаш не е това, което получаваш.

Причината за това е следната: когато се транслира от иРНК, първичната аминокиселинна верига може да бъде силно модифицирана по пътя на превръщането си в зрял протеин. Дори такива прости физико-химични свойства на зрелия протеин като размер, молекулно тегло или изоелектрична точка (ИЕТ) е много трудно да се предскажат, ако ни е известна единствено компютърно генерираната АК секвенция. Сложният процес на **узряване на протеина** (*protein maturation*) се нарича **пост-транслационна модификация** (*post-translational modification).* Обобщена е на Фигура2-1.

Пост-транслационната модификация на един белтък може да се състои от всякаква комбинация от следните процеси:

- Разрязване във вътрешността на АК веригата;
- Премахване на фрагменти от АК веригата (например при узряването на инсулина);
- Химични модификации на специфични аминокиселини (например метилиране);
- Прибавяне на липидни молекули към белтъчната верига (например миристоилиране (myristoylation));
- Прибавяне на гликозидни (захарни) молекули (например гликозилиране).

Основната функция на една протеинова база данни (за разлика от една колекция от ORF секвенции) е да покаже този тип информация, когато такава е достъпна от експериментални данни или е предсказана на базата на различни компютърни методи. Следователно, пост-транслационните модификации представляват значителна част от свойствата на протеините, описани в Swiss-Prot.



Фигура 2-1. Пост-транслационни модификации по време на узряването на протеина.

Индивидуална крайна дестинация за всеки протеин

За да изпълняват нормално своите функции, белтъците трябва да достигнат правилната дестинация в организма и в клетката. По време на транслацията си, пептидната верига може да покаже разнообразие от високо-специфични сигнални секвенции (нещо като маршрутни кодове), които клетката използва, за да насочи новотранслирания протеин към правилния компартмент (в или извън клетката). Това сортиране винаги включва транспорт на протеина през една или няколко мембрани и се нарича още **транслокация** (*translocation).* Най-често срещаните финални дестинации за един белтък могат да бъдат:

- Прикрепване към клетъчната мембрана;
- Секретиране извън клетката;
- Транспортиране към периплазмата (при бактериите);

- Транспортиране в митохондриите или други органели;
- Транспортиране в клетъчното ядро.

Познаването на финалния компартмент, в който функционира всеки протеин е важно за разбирането на неговата функция, и затова тази информация (експериментално доказана или компютърно предсказана) е една от важните характеристики, описани в протеиновите бази данни като Swiss-Prot.

Комбинаторно многообразие от форми и функции

Най-важната стъпка от превръщането на една новосинтезирана пептидна верига във функционален белтък е нагъването на линейната верига в компактна и стабилна 3D структура. С изключение на малките белтъци (под 100 АК), крайната структура на един белтък обикновено се състои от няколко относително независими домена. Можете да си ги представите като основен набор от сравнително твърди тухлички от играта-конструктор LEGO. Природата използва такива домени – «тухлички», за да създаде огромното разнообразие от съществуващи протеини. Например, от начален набор от три домена (A, B, C) можете да получите протеините AAA, ABC, BCC, BAC, CCA и т.н. Повечето природни протеини са съставени от комбинации на един до десет домена. Общият брой на съществуващите базови домени се оценява на няколко хиляди. Въпреки значителните вариации в секвенцията, домените могат да бъдат идентифицирани по специални запазени мотиви в секвенцията (*scaffold sequence signatures)* — висококонсервативни мотиви, които могат да бъдат разпознати и след милиарди години дивергентна еволюция. Разпознаването и дефинирането на белтъчните домени е една от основните изследователски области в биоинформатиката. Доменната архитектура на дадена белтъчна секвенция е важна, защото може да подскаже много за възможната 3D структура на белтъка, а също и за неговата вероятна биохимична и/или клетъчна функция. Затова значителна част от всеки Swiss-Prot запис е посветена на описанието на доменната организация на протеините, или предсказана с всички налични биоинформатични методи и/или установена експериментално.

Четене на Uni-Prot запис

Въпреки сложната терминология на тяхното описание, протеините са много по-прости обекти от гените, защото:

- Протеините кореспондират на сравнително къси секвенции (средна дължина - 350 АК);
- За разлика от гените, началната и крайната точка на протеините са ясни;
- Протеините се дефинират на една верига;
- Каквито и модификации да се случат между ORF секвенцията и зрелия протеин, аминокиселините в нея запазват своята последователност.

По тези причини ще бъдете прави, ако предположите, че записите на Swiss-Prot са по-лесни за четене и разбиране от тези на GenBank!

Подобно на GenBank запис, един Swiss-Prot запис се разделя на няколко основни секции, съдържащи:

- Обща информация/ The general information part;
- Библиографска информация / The bibliographic information part;
- Функционална информация / The functional information part;
- Таблица на свойствата / The feature table;
- Секвенцията на протеина/ The sequence part.

Използвайки като пример рецептора за човешкия епидермален растежен фактор (epidermal growth factor receptor, EGFR), в следващите секции ще ви покажем стъпка по стъпка как да дешифрирате напълно един Swiss-Prot запис.

Дешифриране на Swiss-Prot записа за EGFR

EGF рецепторът е доста сложна молекула и затова много добре илюстрира цялото богатство на информацията, която можем да извлечем от един Swiss-Prot/UniProt запис.

Има и по-сложни начини за търсене в Swiss-Prot, но за нашите цели в момента е достатъчно следното основно търсене:

1. Отидете на адрес http://www.expasy.org
Появява се началната страница на ExPASy .

2. От падащото меню вляво от прозореца за търсене изберете база данни UniProtKB, а в полето за търсене въведете **P00533 (Swiss-Prot ID).**

Това е записът, съответстващ на рецептора за човешкия епидермален растежен фактор (EGFR). Той е подходящ пример, понеже е модерен напоследък – обект на усилени изследвания. EGFR е онкоген (участва в инициирането на рак), и поради това е анотиран много подробно. Понеже голям брой изследователи едновременно работят върху тях, обикновено "модерните" протеини са много по-добре анотирани от останалите.

3. Натиснете бутона search.

Появява се страница с резултатите от търсенето – в случая един запис, съответстващ на UniProtKB/Swiss-Prot номера за достъп P00533.

Записът за EGFR е много сложен и съдържа огромно количество информация, което не е така за повечето други протеини.

Номерът P00533 е хипервръзка, кликнете върху него, за да влезете в основната страница за този запис.

Обща информация за записа

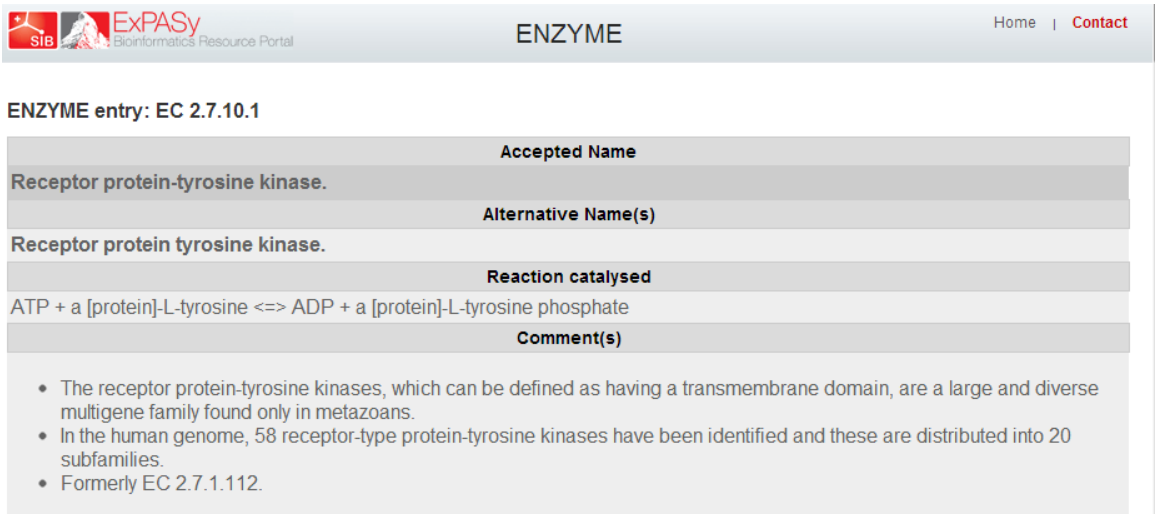
Първата секция е озаглавена **Имена и произход** (**Names and origin)** и съдържа следните полета:

- **Protein names –** имена на протеина (генния продукт)**:** ***Epidermal growth factor receptor***: Това е основната описателна (descriptive) информация за секвенцията, обикновено тя е достатъчна за прецизното идентифициране на протеина. Номерът E.C. (2.7.10.1) обозначава биохимичната реакция/функция (tyrosine protein kinase), която изпълнява протеинът. E.C. е съкращение от Enzyme Nomenclature Committee, на Таблица 4-1 има малко повече информация за това. Оттук може да се стартира увлекателно пътешествие към цялостно разбиране на тази ензимна функция. Натиснете хипервръзката 2.7.10.1, за да видите ENZYME страницата - поглед върху ензима, както е показано на Фигура 2-2. Линковете на тази страница водят до различни типове информация.

Gene namesн – Имена на гена или абревиатури: ***EGFR.*** Това поле е полезно, когато имате нужда от информация от геномни бази данни или такива с нуклеотидни секвенции. То помага търсенията (queries) да бъдат ясни и недвусмислени.

- **Synonyms – Синоними:** ***Receptor tyrosine-protein kinase ErbB-1, EC 2.7.10.1:***

- **Organism** – Организъм. Популярното и официалното (латинско) название**: *Homo sapiens (Human)*** Указанието на вида (species designation) се състои, в повечето случаи, от името на рода и вида на латински, последвани от име на английски (в скоби). За вируси е дадено само обичайното английско име.
- **Taxonomic identifier** – Таксономичен идентификатор. Цифров код, отговарящ на таксономичната позиция на организма. Следвайки линка Taxonomic ID 9606 , ще бъдете въведени в страницата на *H.sapiens* от базата данни за таксономия **UniProt taxonomy,** поддържана от групата UniProtKB (http://www.uniprot.org/taxonomy/9606).
- **Taxonomic lineage** – Таксономична линия.
- ***Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;..., и т.н.*** *:* Това поле изброява таксономичната класификация на организма – източник. Поддържаната класификация е същата като подържаната при National Center for Biotechnology Information (NCBI). Всяко име на група е хипервръзка към списък с всички Swiss-Prot записи от съответната група. Това поле е много удобно, когато ви се налага бързо да съберете всички белтъчни секвенции от даден вид.



Фигура 2-2. ENZYME изглед на UniProt записа P00533.

Секция Protein attributes - атрибути на протеина

Включва следните полета:

- **Sequence length** (Дължина на секвенцията):1210 AA.
- **Sequence status** (Статус на секвенцията): Complete.

Обозначава покритието на секвенцията от проекта по секвениране.

- **Sequence processing** (Процесинг на секвенцията): The displayed sequence is further processed into a mature form.
  Обозначава дали секвенцията се изрязва или претърпява други модификации в процеса на узряване.

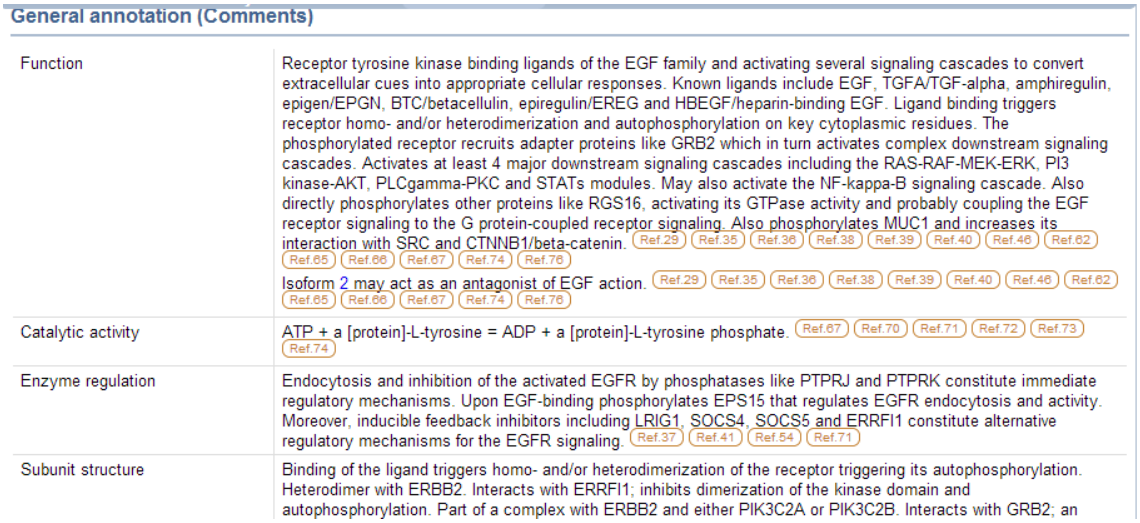
- **Protein existence** (Съществуване на протеина): Evidence at protein level.
  Описва наличните доказателства за съществуването на протеина.

Секция General annotation (Comments) – Основна анотация (Коментари)

Тази секция съдържа основна информация за протеина. Коментарите са групиран в блокове, представени с тематична ключова дума. Кликването върху всяка от тях отваря помощно поле с обяснение на значението на термина.

EGFR е добре изследван протеин и съответно секцията Коментари за записа P00533 е доста богата, както виждате на Фигура2-3. Тя съдържа:

- **Function** (Функция) – общо описание на функцията/функциите на протеина.
- **Catalytic activity** (Каталитична активност) – описание на реакциите, катализирани от този ензим.
- **Subunit structure** (Структура на субединиците) – описва четвъртичната структура на протеина.
- **Subcellular location** (Субклетъчна локализация).
- **Tissue specificity** (Тъканна специфичност) – описва експресията на гена в различни тъкани.
- **Post-translational modification** (Посттранслационни модификации).
- **Involvement in disease** (Участие в заболявания) – описание на заболяванията, свързани с мутации в гена/протеина.
- **Miscellaneous** (Разни) – всякаква полезна информация, която не пасва в никоя друга секция.
- **Sequence similarities** (Подобие на секвенцията) – описание на подобията с други протеини.



General annotation (Comments)

Function	Receptor tyrosine kinase binding ligands of the EGF family and activating several signaling cascades to convert extracellular cues into appropriate cellular responses. Known ligands include EGF, TGFA/TGF-alpha, amphiregulin, epigen/EPGN, BTC/betacellulin, epiregulin/EREG and HBEGF/heparin-binding EGF. Ligand binding triggers receptor homo- and/or heterodimerization and autophosphorylation on key cytoplasmic residues. The phosphorylated receptor recruits adapter proteins like GRB2 which in turn activates complex downstream signaling cascades. Activates at least 4 major downstream signaling cascades including the RAS-RAF-MEK-ERK, PI3 kinase-AKT, PLCgamma-PKC and STATs modules. May also activate the NF-kappa-B signaling cascade. Also directly phosphorylates other proteins like RGS16, activating its GTPase activity and probably coupling the EGF receptor signaling to the G protein-coupled receptor signaling. Also phosphorylates MUC1 and increases its interaction with SRC and CTNNB1/beta-catenin. Ref.29 Ref.35 Ref.36 Ref.38 Ref.39 Ref.40 Ref.46 Ref.62 Ref.65 Ref.66 Ref.67 Ref.74 Ref.76 Isoform 2 may act as an antagonist of EGF action. Ref.29 Ref.35 Ref.36 Ref.38 Ref.39 Ref.40 Ref.46 Ref.62 Ref.65 Ref.66 Ref.67 Ref.74 Ref.76
Catalytic activity	ATP + a [protein]-L-tyrosine = ADP + a [protein]-L-tyrosine phosphate. Ref.67 Ref.70 Ref.71 Ref.72 Ref.73 Ref.74
Enzyme regulation	Endocytosis and inhibition of the activated EGFR by phosphatases like PTPRJ and PTPRK constitute immediate regulatory mechanisms. Upon EGF-binding phosphorylates EPS15 that regulates EGFR endocytosis and activity. Moreover, inducible feedback inhibitors including LRIG1, SOCS4, SOCS5 and ERRFI1 constitute alternative regulatory mechanisms for the EGFR signaling. Ref.37 Ref.41 Ref.54 Ref.71
Subunit structure	Binding of the ligand triggers homo- and/or heterodimerization of the receptor triggering its autophosphorylation. Heterodimer with ERBB2. Interacts with ERRFI1; inhibits dimerization of the kinase domain and autophosphorylation. Part of a complex with ERBB2 and either PIK3C2A or PIK3C2B. Interacts with GRB2; an

Фигура 2-3. Част от основните полета в секцията General annotation (Comments) за Swiss-Prot записа P00533.

Съдържанието на секцията Comments често може да ви е достатъчно да прецените дали даден резултат (получен с BLAST или друга програма), който сте получили за съответен Swiss-Prot запис има смисъл или няма.

Секция Ontologies – Онтологии

Включва две подсекции:

- **Keywords** (Ключови думи). Подбрани ключови думи от йерархичен речник, сумиращи съдържанието на всеки UniProtKB запис и улесняващи търсенето на протеини с определени свойства и характеристики. Подсекцията Keywords включва полетата **Cellular component, Coding sequence diversity, Disease, Domain, Ligand, Molecular function, PTM** и **Technical term.** Срещу всяко от тях има списък с хипервръзки към термини, подчинени на всяка от ключовите думи.
- **Gene Ontology (GO)** – представлява списък с избрани термини, извлечени от проекта Gene Ontology (GO). Този проект има за цел унифицирано и пълно описание на свойствата на всеки протеин в клетката в три аспекта, които присъстват и в Swiss-Prot записа:
    - **Biological process** (Биологичен процес) – клетъчният процес, в който участва съответният протеин.
    - **Cellular component** (Клетъчен компонент) – локализацията на протеина в клетката.

- **Molecular function** (Молекулна функция) – конкретната биохимична функция на протеина.

В края на тази секция има хипервръзка към пълната GO анотация (Complete GO annotation) – пълното описание на вашия протеин по йерархичната система на GO проекта.

Секция Binary interactions (Взаимодействия)

Описва двойните протеин-протеинови взаимодействия, в които участва изследваният от нас протеин. Представени са във вид на Таблица, от която можем да научим името и номера за достъп на протеина, с който взаимодейства нашият протеин, броя експерименти, потвърждаващи взаимодействието, хипервръзки към базата данни за взаимодействия IntAct към EBI и забележки (Фигура 2-4).

Binary interactions

With	Entry	#Exp.	IntAct	Notes
itself		15	EBI-297353,EBI-297353	
AKAP12	Q02952	2	EBI-297353,EBI-2562430	
CALM3	P62158	3	EBI-297353,EBI-397435	
Calm3	P62161	6	EBI-297353,EBI-397530	From a different organism.
CBL	P22681	4	EBI-297353,EBI-518228	
Cbl	P22682	2	EBI-297353,EBI-640919	From a different organism.
CYTH2	Q99418	5	EBI-297353,EBI-448974	
EGF	P01133	3	EBI-297353,EBI-640857	
ERBB2	P04626	9	EBI-297353,EBI-641062	
ERBB3	P21860	6	EBI-297353,EBI-720706	
ERBB4	Q15303	2	EBI-297353,EBI-80371	
ERRFI1	Q9UJM3	2	EBI-297353,EBI-2941912	
ESR1	P03372-4	4	EBI-297353,EBI-4309277	

Фигура 2-4. Част от секцията Binary interactions за Swiss-Prot записа P00533.

Секция Alternative products (Алтернативни продукти)

Дава подробно описание на протеиновите продукти, произведени от съответния ген, получени по едно (или комбинация от повече) биологични събития: алтернативно използване на промотор, алтернативен сплайсинг, алтернативна инициация и рибозомно изместване на рамката (ribosomal frameshifting). Дава се основна информация за всяка от съществуващите

алтернативни изоформи на протеина. Тук можете да научите, че интересуващият ви протеин EGFR има четири известни изоформи, че от тях Изоформа 1 е приета за канонична и с какво останалите изоформи се отличават от каноничната. Кликването върху името на всяка изоформа, което е хипервръзка, ви отвежда до секвенцията, а върху линка [Align] ви показва алайнмънт на всички изоформи.

Секция Sequence annotation (Features) – Анотация и свойства на секвенцията

Тук имаме позиционно-зависими анотации – информацията за протеина е прецизно картирана върху неговата секвенция. Ако сте работили с GenBank, концепцията ще ви е позната, а и нещата са доста по-прости при протеините.

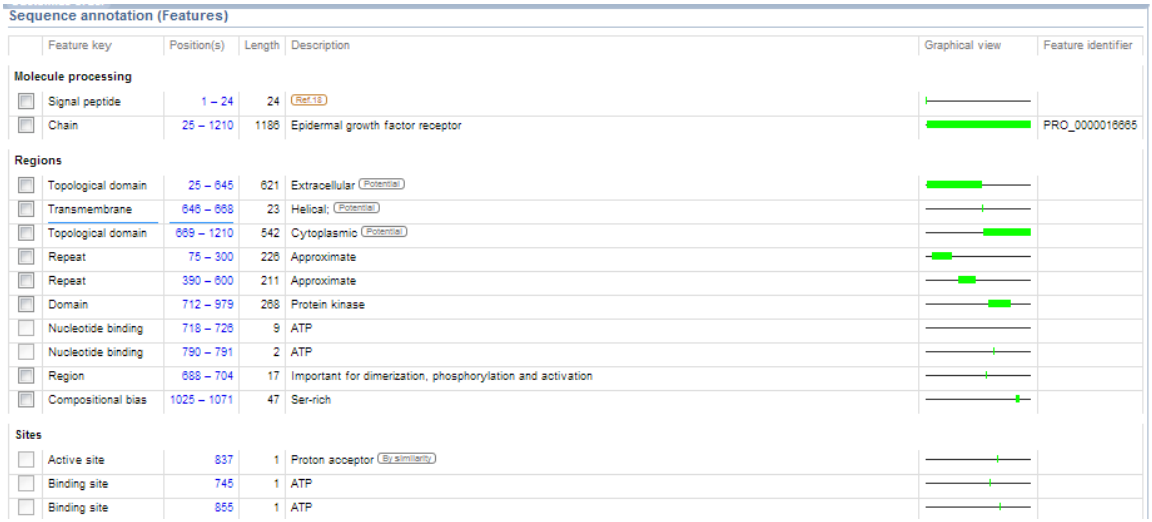
На Фигура 2-5 виждате началото на секцията Features за P00533.

Секцията е организирана като Таблица със следните колони:

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier

В полето **Feature Key** е обозначена характеристиката, която наблюдавате. Цифрите в колоната **Position(s)** обозначават съответно N-крайния и C-крайния амиокиселинен остатък, който се покрива от съответната характеристика, а **Length** ви дава дължината на секвенцията между тях. Полето **Description** ви дава кратко описание на характеристиката/свойството. **Graphical view** ви дава графичен изглед на характеристиките на секвенцията.

Разглеждайки отгоре надолу колоната Feature Key за записа P00533 можем да видим, че характеристиките са обособени в няколко групи: (Molecule processing) – данни за процесинг на протеина; Regions (Региони) – структурни области в протеина, Sites (Сайтове) – функционални участъци; Amino acid modifications (ПТМ), Natural variations (Естествени вариации), Experimental info (Експериментална информация) и Secondary structure (Вторична структура). Началото на списъка със свойства, картирани върху секвенцията на протеина можете да видите на Фигура 2-5.



Sequence annotation (Features)

	Feature key	Position(s)	Length	Description	Graphical view	Feature identifier
Molecule processing						
☐	Signal peptide	1 – 24	24	Ref.18		
☐	Chain	25 – 1210	1186	Epidermal growth factor receptor		PRO_0000016665
Regions						
☐	Topological domain	25 – 645	621	Extracellular Potential		
☐	Transmembrane	646 – 668	23	Helical; Potential		
☐	Topological domain	669 – 1210	542	Cytoplasmic Potential		
☐	Repeat	75 – 300	226	Approximate		
☐	Repeat	390 – 600	211	Approximate		
☐	Domain	712 – 979	268	Protein kinase		
☐	Nucleotide binding	718 – 726	9	ATP		
☐	Nucleotide binding	790 – 791	2	ATP		
☐	Region	688 – 704	17	Important for dimerization, phosphorylation and activation		
☐	Compositional bias	1025 – 1071	47	Ser-rich		
Sites						
☐	Active site	837	1	Proton acceptor By similarity		
☐	Binding site	745	1	ATP		
☐	Binding site	855	1	ATP		

Фигура 2-5. Горната част на секцията Features за Swiss-Prot записа P00533.

- **Signal Peptide:** АК остатъци от 1 до 24 представляват сигнален пептид (нормално свойство на трансмембранните белтъци). Ако кликнете върху линка **Signal Peptide** (това се отнася и за другите линкове в тази колона), се отваря страница с точното значение (дефиниция) на съответния термин. Ако кликнете върху подчертания сегмент от секвенцията (примерно 1 - 24), този сегмент се оцветява в полето за графично показване (display) на секвенцията.
- **Chain**: продължава от N-краен остатък 25 до C-краен остатък 1210 (което означава, че сигналния пептид се изрязва, за да се получи зрелия белтък.
- **Topological Domain (Топологичен домен)**. Обозначава коя част от секвенцията съответства на частта, подаваща се извън клетката (извънклетъчния домен) и коя остава отвътре (вътреклетъчен домен).
- **Transmembrane**: Тук можем да видим, че 23 остатъка (646–668) представляват трансмембранния сегмент от протеина.
- **Domain**: Тук остатъци от 669 до 1210 са обозначени като вътреклетъчен домен на протеина. Забележете, че това, което Swiss-Prot нарича Domain тук, съответства на по-широка топологична дефиниция от това, което обикновено се има предвид като домен в бази данни като Inter-Pro или Pfam. Например, извънклетъчния домен на белтъка EGFR съдържа няколко Inter-Pro домена (проследете линка, за да се убедите сами). По-надолу срещаме ключовата дума Domain да се използва пак в по-тесен смисъл, обозначавайки богат на серин сегмент и областта от секвенцията, отговорна за протеин-киназната активност. Така че използването на този термин в Swiss-Prot е доста многозначно.

- **Nucleotide Binding:** Тук е обозначен сегмента за свързване с нуклеотид фосфат (**n**ucleotide **p**hosphate) – АК позиции от 718 до 726, където се свързва АТФ.
- **Binding site**: Тук можем да видим точния сайт на свързване на АТФ, а също и на лизин (АК позиция 745).
- **Active Site**: Тук е даден сегмента от АК остатъци, включени в активността на ензима.
- **Compositional bias:** Указва сегмент/и от веригата с отклонения в състава (compositionally biased - в случая богати на серин) регион/и в секвенцията на протеина.

Следва информация за АК остатъци, които претърпяват химична модификация след транслацията (Фигура 2-6). Съответните ключове са:

- **Modified residue**: Указва остатъците, подложени на модификация (фосфорилиране).
- **Glycosylation**: Тук можем да видим многобройните АК остатъци, към които е прикрепена въглехидратна молекула, както и типа на връзката (C–, N–, или O–). Забележете, че те всички са в извънклетъчния домен, както и се очаква.
- **Disulfide bond**: Числата тук показват множеството цистеинови двойки, формирани чрез дисулфидни мостове в зрелия протеин.
- **Cross-link** – В тази под-секция на 'Sequence annotation (Features)' се описват различните типове ковалентни връзки, формирани между два протеина (междуверижни връзки, interchain cross-links) или между две различни части на същия протеин (вътреверижни връзки, intrachain cross-links), с изключение на дисулфидните мостове, описани в предната група 'Disulfide bond'.

Секцията Features завършва със серия от по-специализирани характеристики:

Natural variations:

- **Alternative sequence**: Това поле показва промените в секвенцията от аминокиселини, дължащи се на транслацията на различни иРНК (изоформи) на същия протеин, получени чрез алтернативен сплайсинг.
- **Natural variant**: Тук са показани естествените вариации (полиморфизми) в секвенцията на протеина EGFR. Както виждаме на Фигурата, повечето от тях са асоциирани с рака на белите дробове.

Experimental info:

- **Mutagenesis**: За разлика от предишните две, това поле показва промените в секвенцията на протеина, които са били причинени изкуствено в различни експерименти.
- **Sequence conflict**: Тук са посочени разминаванията между различни източници, засягащи една и съща част от секвенцията. Най-често това са грешки или неописани полиморфизми.

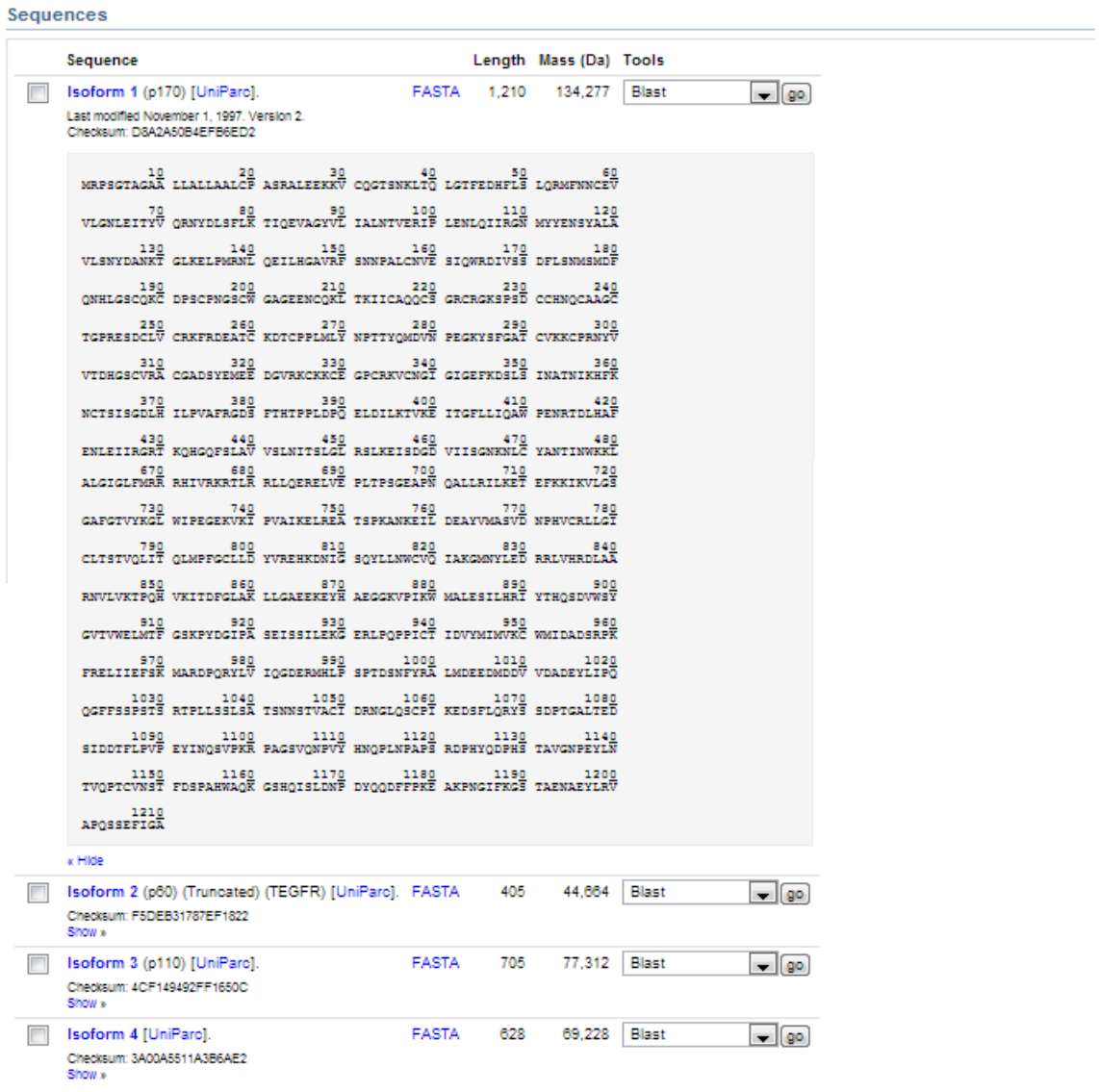
Фигура 2-6. Долната част на секцията Features за запис P00533.

След описаните характеристики, навлизаме в детайлни описания на локалните 3D форми (вторичната структура) на протеина. С цветни ивици е показано графичното им изображение върху секвенцията. Вторичните структури са три разновидности: STRAND (удължена, нишковидна форма (β-нишка), TURN (извивка) и HELIX (спираловидна верига, α-спирала). Кликнете върху линка Details под изображението на вторичните структури, за да видите подробна информация за тях.

Секция Sequences

По подразбиране тук са показани в детайли „каноничната" протеинова секвенция и при поискване - всички други изоформи, описани в записа. Тя включва и друга информация свързана със секвенциите, например дължина и молекулно тегло (Фигура 2-7).

Можете да изтеглите секвенцията в по-удобния FASTA формат, като кликнете върху линка FASTA, разположен на заглавния ред на всяка изоформа. Използвайте опцията File⇨Save As на вашия браузър, за да го запишете, а можете да го копирате и поставите в текстообработваща програма.



Фигура 2-7. Секцията Sequence.

Секция References

Това е списъкът с библиографски източници, използвани в настоящия запис – т.е. източниците на данните, използвани за анотирането на записа. Всеки източник е номериран и съдържа прецизно описание в няколко под-секции. Тук можете да се запознаете с различни изследвания, допринесли за секвенирането, експресията, определянето на функциите, изоформите, посттранслационните модификации и т.н. Важността на този протеин се вижда от големия брой публикации за него.

За всяка публикация е даден NCBI/PubMed идентификатор, а за повечето-директна връзка към абстракта.

Cross-References

Секцията Cross-References съдържа връзки към записи в други бази данни, които съдържат някаква информация за нашия протеин.

Повечето връзки в тази секция ви водят до други бази данни. Прескачайки от база в база, лесно може да се изгубите из мрежата и да изгубите първоначалния си Swiss-Prot запис. За да избегнете това, винаги отваряйте връзките в нов прозорец: кликвате върху линка с десния бутон на мишката и избирате от контекстното меню опцията Open in a New Window. В секцията Cross-References има множество полета. Тук ще споменем само някои бази данни от дългия списък:

- **EMBL**: съдържа всички необходими връзки със света на нуклеотидните секвенции. Много записи на GenBank, EMBL и DDBJ могат да са свързани с една белтъчна секвенция.

- **UniGene:** Линк към базата данни за генна експресия на NCBI (вижте и полето CleanEx за връзки към експериментални изследвания с ДНК чипове).

- **PDB**: Това поле съдържа линк към секвенция, хомоложна на изследваната от нас, за която има достъпна експериментална 3-D

структурна информация (рентгенографски кристални структури, X-ray crystal structures).

- **ModBase**: Тук има връзка към ModBase – база данни с теоретично построени модели, а не експериментално определени структури. Тези модели понякога съдържат значителни грешки.

Полето PDB не присъства, когато за съответния протеин няма налична директна или индиректна структурна информация. Такъв е случаят с голяма част от Swiss-Prot записите.

- **DIP и IntAct**: Линкове към бази данни за междубелтъчни взаимодействия. DIP е съкращение на Database of Interacting Proteins, база данни поддържана от UCLA. Този интересен ресурс дава списък с всички протеини, за които е експериментално установено, че взаимодействат с нашия EGF-рецепторен протеин. Базата данни IntAct е изградена от информация за взаимодействия, открита в литературата.

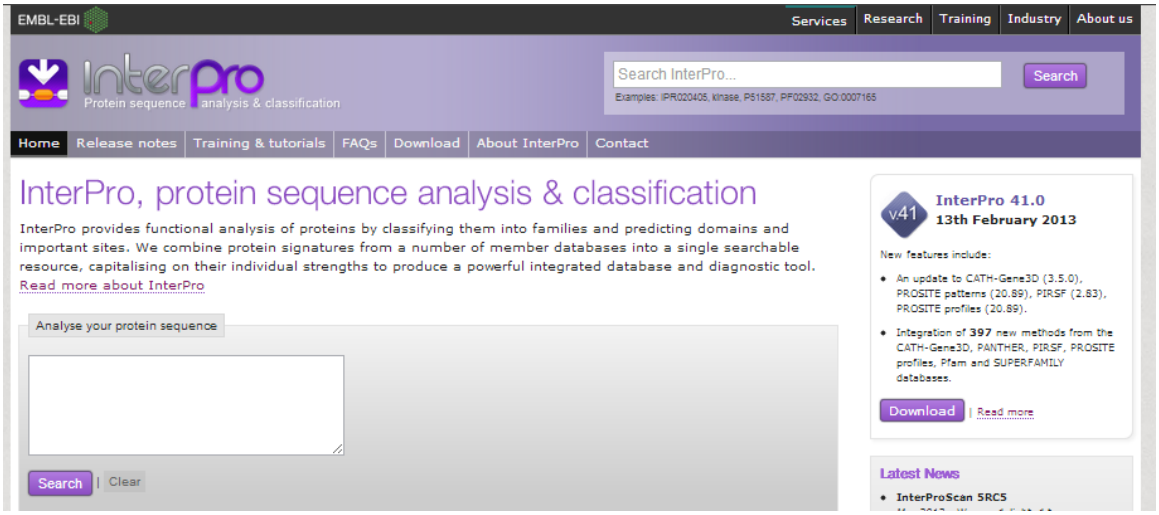
- **GlycoSuiteDb:** показва моделите на гликозилиране (ако има такива). Ресурса, към който води тази връзка е с ограничен достъп (изисква персонална регистрация).

- **SWISS-2DPAGE:** Тук е линка към света на протеомиката. Кликването върху SWISS-2DPAGE ви отвежда към колекция от изображения на 2-D ЕФ гелове от различни човешки нормални и туморни тъкани. Ако вашият протеин е бил идентифициран върху 2-D гел, както е в случая на P00533, съответстващите му петна са осветени. Ако протеина, който търсите не е идентифициран на гел, се появява бокс с информация за предполагаемата му локализация върху гела (базирана на теоретичната му ИЕТ (pI) и предсказаното молекулно тегло). Това много улеснява истинската експериментална работа.
- **HGNC**: Това поле е само за човешки протеини и дава линк към официалната номенклатура на човешките гени, съставена от HUGO номенклатурния комитет/ (HUman Genome Organization) Gene Nomenclature Committee.

- **GeneCards**: Линк към съответния запис в базата данни GeneCards поддържана от Weizmann Institute of Science, базиран в Израел.
- **GeneLynx**: Линк към съответния запис в GeneLynx, колекция от хипервръзки към публично достъпни ресурси, асоциирани с всеки човешки ген.
- **MIM**: Това поле (предимно за протеини от човек и мишка) дава линк към съответния запис в базата данни за генетични заболявания OMIM (Online Mendelian Inheritance in Man).
- **Family and domain databases.** Тази секция към Cross-References съдържа серия полета, свързани с класификацията на протеините на семейства от свързани секвенции на базата на основните домени, които те съдържат. Има много класификации на белтъчни семейства, всички те използват различни методи (алгоритми) за откриване и дефиниране на структурни блокове или мотиви в секвенциите, както и за групиране на протеините в семейства. Съответните бази данни (с линкове) са:
    - Inter-Pro (тя обобщава почти всички следващи);
    - Pfam;
    - Prodom;
    - PRINTS;
    - SMART;
    - PROSITE;
    - BLOCKs.

Ако имате малко време, използвайте линка Inter-Pro, който е сумарен от всички останали (без BLOCKs). В полето срещу Inter-Pro, кликването върху линка Graphical view structure ви показва, че в EGFR белтъка P00533 са били идентифицирани множество повтарящи се домени.

- Ако сте се разходили из повече хипервръзки в тази секция (доста дълга за P00533), сигурно сте открили почти всичко, което е известно за въпросния протеин.



Фигура 2-8. Началната страница на ресурса Inter-Pro към EBI.

- **Entry information**

Този идентификатор ви дава бърза идея за същността на записа.

Entry name EGFR_HUMAN

Accession

- **Първичен номер за достъп (Primary Accession Number): P00533** това е истински уникалния и стабилен идентификатор за тази секвенция. Когато цитирате запис в своя статия, трябва да използвате този номер. Номерата за достъп са много полезни – те позволяват да се проследи историята в базата данни на даден протеин. Например, ако един протеин е твърде дълъг, той може да е бил описан първоначално като частична секвенция (partial sequence), след това са били описвани по-големи части от него и т.н. На всяка стъпка може да е публикувана статия и да е депозирана нова версия на протеина. За разлика от GenBank, Swiss-Prot пази само една (последната) версия за всеки протеин, която обобщава съвременното състояние на знанията за него. Старите номера за достъп не умират, а се съхраняват в полето за вторични номера за достъп (secondary accession number field).
- **Вторични номера за достъп (Secondary accession numbers): O00688, P06268, Q14225, и т.н.** (пълният списък с номера може да се скрива и показва по ваше желание, като кликнете върху иконката). Те позволяват, когато четете стара статия, споменаваща Q12225 или P06268 като номера за достъп за EGFR, все още да можете да използвате този стар източник, за да намерите и по-нови статии. Можете да проверите това, като напишете един от вторичните номера за достъп в прозореца

за търсене. Който и от тях да използвате, винаги ще попадате на страницата на последния – P00533. Това е прост и ефикасен начин да се премахнат старите записи, като все пак те остават еднозначно свързани с последната версия (state-of-the-art version).

Следващото поле **Entry history** съдържа важни дати от работата върху протеина:

- **Integrated into Swiss-Prot on: July 21, 1986.**
- **Last sequence update: November 1, 1997**
- **Last modified: March 6, 2013**

Оттук се вижда, че са били необходими почти 10 години, за да се достигне до цялостната секвенция!

Кликването върху **[Complete history]** ви отваря подробна история на анотирането и интегрирането на протеина.

В целия запис, дефиницията на всяко поле (лявата част на всеки ред) е активна хипервръзка, която ви насочва към съответната част на справочника с помощна информация (Help) на UniProt.

Открийте повече за вашия протеин

Ако внимателно сте разгледали Swiss-Prot записа за протеина, който ви интересува и сте проследили всички хипервръзки, вероятно вече ви е известно 90% от всичко, което светът знае за този протеин.

Следващият ви проблем е да намерите смисъл, да интегрирате или просто да разберете всичко това. Идва момент, когато ще ви е нужна допълнителна информация за по-добро разбиране на всичко, което сте прочели в Swiss-Prot. В следващите секции ще ви покажем някои много полезни ресурси в тази връзка.

Научете повече за биохимичните и сигнални пътища

Таблица 2-1. Някои онлайн ресурси за биохимични и сигнални пътища.

Адрес	Описание
http://www.genome.jp/kegg/	Известната Kyoto Encyclopedia of Genes and Genomes (KEGG).
http://www.brenda-enzymes.org/	Информационната система за ензимна информация BRENDA – много полезна за реални експерименти.
http://www.chem.qmul.ac.uk/iubmb/	Официалния сайт за ензимна номенклатура към International Union of Biochemistry and Molecular Biology (IUBMB).
http://www.ecocyc.org/	Енциклопедията за гени и метаболизъм на *E. coli*, сега прогресивно се разпростира и обхваща и други бактерии.

Научете повече за белтъчните структури

Ако сте анализирали своя протеин на ниво секвенция по всички възможни начини, следващия ви въпрос вероятно ще е как са разположени определени аминокиселинни остатъци в 3-D структурата на вашия протеин – дали са на повърхността или във вътрешността; дали са близо или далеч от активния център, или от други определени АК остатъци? Списъкът с такива въпроси може да е доста дълъг. Всички те изискват структурна информация

Таблица 2-2 ви дава начална информация за някои основни ресурси на тази тема.

Таблица 2-2. Бази данни за структурна информация.

Адрес	Описание
http://www.rcsb.org/pdb/home/home.do	PDB, официалната база данни – хранилище за белтъчни 3-D структури.
www.ncbi.nlm.nih.gov/Structure/	Тук има и инструменти за визуализация и сравнителен анализ.
http://scop.mrc-lmb.cam.ac.uk/scop/	SCOP - Structural Classification Of Proteins, осигурява описания на структурните и еволюционни връзки между всички протеини с известна 3-D структура.
www.cathdb.info	CATH (Class, Architecture, Topology, Homologous superfamily) – също съдържа йерархична класификация на белтъчни доменни структури.
swissmodel.expasy.org	Swiss-Model е напълно автоматизиран сървър, който моделира белтъчни структурни хомологии, достъпни чрез сървъра ExPASy.

Научете повече за белтъчните семейства

Някои белтъчни семейства са толкова огромни и за тях е събрана толкова много информация, че са се превърнали в цели отделни светове. Винаги идва момент, когато общите протеинови бази данни просто не могат да се справят с огромното количество детайлна информация, необходима на специализираната научна общност. Затова са създадени високо специализирани информационни ресурси, които да запълнят тази ниша. Ако вашият любим протеин принадлежи към някоя от следващите категории, може да изпробвате някои от ресурсите, изброени в Таблица 2-3.

Таблица 2-3. Някои специализирани протеинови бази данни.

Адрес	Описание
http://rebase.neb.com/rebase/rebase.html	REBASE – restriction/modification enzyme database - база данни за рестриктази и модифициращи ензими.
www.cazy.org/CAZY/	CAZy – информационен ресурс за ензими, които разграждат, създават или модифицират гликозидни връзки.
merops.sanger.ac.uk	MEROPS - база данни за протеазите.
www.nursa.org	Nursa (Nuclear Receptor Signaling Atlas) – чудесен информационен ресурс за ядрени рецептори, а също и за техните коактиватори, корепресори и лиганди.
senselab.med.yale.edu/senselab	База данни за човешкия мозък – съдържа информация за протеини включени в неврони процеси – йонни канали, рецептори, невротрансмитери, невромедиатори, обонятелни рецептори и др.
www.ncbi.nlm.nih.gov/COG	Базата данни COG (Clusters of Orthologous Groups) прегрупира протеините в ортоложни групи според съдържащи се в тях древни консервативни домени, споделени от поне 3 основни филогенетични линии.

Някои биохимични сайтове за напреднали

Ако искате да освежите познанията си по биохимия, може да посетите някои от следните сайтове:

- Богата база данни с информация за липиди: **Lipid Bank –** lipidbank.jp
- База данни за биокатализа и биодеградация: http://umbbd.ethz.ch/index.html
- **Бази данни за лекарствени, токсични и други молекули:**
  - **ChemIDplus –** chem.sis.nlm.nih.gov/chemidplus/ и
  - **TOXNET -** http://toxnet.nlm.nih.gov/index.html

Сравняване на двойка секвенции (pairwise alignment)

В тази част ще разгледаме следните теми:

- Извършване на dot plot в Интернет;
- Идентифициране на повторения в протеинова секвенция;
- Локален и глобален алайнмънт;
- Намиране на най-добрите локални алайнмънти между две секвенции;
- Управление на параметрите за разкъсване на секвенция и матрици на заместване.

Истинската стойност на нещата се определя изцяло от това, с какво я сравнявате.

Боб Уелс

В тази част ще покажем някои от методите за сравняване на две протеинови или ДНК/РНК секвенции. биоинформатиците наричат тези анализи сравняване по двойки, тъй като те се отнасят за двойка секвенции. Тези анализи са изключително полезни и най-често се използват в следните случаи:

- Дали две секвенции са в действителност хомоложни;
- Дали вашите секвенции притежават общи домен или регион;
- Определяне на точното местоположение на общи характеристики, като например бисулфидни мостове или каталитично-активни места;
- Сравняване на два гена и/или неговите продукти.

Обикновено биолозите използват сравняване по двойки (pairwise alignment), за да усъвършенстват резултатите от конкретно търсене в база данни и да направят подробен анализ. Използват се техники отнасящи се за двойки секвенции в базата данни, за да се провери дали някои от докладваните съвпадения са наистина интересни или само изглеждат такива. В тази глава ще покажем кой метод е най-подходящ за вашите конкретни цели и как може да използвате интернет-базирани приложения, за да стартирате даден метод със секвенциите, които ви интересуват.

Уверете се, че имате правилните секвенции и подходящите методи

Когато правите анализ с двойка секвенции, най-деликатният момент е изборът на самите секвенции за сравнение. Избирането на секвенциите е нещо като организиране на боксов мач между двама опоненти. Целта е да получите най-вълнуващата битка.

Не използвайте методите за сравнение на двойки, за да откриете секвенция, която би могла да бъде хомоложна на секвенция, която вече имате. Това отнема твърде много време. Ако искате да сравните вашата секвенция с всяка една налична в базата данни е по-добре да използвате BLAST (за повече информация относно BLAST, виж част “Търсене на хомоложни секвенции в бази данни”). Ако вече сте прегледали тази част може да твърдите, че програмите за търсене в база данни просто правят сравнение между секвенцията заявка и всички секвенции в базата данни, така че няма реална нужда да се прави допълнително сравнение по двойки! Вярно е, че програми като BLAST търсят в базите данни чрез подобни методи, но те са оптимизирани за скорост при търсенето, а не за точност при алайнмънта.

Програмите, които описваме тук са с точно обратните характеристики. Те са оптимизирани за предоставяне на възможно най-точен резултат, което е причина да се прилагат към внимателно подбрани последователности.

Избор на подходящи секвенции

Добра причина да се прави сравняване по двойки между две секвенции е предположението, че тези секвенции са хомоложни и имат общ произход. Тези последователности често имат сходни 3-D структури и функции. Най-добрият начин да се намери последователност, която е хомоложна на ваша секвенция е да се търси в базата данни с BLAST програма. Избирайте вашите секвенции по следните (консервативни) критерии:

- ДНК секвенция: Най-малко 70% идентичност по дължината на повече от 100 бази между заявката и хита или наличие на E-стойност по-ниска от 10^{-4} .

- Протеинова секвенция: Повече от 25% идентичност по дължината на повече от 100 аминокиселини между заявката и хита или наличие на E-стойност по-ниска от 10^{-4}.

За съжаление тези критерии са само индикация за съвпадение, но не са гаранция за хомология. Ако дадено попадение се намира на границата, то може би не е хомоложно на вашата заявка или може да бъде толкова далечен родственик, че осъществяването на коректен алайнмънт би било много трудно. Това е точката, в която методите за сравняване по двойки отговарят на следното - да решите колко значимо е попадението.

Търсенето в база данни е като сканиране на обяви за нов апартамент ☺. Може да намерите интересни неща, но не може да бъдете сигурни, че сте открили перфектното (или може би просто адекватно) място за живеене, докато не го посетите. "Помещение с климатик и изумителна гледка" може в действителност да бъде порутено помещение с дупки в стената, разположен на върха на промишлена сграда. Избор на секвенция от BLAST резултат е като да направите компромис между нова информация и сигурност. Вие искате секвенция достатъчно подобна на вашата заявка, за да сте сигурни, че е хомоложна. От друга страна искате тази последователност да бъде интересна от биологична гледна точка, така че сравнението да разкрива нещо ново за вас. Или накратко казано, обекта на желанието ви (например протеинова секвенция) трябва да е широко анотирана Swiss-Prot последователност, която отговаря на вашата заявка с E-стойност много по-ниски от 10^{-4}. В реалният живот често се налага да правим компромиси. Една от възможностите е да изберете вашите секвенции от резултата, получен от базата данни. Друга възможност е да изберете вашите две секвенции с помощта на експериментални критерии, например да имат еднаква функция или други подобни свойства. За тази цел може да използвате Sequence Retrieval System (SRS), която ви дава възможност да идентифицирате секвенции по ключови думи. Ако имате само една секвенция е възможно (а понякога и препоръчително) да се направи сравнение по двойки със самата нея. Всичко, което трябва да направите е да се преструвате, че втората последователност е идентична с първата. Това може да изглежда малко глупаво, но е добър начин да се открият интересни свойства в дадена последователност, като например:

- Повторени домени;

- Региони с многократно повтарящ се малък мотив (ниско ниво на сложност);

- Палиндроми (ДНК фрагменти, които се повтарят в противоположни посоки) и потенциални вторични структури в РНК.

Избор на подходящ метод на сравняване

Една от най-трудните задачи за биолозите е сравняването на секвенции. Реално този проблем все още не е напълно решен. Таблица 6-1 показва три метода и подходящите решения за всяка една от ситуациите.

Таблица 6-1. Различни видове сравнения по двойки секвенции.

Име на метода	**Положение**
Dot plot	Общо проучване на вашата секвенция. Откриване на повтори. Намиране на дълги инсерции/делеции. Извличане на части от секвенции, с цел осъществяване на множествен алайнмънт.
Локален алайнмънт	Сравняване на секвенции с частична хомология. Сравняване на части от секвенция. Осъществяване на качествени алайнмънти. Осъществяване на анализи от типа остатък спрямо остатък.
Глобален алайнмънт	Сравняване на две секвенции по цялата им дължина. Идентифициране на дълги инсерции/делеции. Проверка качеството на вашите данни. Идентифициране на всяка една мутация в секвенцията.

Добра стратегия е да се започва с т.н. dot plot (ще дадем всички подробности за dot plot метода в следващия раздел). Dot plot са много мощни инструменти, които ви дават мигновена обща картина при едно сравняване. При наличие на опит, първоначален поглед към dot plot показва какво реално се случва. Това е идеален инструмент за вземане на решение за следващата стъпка от вашия анализ. И все пак, dot plot не е достатъчен за по-прецизно проучване. Dot plot наистина не провежда алайнмънт. Просто дава общи указания. Той не показва

еднозначно как аминокиселините или нуклеотидите си съответстват едни с други между двете секвенции. За да извлечем тази важна информация, трябва да направим алайнмънт (сравняване). Има два вида алайнмънт: глобален алайнмънт (където двете секвенции се сравняват по цялата си дължина) и локален алайнмънт (където програмата сравнява само най-подобните части от двете секвенции). Глобалният алайнмънт е по-лесен за тълкуване от локалния. Въпреки това, има смисъл да се извършва глобален алайнмънт само, ако вашите две секвенции са свързани по цялата си дължина и ако те не съдържат много дълги инсерции или делеции. Ако не знаете как да изберете между глобален и локален алайнмънт, винаги е по-добре да ползвате локални методи. Когато сравнявате две секвенции, започнете с dot plot за всяка секвенция срещу нея самата. Този подход прави възможно по-лесно да се идентифицират потенциални повтори в рамките на всяка секвенция.

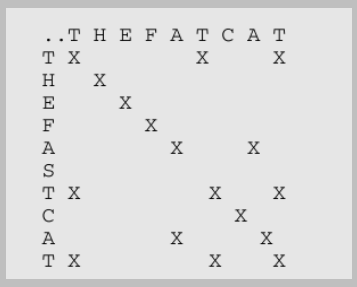
DIY (do it yourself) ръководство за dot plot

Осъществяването на dot plot много прилича на игра на вашите секвенции една срещу друга. Например, представете си, че искате да сравните тези две последователности:

Секвенция 1: THEFATCAT

Секвенция 2: THEFASTCAT

Те са почти идентични, с изключение на допълнителната S в секвенция 2 (разбира се, сложността и трудността е значително по-голяма с реални ДНК или протеинови секвенции). За да се направи dot plot вземете лист хартия, начертайте решетка и напишете отделните букви от секвенция 1 над всяка една колона. След това напишете секвенция 2 от лявата страна с по една буква на всеки ред. Това е моментът, в който започва играта. Разглеждайте всяка клетка една по една и я зачеркнете с кръстче, ако горната секвенция и съответната по реда съдържат еднакъв символ. Когато това е направено, трябва да получите картина подобна на тази показана тук.



Разбира се, както във всяка тик-так-игра, най-дългият диагонал печели! В този случай, те показват сегменти на сходство между двете серии. От матрицата по-горе се вижда много ясно, че двете секвенции са хомоложни по цялата си дължина, с изключение на допълнителния остатък във вертикалната секвенция.

В списък в Таблица 6-1 показваме основните трудности при използване на тези методи, както и подбора на подходящи параметри. За съжаление няма абсолютно правило. Провеждане на точно сравнение е преди всичко процес на опит-грешка. В останалата част от тази глава ще покажем как да променяте тези параметри, докато не получите резултат, който да ви удовлетворява.

Сравняване чрез dot plot

Dot plot е най-лесният начин за сравняване на две секвенции. Всъщност, dot plot, осъществяван само с молив и хартия е бил единственият начин за сравняване на последователности в дните без достъп до компютър (виж "DIY ръководство за dot plot"). Dot plot техниките са тясно свързани с методите на плъзгащите се прозорци, където прозорецът е част от вашата секвенция, която сравнявате с част от друга поредица с подобен размер. Dot plot има две основни предимства:

- Лесно изпълним е и не се изисква никаква биологична хипотеза;
- Трудно е да се направи грешка, когато използвате dot plot.

Ако не сте сигурни какво представлява в действителност dot plot, идея би могло да ви даде бърз поглед към матрицата, показана в "DIY ръководство за dot plot". Тълкуването на dot plot разчита на най-финното устройство за детектиране, с което разполагате - вашите очи.

Когато става дума за определяне на модели (pattern recognition) като диагонали и др., нищо не превъзхожда човешкото око, дори и когато сигналът има шумове. Всеки път, когато видите добър сигнал от dot plot, може да заложите живота си, че нещо интересно става, дори и когато статистиката е твърде "умна", за да види нещо полезно!

Избор на подходящ dot plot

В този раздел ще представим един много мощен инструмент, наречен Dotlet. Dotlet е много удобен, защото е лесен за използване, безплатен, не се нуждае от инсталация и работи на (почти) всеки компютър, който има достъп до Интернет. Dotlet е идеален за последователности с дължина по-малка от 10 000 аминокиселини или нуклеотиди. Така, че може да бъде от полза за повечето протеини, но е с ограничено приложение за малки ДНК секвенции. Ако искате

да разглеждате по-дълги секвенции онлайн, ще трябва да използвате Dnadot (Таблица 6-2), инструмент, предназначен за dot plot между дълги секвенции.

Ако вашите секвенции са по-дълги от 100 000 знака, няма на разположение подходящ онлайн инструмент. Налага се да инсталирате по-мощна версия на Dotlet на вашия компютър. Най-малко два такива пакети са достъпни свободно: Dotter и Dottup. Можете да намерите съответните им спецификации (и URL адреси) в Таблица 6-2.

Таблица 6-2. Различни Dot plot програми.

Име:	**Използва се за:**	**Диапазон:**	**URL:**	**Платформа:**
Dotlet	протеини, ДНК	10000	**http://www.ch.embnet.org/**	всички (Java)
Dnadot	протеини, ДНК	100000	**http://arbl.cvmbs.colostate.edu/molkit/dnadot/**	всички (Java)
Dotter	протеини, ДНК	100000	**http://arbl.cvmbs.colostate.edu/molkit/dnadot/**	Unix, Linux, Windows
Dottup	цели геноми, ДНК	> 100000	**http://emboss.sourceforge.net/**	Unix, Linux

Използване на Dotlet online

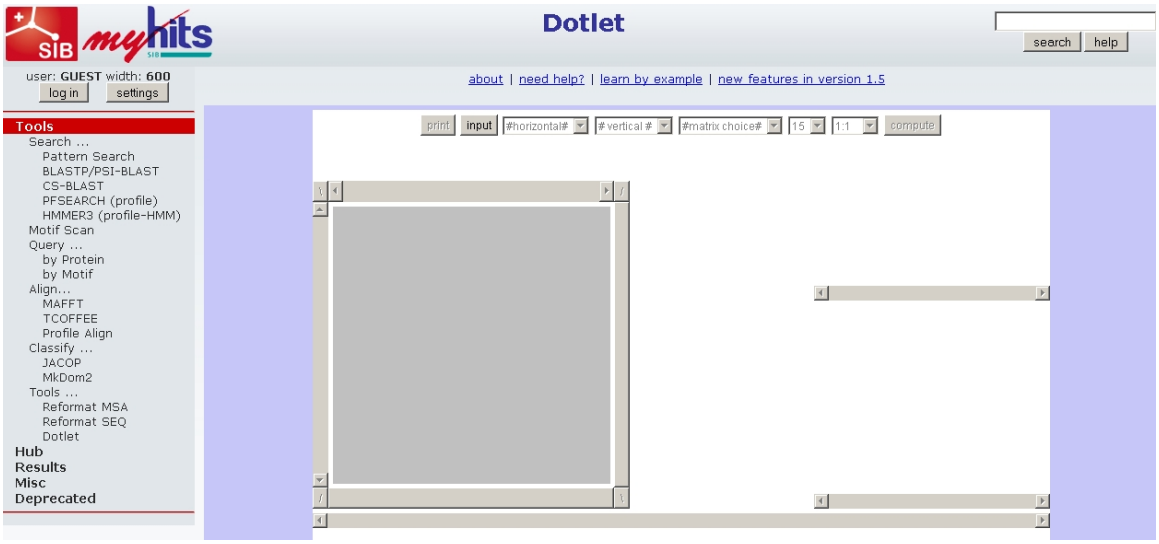
Dotlet е един от най-лесните за dot plot програми достъпни в Интернет. Това е Java аплет, до който имате достъп на сървъра EMBnet. Аплета е малка програма, която се зарежда автоматично във вашия Интернет браузър и след това работи на вашия компютър. Хубавото е, че не е необходимо да инсталирате нищо, стига да имате Java (което повечето интернет браузъри правят сами по подразбиране).

Dotlet не е сложна програма, но притежава много функции. Ако не използвате правилно тези функции, може да преминете много близо до резултата, който ви трябва, но да го пропуснете. Основно правило при използване на Dotlet е, че вие искате да намерите един ясен сигнал и ще трябва да променяте параметрите, докато не го получите. В списъка от стъпки, който ви даваме малко по-долу, ще покажем какво разбираме под термина ясен сигнал и ще ви дадем съвети за това как да се получи такъв сигнал.

Стартиране на Dotlet

Може да стартирате Dotlet като позиционирате вашия браузър на адрес: http://myhits.isb-sib.ch/cgi-bin/dotlet

Страницата, която се показва, изглежда като показаната на Фигура 6-1. Ако браузърът ви не показва правилно тази страница, това означава, че не може да стартира Java. За да работи Java на вашия компютър, може да се наложи актуализация на браузъра с най-новата версия.



Фигура 6-1. Страницата на Dotlet.

Въвеждане на секвенция в Dotlet

В списъка от стъпки по-долу сравняваме две секвенции, които съдържат специфичен киназен домен:

1. Отворете нов прозорец на браузъра, като натиснете Ctrl+N от клавиатурата.

Имате нужда от едновременен достъп до два прозореца на вашия браузър, което е причината да започнем с тази стъпка.

2. Въведете адрес: http://www.uniprot.org/uniprot/P05049.fasta **в новия прозорец, който отворихте в стъпка 1 и след това натиснете бутон Enter или Return.**

Въведете точният URL адрес. Този URL ви дава директен достъп до Swiss-Prot, ако знаете идентификационния номер (номер за достъп) на вашата секвенция. За нашия пример, идентификационния номер е P05049. След кратко изчакване се появява нов прозорец, показващ секвенцията, свързана с въведения от вас идентификационен номер.

3. Копирайте секвенцията показана във вашия браузър, като използвате клавишната комбинация Ctrl+C.

4. Върнете се до първия прозорец на вашия браузър (прозореца Dotlet, показан на Фигура 6-1) и щракнете върху бутона Input в горния ляв ъгъл. Ще се появи прозорец като този показан на Фигура 6-2.

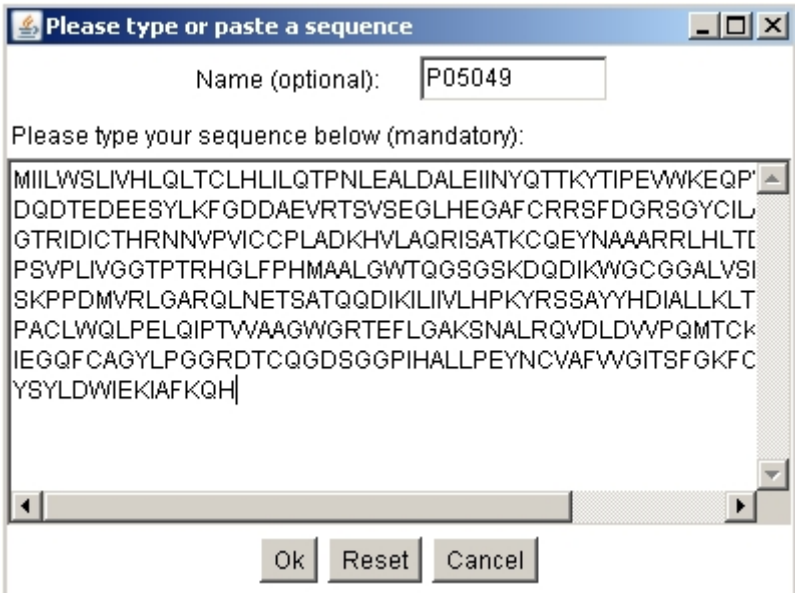
5. Използвайте Ctrl+V, за да поставите вашата последователност в изскачащия прозорец и след това въведете име.

Трябва да въведете секвенцията си в суров формат (само секвенцията), както е показано на Фигура 6-2.

Ако желаете, в полето Name може да въведете името на вашата секвенция за dot plot.

6. Натиснете бутона OK в Sequence прозореца.

Когато натиснете бутона OK, прозорецът изчезва. Това означава, че вашата секвенция вече е заредена в Dotlet. Ако искате да сравните една секвенция със самата себе си, отидете направо на стъпка 13.



Фигура 6-2. Прозореца за секвенция в Dotlet.

7. Заредете адрес: http://www.uniprot.org/uniprot/P08246.fasta **в новия прозорец, който отворихте в стъпка 1 и натиснете Enter или Return. Това е втората секвенция, която искате да заредите в Dotlet.**

8. Копирайте секвенцията, показана във вашия браузър.

9. В прозореца Dotlet (виж Фигура 6-2) щракнете върху бутона Input.

10. Поставете вашата секвенция използвайки Ctrl+V в изскачащия прозорец и въведете, ако желаете име.

11. Щракнете върху бутона OK в прозореца.

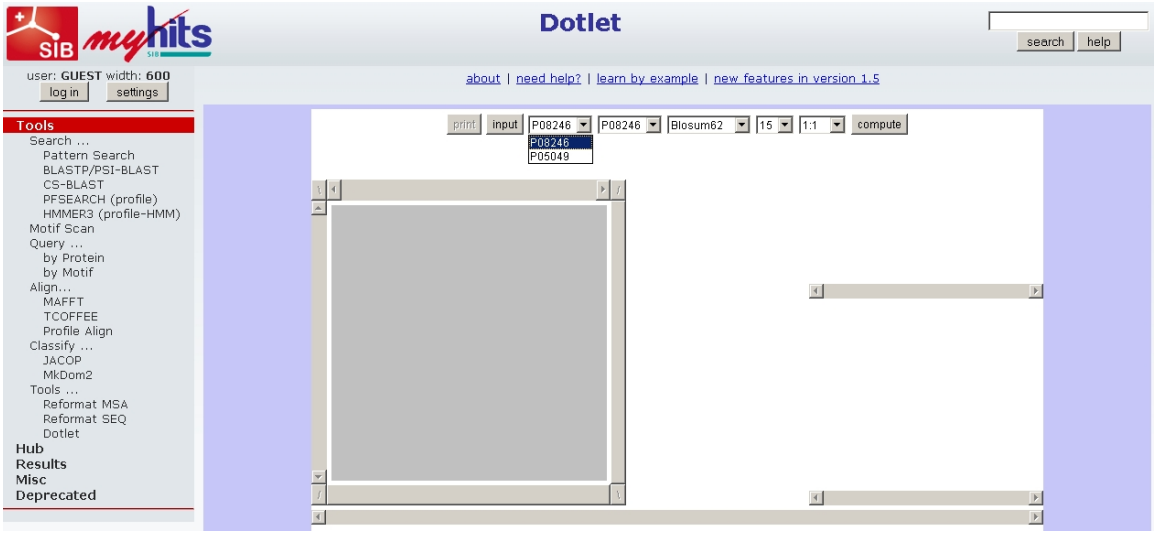
Когато щракнете върху бутона ОК, прозорецът изчезва, което означава, че вашата последователност е заредена в Dotlet!

12. Върнете се обратно в оригиналния прозорец на браузъра (прозореца Dotlet), изберете двете секвенции, които искате да сравните от двата падащи прозореца разположени в близост до Input бутона (Фигура 6-3).

Фигура 6-3 показва как да използвате падащите менюта, за да изберете секвенциите, които искате да сравните. Имената на менютата са посочените от вас в полето Name в стъпки 5 и 10.

• Секвенцията, която избирате в първото падащо меню е хоризонталната последователност (отгоре на dot plot, от ляво на дясно).

• Секвенцията, която избирате във второто меню е вертикалната последователност (лявата страна на dot plot, от горе на долу).



Фигура 6-3. Падащи менюта за секвенцията в Dotlet.

13. В прозореца Dotlet щракнете върху бутона Compute, намиращ се най-вдясно.

Dotlet не предприема никога автоматични действия (изключение е само, когато промените дадена прагова стойност). Всеки път, когато променяте параметри, трябва да уведомите Dotlet, като щракнете върху бутона Compute.

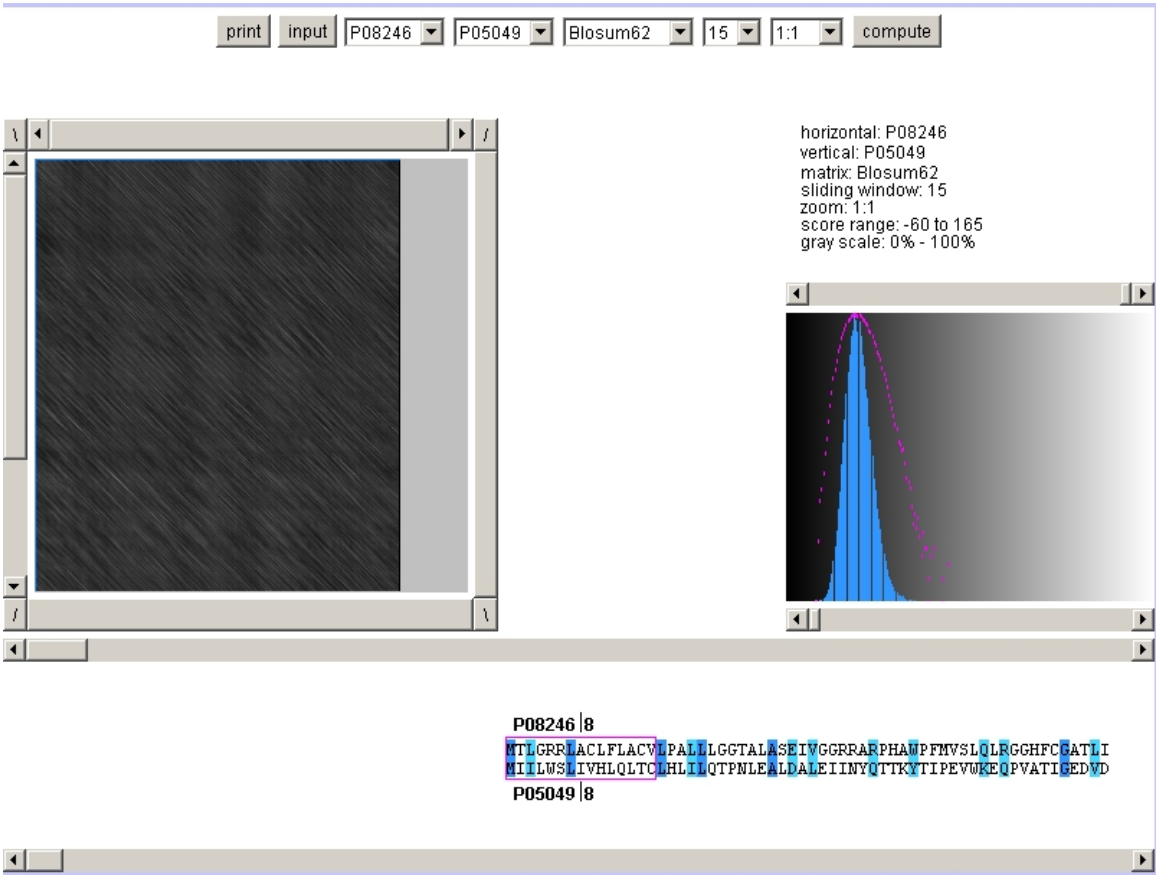
Резултатът изглежда, като този показан на Фигура 6-4. От нея се виждат трите Dotlet прозореца:

• **dot plot прозорец:** Прозорецът вляво, съдържащ вашия dot plot.

• **прагов прозорец:** Малкият прозорец вдясно, който трябва да използвате, за да настроите чувствителността на вашия dot plot.

• **прозорец на алайнмънта:** Дългият тесен прозорец по-долу, който показва кои двойки остатъци от двете секвенции съответстват на всяка точка в dot plot прозореца.

Резултата от Dotlet е труден за интерпретиране. За да се извлече някакъв смисъл, трябва да се усъвършенства. Ще покажем как може да се направи това в следващия раздел.



Фигура 6-4. Резултат в Dotlet.

Детайлна настройка на Dotlet

Инструкциите, които ще представим не винаги водят до оптимален резултат, но следват подхода "отгоре – надолу". Ще започнем с настройките, които са най-общи и ще завършим с настройките, които трябва да прецизират информацията, която получавате от dot plot.

1. Определяне на фактор за мащабиране.

По подразбиране Dotlet има фактор за мащабиране 1:1, което означава, че 1 пиксел съответства на 1 остатък.

Ако вашите секвенции са дълги, не може да видите целия dot plot. Това е показано на Фигура 6-4. Може да разберете, че и във вашия случай е така по наличието на активен плъзгач от лявата страна. Може да използвате правоъгълното квадратче - плъзгач за превъртане надолу, за да видите останалата част от дисплея. За да видите целия dot plot направете следното:

А. При избор на мащабиране от падащото меню в ляво на бутона "Compute", изберете 1:2, както е показано на Фигура 6-5.
Б. Щракнете върху бутона Compute.

Dotlet вече показва пълния dot plot на вашите две секвенции. Полезно е да виждате изцяло резултата, дори ако изглежда доста малък. След като получите обща представа за регионите в dot plot, винаги може да ремащабирате.

2. Изберете размер на прозореца.

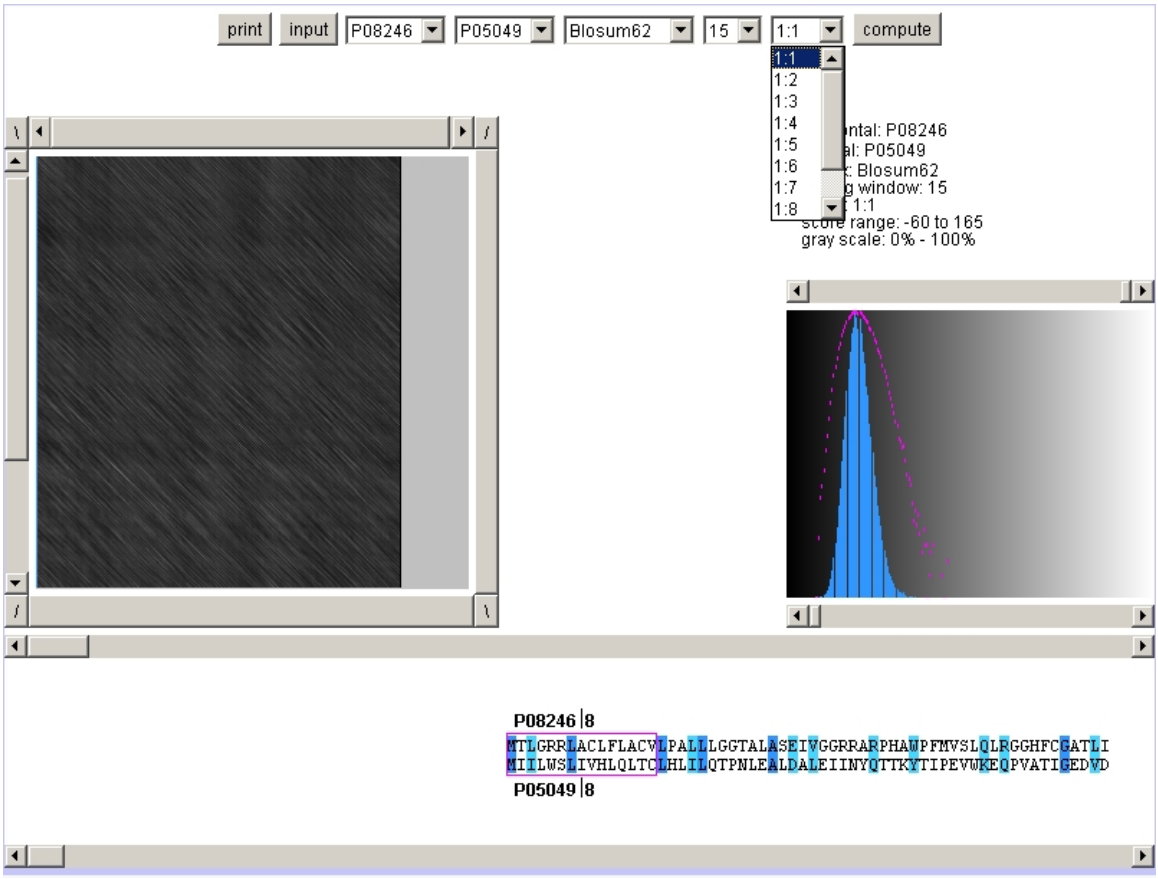
Когато Dotlet сравнява две аминокиселини или два нуклеотида, то той също сравнява и аминокиселините или нуклеотидите в непосредствена близост. Размерът на прозореца контролира размера на тези области (за някои полезни съвети, вижте раздела "Извличане на прозорец с подходящ размер"). За да използвате нашия пример, трябва да промените стойността по подразбиране, както следва:

А. Изберете 51 от падащото меню Window (менюто, намиращо се в ляво на сектора за мащабиране).
Б. Щракнете върху бутона Compute.

След промяна на размера на прозореца трябва да настроите и прага (вижте стъпка 3). Има смисъл да променяте размера на прозореца, ако не коригирате прага.

3. Регулиране на прага.

Прагът е стойност, която определя цвета на всяка точка в dot plot прозореца. Точките с резултат над прага са бели, а тези с резултат под прага са черни. Прагът е най-мощният и най-деликатният параметър в dot plot. За щастие, Dotlet ви позволява да настройвате прага динамично и да видите ефекта от това в реално време (дори не трябва да кликвате върху бутона Compute!).



Фигура 6-5. Избор на фактор на увеличение в Dotlet.

Получаване на рамка с подходящ размер

Ето някои полезни правила за определяне размера на вашия прозорец:

• Дългите прозорци правят вашия dot plot ясен. Дават ви увереността, че две подобни аминокиселини (или два нуклеотида) дават точка само, ако аминокиселините (или нуклеотидите) около тях също са сходни.
Специалистите казват, че дългите прозорци правят dot plot по-строг. С цел да се постигне равновесие в тази насока, може да използвате следните прости правила:

• Размерът на прозореца трябва да бъде в границите на размера на елементите, които търсите. Например е подходящо, ако търсите консервативни домени в протеин с размер от 50 аминокиселини или повече.
• По-късите прозорци са по-чувствителни, но има доста шум, който ги съпътства. Те са удачни, ако търсите домени в протеини с отдалечено родство.

• Удобно е да се започне с голям прозорец и да го мащабирате, докато сигналът, който търсите се появи.

Отгоре на прозореца за определяне на прага (Фигура 6-4), има малък хоризонтален плъзгач с маркер. Това е горния курсор. Отдолу на същия прозорец има друга хоризонтална лента за превъртане. Това е долния маркер.

Хоризонталната ос на прозореца за прага представя резултата. Ниските са наляво, високите са надясно. Големият пик на кривата показва най-често срещаният резултат за два остатъка, избрани на случаен принцип.

Може да си представите прага, като две вертикални линии, преминаващи през горната и долната част на маркера. Неговата позиция контролира външния вид на dot plot прозореца (големият прозорец в ляво на Фигура 6-4). В този dot plot прозорец, може да видите:

• Точки с резултат под долния праг в черно;

• Точки с резултат над горния праг в бяло;

• Точки с резултат между двата прага в сиво.

Всичко се свежда до определяне на прага. По този начин интересните остатъци се появяват като бели точки на черен фон в dot plot прозореца в ляво. Може да използвате следната процедура за постигане на тази цел:

А. Преместете долния курсор на прозореца за прага изцяло в дясно.

Това прави dot plot в ляво изцяло черен.

Б. Хванете дясната страна на прозореца за прага до средата и бавно го плъзнете наляво, като държите десния бутон на мишката натиснат.

Докато плъзгате, следете dot plot, за да видите дали се появява консервативния сегмент, както е показано на Фигура 6-6.

Когато горният и долният курсор са най-отгоре, точките в dot plot са или черни, или бели. Ако плъзнете един от двата курсора/маркера наляво или надясно, точките, чиито резултати са между тези два прага се появяват в сиво. Това може да помогне за по-добрата визия на dot plot и да стане по-подходящ за публикуване.

Ако искате да видите черни точки на бял фон, плъзнете горния курсор леко в дясно спрямо долния курсор. Това води до инверсия на цвета на Таблицата.

4. Записване на dot plot.

Записване на резултата е най-деликатната част от Dotlet. Не всички браузъри позволяват да отпечатвате резултатите. Не можем да намерим браузър, който да съхрани информацията на дисплея като файл. Доколкото знаем, има само два начина да се запази информацията на дисплея:

Принтиране на съдържанието на браузъра във файл в PDF формат. За да направите това, трябва да имате PDF принтер, инсталиран на вашия компютър.

Използване на screen-capture техниката. Можете да я използвате винаги при опит да съхраните графична информация, която е проблемна. Следвайте следната процедура, за да съхраните информацията на екрана:

A. Натиснете бутона PrntScrn на клавиатурата (обикновено разположен в горния десен ъгъл).

B. Стартирайте Power Point презентация.

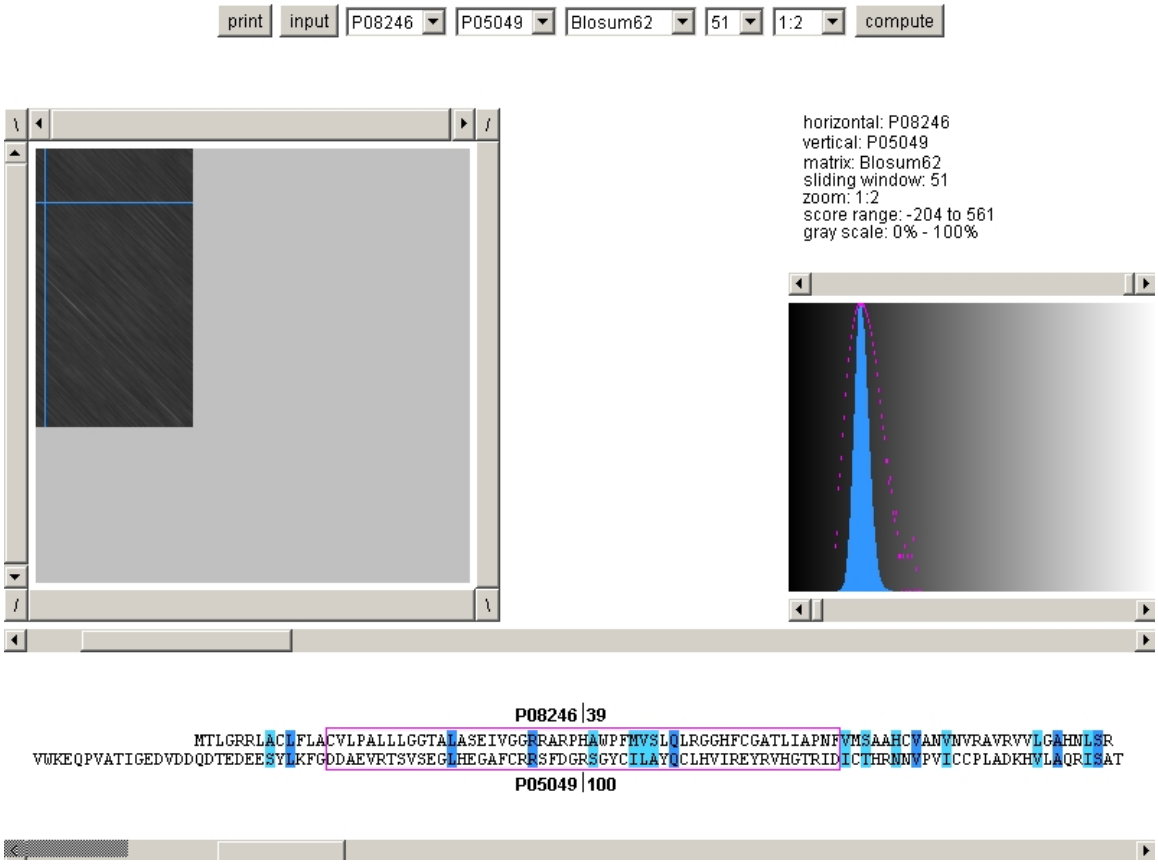
C. Натиснете Ctrl+V, за да изрежете и поставите информацията в презентацията.

Сега може да отпечатате dot plot или да го вмъкнете във всеки тип документ.

Тълкуване на резултатите

Двата протеина, които избрахме (P05049 и P08246) в предходния списък от стъпки (в раздела "Въвеждане на последователност в Dotlet") са силно отдалечени родствено. Ако проведете BLAST между тях, получавате съвпадение с E-стойност 10^{-4}. Ако сте се запознали със съдържанието в главата за търсене на секвенции в БД (там се разглежда подробно BLAST) сте разбрали, че не е добра идея да се радвате от такова попадение. Този резултат дори може да бъде фалшиво положителен.

Знаейки, че P05049 е серинова протеаза, би било интересно да се провери функцията на протеаза P08246 в лабораторни условия. Проблемът е, че когато започнете такъв скъп експеримент на база на такова слабо BLAST попадение, то единствено може да разчитате на шанс. Ако не успеете в този експеримент, има голяма вероятност вашият шеф да ви зададе неудобния въпрос: "Мислиш ли въобще какво правиш!?"



Фигура 6-6. Dotlet с правилно подбрани параметри.

Преди да се направи такъв лабораторен експеримент трябва да се уверите, че P05049 и P08246 са действително хомоложни в определена област.

Провеждане на dot plot е подходящ, за да се отговори на този въпрос. Фигура 6-6 показва съществуване на консервативни области между двата протеина. Може да се убедите в това, като погледнете тяхната анотация в Swiss-Prot (достъпна на адрес www.expasy.ch). Swiss-Prot показва, че и двата протеина съдържат такъв серинов протеиназен домен в областта, която определя и dot plot.

Разбира се, в друг случай може да няма анотации в Swiss-Prot. В действителност може да няма никъде анотация! Това не трябва да ви спира да направите dot plot. Идентифициране на консервативни домени между два протеина (дори и ако не знаете нищо за функциите им) е добър подход, за да се намери конкретен регион, отговорен за дадена функция.

Биологичен анализ с помощта на dot plot

Предходният пример показва как да се открият родствено свързани домени в два протеина. Dot plot дава възможност за представяне на важни биологични връзки. Използването им е въпрос на познаване на няколко графични модела и на кои ситуации отговарят те.

В следният пример ще покажем някои от тези ситуации. Предполагаме, че сте въвели вашите секвенции в Dotlet по процедурата описана в раздела "Използване на Dotlet в Интернет".

Определяне на тандемни повтори

Протеините много често съдържат къси многократно повтарящи се домени. Вътрешното дублиране е инструмент, който често се използва от еволюцията за създаване на нови протеини или такива, които функционират по-ефективно. Домени, които често може да откриете дуплицирани са: EF-hands (свързват калций) или Zn-finger (участват в ДНК свързване).

Dot plot е най-добрият начин да се идентифицират повторени домени. На Фигура 6-7 показваме последователността на YE73 - човешки протеин. YE73 е потенциален транскрипционен фактор (участващ в транскрипцията на РНК), който съдържа 13 домена с повторени Zn-finger. За да анализирате тази последователност, следвайте следните стъпки:

1. **Заредете следният адрес във вашия браузер:**
http://www.uniprot.org/uniprot/Q9P255.fasta

Q9P255 е номера за достъп на YE73 човешката протеинова секвенция.

2. **Въведете секвенцията в Dotlet, следвайки процедурата, представена в раздел "Използване на Dotlet в Интернет."**
3. **Задайте праг и настройки на прозореца, така че да съвпадат с тези на Фигура 6-7.**
Задайте размер на прозореца 21, мащабиране 1:2 и 39% за сивата скала, за горния и долния курсор на прозореца за прага.

Фигура 6-7 е типична за тандемните повтори. Обърнете внимание на следните характеристики:

• Основният диагонал представлява секвенцията срещу самата себе си;

• Повторенията се появяват като дълги непрекъснати диагонали над и под главния диагонал;

• Диагоналите са равномерно разпределени;

• Диагоналите са ограничени от квадратна форма.

Ако видите този вид графичен модел, трябва да знаете, че гледате тандемни повтори. Ако е така, може да използвате тази колекция от диагонали, за да направите три интересни извода:

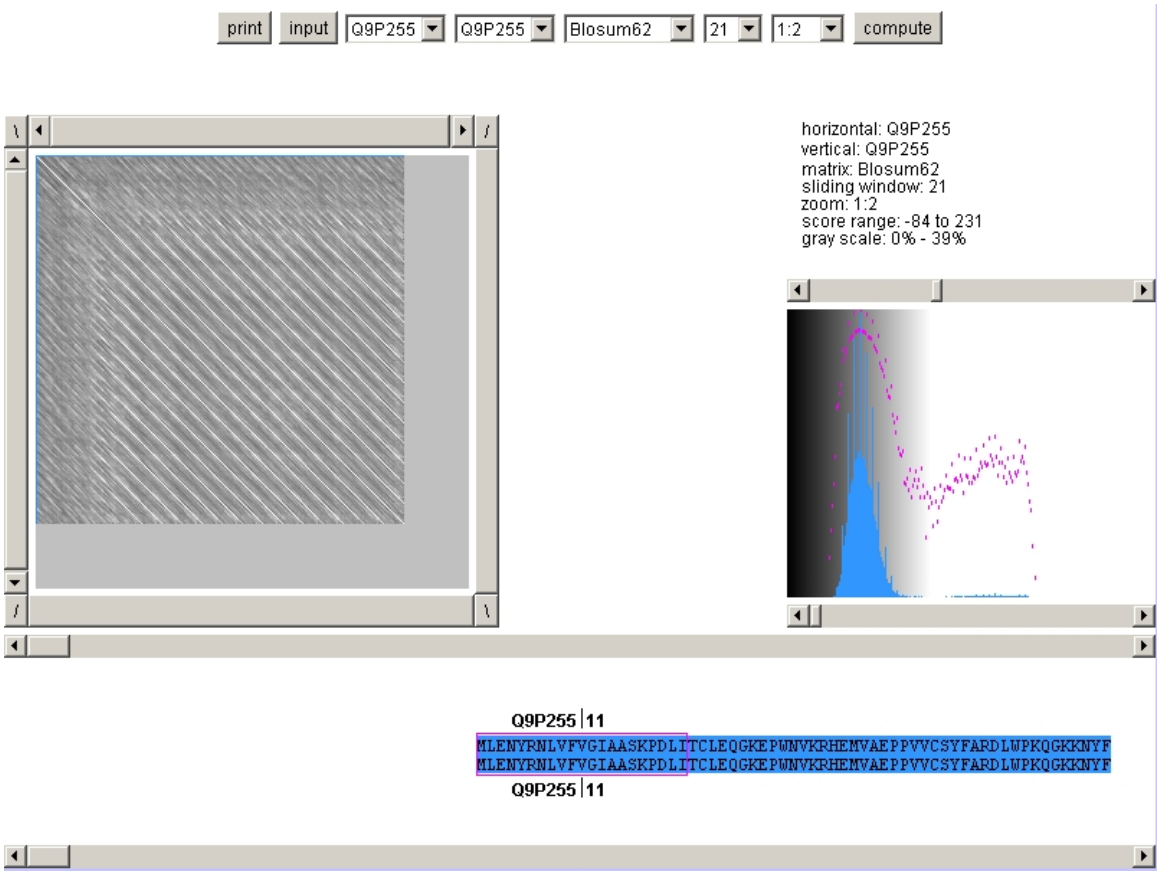
- Броят на повторенията е равен на броя на диагоналите по-горе (или по-долу) на главния диагонал (включително и на самия главен диагонал);
- Разстоянието между два съседни диагонала показва размера на повтора;
- Най-късият диагонал ви дава координатите на еднократно повторен елемент.

YE73 е краен пример, защото съдържа силно консервативни повтори. Нещата не винаги са толкова ясни и в повечето случаи многократно повторените домени не са толкова идентични (поради натрупани мутации в тях). Протеини, които съдържат родствено отдалечени повтори, продуцират по-сложни dot plot модели. В тези протеини, всяка повторена единица не е задължително да разпознава всяка друга единица.

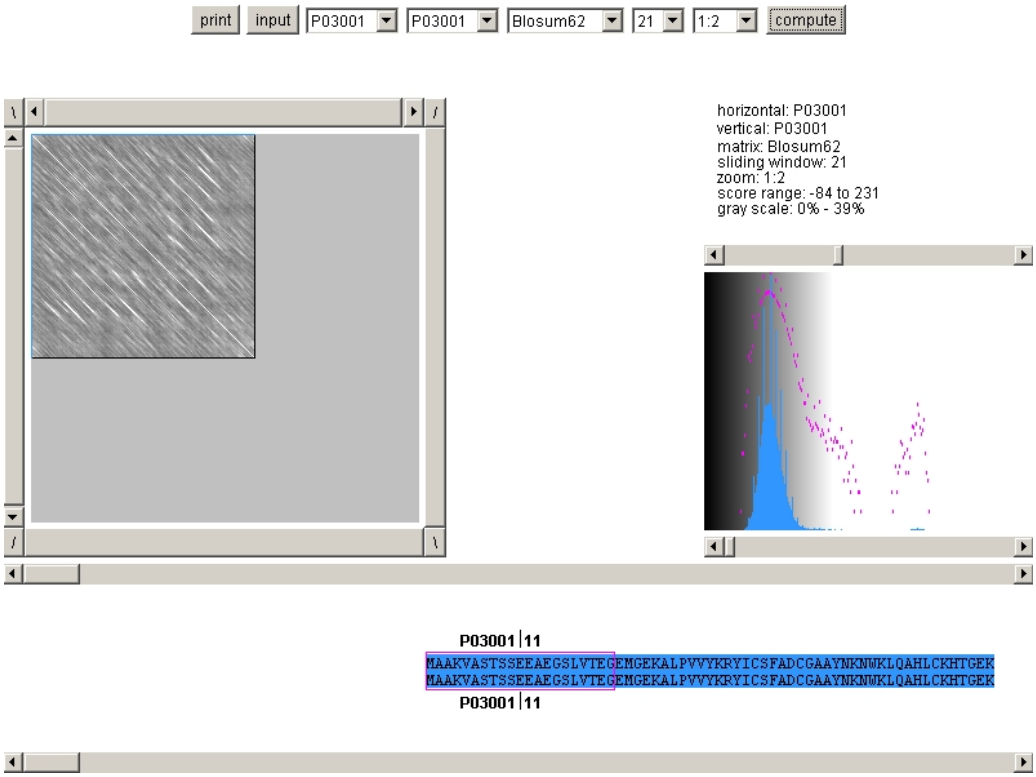
Фигура 6-8 показва един такъв пример. Tf3a е транскрипционен фактор, който също съдържа тандемни Zn-finger домени. Може да намерите секвенцията на адрес:

http://www.uniprot.org/uniprot/P03001.fasta

Tf3a разкрива историята на домени, съдържащи Zn-finger, които първоначално са дуплицирани и след това дивергирали. Тези Zn-finger са по-малко консервативни и показват модел по-малко симетричен от този, получен при YE73.



Фигура 6-7. Разглеждане на тандемните повтори с Dotlet.



Фигура 6-8. Разглеждане на слабо консервативни тандемни повтори на Tf3a с Dotlet.

Тандемните домени не са единствената форма на повтарящи се домени при протеини и ДНК. Описанието на повтори е винаги едно и също: Диагонал, който се намира встрани от главния диагонал, когато сравнявате дадена секвенция с нея самата.

Повторените домени могат да ви помогнат при изясняване функциите на вашия протеин. Ако откриете повторен домен с неизвестна функция и никаква прилика с характеризираните вече протеини, то тогава е добре да направите следното:

1. Извлечете всяка повторена единица;

2. Направете множествен алайнмънт на домените;

Виж следващата глава за повече информация как може да направите това.

3. Определете консервативните места в домена;

4. Включете вашият домен в PROSITE модел или профил.

Виж предната глава за повече информация относно PROSITE.

5. Сканиране на Swiss-Prot, за да се провери дали този модел е свързан с определена функция.

Намиране в протеини на области с ниска сложност

Регионите с ниска сложност (low complexity) са фрагменти, които съдържат някои аминокиселини по-често, отколкото се очаква в нормални протеини. Например, един сегмент от 100 аминокиселини, който съдържа 45 пролина, определено е фрагмент с ниска сложност.

Региони с ниска сложност често имат биологични функции като протеин-протеиново взаимодействие (Leucine Zipper) или неспецифично ДНК/РНК свързване (ARG-богатите области).

Хубавото за тези региони с ниска сложност е, че се виждат много лесно при dot plot. При сравняване на секвенция със самата себе си, регионите с ниска сложност се появяват като квадрати. Може да видите един от тях на Фигура 6-9.

За да се получи резултат като този, който виждате на Фигура 6-9, следвайте следните стъпки:

1. **Заредете на вашия браузър следния адрес:**

http://www.uniprot.org/uniprot/P21997.fasta

P21997 е номер за достъп на сулфатен гликопротеин 185.

2. Въведете секвенцията в Dotlet, следвайки процедурата описана в раздел "Използване на Dotlet online."

3. Задайте настройките така, че да съвпадат с тези показани на Фигура 6-9.

Имате нужда от размер на прозореца 7, мащабиране 1:2 и 44% за сивата скала, за горния и долния курсор на прозореца за прага.

Анализиране на нуклеинови киселини с Dotlet

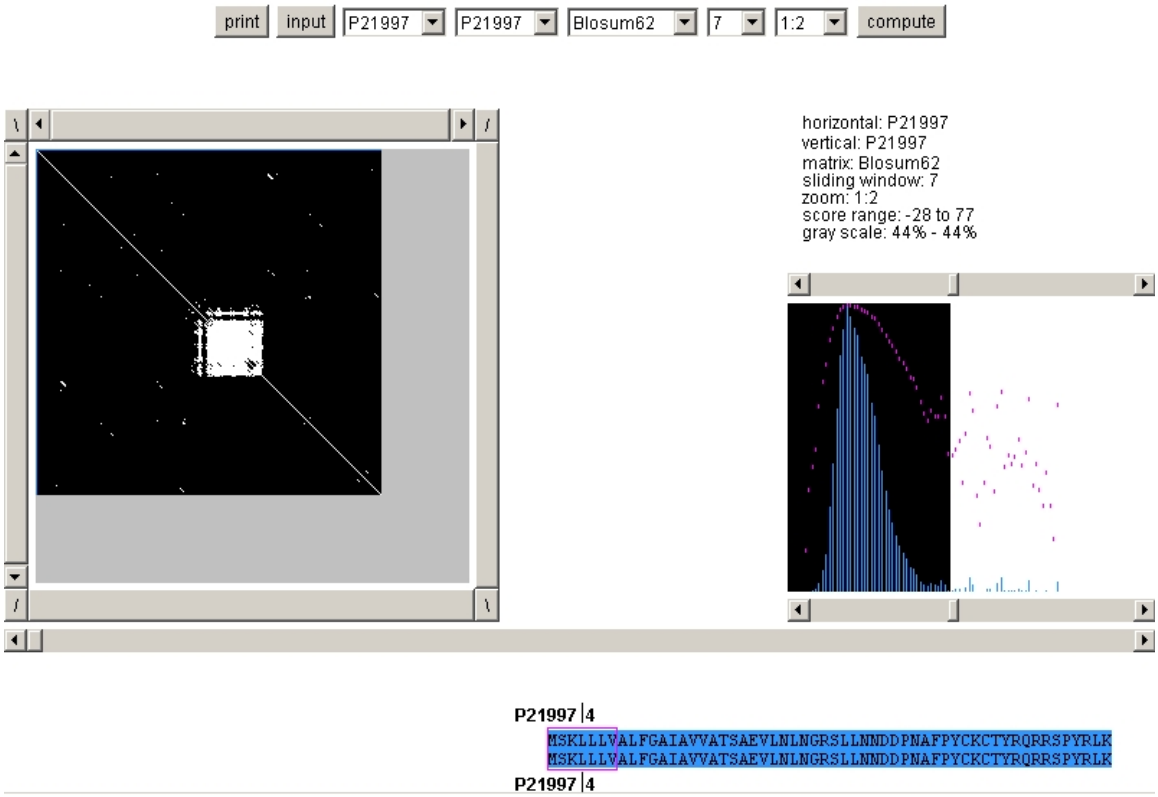
По принцип, dot plot е много подходящ за картиране на гени. За съжаление, Dotlet може да бъде доста бавен за приложения, които изискват рутинно сравняване на последователности с дължина по-голяма от 10000 остатъка.

При анализ на нуклеотидни последователности, Dotlet съдържа две полезни функции:

- Ако сравнявате протеинова секвенция с нуклеотидна последователност: Dotlet автоматично превежда нуклеотидната секвенция в трите и възможни рамки на четене. По този начин, могат да се определят точно екзон/интрон границите.
- Ако сравнявате нуклеотидна секвенция със самата себе си, Dotlet автоматично замества една от поредиците с нейната комплементарна верига. По този начин, ако вашата секвенция съдържа две комплементарни вериги (като фуркетната бримка в структурата на РНК), то те ще се появяват като диагонал в горния десен ъгъл към долния ляв (перпендикулярно на главния диагонал).

Ако искате да използвате Dotlet за по-прецизен анализ на секвенции, препоръчваме отлично онлайн ръководство, създадено от авторите на Dotlet, което е достъпно на адрес:

http://myhits.isb-sib.ch/util/dotlet/doc/dotlet_help.html



Фигура 6-9. Идентифициране на сегменти с ниска сложност с Dotlet.

Извършване на локален алайнмънт

Ако сте се запознали вече с dot plot в тази глава, вие знаете, че провеждането на dot plot е идеален начин да се изследват отношенията между секвенции. Въпреки това, dot plot не са предсказуеми методи. Те ви дават информация, но не и какво всъщност означава тя. За да тълкувате резултатите имате нужда от метод, който осъществява алайнмънт.

Има два вида алайнмънт: глобален (когато двете секвенции са подравнени по цялата си дължина) и локален (където програмата сравнява само най-подобните части от двете секвенции и игнорира останалата част от тях).

Ако имате dot plot, показващ че вашите две секвенции са свързани по цялата си дължина, не е необходимо да правите локален алайнмънт, а може да преминете директно към глобален. Ако не сте сигурни точно какъв алайнмънт да проведете, в този раздел ще ви обясним подробно!

Има две основни причини, за да анализирате вашите секвенции с помощта на локален алайнмънт:

- Да се сравнят две родствено отдалечени последователности, които споделят само няколко несъседни домени/региона;
- Да се анализират повторени елементи в рамките на една и съща последователност.

Методите за локален алайнмънт правят това, което подсказва името им: Вие задавате две секвенции, в резултат се връща алайнмънт на най-подобните части от тези последователности.

Хубаво на методите за локален алайнмънт е, че автоматично премахват регионите от аминокиселините и нуклеотидите, които не могат да бъдат сравнени.

На теория, локалният алайнмънт съответства на един от диагоналите, които се появяват при dot plot. На практика, нещата не са толкова прости. Може да видите сигнал от dot plot, който не съответства на нито един локален алайнмънт, а понякога програмите за локален алайнмънт докладват за алайнмънти, които не фигурират в dot plot. Само когато dot plot и алайнмънта съвпадат напълно, тогава може да сте сигурни.

Избор на подходящ локален алайнмънт

Съществуват два основни метода за провеждане на локален алайнмънт: бърз, евристичен метод, интегриран в програмата BLAST и по-бавен, и по-точен, програмата Lalign. Таблица 6-3 изброява предимствата и недостатъците на тези два метода и как да изберете този, който е най-подходящ за вашите цели.

BLAST, при който се сравняват две секвенции (bl2seq), може да използвате и за търсене в дадена база данни. Той е адаптиран така, че да може да го ограничите само до две секвенции. Той не генерира алайнмънти, различни от тези BLAST резултати, които получавате, когато търсите в дадена база данни. Накратко, ако искате да проверите съвпадение, BLAST-ването на две секвенции не може да ви каже нищо повече от търсене в база данни. Имате достъп до този вид BLAST на адрес:

http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome .

Таблица 6-3. Използване на BLAST или Lalign при сравняване по двойки.

Особеност	**BLAST**	**Lalign**
Скорост	Много бърз	По-бавен

Дължина на секвенциите	Много дълги секвенции	По-къси секвенции
Резултат	E-стойност	E-стойност
Алайнмъмнти	Докладва най-добрият	Докладва най-добрите десет (или повече)
Тип на секвенцията	Най-добър с ДНК	Най-добър с протеини

Използване на Lalign, за да намерите десетте най-добри локални алайнмънта

Основен проблем при използване на BLAST (bl2seq) е, че тази програма връща като резултат само един алайнмънт - най-добрият. Това е полезно при търсене в бази данни, но не може да е толкова интересно, ако региона от секвенцията, който ви интересува, принадлежи към добър локален алайнмънт, но не и към най-добрия. Това е малко, като в музикален магазин. Ще ви продадат само едно CD, което е най-добре продаваното за седмицата, но на вас може да не ви хареса.

Lalign е идеално допълнение на BLAST. Той е по-бавен, но по-точен и ви дава като резултат, колкото искате на брой локални алайнмънти (най-добрият, следващият и т.н. до броя, който сте пожелали и указали). Не се притеснявайте за скоростта. За секвенции по-къси от 1000 символа, може и да не забележите разликата между bl2seq сървъра и Lalign.

Като цяло Lalign е страхотен, когато става въпрос за анализ на сложни протеини с много повтори. В списъка от стъпки по-долу ще стартираме Lalign, за да извлечем локални алайнмънти от две родствено отдалечени последователности, съдържащи серин-протеиназни домени:

1. Заредете във вашият браузър адрес:
www.ch.embnet.org/software/LALIGN_form.html

Появява се страницата Lalign на ch.EMBnet.org.

2. Изберете локален от опциите за алайнмънт методи, както е показано на Фигура 6-10.
Може да използвате Lalign интерфейса, както за локален, така и за глобален алайнмънт.

3. Изберете броя на показваните суб-алайнмънти.

Трябва да изберете номер, който съответства приблизително на броя на диагоналите, които сте наблюдавали в dot plot. Ако искате Lalign да показва 10 локални алайнмънта, изберете цифрата 10.

4. Изберете по подразбиране матрица на заместване.

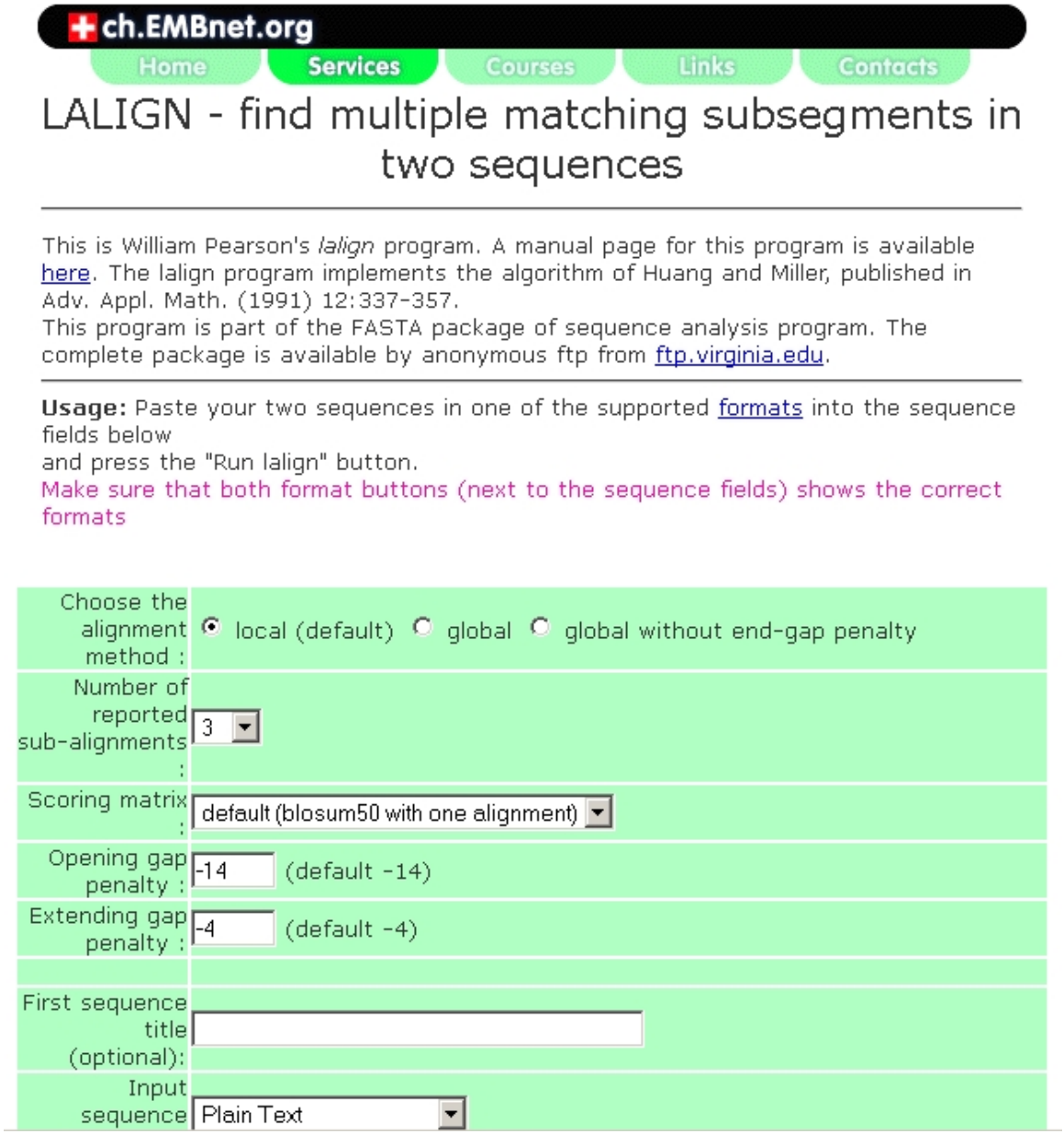
Матрицата на заместване контролира стойността на мутациите, когато Lalign сравнява секвенциите.

Матрицата по подразбиране е identity за ДНК и blosum 50 за протеини. В повечето случаи, настройките по подразбиране са напълно подходящи. Ако искате да промените матрицата, имайте предвид, че високите стойности на BLOSUM правят Lalign по-строг и като резултат се получават по-къси алайнмънти. Ниските PAM индекси имат противоположния ефект.

5. Задайте размер на gap-opening penalty.

Gap-opening penalty (стойност за отваряне на „дупки“ – създаване на разкъсване) дефинира възможността за отваряне на “дупка” в една от секвенциите, когато се сравняват.

Ако сте задали на gap-opening penalty стойност по-висока от тази по подразбиране, локални алайнмънти съдържащи дупки могат да бъдат разделени на няколко по-къси. Няма просто правило за дефиниране на стойността на gap-opening penalty.



Фигура 6-10. Началната страница на Lalign.

Стойностите по подразбиране са настроени за матрицата по подразбиране. Ако промените матрицата, Lalign автоматично не коригира gap-opening penalty. Така например, ако използвате BLOSUM с индекс по-нисък от 45, трябва да намалите gap-opening penalty до стойност по-ниска от 14.

6. Задайте gap-extension penalty.

Gap-extension penalty (стойност за удължаване на разкъсването) заедно с gap-opening penalty характеризират разкъсването при сравняване на секвенции. Стойността на gap-extension penalty трябва да бъде поне десет пъти по-ниска от стойността на gap-opening penalty.

Когато се сравняват отдалечени секвенции, високи стойности на gap-opening penalty и много ниски стойности на gap-extension penalty, често водят до най-добри резултати. Те показват, че разкъсванията в една секвенция трябва да се характеризират въз основа на съществуването ѝ, а не на нейната дължина.

7. Изберете Swiss-Prot ID или AC от падащото меню Input Sequence Format.
Трябва да превъртите надолу Lalign страницата, за да стигнете до това меню.

8. Въведете P05049 в първото поле.
P05049 е номера за достъп на серинова протеаза при някои видове змии. Ако използвате секвенция, която няма идентификационен номер, поставете я в прозореца и се уверете, че сте избрали правилния формат в Стъпка 7.

9. Изберете Swiss-Prot ID или AC от второто падащо меню Input Sequence Format.

10. Въведете P08246 във второто поле.
P08246 е номера за достъп на човешката еластаза.

11. Кликнете върху бутона Run Lalign.
Изчислението трябва да се извърши за сравнително кратко време (по-малко от една минута). Ако програмата не даде никакъв резултат, проверете дали сте въвели правилно параметрите. Lalign връща резултата под формата на HTML документ.

12. Запишете резултатите.
За съхранение на резултата във файл, използвайте командите File□Save As като опция на вашия браузър.

Тълкуване на Lalign резултати

Lalign дава като резултат определения от вас брой алайнмънти (зададени, както е описано в стъпка 3) подредени по техния рейтинг. Интересна особеност на Lalign е, че той отчита само не препокриващи се алайнмънти. Това означава, че в резултата на Lalign може да намерите две аминокиселини или два нуклеотида сравнени по между си само веднъж. Те могат да се наблюдават в няколко алайнмънта, но се появяват един пред друг само веднъж. Това предпазва Lalign от появата в резултатите на незначителни вариации от даден важен локален алайнмънт.

Фигура 6-11 показва резултата от Lalign на двете секвенции, използвани в предходния списък от стъпки. Lalign алайнмънтите са в BLAST-подобен формат, подредени според техните E-стойности. На първият ред, свързани с всеки алайнмънт, са следните характеристики:

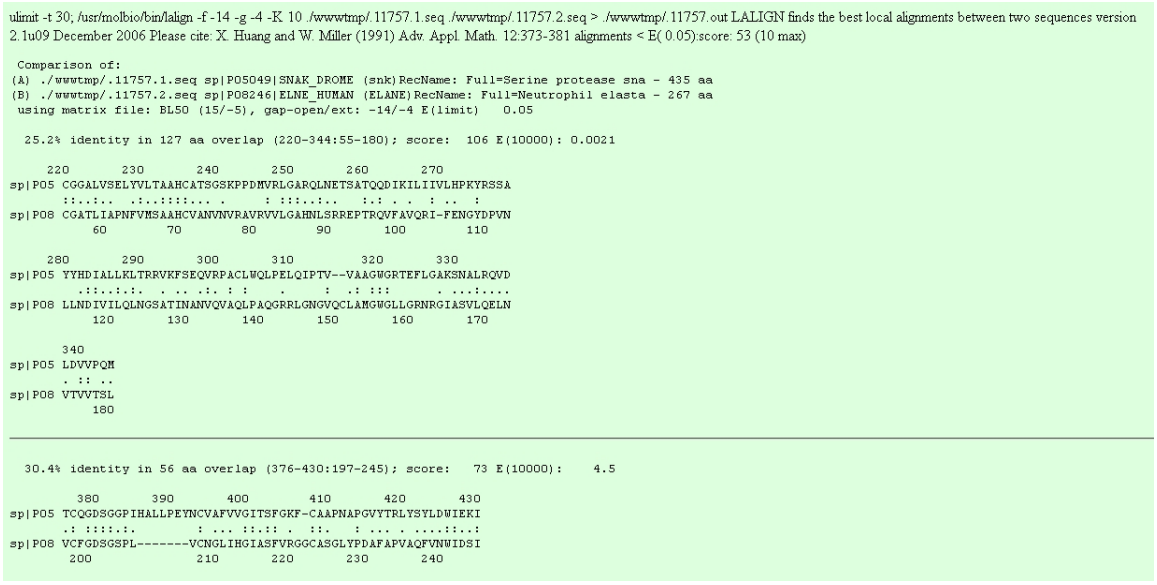
- **Процент на идентичност:** Частта на идентичните остатъци сравнени един с друг. На Фигура 6-11 може да видите, че най-добрият локален алайнмънт е с процент на идентичност 25,2%, а следващият е със стойност 30,4%.
- **Дължина на локалния алайнмънт (припокриване):** Това е общата дължина на локалния алайнмънт.
- **Score:** Оценка на сравнението. Колкото е по-голяма тази стойност, толкова е по-добър алайнмънта, и все пак: Е-стойността е по-добър индикатор за качеството на даден алайнмънт.
- **Е-стойност:** Тази стойност показва колко пъти случайно може да се очаква да се намери такъв добър алайнмънт за вашите две секвенции. Имайте предвид, че Е-стойността при този метод е с много по-малко значение, отколкото при BLAST, свързан с търсене в база данни. Добрата Е-стойност трябва да бъде под 10^{-4}.

Самият алайнмънт съдържа три типа информация:

- **Индекс на остатъците - намира се на линията над секвенцията.**
- **Самият алайнмънт, при който дупките в секвенцията са представени с помощта на тирета.**
- **Идентичност и подобие -** по линията между двете сравнявани секвенции. (_) символ означава идентичност, докато (.) означава подобие. **Идентичност може да наблюдавате при сравняване на секвенции на протеини и на ДНК, докато подобие може да има само при алайнмънта на протеини. Подобието ви дава частта на аминокиселините които са със сходни физико-химични характеристики.**
- Два остатъка са подобни, когато техния substitution score е по-голям от 0.

Първият алайнмънт съответства на консервативните домени на сериновата протеаза. Този алайнмънт, обаче не изглежда убедителен, защото съдържа по-малко от 26% идентичност за около 200 остатъка.

За да сме наистина сигурни се нуждаем от много по-голяма прилика (най-малко 30%) и по-добра (по-ниска) Е-стойност. Причината да се доверяваме на резултата от този алайнмънт е, че той е в съответствие със сигнала, който получихме от dot plot (виж Фигура 6-7). Фактът, че тези два различни анализа дават подобни резултати, (Dotlet и Lalign) дава добра предпоставка да приемем съществуване на консервативен серин-протеазен домен при нашите два протеина.



Фигура 6-11. Резултат от Lalign.

По-ниски стойности на gap-opening или gap-extension penalty, може да доведе до сливане на диагоналите. Какво ще се случи, зависи от дължината на дупките, които Lalign е необходимо да вмъкне при локалния алайнмънт.

Когато трябва да оцените качеството на вашия алайнмънт, обърнете внимание на зададените размери. Не е добре да виждате къси фрагменти с много добри показатели. Това може да е резултат от следните причини: ниска сложност на фрагментите, ниски нива на повторяемост в протеиновите секвенции, повтарящи се модели на хидрофобно/хидрофилни остатъци и т.н. Например, на Фигура 6-11, втори и трети локален алайнмънт са очевидно безсмислени, защото са твърде къси. Въпреки това, техните E-стойности са по-добри от тази на най-високо разположения алайнмънт. Това ясно показва, че къс алайнмънт може да води до случайно получаване на добър резултат.

Биоинформатичните програми не са перфектни. Те могат да правят грешки и да дават некоректни резултати. За съжаление, доброто и злото се подчиняват на същите правила, както в живота, така и в биоинформатиката.

Най-горния алайнмънт, който ни дава Lalign, илюстрира добре това ограничение. Заради информацията, която ни дава dot plot, ние вярваме на този алайнмънт. От друга страна, двете секвенции са толкова отдалечени и малко сходни, че трябва да сме доста предпазливи в тълкуванията, които правим. Много аминокиселини могат да бъдат неправилно сравнени, което означава, че този алайнмънт е както правилен, така и грешен. Ако искате да задълбочите този анализ, ще трябва да направите множествен алайнмънт или да използвате информация за структурите (За повече информация относно множествен алайнмънт - вижте следващата глава).

Извършване на глобален алайнмънт

Глобален алайнмънт е това, което името предполага: алайнмънт на всяка аминокиселина или нуклеотид във вашите секвенции. Когато решите да правите глобален алайнмънт, трябва да сте сигурни, че вашите две секвенции са малко или много близки по цялата си дължина.

Глобалният алайнмънт е много важен при множествените алайнмънти. Извършване на такъв алайнмънт само между две секвенции е от малък интерес за биологичните анализи. Когато имате две тясно свързани секвенции, може да се получи много по-информативен резултат, ако съберете още няколко свързани с тях последователности и проведете множествен алайнмънт (следващата глава представя подробна информация за осъществяване на множествен алайнмънт).

Глобалните алайнмънти не са полезни при откриване на прилики между две секвенции, тъй като статистическият метод за оценяване на E-стойности не се прилага към тях. Дори и така, начинаещите често предпочитат глобален аллайнмънт, защото е по-лесен за разбиране.

При глобален алайнмънт няма аминокиселини или нуклеотиди, които мистериозно да изчезват. Ще намерите в резултата точно това, което сте поставили в полето, плюс няколко допълнителни разкъсвания. Има основно три причини, за да предпочетете да използвате глобален алайнмънт:

- Проверка на малки разлики между две секвенции. Това може да се случи със секвенции, които сте манипулирали и вероятно променили. Глобалният алаймънт е най-добрият начин за локализирането на потенциални проблеми.

- Анализиране на полиморфизми (например SNPs) между тясно свързани секвенции.

- Сравняване на две секвенции, които частично се припокриват.

Използване на Lalign за глобален алайнмънт

По отношение на използваните алгоритми, няма големи различия между локален и глобален алайнмънт. Ако се налага да пишете програми за двата, няма да има повече от един ред разлика между тези два алгоритъма!

Може да използвате Lalign за провеждане на глобален алайнмънт, но има и други по-добри решения за това (Таблица 6-4 в края на тази глава).

Използване на Lalign за провеждане на глобален алайнмънт е като да използвате пак него за провеждане на локален алайнмънт. Единственото нещо, което допълнително трябва да направите е да кликнете върху бутона Global without End-Gap Penalty, разположен близо до горната част на формуляра Lalign на адрес:

http://www.ch.embnet.org/software/LALIGN_form.html

Алайнмънт на протеини с ДНК секвенции

Инструментите на алайнмънт, които ви представяме по-горе са истински ефективни само, когато искате да сравните последователности от подобен тип - протеин с протеин или ДНК с ДНК. Понякога, се налага да се сравни протеин с фрагмент ДНК (например неговият оригинален ген). Провеждането на такъв алайнмънт изисква специални инструменти, които могат да вмъкнат по дълги разкъсвания, съответстващи на интроните. Има най-малко две места, където може да направите това онлайн. Едното е в Института Пастьор (http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms), а другото се поддържа от д-р Пеер Борк и неговата група от Европейската лаборатория по молекулярна биология в Хайделберг (http://www.bork.embl.de/pal2nal/). И двата сървъра изискат предварително да имате секвенциите на протеините и ДНК. Ако имате секвенцията само на протеина, може да извършите blastx срещу целия геном (виж главата за търсене в БД) или да използвате web сървъра, наречен Protogene (достъпен на адрес: http://www.tcoffee.org/). Protogene автоматично извлича ДНК секвенцията, отговаряща на даден протеин.

Ресурси за сравняване на секвенции

Много от съществуващите сървъри предлагат възможност за сравняване по двойки. В Таблица 6-4 е представен кратък списък на някои от по-стабилните ресурси за това.

Таблица 6-4. Програми за онлайн алайнмънт.

Име	Адрес	Тип алайнмънт
Lalign	**www.ch.embnet.org/software/LALIGN_form.html**	глобален/локален
Lalign	**http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=lalign**	глобален/локален
USC	**http://www.usc.edu/its/software/**	глобален/локален
Alion	**http://motif.stanford.edu/distributions/alion/**	глобален/локален
xenAliTwo	**http://users.soe.ucsc.edu/~kent/xenoAli/xenAliTwo.html**	локален за ДНК
Blast2seqs	**http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome**	локален BLAST
Protal2dna	**http://mobyle.pasteur.fr/cgi-bin/portal.py**	протеини срещу ДНК
Pal2nal	**http://www.bork.embl.de/pal2nal/**	протеини срещу ДНК

Генерирането на алайнмънт не винаги е достатъчно. Може да се наложи да се визуализира този алайнмънт и да се прецени, дали статистически са значими резултатите. Таблица 6-5 изброява някои сайтове, които ви позволяват да направите това онлайн.

Таблица 6-5. Анализи за онлайн алайнмънт.

Име	Адрес	Тип алайнмънт
lalnview	**http://www.expasy.ch/tools/sim-prot.html**	визуализация
prss	**http://www.ch.embnet.org/software/PRSS_form.html**	изчисление
prss	**http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=shuffle**	изчисление
graph-align	**http://www.itu.dk/~sestoft/bsa/graphalign.html**	изчисление

Сравняване на множество секвенции (Multiple Alignment)

В тази част ще се запознаем със следното:

- Подбиране на секвенции за множествен алайнмънт;
- Провеждане на множествен алайнмънт с помощта на ClustalW и MUSCLE;
- Множествен алайнмънт с Tcoffee;
- Сравняване на секвенции, които не могат да се приложат в множествения алайнмънт.

Човек с два часовника никога не знае кое време е, а човек с един само си мисли, че го знае.

Мисъл на човек с много часовници

В тази глава ще покажем как може да сравните едновременно много протеини или нуклеотидни секвенции. Това бихте направили, ако имате секвенция и искате да изясните ролята на всяка една аминокиселина или нуклеотид, от които са изградени.

При сравняване на множество секвенции, първо трябва да подберете подходящи секвенции и след това да им направите алайнмънт. В тази част ще покажем как да изберете най-подходящите секвенции и как да генерирате множествен алайнмънт. Представяме три различни метода за множествен алайнмънт, които покриват широк диапазон от нужди: ClustalW, една от най-известните програми; MUSCLE, защото е много бърз; Tcoffee, защото е много точен и позволява да комбинирате секвенции и структури. Лесно е да се генерират лоши алайнмънти, които изглеждат добре. Ще покажем как да оцените качеството на конкретен алайнмънт.

Понякога, един множествен алайнмънт не е достатъчно добър, за да бъде полезен. Ако това се случи ще ви покажем алтернативи за сравняване на секвенции, без да ги подлагате на алайнмънт. Тогава ще ви представим два от тези метода: Gibbs sampler (за идентифициране на свързани сегменти с еднаква дължина) и Pratt (метод за откриване на мотиви и установяване на PROSITE модели).

Множественият алайнмънт е може би най-полезният изследователски инструмент в биоинформатиката. Трудно ще измислим ситуация, в която не може да помогне. Ако не сте сигурни дали множествения алайнмънт може да помогне на вашата работа или не знаете точно как да го приложите за вашия експеримент, ние предлагаме да започнете четенето на първата част от тази глава, "Как множественият алайнмънт може да ви бъде полезен".

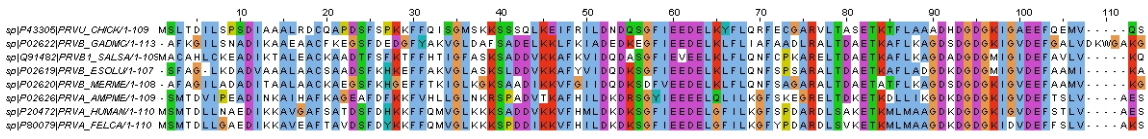
Не забравяйте, че в някои случаи множественият алайнмънт е полезен за предвиждане на протеинови структури, важен за предсказване функцията на протеини и незаменим за филогенетични анализи. Разбира се, при по-добър множествен алайнмънт ще имате по-добри възможности да извлечете полезни резултати, по-добри структурни модели, по-добро предсказване на функции и филогенетични дървета.

Изграждането на множествен алайнмънт далеч не е точна наука. В действителност това е повече изкуство, отколкото наука, изискващо използване на всичко, което знаете в биоинформатиката и биологията. Тази глава дава тайните за създаване на множествен алайнмънт, който да обслужва вашите нужди, дори и да намерите собствени рецепти за това.

Случаи, в които множествения алайнмънт може да ви помогне

Фигура 7-1 показва как изглежда един множествен алайнмънт. Всичко изглежда ясно и просто.

Множествения алайнмънт е идеален, ако искате да изучите семейство от секвенции с общ прародител. Не се притеснявайте, ако имате само един член на този първоначален етап. В тази глава ще ви покажем как да съберете членове на дадено семейство гени/протеини!



Фигура 7-1. Множествен алайнмънт.

Определяне на ситуации, при които множествен алайнмънт не може да ви помогне

Част от това да разберем кога да използвате множествен алайнмънт включва да си изясните кога не е удачно той да се прилага. Методите, отнасящи се за множество секвенции, не работят добре при сглобяване на секвенции в секвенционни проекти (описано подробно в тази глава). Използването на множествен алайнмънт не е подходящо, ако имате набор от къси, частично припокриващи се секвенции (получени от секвениране), или ако искате да превърнете EST в генна последователност. В тези случаи биха се получили много разочароващи и не информативни резултати.

Ако сте изправени пред този конкретен проблем, може да използвате специализирани инструменти за асемблиране на секвенции, като Phred и Phrap (http://www.phrap.org/) или Fault-Tolerance Synthesizer (http://genome.cs.mtu.edu/Tools.html). За съжаление, комплицираността на специализираните инструменти за сглобяване/асемблиране на секвенции са далеч извън обхвата на тази книга.

Друга ситуация, която не изисква провеждане на множествен алайнмънт е, когато секвенцията, от която се интересувате, няма хомология с нито една секвенция в базата данни. В такъв случай, вие нямате късмет и няма как да осъществите множествен алайнмънт на тази секвенция. Може да се опитате да намерите няколко последователности, като използвате функционални критерии и проведете търсене с Platt, но не разчитайте на чудеса.

Подпомагане на научните изследвания с помощта на множествен алайнмънт

Може да създадете коректен множествен алайнмънт само при наличие на подходящи секвенции. За да бъдем честни, може да проведете този алайнмънт и ръчно, а не само с помощта на компютър. Ръчно изработен алайнмънт е като домашно приготвена храна: Когато е приготвен от добър готвач, нищо не може да го победи. За повечето хора обаче, автоматизираната обработка обикновено е много по-добър вариант.

Основната идея зад един множествен алайнмънт (като този представен на Фигура 7-1) е поставяне на аминокиселините или нуклеотидите в една и съща колона, защото са подобни според някакъв критерий. Може да използвате три основни критерии за

изграждане на множествен алайнмънт, като всеки един от тях има различни свойства, както е описано в Таблица 7-1.

Таблица 7-1. Основни критерии при провеждане на множествен алайнмънт.

Критерий	**Значение**
Структурно сходство	Аминокиселини с една и съща роля във всяка структура са в една колона. Програмите за прецизно позициониране са единствените, които използват този критерий.
Функционални прилика	Аминокиселини или нуклеотиди с една и съща функция са в една колона. Няма автоматична програма, която изрично да използва този критерий, но ако информацията е налична, може да принудите някои програми да го спазват или може да редактирате вашия алайнмънт ръчно.
Секвенционна прилика	Аминокиселини в една колона са тези, които дават алайнмънт с максимално сходство. Повечето програми използват секвенционната прилика, защото това е най-лесният критерий. Когато последователности са тясно свързани техните структурни и функционални прилики са еквивалентни на секвенционната прилика.

Първите два критерия имат ясно биологично значение, докато третият - няма. И все пак, когато секвенциите са достатъчно сходни, може да използвате съществуващата прилика, за да продуцирате множествен алайнмънт, който да отразява еволюцията, структурни и функционални връзки, съществуващи между вашите секвенции. Разбира се, за да направите това, вие се нуждаете от алайнмънт, на който може да се доверите. В тази глава ще покажем как да сте сигурни, че влагате парите си в правилния алайнмънт.

В Таблица 7-2 са изброени основни приложения на множественият алайнмънт.

Таблица 7-2. Основни приложения на множествен алайнмънт.

Приложение	Процедура
Екстраполация	Добър множествен алайнмънт може да ви убеди, дали дадена неизвестна секвенция е истински член на дадено протеиново семейство.
Филогенетичен анализ	Ако внимателно подберете секвенциите за даден множествен алайнмънт, може да пренапишете историята на тези протеини.
Идентифициране на мотиви	Чрез откриване на много консервативни позиции, може да определите регион, който е характерен за определена функция.
Идентифициране на домени	Възможно е множествен алайнмънт да се превърне в профил, който описва протеиново семейство или протеинов домен (PSSM). Може да ползвате този профил за сканиране на бази данни за нови членове на семейството.
ДНК регулаторни елементи	Може да превърнете един множественДНК алайнмънт в матрица за транскрипционни фактори и да сканирате други ДНК секвенции за потенциално подобни места за свързване.
Предвиждане на структури	Добър множествен алайнмънт може да ви даде перфектна прогноза за вторична структура, както на протеини, така и на РНК. Понякога може да помогне в изграждане на 3-D модел.
SNP анализ	Редица генни алели често имат различни аминокиселинни последователности. Множествен алайнмънт може да ви помогне да предскажете дали несинонимни полиморфизми в единични нуклеотиди са полезни или вредни.
PCR анализ	Един множествен алайнмънт може да ви помогне да определите по-малко дегенеративните части от дадено протеиново семейство, което би позволило откриване на нови членове на това семейство чрез PCR (полимеразна верижна реакция). Или поне да направите своите праймери по-оптимални.

Независимо какъв е вашия избор или мотивация, трябва да имате предвид, че причината да използвате множествен алайнмънт е, че дава мигновена картина на силите, които оформят еволюцията! Когато погледнете протеини или ДНК секвенции правилата са безмилостни:

- Важни аминокиселини (или нуклеотиди) не могат да мутират. Например, активните центрове на ензимите са много консервативни.

- По-малко важно остатъци се променят по-лесно, понякога случайно, а понякога, за да се адаптира генът/протеинът към дадена функция.

Всичко казано до тук води до едно просто заключение: Когато разглеждате множествен алайнмънт, може да изкажете хипотезата, че консервативните позиции (колоните, където всички секвенции съдържат една и съща аминокиселина или нуклеотид) са по-важни за функцията, отколкото не консервативните (колоните, където секвенциите съдържат различни аминокиселини или нуклеотиди). Разбира се, може да кажете всичко това от провеждане на алайнмънт само между две секвенции, но когато се използват повече, по-лесно и по-сигурно е да направите разлика между важни и по-малко важни позиции.

Множественият алайнмънт е страхотен начин да представите своите резултати. Той ви позволява да поставите много информация в един единствен модел и изключително лесно да се локализират несъответствия или потенциални проблеми.

От всичко казано до тук следва, че множествения алайнмънт НЕ лекува астма, артрит, лумбаго или плешивост. Не ви прави по-силни или увеличава някои ваши способности (с изключение може би на изследователския ви потенциал). Ако имате нужда от повече аргументи, за да се убедите в гениалността на множествения алайнмънт, то по-скоро нямате нужда да провеждате този вид алайнмънт.

Избор на подходящи секвенции

Всеки, който някога е работил в лаборатория по молекулярна биология знае, че много наподобява готвене: Всичко се свежда до правилен избор на съставки и поставянето им в тенджерата в точното време, правилните пропорции и в правилния ред.

Провеждането на множествен алайнмънт се подчинява на същите правила. Преди започване на вашия алайнмънт трябва много внимателно да подберете секвенциите, които искате да подложите на алайнмънт. Тези секвенции трябва да са членове на едно и също протеиново семейство и да имат общ прародител. Обикновено семействата са твърде големи, за да бъдат изцяло включени в множествен алайнмънт и затова изборът на правилни секвенции си е цяло изкуство. Ако искате да сте добър в тази игра, трябва много добре да знаете какво искате да покажете с вашия алайнмънт и точно как работят програмите, с помощта на които може да проведете един такъв алайнмънт.

За да се разберат добре тези програми, трябва да ги видите в действие и за това са необходими няколко секвенции! Това е типичната дилема за кокошката и яйцето, за която не съществува лесно решение. Препоръчваме да разгледате следващия раздел

веднъж и да се върнете отново, когато решите да провеждате множествен алайнмънт. След преглеждане на съдържанието на тази глава, един или два пъти нещата трябва да си дойдат на мястото.

Видове секвенции, с които може да работите

Да предположим, че започнете тази процедура с любимата си секвенция. Искате старателно да проучите тази последователност и знаете, че за да направите това, трябва да проведете и множествен алайнмънт. Таблица 7-3 обобщава основните неща, които трябва да вземете под внимание при избора на тези допълнителни секвенции.

Таблица 7-3. Насоки при избор на секвенции.

Проблем	**Решение**
Протеини или ДНК	Използвайте протеини, когато е възможно. Може да ги превърнете отново в ДНК след провеждане на множествен алайнмънт.
Много секвенции	Започнете с 5-10 секвенции, избягвайте алайнмънт с повече от 30.
Много различни секвенции	Често има проблеми при секвенции, които са по-малко от 30 на сто идентични с повече от половината от другите секвенции в групата.
Идентични секвенции	Никога няма полза от тях. Ако имате причина да работите с такива секвенции, изключете тези, които са повече от 90% идентични от другите секвенции в групата.
Непълни секвенции	Програмите за множествен алайнмънт предпочитат секвенции, които са с относително еднаква дължина. Те често имат проблеми при сравняване на цели секвенции и непълни фрагменти.
Многократното повторени домени	Секвенции с многократно повторени домени обикновено предизвикат проблеми при работа на програмите за множествен алайнмънт (особено, ако броят на домените е различен). В такъв случай, може би е по-добре да се извлекат домените с Dotlet или Lalign (виж предната глава) и направете множествен алайнмънт на тези фрагменти.

Използване на ДНК или белтъчни секвенции

Ако се интересувате от некодиращи секвенции, очевидно нямате избор и трябва да използвате ДНК. Внимавайте, некодиращите ДНК секвенции са трудни за алайнмънт! Ако не можете да генерирате правилен алайнмънт между секвенции, за които знаете със сигурност, че са свързани, може да използвате методи за локален множествен алайнмънт, като Gibbs sampler или Pratt (виж "Сравняване на секвенции, които не могат да се подложат на алайнмънт". За описание на тези два метода - по-нататък в тази глава).

Методите за множествен алайнмънт са по-подходящи за протеини. Причината е, че протеиновите секвенции са три пъти по-къси от съответната им ДНК и използват по-информативна азбука от 20 аминокиселини.

Ако искате да извършите филогенетичен анализ на набор от кодиращи ДНК секвенции, следвайте следните стъпки:

1. Транслирайте вашите ДНК секвенции в протеини.
2. Извършете множествен алайнмънт на тези протеини.
3. Превърнете протеините, използвани при множественния алайнмънт отново в ДНК с помощта на pal2nal (http://www.bork.embl.de/pal2nal/) или Protogene, ако нямате оригиналната секвенция на ДНК (http://www.tcoffee.org/).

Ако е трудно да се проведе алайнмънт с вашите протеини, защото са с малко подобие, информацията от ДНК не би помогнала. Трябва да се използва дегенеративността на генетичния код и факта, че има 20 аминокиселини и само 4 нуклеотида. Ако има слаб сигнал на протеиново ниво, може да сте сигурни, че няма да има по-добър сигнал на ниво ДНК.

Избор на правилен брой секвенции

Разбира се, няма абсолютен отговор на този въпрос, като например 16 или 7. Преди няколко години отговорът беше лесен. Вземете всичко, което може да намерите и отидете в лабораторията, ако няма достатъчно секвенции в базите данни! Но днес това вече не е вярно. Днес не е така, като се има предвид размера на базите данни и новите пълни геномни секвенции, появяващи се поне два пъти месечно. Може лесно да намерите стотици или хиляди секвенции, които биха били подходящи за включване

във вашия множествен алайнмънт. Но това не означава, че трябва да ги използвате всичките!

По наше мнение, трябва да започнете с относително малък брой секвенции. В повечето случаи ще бъде подходящо да започнете с 5 до 10. Ако се случва нещо интересно с тази малка група, винаги може да увеличите броя на секвенциите. Във всеки случай, няма основателна причина да провеждате множествен алайнмънт с над 50 секвенции, освен ако не сте заинтересувани от изграждане на обширни филогенетични дървета.

Ако започнете със стотици секвенции, веднага може да попаднете на неприятности. Причините за това са:

- **Изчисляването на голям алайнмънт е трудно.** Обществените сървъри не са с безкрайни ресурси. Вашата работа може да отнеме много време (ако въобще стартира). Това води до трудности при настройка на параметрите и проверка на алтернативите.
- **Създаването на голям алайнмънт е също трудно.** Програмите за множествен алайнмънт не са много добри при манипулиране с голям набор от секвенции.
- **Показването на голям алайнмънт е неудобно.** Не може да ги отпечатате и да ги визуализирате, когато искате, защото ще блокират вашия компютър. Ако колоните са по-дълги от една страница, тълкуването на резултатите става невъзможно.
- **Използването на голям алайнмънт е неефективно.** Програмите за построяване на дървета и предвиждане на структури не могат да манипулират лесно с тях.
- **Осъществяването на точен и голям алайнмънт е почти невъзможно.** Искате вашият множествен алайнмънт да бъде достоверен и изключително надежден, това може да стане само ако сте сигурни, че всички секвенции, които участват в него са истински членове на даденото семейството. Основна причина за безпокойство е, че програмите за множествен алайнмънт правят грешки. Проклятието е, че тези грешки не се добавят, те се размножават. Така малък брой неподходящи секвенции, използвани при провеждане на множествен алайнмънт могат да разрушат целия алайнмънт. Разбира се, с колкото повече секвенции работите, толкова по-вероятно е това да се случи. Най-добрият начин да се избегне такова бедствие е да се започне с малко и постепенно да увеличавате броя на секвенциите, с които провеждате множествен алайнмънт, докато включите всички секвенции, от които се интересувате.

Вече знаете от колко секвенции имате нужда и последния въпрос, който трябва да решите е колко близки трябва да са тези секвенции. Трябва да изберете секвенции, които са много сходни или много различни?

Компромис между сходство и нова информация

Ако мислите, че много подобни секвенции дават много добри алайнмънти, то вие сте прави! Обаче, един множествен алайнмънт трябва да бъде освен акуратен, също и полезен.

Например, алайнмънт на много близки секвенции носи малко информация. Такъв алайнмънт може да го използвате за екстраполиране на анотации, но не и да направите филогенеза, структурни прогнози или някое от другите полезни приложения, изброени в Таблица 7-2. Тези и други задачи изискват да се наблюдават мутации в моделите на всяка колона, което не е възможно, ако имате алайнмънт, в който колоните са изцяло консервативни.

Работа със силно отдалечени в родствено отношение секвенции също не е подходящо. Програмите за множествен алайнмънт не могат да използват секвенции, които са твърде различни, дори ако тези секвенции са хомоложни. В действителност има две неща, които програмите за множествен алайнмънт наистина не харесват:

- Секвенции, които много се различават от всички останали в групата;
- Секвенции, които изискват дълги инсерции/делеции, за да бъде правилно проведен един алайнмънт.

Избиране на секвенции за множествен алайнмънт е като упражнение за изграждане на отбор. Искате всички да бъдат малко различаващи се, но не искате членовете на вашия екип да бъдат силно различни, че да не могат да общуват помежду си.

Когато избирате секвенции общо правило е, че те трябва да бъдат родствено свързани възможно най-силно, без да се изискват прекалено много разкъсвания при провеждане на алайнмънт. Тези два критерия са взаимно изключващи се, така че намирането на компромис между тях е въпрос на правилна стратегия. Показваме основните стъпки, които е добре да следвате при подбор на секвенции:

1. **Изберете няколко секвенции.**
   Вижте раздела "Подбиране на секвенции с помощта на онлайн базирани BLAST сървъри" по-нататък в тази глава.
2. **Изчислете множествен алайнмънт с помощта на един от сървърите, представени в тази глава.**
   Вижте раздела "Избор на подходящ метод за множествен алайнмънт" по-нататък в тази глава.
3. **Визуално оценяване качеството на вашият алайнмънт.**
   Вижте раздела "Тълкуване на множествен алайнмънт" по-нататък в тази глава.

4. **Ако вашият алайнмънт изглежда добре, запазете секвенциите.**
   Един добър алайнмънт съдържа консервативни региони, разделени от региони с делеции и инсерции. Ако имате такъв алайнмънт, то вашият набор от секвенции е вероятно подходящ и може да се опитате да го увеличите с допълнителни нови серии.
5. **Ако вашият алайнмънт е труден за интерпретиране:**
   A. Разгледайте секвенциите по-подробно - опитайте се да премахнете секвенциите „нарушители", които са най-отдалечени родствено или тези, които предизвикват дълги инсерции/делеции.
   B. Проведете алайнмънта наново с по-малък набор от секвенции.
   C. Съхранявайте първоначалния набор от секвенции, преди да получите задоволителни, лесни за тълкуване резултати.

Ако можете уверете се, че всяка секвенция е между 30 и 70 процента идентична с повече от половината секвенции в комплекта. По този начин се прави разумен компромис между нова информация и качество на алайнмънта.

Преди да се добавят още секвенции към даден множествен алайнмът, може да се опитате да разберете, дали това е удачно като сравнявате по двойки с някои от инструментите описани в предишната глава. Все пак, ние не препоръчваме да го правите изцяло. Това е загуба на много време.

Задаване на име на секвенциите

Задаването на име на секвенциите може да звучи тривиално, но не е така. Програмите за множествен алайнмънт нямат стандартни начини за манипулиране с имената на секвенциите. Ако се придържате към следните четири правила, няма да изпаднете в неприятности:

- Не използвайте празни позиции в имената на вашите секвенции;
- Не използвайте специални символи. Придържайте се към обикновени букви, цифри и долно тире (_), за да замените празно пространство. Избягвайте всички други символи, особено тези, които са най-примамливи за специални секвенции (като @, #, _, ^ и т.н.);
- Не използвайте имена по-дълги от 15 символа;
- Не давайте едно и също име на две различни секвенции във вашето множество.

Въпреки, че някои програми го приемат, други (като ClustalW) не.

Ако не се подчинявате на тези правила, някои програми за множествен алайнмънт може автоматично да променят името на вашите секвенции, без да ви информират дори.

Събиране на секвенции с помощта на BLAST сървъри

Има два вида секвенции, които може да интегрирате в един множествен алайнмънт:

- **Характеризирани/Анотирани секвенции:** Това са последователности, за които има добри анотации и експериментални данни, които ги подкрепят. Определено се предпочитат такъв тип секвенции, защото те носят със себе си определена биологична информация и позволяват бъдещо развитие.
- **Непознати секвенции:** Тази категория включва, както ваши последователности, така и от базите данни. Нехарактеризираните секвенции трябва да бъдат членове на едно семейство. Основен мотив за включването им в даден множествен алайнмънт е да се направи разграничение между консервативни позиции, които не могат да мутират и другите, по-малко важни колони. Те съдействат за постигане на контраст между секвенциите от интерес.

В този раздел ще покажем как се събират тези секвенции с помощта на BLAST програми за търсене в бази данни. Ако искате да знаете всички подробности за BLAST, вижте следващата глава.

С помощта на BLAST може да търсите в базата данни за последователности, които са хомоложни (подобни) на вашата заявка. Заявката може да бъде всяка протеинова или ДНК последователност. Може да използвате BLAST, за да търсите в протеинови и ДНК бази данни. Основна причина за използването на BLAST е да се идентифицират от базата данни последователности, които са толкова сходни на вашата заявка, че те най-вероятно са хомоложни. Често се позоваваме на такива резултати или съвпадения.

Има много BLAST сървъри по света (виж следващата глава). Вие сте свободни да опитате всеки (или всички) от тях. Въпреки това, те не са еднакво удобни за изтегляне на секвенциите, от които се интересувате. В Таблица 7-4 са изброени три BLAST сървъра, които са полезни за генериране на списъци с последователности във формат FASTA (най-добрия формат за използване на дадена секвенция, при работа с различни програми) или за изпращане на последователност на сървър за множествен алайнмънт.

Таблица 7-4. BLAST сървъри с интегрирани методи за множествен алайнмънт.

Адрес:	**Възможности:**
http://www.expasy.ch/tools/blast/	извличане на цели секвенции, експорт на секвенции във FASTA формат, заявяване на секвенции към ClustalW, Tcoffee или MAFFT.
http://blast.ncbi.nlm.nih.gov/	извличане на цели секвенции, извличане на фрагменти, експорт на секвенции във FASTA формат, заявяване на секвенции към ClustalW.
http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+srsq2+-noSession	заявяване на секвенции към ClustalW.

Избиране на секвенции на ExPASy сървъра

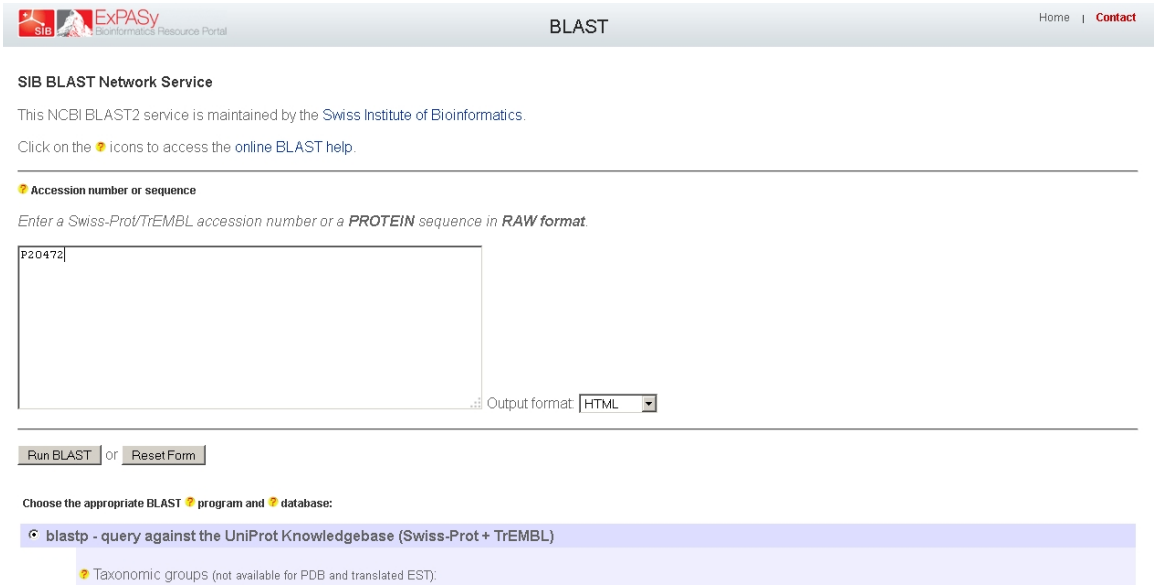
В списъка от стъпки по-долу, подбираме подходящи секвенции, за да направим множествен алайнмънт на калций-зависима протеин киназа.

Може да ползвате този сървър само за изтегляне на протеинови последователности. Ако се интересувате от събиране на ДНК секвенции, използвайте NCBI, BLAST (http://blast.ncbi.nlm.nih.gov).

BLAST сървърът на PBIL сайта е много подобен на ExPASy. Ако не намерите интересуващата ви база данни на ExPASy сървъра, опитайте PBIL.

1. **Въведете във вашият браузър адреса:**
   http://web.expasy.org/blast/
   Появява се BLAST страницата на ExPASy сървъра.
2. **Въведете идентификационен номер на секвенцията P20472, както е показано на Фигура 7-2.**
   Това е номера за достъп на човешкия *parvalbumin*.

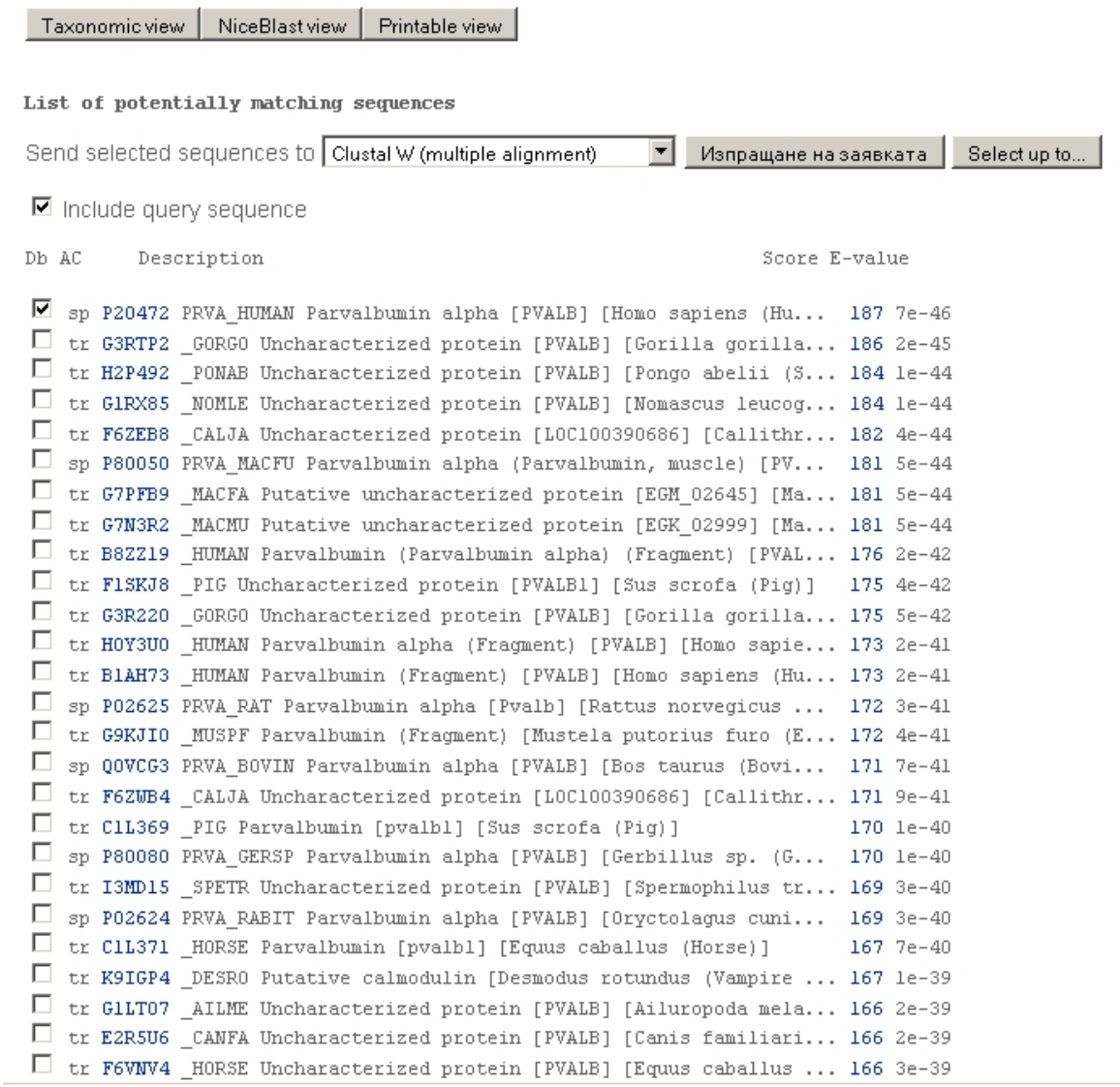
Ако предпочитате, може да поставите секвенцията в суров вид (само последователността, без заглавен ред). Програмата игнорира празни позиции и цифри.



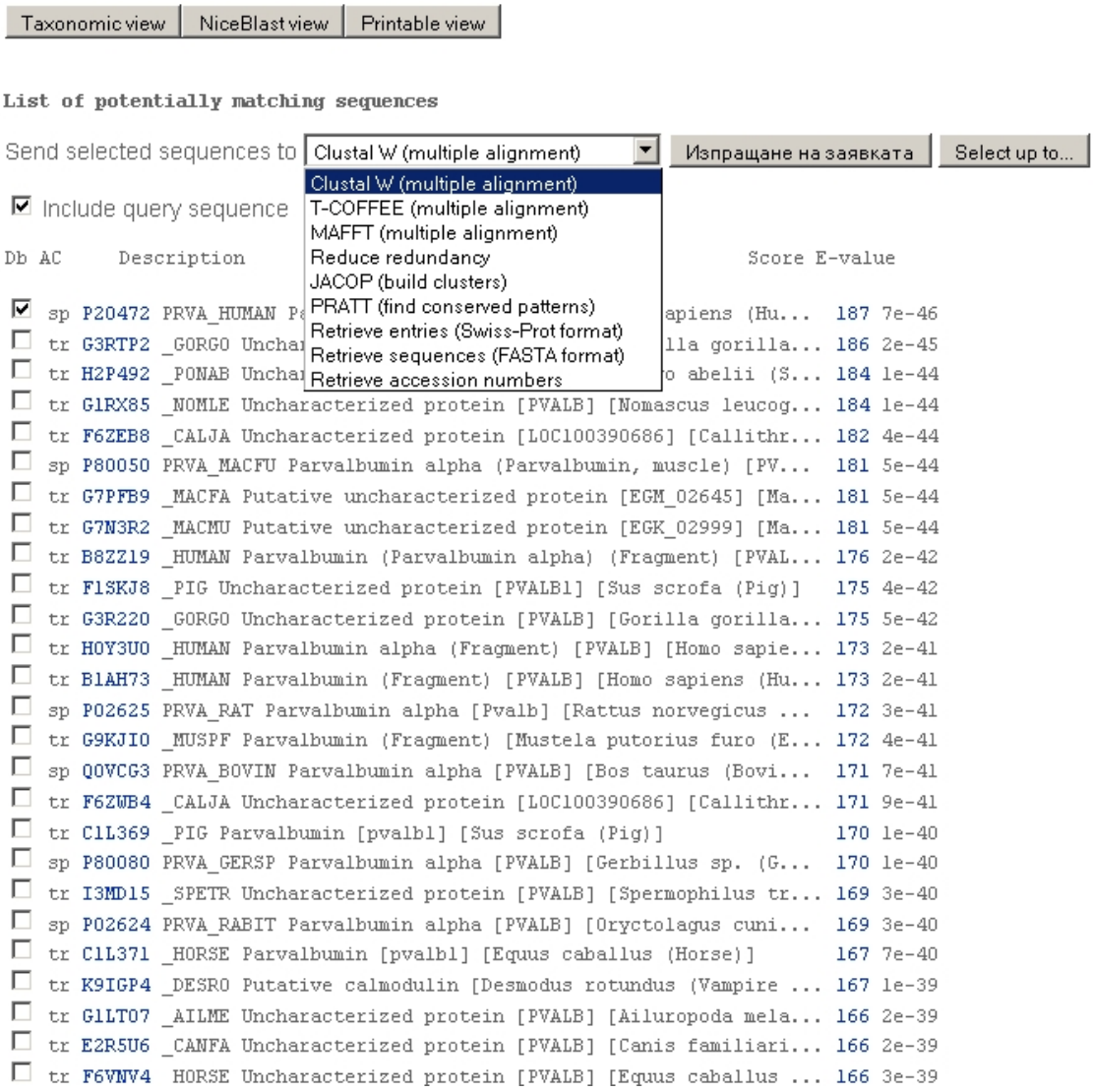
Фигура 7-2. Използване на BLAST ExPASy сървъра за събиране на секвенции.

1. **Изберете вида BLAST, от който се интересувате.**
   Ако от стъпка 2 имате протеин, изберете blastp.
   Ако от стъпка 2 имате ДНК, изберете tblastn.
2. **Запазете опциите по подразбиране (пълна база данни) от падащото меню.**
   Това води до едновременно търсене в Swiss-Prot + TrEMBL + TrEMBL_NEW. Ако в резултата от търсенето се докладват твърде много секвенции, които са сходни с вашата, може да намалите броя на идентичните попадения, като изберете по-малка база данни от падащото меню за избор на база данни, Swiss-Prot или база данни само за микроорганизми.
   За множествен анализ не избирайте транслирани ESTs от падащото меню. Тези последователности са предимно непълни протеини, които може да объркат процедурата на множествен алайнмънт.
3. **Превъртете надолу до раздела с опции и задайте броя на показваните най-добри секвенции на 1000.**
   Това прави по-вероятно намирането на подходящи секвенции в BLAST резултата на множествения алайнмънт.
4. **В същия раздел с опции, задайте броя на показваните най-добри алайнмънти на 1000.**
   Този избор дава възможност да се прецени качеството на алайнмънтите, преди да изберете секвенция.
5. **Кликнете върху бутона Run BLAST.**
   След кратка пауза се появява страницата с резултати.
6. **Превъртете надолу страницата, за да изберете последователностите, които искате.**
   Вие избирате дадена секвенция, като отбелязвате квадратчето намиращо се от лявата и страна.
   Това е най-деликатната част от целия процес. Няма абсолютно правило при избора на секвенции, но може да използвате следните насоки:
   A. **Изберете най-горната секвенция.**
      Обикновено това е секвенцията, която ви трябва. Ако това не е тя, може да се наложи да я добавите в списъка по-късно.
   B. **При първи анализ трябва да изберете десет последователности или по-малко.**
      В идеалния случай десетте последователности трябва да са с E-стойности между 10^{-40} и 10^{-5}.
      Фигура 7-3 ви дава представа за това как изглежда такъв избор.
      Изберете (от горе на долу) секвенциите със следните идентификационни номера:
      P20472, P80079, P02626, P02619, P43305, Q91482, P02620, P02622, P02627.
   C. **Преди да изберете последователност проверете, дали тя е подобна на вашата заявка по цялата дължина.**
      Разделът с алайнмънт е в долната част на BLAST резултата. Трябва да сте особено внимателни с попаденията, които са с по-високи от 10^{-10} E-стойности. Те може да са действителни добри попадения на целите секвенции или на отделни фрагменти, може да бъде също и съвпадение между протеинов фрагмент и вашата секвенция. Проверка на алайнмънта е единственият начин да се направи разграничение между тези ситуации.

7. **Изберете метода, по който искате да експортирате избраните секвенции от падащото меню – Send Selected Sequences, както е показано на Фигура 7-4.**
   Има няколко начина да експортирате секвенции:
   A. **FASTA:** Генерира файл, който съдържа вашата последователност във FASTA формат. Може да запишете този файл с командите File□Save As от вашия браузър. При нужда, може да отворите отново този файл с браузъра, за да изрежете и поставите съдържанието му в друг сървър (например MUSCLE на адрес: www.drive5.com/muscle/).
   B. **ClustalW, Tcoffee и MAFFT:** Това са пакети за множествен алайнмънт, работещи на сървъра EMBnet. Изберете един от тях, за да проведете алайнмънт на избраните от вас секвенции.
   C. **Намаляване на излишни повторения:** С тази опция ще извлечете най-значимите секвенции от вашия набор данни - подходяща опция, ако имате прекалено много секвенции и не знаете кои да изберете.
   D. **Pratt:** Сканира за консервативни мотиви във вашата последователност, без да ги включва в алайнмънта.



Фигура 7-3. Избиране на секвенции от BLAST резултат.



Фигура 7-4. Резултат на ExPASy BLAST страница.

Събиране на известна колекция от секвенции от Swiss-Prot

Ако вече знаете името или номера за достъп на всяка една от секвенциите, които искате да включите в множествения алайнмънт и ако те са в Swiss-Prot или TrEMBL, директно имате достъп до тях с помощта на специални онлайн ExPASy инструменти.

Това може да бъде полезно, ако използвате BLAST сървър, който няма инструменти за екстрахиране на секвенции. Може само да копирате номера за достъп на избраните от вас секвенции и да ги екстрахирате тук.

Тази процедура работи само ако интересуващите ви секвенции принадлежат на базите данни Swiss-Prot или TrEMBL. Списъка със стъпки по-долу показва как да използвате този сървър:

1. Задайте на вашия браузър адрес: http://www.uniprot.org/batch/

2. Въведете номера за достъп на вашата секвенция в Sequence прозореца, както е показано на Фигура 7-5.

Въведете един номер за достъп (или име на секвенция) на ред. За нашия пример въведете: P20472, P80079, P02626, P02619, P43305, Q91482, P02620, P02622.

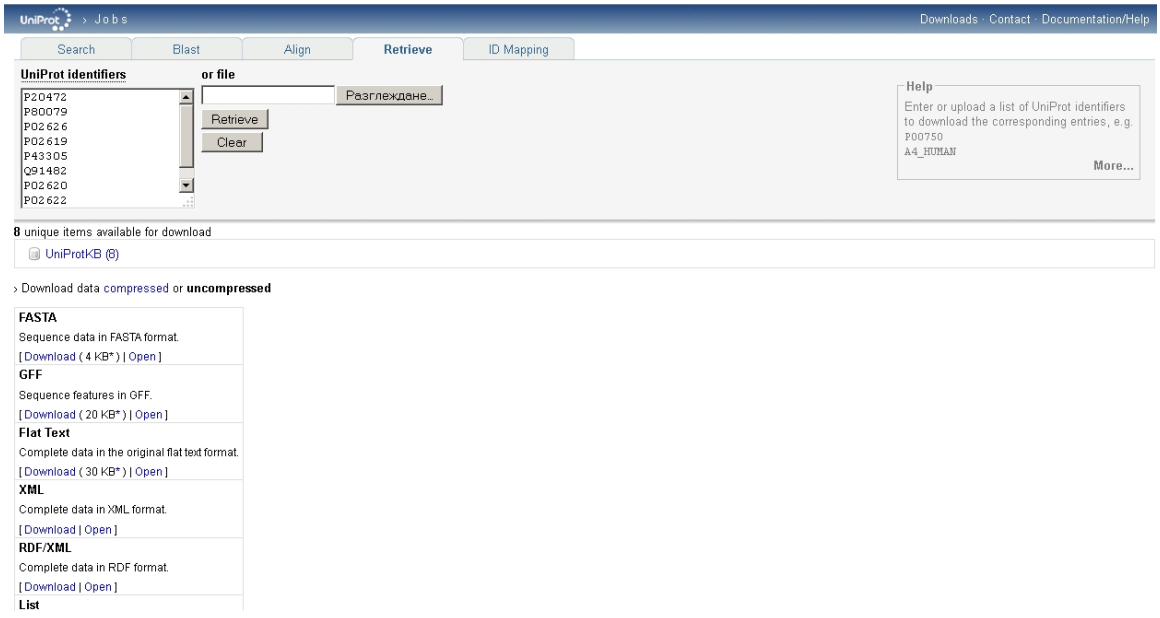
Може да качите файл с информация като напишете името на файла в полето или използвате бутона Browse и изберете вашия файл за качване.

3. Под полето, където сте въвели номера за достъп, щракнете върху бутона Create FTP File.

Така се изпраща вашата заявка към сървъра ExPASy.

4. Запазване на резултатите в текстов файл с командите File□Save As.

Този сървър дава имена на секвенции, които са по-дълги от 15 символа, което може да ви създаде проблеми при някои сървъри.



Фигура 7-5. Събиране на секвенции от ExPASy сървъра.

Избор на подходящ метод за множествен алайнмънт

Преди да започнете да провеждате множествен алайнмънт, трябва да знаете, че няма перфектен метод. Всички използват приближения.

Изграждането на добър множествен алайнмънт изисква практика. Обичайната стратегия изисква сравняване на няколко резултата и търсене на алтернатива за солидност и стабилност.

В този раздел ще покажем как да използвате ClustalW - най-широко използвания пакет за множествен алайнмънт. Ще покажем също как да използвате Tcoffee - една от последните създадени програми за множествен алайнмънт. С Tcoffee може да комбинирате секвенции и структури, да оценявате алайнмънт или да сливате няколко алтернативни множествени алайнмънта в единен резултат. Ще ви запознаем с MUSCLE, един от най-бързите алайнмънт методи.

Използване на ClustalW

ClustalW (най-новата версия е наречена ClustalOmega) със сигурност е най-често използваната програма за извършване на множествен алайнмънт. Ако видите в научна публикация проведен множествен алайнмънт, може да сте сигурни, че авторите използват ClustalW, за да го генерират.

ClustalW използва постепенни нарастващи методи при изграждане на алайнмънта. Вместо да сравнява едновременно всички секвенции, той ги добавя една по една. Разбирането на неговите основни принципи помага изключително, ако искате да получите най-добрите възможни резултати.

Съществуват много ClustalW сървъри. Те обикновено работят на една и съща версия на тази програма, но техният интерфейс ви дава достъп до различни опции. В края на тази глава даваме списък със сървъри, които работят с ClustalW. Изберете сървър, който съчетава бързина и надеждност.

Стартирайте EBI ClustalW сървъра

ClustalW е типичен пример за програма, която може да доведе до добър резултат с настройките си по подразбиране. Трябва да променяте настройките, само ако искате да промените изходния формат на програмата.

Как работи ClustalW?

Методът, който използва програмата е бърз и опростен – т.н. ***прогресивен алайнмънт****. Първоначално, всички секвенции се сравняват 2 по 2, и програмата ги подрежда по подобие. Това групиране изглежда като филогенетично дърво (но не е такова наистина!), нарича се водещо дърво (guide tree) и се съхранява във файла с разширение .dnd на изхода на ClustalW. Когато прави прогресивния алайнмънт, ClustalW се ръководи от топологията на водещото дърво. Основният трик е, че когато направи първичните алайнмънти, след това ClustalW сравнява алайнмънтите така, сякаш не са алайнмънти, а единични секвенции. Това може да причини известни проблеми – някои важни мотиви може да не се покажат като толкова важни, в зависимост от това в коя сравнявана двойка са попаднали. Също, има значение порядъка, в който са подредени секвенциите във входния файл. Но като цяло, почти винаги дава добри резултати.*

W в ClustalW идва от Weights – на всяка секвенция в алайнмънта се дава „тежест" пропорционална на новата информация, която допринася.

Стартиране на ClustalW сървъра към EBI

ClustalW е добър пример за програма, с която може да получите смислени резултати, използвайки параметрите по подразбиране. Вероятно ще ви се наложи да ги променяте само за изходния формат на резултатите.

ПОЛЕЗЕН СЪВЕТ: Преди да започнете работа с ClustalW, трябва да съберете всички секвенции, с които искате да работите (по-горе в тази глава е описано как).

Най-удобният начин е да въведете секвенциите във FASTA формат, но ClustalW приема и други формати, включително Swiss-Prot и PIR, както и вече сравнени секвенции в най-разпространените формати след множествен алайнмънт.

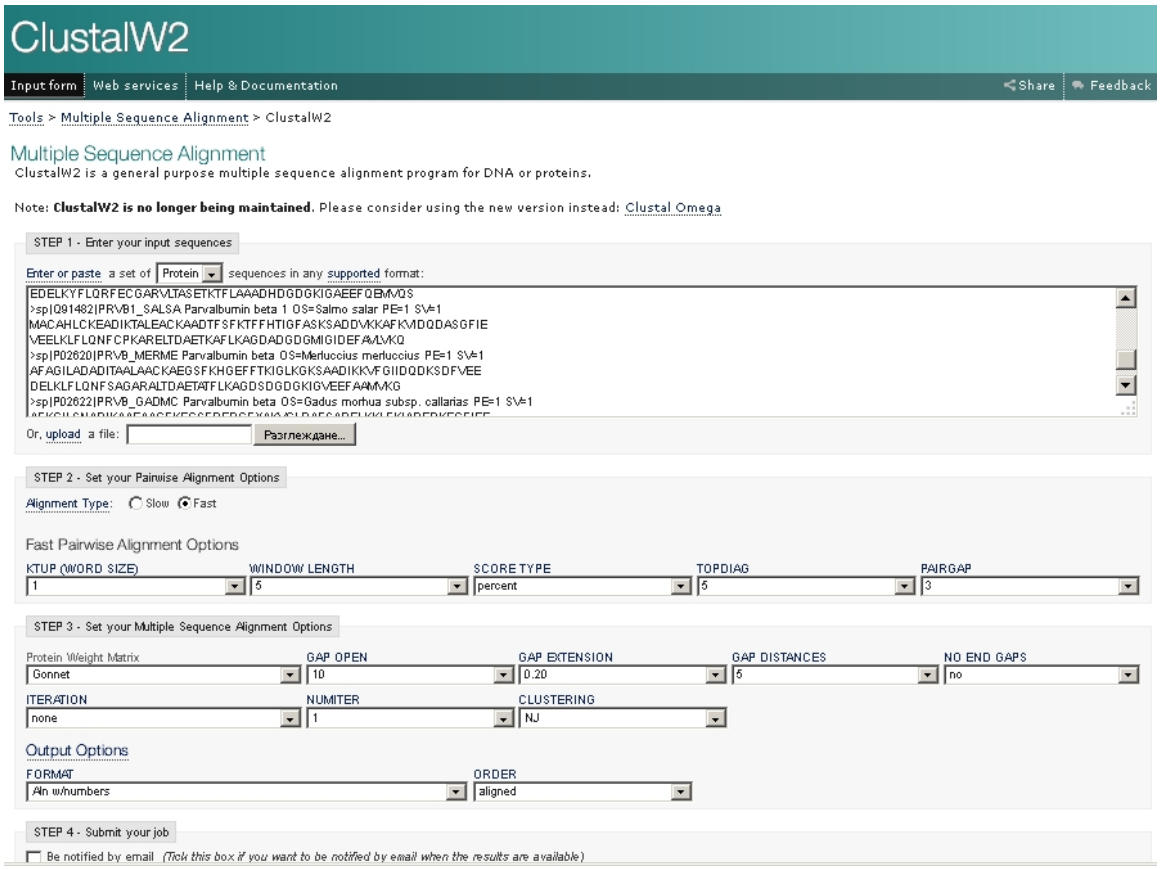
ВНИМАНИЕ! Ако искате да използвате ClustalW ефективно, имайте предвид следното:

- Ако му въведете група секвенции, които вече са преминали множествен алайнмънт, ClustalW не премахва съществуващите „дупки" (gaps), т.е. **входният алайнмънт, който въвеждате ще повлияе върху алайнмънта, произведен от ClustalW.**
- **Редът на секвенциите във файла, който подавате за обработка, понякога може да повлияе върху крайния алайнмънт**. Ако го промените, може да получите различен краен резултат дори и със същите секвенции.

Работа с ClustalW сървъра

1. **Отидете на страницата на EBI ClustalW сървъра на** http://www.ebi.ac.uk/Tools/msa/clustalw2/
   Появява се входната страница на **ClustalW2**.
2. **Въведете секвенциите, които сте събрали в съответното поле или качете файла, който ги съдържа, в полето Upload a file.**
3. **Като опция за „Alignment" изберете „fast" от спускащото се меню (Фигура7-6).**
4. **Изберете изходен формат (Output Format).**
   Различните изходни формати имат различни плюсове и минуси. Най-сигурно е да използвате „Aln Without Numbers" – подразбиращия се ClustalW формат.
   ПОЛЕЗЕН СЪВЕТ: Никога не е късно да промените изходния формат. За да изберете най-подходящия за вас, НЕ стартирайте наново целия алайнмънт! Можете лесно да преформатирате алайнмънтите си с помощта на някой от онлайн инструментите за преформатиране (например Fmtseq на: http://www.bioinformatics.org/JaMBW/1/2/index.html).

5. **Изберете „Input” от менюто „Output Order” (Фигура 7-6).**
   Когато изберете за „output order” Input, във файла с алайнмънта ClustalW показва вашите секвенции в оригиналния ред, в който сте ги въвели.
   ВНИМАНИЕ! Ако изберете за „output order” „Aligned”, секвенциите се появяват в порядъка, в който ги е подредил ClustalW, когато е съставял водещото дърво (guide tree) – близкородствените секвенции са близо една до друга. Въпреки, че опцията „Aligned” е по-информативна от „Input”, ако изберете „Aligned” ще срещнете трудности, когато сравнявате алайнмънти, генерирани с различни методи или с различни параметри. Най-добре е предварително да подредите секвенциите си по най-информативния начин във файла, който ще подадете и след това да изберете опцията „Input”.

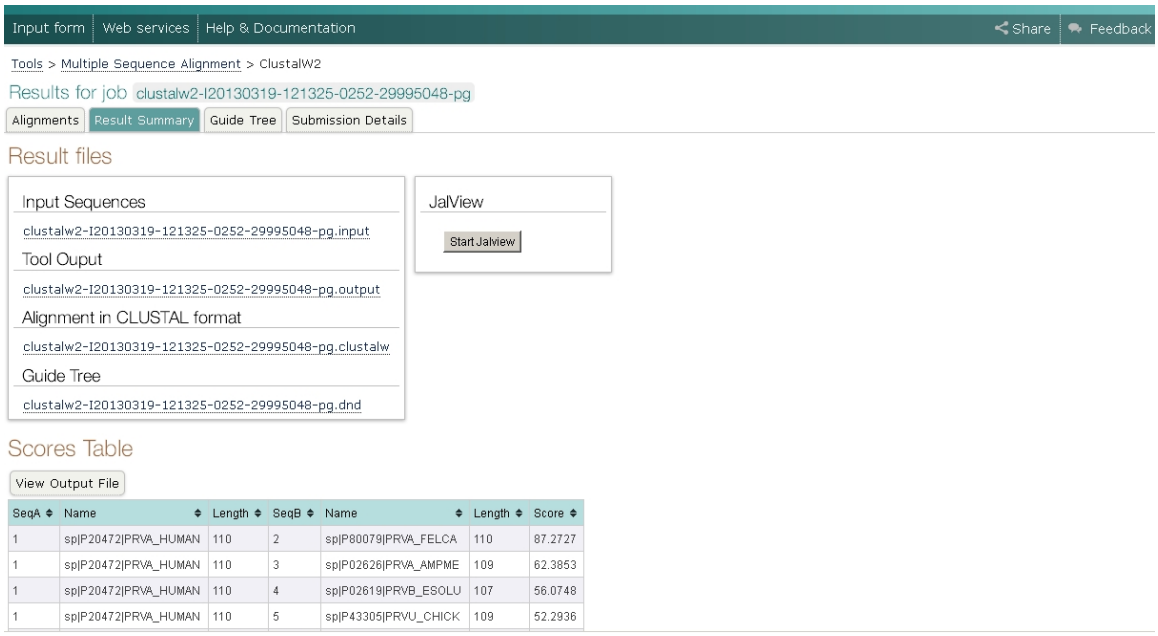


Фигура 7-6. Входната страница на ClustalW2 сървъра към EBI.

6. **Натиснете бутона Submit.**
   Изчакайте, докато браузърът ви покаже страницата с резултатите.

7. **Запазете резултатите си.**
   Резултатите ви се появяват в Таблица с няколко секции с хипервръзки. При проследяването им се отварят:
   A. **Output file - резултати от сравняването по двойки (Pairwise scores).** Можете спокойно да подминете съдържанието на тази секция. Тя е описание на алайнмънтите по двойки, които ClustalW генерира, за да си състави водещото дърво.
   B. **Alignment file – вашият множествен алайнмънт.** Тази секция можете да видите на екрана или да запазите като текстов файл.
   C. **Guide tree file** – Повечето ClustalW Web сървъри (включително този към EBI) показват водещото дърво на ClustalW (като .dnd файл).
   ВНИМАНИЕ! Не забравяйте, че водещото дърво НЕ Е филогенетично дърво!
   D. **Your input file** – файла със секвенциите, които сте въвели в ClustalW.

Съдържанието на Output file, Alignment file и Guide tree file можете да видите и направо на страницата с резултати под Таблицата с хипервръзки (Фигура 7-7).



Фигура 7-7. Страницата с резултати на ClustalW.

Промяна в параметрите на ClustalW

Има три параметъра на ClustalW, които може да променяте във вашия алайнмънт – матриците на заместване (substitution matrices), стойност за „отваряне на дупка/разкъсване” (gap-opening penalties) и стойност за „големина на разкъсване”

(gap-extension penalties). Можете да видите падащите менюта за тези параметри на Фигура 7-6.

Таблица 7-5 описва тези параметри и евентуалните последствия от промяната им.

Таблица 7-5. Ефекти от промяната в настройката на параметрите в ClustalW.

Параметър	**Ефект**
Матрица на заместване (Substitution matrix)	Тези матрици оценяват стойността на мутациите в алайнмънта. Ако изберете категория матрици като PAM или BLOSUM, ClustalW автоматично избира най-адаптирания индекс. Трудно е да се предскаже ефекта от промяната на матрицата и няма идеална матрица. Ако вашите секвенции са с голяма идентичност, промяната на матрицата няма да има ефект. Ако алайнмънта, който получите е труден за интерпретиране, може да е полезно да промените матрицата от BLOSUM на PAM. Тези матрици измерват вероятността за промяната/мутацията на базите в хода на еволюцията, като така дават оценка при сравняването им между секвенциите в един алайнмънт.
Gap-opening penalty (GOP)	Този параметър оценява отварянето на „дупки”/разкъсвания на секвенция в алайнмънта. Колкото по-голяма е зададената стойност, толкова по-трудно е да се вмъкне „дупка” във вашия алайнмънт. GOP се прилага по един път за отварянето на всяка дупка. Ако го промените, това няма да има голям ефект, понеже ClustalW пренастройва тези стойности автоматично.
Gap-extension penalty (GEP)	Този параметър контролира размера на дупките в алайнмънта. Не е възможно да се предвиди оптималната двойка GOP/GEP, но за всяко семейство протеини например, съществува такава и само с опита може да се установи.

ВНИМАНИЕ! Никога не променяйте параметрите само, за да накарате ClustalW да произведе алайнмънта, за който знаете, че е правилен! Евентуално можете да го редактирате след това с някоя от програмите за тази цел. Има смисъл да променяте параметрите само, ако искате да разберете как малките промени могат да подобрят цялостния вид на вашия алайнмънт като блокове или слабо консервативни позиции.

Сравняване на секвенции и структури с Tcoffee

Tcoffee е съвременен метод за множествен алайнмънт на секвенции. Той използва принцип, донякъде подобен на този на ClustalW, но дава много по-точни алайнмънти за сметка на малко по-дълго работно време. Tcoffee прави прогресивен алайнмънт като ClustalW, но сравнява сегменти от целия набор от секвенции. В Таблица 7-6 са изброени основните приложения на Tcoffee.

Освен точността, други големи предимства на Tcoffee е способността му да сравнява секвенции и структури (EXPRESSO), възможността за оценяване на точността на алайнмънта (CORE), както и възможността да комбинира много алтернативни множествени алайнмънти в един (Mcoffee).

Таблица 7-6. Програмите, достъпни на [www.tcoffee.org.](http://www.tcoffee.org)

Програма	**С нея можете да**
TCOFFEE	правите множествени алайнмънти с Tcoffee.
CORE	оценявате надеждността на съществуващ множествен алайнмънт.
MCOFFEE	стартирате всякакви анализи за множествен алайнмънт и комбинирате всички резултати в един финален алайнмънт.
EXPRESSO	включите всяка достъпна структурна информация във вашия алайнмънт. Ако има известни структури, прави най-добрия алайнмънт на секвенции.

Множествени алайнмънти с Tcoffee

За правенето на множествен алайнмънт на обичайния (Regular) Tcoffee сървър ще трябва само да поставяте вашите секвенции в подходящия прозорец (по-рано в тази глава обяснихме как да си направите колекция от секвенции).

1. Отидете на главната страница на Tcoffee сървъра на www.tcoffee.org

2. Натиснете бутона T-Coffee.

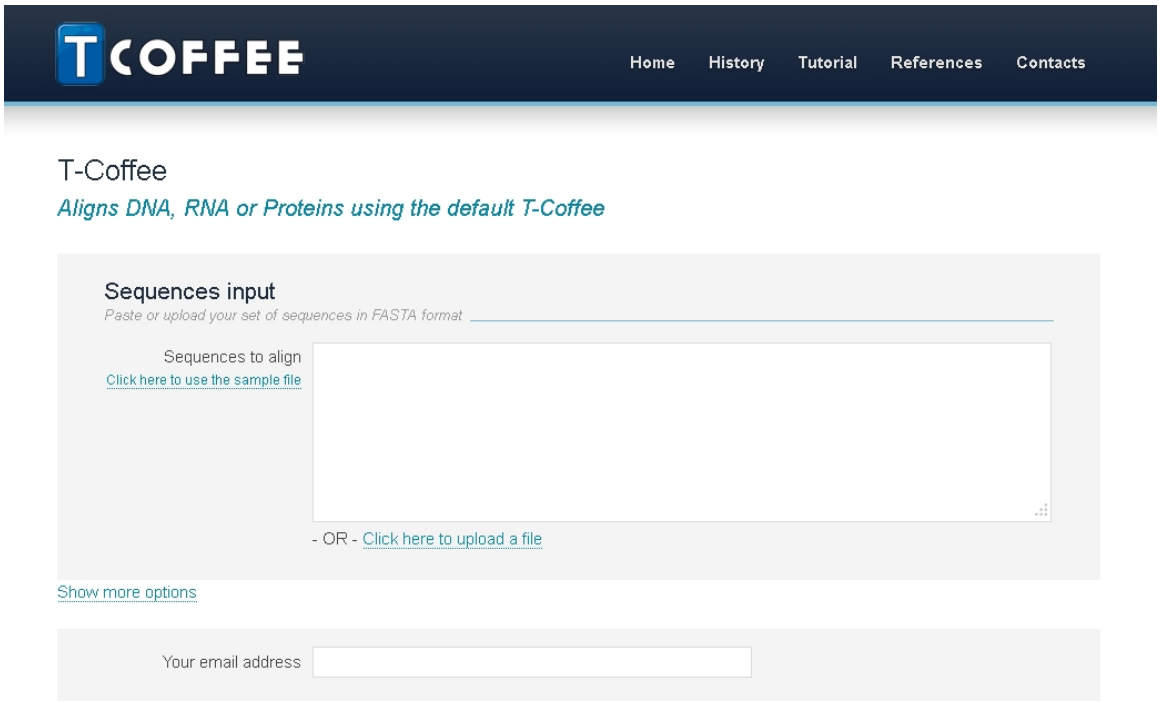
Появява се страницата за множествен алайнмънт (Фигура 7-8).

3. Поставете секвенциите си в големия прозорец.

Можете да използвате повечето формати. Ако секвенциите ви са в текстов файл (.txt), можете да го качите използвайки бутона Browse (Разглеждане).

4. Натиснете бутона Submit.

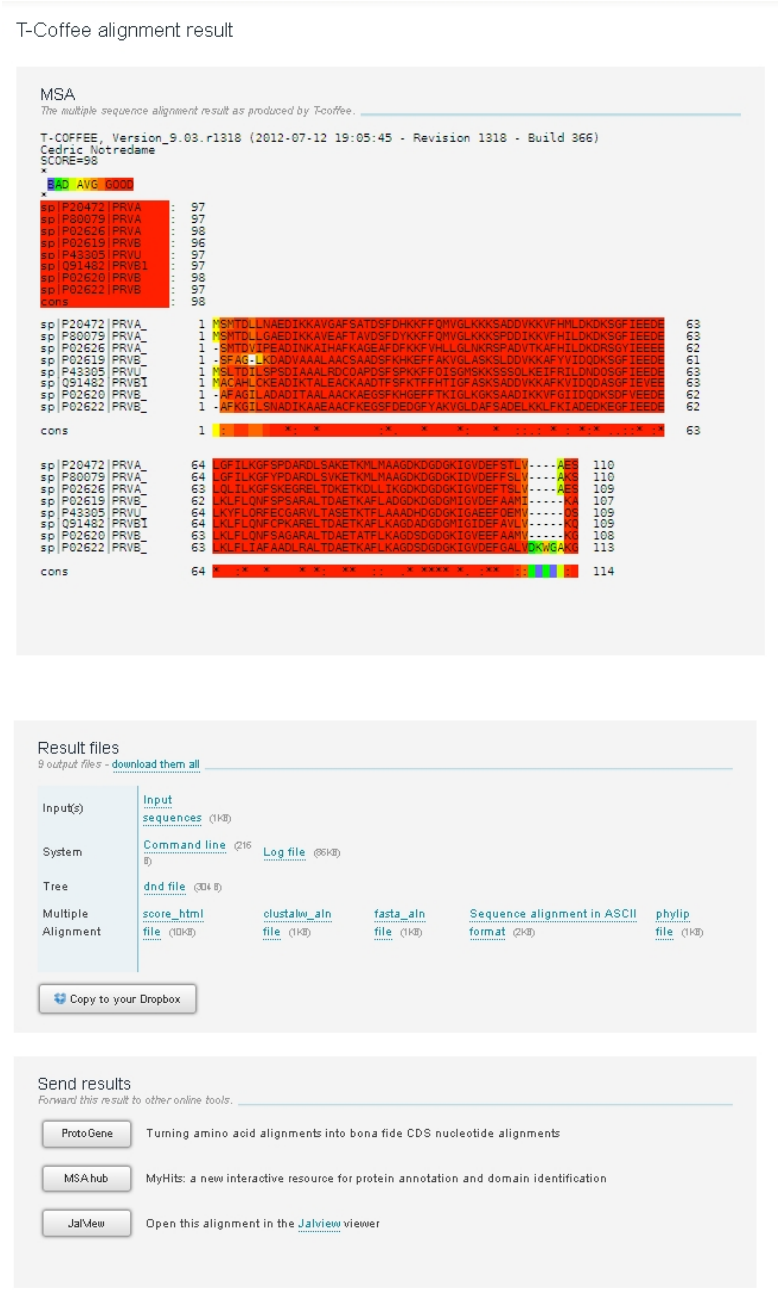
Tcoffee понякога е доста бавна. Можете да въведете e-mail във Web формата, за да ви уведомят, когато резултатите ви са готови.



Фигура 7-8. Началната страница за множествен алайнмънт на Tcoffee.

5. Разгледайте резултатите си.

Tcoffee дава резултатите в Таблица, съдържаща хипервръзки към вашите резултати, като на Фигура 7-9.



Фигура 7-9. Изходната страница с хипервръзки към резултатите на Tcoffee.

Първият ред от Таблицата е посветен на множествения алайнмънт и включва:

- **msf_aln, clustalw_aln, fasta_aln**: Текстови файлове с вашия алайнмънт в различни формати.

ПОЛЕЗЕН СЪВЕТ: Запазете тези файлове, ако искате да използвате вашият алайнмънт като входен файл за друга програма.

- **score_html, score_pdf:** Оцветен алайнмънт, като цвета на фона на всеки нуклеотид/АК показва качеството на алайнмънта в тази област. Червено означава висококачествен сегмент, синьо – областите, на които нямате основание да се доверите. **score_pdf** е **pdf** версия на **html** файла.
  ВНИМАНИЕ! Тези два файла са предназначени само за разглеждане – не можете да ги използвате като входни файлове за други програми за анализ на секвенции.

Вторият ред се отнася за **.dnd** файла - водещото дърво (дендрограма), генерирана от Tcoffee в Newick формат.

Комбиниране на секвенции и структури с EXPRESSO

EXPRESSO е последното изобретение към Tcoffee, което замества предишното 3D-Coffee. Когато стартирате Expresso, програмата използва BLAST за да търси в PDB (структурна база данни) за структури, чиито секвенции са подобни на вашите секвенции. След това използва тези структури за подпомагане на алайнмънта.

За алайнмънтите, основани на структури се очаква да са много по-точни от обикновените, направени само по секвенции.

EXPRESSO е по-бавна от Tcoffee, но ако намери достатъчно структури, тя произвежда най-точните алайнмънти на секвенции известни в момента, така че си струва да я опитате.

EXPRESSO сравнява структурите със SAP – програма на Taylor & Orengo, и сравнява секвенции и структури с помощта на FUGUE, софтуер разработен в университета в Кеймбридж.

За да стартирате EXPRESSO, натиснете бутона **Regular** на линията EXPRESSO на www.tcoffee.org . Останалото е напълно идентично с Tcoffee — дотолкова, че дори няма да разберете, че сте използвали структури!

ПОЛЕЗЕН СЪВЕТ: В секцията с резултатите потърсете файла, наречен **template_list**. Той съдържа списък на всички структури, които програмата е успяла да асоциира със секвенциите, които сте и подали. Ако този файл е празен, значи никаква структура не е

съвпаднала с вашите секвенции. Тогава вашия EXPRESSO алайнмънт просто е бил стандартният алайнмънт с Tcoffee.

Оценка на качеството на алайнмънт с CORE

Tcoffee може да оценява качеството на вашия алайнмънт, с което да ви помогне да разберете на кои части от него можете да вярвате и кои е по-добре да изхвърлите.

Изрежете и поставете вашия алайнмънт в CORE сървъра на www.tcoffee.org. Може да използвате всеки от по-често използваните формати (MSF, ALN, FASTA и PIR).

ПОЛЕЗЕН СЪВЕТ: Тези оценки са само емпирични – те не заместват E-стойностите. Все пак е полезно да знаете, че всички бази или АК показани на жълт/оранжев/червен фон имат индекс над 5 (до 10) — това е над 80% вероятност да са били сравнени коректно.

Как работи Tcoffee?

Tcoffee е метод за прогресивен алайнмънт и в много отношения е по-усъвършенстван „роднина" на ClustalW. Основната разлика между тях е, че Tcoffee не използва директно матрици на заместване при алайнмънта на секвенциите. В процеса на сравняване на секвенциите, Tcoffee също започва със сравняване по двойки като прави глобален алайнмънт с ClustalW, но също така прави и локално сравняване на същите двойки с програмата Lalign, която генерира 10 най-добри локални алайнмънта за всяка двойка секвенции. Цялата колекция от локални и глобални алайнмънти се нарича библиотека (library). След като завърши библиотеката, Tcoffee построява множествен алайнмънт, който е с най-голямото възможно съответствие с всички алайнмънти на двойки секвенции, налични в библиотеката, използвайки отново прогресивен алгоритъм като ClustalW. Процесът прилича на политически избори, в които всички битове информация налични в библиотеката се състезават за място в крайния алайнмънт. В библиотеката може да има много неща – двойни алайнмънти, множествени алайнмънти, структурни алайнмънти на EXPRESSO и дори „ръчно" направени алайнмънти на базата на експериментална информация.

Обработка на големи масиви данни с MUSCLE

MUSCLE е нов член на семейството на програмите за множествен алайнмънт, но е много ефикасен софтуер, едновременно бърз и висококачествен. MUSCLE е идеален, ако искате да направите алайнмънт едновременно на няколко стотин секвенции. Можете да ги стартирате от много сървъри, включително от „домашната" му страница на www.drive5.com/muscle/. Тази страница съдържа работната документация и обяснения, а също и линк към самата програма, която е на адрес http://www.ebi.ac.uk/Tools/msa/muscle/.

Работата с MUSCLE е много опростена – трябва само да поставите секвенциите си в съответния прозорец.

Интерпретиране на вашия множествен алайнмънт

Да интерпретирате един множествен алайнмънт си е до голяма степен изкуство. Е-стойностите, които ви казват директно колко надеждно е например вашето търсене на подобие, все още не съществуват за множествените алайнмънти, така че оценката на коректността им е често въпрос на опит и досещане.

Алайнмънтите на ДНК секвенции са най-трудни за интерпретиране. Ако анализирате ДНК секвенции, ще ви трябва високо ниво на консервативност, понеже единични запазени колони най-вероятно ще са безсмислени. Един ДНК блок е информативен, само ако съдържа група от няколко идентични колони. Това е трудно понякога дори с ДНК от близкородствени видове. Затова повечето биолози, когато имат избор предпочитат множествените алайнмънти на протеини.

Разпознаване на добрите участъци в един протеинов алайнмънт

Най-убедителна оценка за един алайнмънт на протеинови секвенции можем да получим от знанията за белтъчните структури. В структурите на протеините има по-„меки" и бързо изменящи се участъци – „бримки", които се намират на повърхността и свързват по-„твърди" области с различна функция, които се изменят по-бавно от

бримките. Така че в един множествен алайнмънт можем да очакваме да има хубави блокове без дупки, съответстващи на важни структури, както и по-изменчиви региони, съдържащи дупки, замени и т.н. – съответстващи на „бримките". Алайнмънтът, показан на Фигура 7-10 (в следващият раздел) добре показва този принцип.

Все пак, как да разберем дали един блок е добър? Когато разглеждате един алайнмънт направен с ClustalW, MUSCLE или Tcoffee, ще забележите че последния ред съдържа странни символи - (*), (:) или (.). Тези три символа имат много точни значения:

- (*) звездичката означава напълно консервативна колона.
- (:) двете точки означават колона, в която всички АК остатъци имат приблизително еднакъв размер и свойства хидрофилност / хидрофобност.
- (.) Точката обозначава колона, в която размера и свойствата хидрофилността/хидрофобността са били запазени в еволюцията.

Тези означения ще са ви много полезни. С тяхна помощ много бързо и лесно ще преценявате алайнмънтите си. Например, един средно добър блок е компактна единица от поне 10-30 АК без разкъсвания, в която има поне 1-3 (*), малко повече (:), разположени близо до звездичките, а също и няколко точки, разпръснати из него. Точно това се вижда на Фигура 7-10.

Магията на множествените алайнмънти е в това, че 4 или 5 консервативни позиции от над 50 сравнявани АК може да са достатъчни да получим истински сигнал. Забележете, това е под 10% идентичност! Ако си спомняте, за да получим някаква интересна информация от алайнмънт на двойка АК секвенции ни трябват поне 25% идентичност между тях. Ето защо, множественият алайнмънт е толкова мощна техника! Той може да ви помогне да намерите истински съкровища в „зоната на здрача", където другите методи не могат да ви кажат нищо за потенциална хомология.

Друг критерий за полезен множествен алайнмънт е да знаете кои типове аминокиселини трябва да очаквате, че ще бъдат консервативни. Аминокиселините не са равностойни – всички те имат много характерни модели на изменение/консервативност в един множествен алайнмънт.

На Таблица 7-7 са изброени най-често срещаните свойства, асоциирани с някои консервативни колони, които ще срещнете във вашия алайнмънт.

ЗАПОМНЕТЕ! Консервативните колони в един множествен алайнмънт имат значение само когато обкръжаващите ги колони не са консервативни.

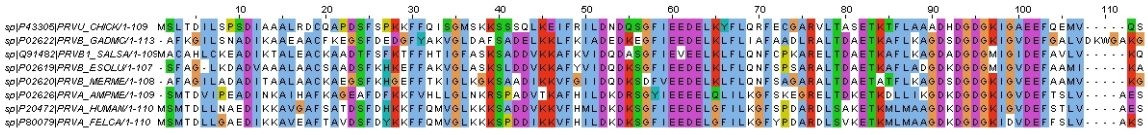
Таблица 7-7. Значения на консервативните АК в множествените алайнмънти.

Амино-киселина	Характеристика
W,Y, F	Консервативни колони с триптофан се срещат много често. Той има голям хидрофобен остатък, който се намира дълбоко в сърцевината на протеините и е много важен за тяхната стабилност. Когато триптофанът мутира (което се случва рядко), той обикновено се заменя от друга хидрофобна АК, като фенилаланин или тирозин. Консервативни колони с ароматни АК са най-често срещаните знаци за разпознаване на белтъчни домени.
G, P	Консервативни колони от глицин или пролин са често явление в множествените алайнмънти. Тези две АК често съвпадат с краищата на добре структурирани бета вериги или алфа спирали.
C	Цистеинът е известен със свойството си да прави C-C (дисулфидни) мостове. Консервативни колони с цистеини се срещат често и обикновено обозначават такива мостове, а когато са на определено разстояние дават добър знак за разпознаване на белтъчни домени и нагъвания.
H, S	Хистидинът и серинът често са участници в каталитични сайтове, особено на протеази. Тези АК са добри кандидати за участници в активен сайт.
K, R, D, E	Заредените АК лизин, аргинин, аспарагинова киселина и глутаминова киселина често участват в свързване на лиганди. Високо консервативни колони с тяхно участие може да означават и солеви мост в сърцевината на протеина.
L	Левцините рядко са много консервативни, освен ако не са включени в протеин-протеинови взаимодействия, например „левцинов цип” (leucine zipper).

Продължението на историята с нашия множествен алайнмънт

Това, което обикновено искате от вашия множествен алайнмънт е да ви покаже важните позиции в протеините, като разберете кои аминокиселини са се запазили – не им е било позволено да мутират дори при далекородствени протеини. Разгледайте например алайнмънта на Фигура7-10 (подобен можете да генерирате със всяка от разгледаните програми, а също и с други подобни). Това е добър алайнмънт. Той

съдържа далекородствени протеини и е начало на интересна история за това протеиново семейство. Виждаме например, че неговата N-крайна област изглежда по-консервативна от С-крайната. В N-края виждаме също къса верига от високо консервативни АК, което прави този регион добър кандидат за свързващ сайт или активен център. Това е интересно, но още не е достатъчно. Този алайнмънт съдържа твърде много консервативни позиции, за да му се направи детайлен анализ. Това, което можем да направим е да добавяме към него още няколко далекородствени секвенции една по една и внимателно да следим как ще се отрази това на качеството на алайнмънта. Най-вече трябва да следим дали тези далечни секвенции ще засилят сигналите, които вече имаме или напротив, напълно ще разрушат запазените блокове.



Фигура 7-10. Един добър множествен алайнмънт.

Една възможна стратегия да разкрием еволюционните ограничения върху нашия протеин е да включим в алайнмънта точно тези секвенции, които BLAST търсенето е извадило като маргинални попадения, когато сме търсели хомолози. Примерът по-долу е с белтъка парвалбумин (*parvalbumin*). Тук не е нужно да стартирате BLAST отново, за да видите този пример – по-долу е нужната ви информация.

Първо, съберете си секвенции по следния начин:

1. **Отидете на** http://www.uniprot.org/batch/

Появява се страницата на Swiss-Prot/TrEMBL **Retrieve a List of Entries**.

2. **На полето за формат (Format line), изберете радио-бутона FASTA.**
3. **Въведете номерата за достъп на вашите секвенции в съответния прозорец Sequence.**

Въвеждайте по един номер на ред. За този пример, ние сме въвели:

P20472, P80079, P02626, P02619, P43305, Q91482, P02620,P02622, P02586.

P02586 е TPCS_RABIT – тропонин С от заек. При BLAST търсенето с човешкия парвалбумин срещу Swiss-Prot, което направихме, за да съберем останалите секвенции, BLAST даде за това попадение E-стойност 5! Сам по себе си, този резултат

не е интересен, но след като вече имаме множествен алайнмънт можем да разберем, дали този заешки белтък може да ни каже нещо ново за човешкия парвалбумин.

4. **Натиснете бутона Create FTP File.**

5. **Копирайте секвенциите в Clipboard.**

Когато съберете всички секвенции, отново е време за алайнмънт.

1. **Отидете на ClustalW сървъра на EBI**

http://www.ebi.ac.uk/Tools/msa/clustalw2/

2. **Поставете събраните секвенции в съответния прозорец Sequence.**

3. **Изберете Fast от менюто за метод на алайнмънт.**

4. **Изберете Output Format, какъвто искате.**

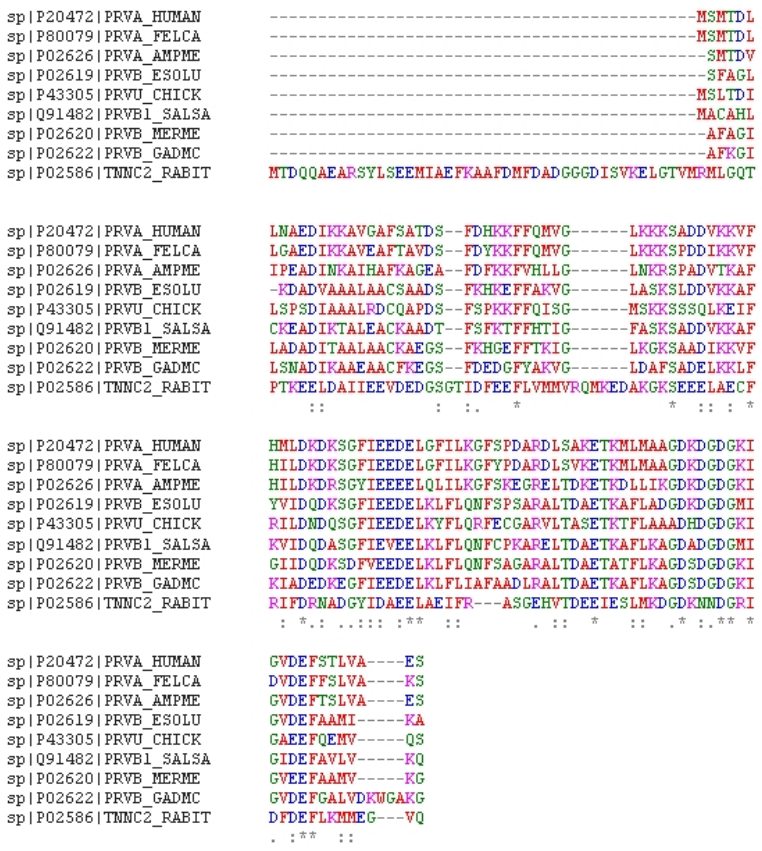
5. **Изберете Input от менюто за Output Order.**

6. **Натиснете бутона Submit.**

7. **Запазете резултатите си.**

Както виждате на Фигура 7-11, новата секвенция „се съобразява" с блоковете, които вече съществуват, макар и да са изчезнали някои консервативни позиции. Тя показва и области, в които са станали инсерции и делеции – добри кандидати за бримки, които по принцип са по-изменчиви.

Следващата Фигура 7-11 показва и съществуването на два консервативни региона в N-края на човешкия *parvalbumin*. Това е интересен резултат, който стеснява полето от възможности, ако търсим аминокиселините, които са отговорни за функцията на нашия протеин.



Фигура 7-11. Множествен алайнмънт на далекородствени протеини.

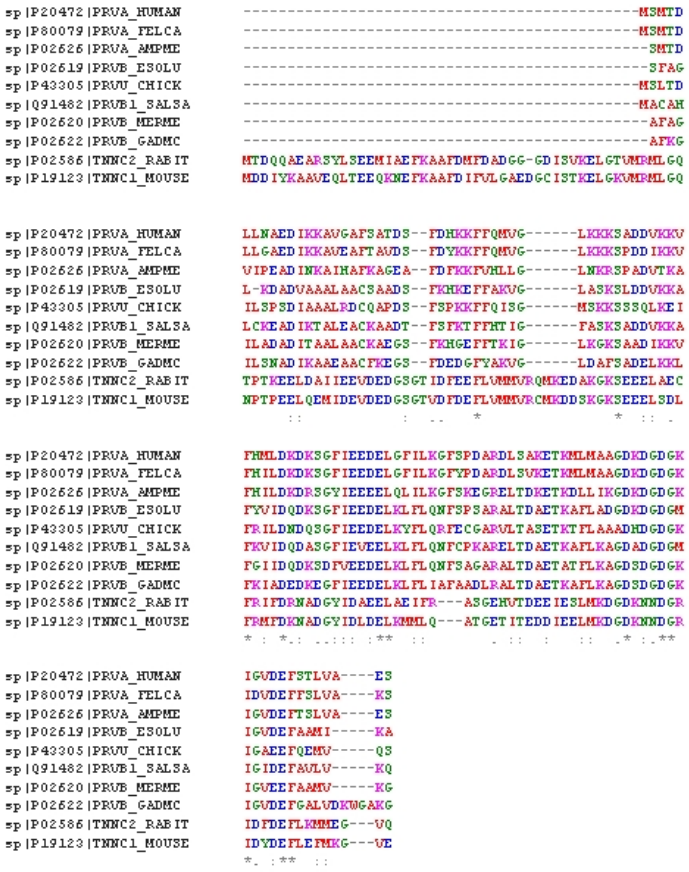
На този етап ще е добра идея, ако добавим още една далечна секвенция към нашия алайнмънт. Целта ни е да проверим, дали тези малки високо консервативни региони наистина са консервативни в цялото белтъчно семейство.

За да направите това, трябва пак да преминете през двата предишни алгоритъма, като този път използвате още една секвенция:

P20472, P80079, P02626, P02619, P43305, Q91482, P02620, P02622,

P02586, P19123

P19123 е TPCC_MOUSE – тропонин С на мишка, много далечен хомолог на човешкия парвалбумин. Фигура 7-12 показва резултатите от включването на този нов протеин в нашия множествен алайнмънт. Ясно се вижда, че най-консервативните колони се запазват.



Фигура 7-12. Включване на още по-далечни секвенции в множествения алайнмънт.

С тези резултати можем да сме сигурни, че двата консервативни региона със сигурност имат биологична функция. Знаем, че повечето от тези протеини свързват калций, така че има голяма вероятност калций-свързващият сайт да е в някоя от тези консервативни области. Това наистина е така, както се разкрива от Swiss-Prot анотацията.

Интернет ресурси за множествен алайнмънт

Има наистина огромен брой такива ресурси. Когато си избирате с кой да работите, имайте предвид следните съвети:

- Използвайте стабилни ресурси и винаги правете някакъв прост тест, за да разберете дали правят това, което очаквате да правят.
- Никога не се доверявайте сляпо на ресурси, над които нямате контрол.

- Ако искате да правите много алайнмънти, или по някаква причина не искате да „пускате” секвенциите си в Интернет, ще трябва да си инсталирате такива програми локално на вашия компютър.

Как да използваме ClustalW денонощно

ClustalW е най-популярната онлайн програма. Таблица 7-8 ви дава списък с достъпните сървъри. Последният ред не е сървър, а FTP сайт, откъдето можете да изтеглите ClustalW и да го инсталирате на своя компютър.

Таблица 7-8. Списък с ClustalW сървъри.

Име	Място	Уеб адрес
EBI	Европа	http://www.ebi.ac.uk/Tools/msa/clustalw2/ (от този адрес може да се пренасочите към новата версия - http://www.ebi.ac.uk/Tools/msa/clustalo/)
EMBnet	Европа	http://www.ch.embnet.org/software/ClustalW.html
PIR*	САЩ	http://pir.georgetown.edu/pirwww/search/multialn.shtml
BCM	САЩ	http://searchlauncher.bcm.tmc.edu/multi-align/multi-align.html
GenomeNet	Япония	http://align.genome.jp/
DDBJ	Япония	http://clustalw.ddbj.nig.ac.jp/top-e.html
Страсбург**	Европа	ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalW/

* От този сървър можете да стартирате по избор и трите програми – ClustalW, T-coffee и MUSCLE.

** Това е адресът, от който може да свалите ClustalW за локално инсталиране.

Открийте своя предпочитан метод за множествен алайнмънт

Ако ClustalW не е подходящ за секвенциите, които искате да анализирате, в Таблицата по-долу са изброени адреси на някои други сървъри.

Таблица 7-9. Ресурси за множествен алайнмънт в Интернет.

Метод	Описание / Предимство	Уеб адрес
Tcoffee	Точна комбинация на секвенции и структури	www.tcoffee.org http://tcoffee.vital-it.ch/cgi-bin/Tcoffee/tcoffee_cgi/index.cgi http://www.ebi.ac.uk/Tools/msa/tcoffee/
MUSCLE	Бърз и точен метод	www.drive5.com/muscle/ http://www.ebi.ac.uk/Tools/msa/muscle/
Kalign	Бърз метод	http://msa.sbc.su.se/cgi-bin/msa.cgi
MAFFT	Бърз и точен метод, използващ – Fast Fourier Transforms	http://align.bmr.kyushu-u.ac.jp/mafft/software/
Dialign	Идеален за секвенции с локална хомология	http://bibiserv.techfak.uni-bielefeld.de/dialign/

Как да сравняваме секвенции, на които не можем да направим алайнмънт

Понякога се налага да сравняваме секвенции, които не е задължително да имат общ произход или пък имат толкова далечно родство, че е трудно да бъдат разпознати като хомолози. Програмите за множествен алайнмънт обикновено са безсилни в такива ситуации. Проблеми има и когато неродствени протеини съдържат хомоложни домени, както често се случва при химерните протеини. Някои сегменти от ДНК пък могат да съдържат сходни регулаторни мотиви, въпреки че множественият алайнмънт не показва нищо общо между тях.

В подобни ситуации, не очаквайте чудеса от биоинформатиката. В повечето случаи първото ви впечатление, че няма нищо, което да извлечем от тези секвенции се оказва

правилно. Все пак преди да се предадете окончателно си струва да опитате двата описани по-долу метода. Първият - Gibbs sampler, изследва едновременно всичките ви секвенции за къси, частично запазени сегменти без разкъсвания. Вторият метод - Pratt, търси гъвкави модели – специална категория сегменти, които може да съдържат разкъсвания и са консервативни само по някои позиции.

Как да правим множествени локални алайнмънти с Gibbs sampler

Gibbs sampler е стохастичен метод: той първо разбърква секвенциите ви, а след това ги сравнява на случаен принцип, докато не се появи добър резултат, след което продължава с разбъркването, за да подобри още този резултат. Интересното, когато използваме случайността е, че можем да решаваме сложни проблеми, без да трябва да изследваме всички възможни решения.

Тези методи са наречени стохастични, защото те включват елемент на шанс. Стохастичните методи са засега най-мощните в биоинформатиката. За съжаление те не са много популярни, защото с тях е трудно един и същ резултат да се възпроизведе два пъти. Например, ако анализирате с Gibbs sampler един и същ набор от секвенции два пъти, може да не получите точно същите резултати. За повечето биолози възпроизводимостта на резултатите е основна – а и никой не обича, когато компютърът му си променя мнението през минута.

Все пак, ако допускате малко по-необичайна логика може да откриете, че Gibbs sampler дава доста добри решения за изключително сложни проблеми. Например, той е много добър за откриване на HTH (Helix-Turn-Helix) домени в дадено белтъчно семейство. Той е много добър и за откриване на общи регулаторни елементи между иначе несвързани ДНК секвенции.

Можете да използвате Gibbs sampler от неговата титулна-първа страница:

http://ccmbweb.ccv.brown.edu/gibbs/gibbs.html

Когато използвате Gibbs sampler не забравяйте, че той се нуждае от максимален брой секвенции, за да е достатъчно точен. Няма смисъл да го използвате с по-малко от 20 секвенции.

Търсене на консервативни мотиви

Gibbs sampler е полезен само ако секвенциите, които търсите имат еднаква дължина, както е при Helix-Turn-Helix домените. Но повечето от интересните мотиви, които бихте могли да срещнете имат различна дължина и често са слабо консервативни. В такива случаи единствената ви надежда е да намерите няколко високо консервативни АК или нуклеотиди, които са в основата на мотива. Ако секвенциите ви съдържат подобен мотив, но не са или са много слабо свързани, единственият начин да ги анализирате е да използвате pattern-finding motif. Има няколко инструменти за откриване на такива общи слабо запазени мотиви. Един от най-мощните от тях е Pratt, защото позволява известна гъвкавост на пространството между консервативните позиции. Можете да използвате също TEIRESIAS, MEME или SMILE. (Разгледайте Таблица 7-10, където е даден непълен списък с такива ресурси).

Ресурси за търсене на мотиви и модели

В Таблицата по-долу са изброени някои от най-използваните онлайн ресурси за търсене на общи мотиви в секвенции, твърде слабо свързани, за да им се направи множествен алайнмънт.

Таблица 7-10. Методи за търсене на мотиви в Интернет.

Метод	**Адрес**
Gibbs Sampler	http://mobyle.pasteur.fr/cgi-bin/portal.py http://ccmbweb.ccv.brown.edu/gibbs/gibbs.html
Pratt	http://www.ebi.ac.uk/Tools/pfa/pratt/
eMotif	http://motif.stanford.edu/projects.html
MEME	http://meme.sdsc.edu/meme/
TEIRESIAS	http://cbcsrv.watson.ibm.com/Tspd.html
Improbizer	http://users.soe.ucsc.edu/~kent/improbizer/improbizer.html

Търсене на хомоложни секвенции в бази данни

В тази част:

- Откриване на подобие и хомоложност и с какво това може да е полезно;
- Разясняване на всичко, което трябва да се знае за хомологията;
- Стартиране на BLAST online, за да се намерят секвенции, подобни на вашите;
- Тълкуване на резултатите, които се получават;
- Избор на вид BLAST;
- Търсене на отдалечена хомология с помощта на PSI-BLAST;
- Откриване на нови протеинови домени с помощта на PSI-BLAST.

"Когато оптимистите търсят игла в купа сено, носят ръкавици."

Из малка книжка на полезните неща

Ако имате протеин или ДНК секвенция и искате да намерите секвенции подобни на тях, то вие сте попаднали на правилното място. В тази глава ще ви покажем как да използвате BLAST, за да сравните вашата секвенция с всяка една секвенция, съдържаща се в базата данни и възможността да се получат най-добрите резултати буквално с две кликвания на мишката. Ще ви покажем случаи, в които можете да използвате PSI-BLAST, разновидност на BLAST, с много повече възможности при редица биологични проблеми.

От друга страна, ако знаете името на вашата секвенция или може да я опишете само с думи (например миши серин, протеаза и т.н.), а не с реалната и аминокиселинна последователност, следваща стъпка е да се влезе в базата данни и да се провери за тази секвенция. В първата глава обясняваме как да търсите секвенции, като използвате имена или характеристики.

Подобието – най-важния признак!

Сходни секвенции често водят началото си от една и съща последователност - предшественик. Това означава, че ако вашите секвенции са сходни, те най-вероятно

имат общ произход, притежават една и съща структура и имат подобна биологична функция. Този принцип е валиден, дори когато секвенциите произхождат от много различни организми. Това означава, че това което се знае за определена ДНК секвенция или протеин, може да се екстраполира за всички подобни ДНК секвенции и протеини.

Например, представете си, че вашата секвенция много прилича на секвенция, която се изучава в друга лаборатория. Тъй като тези две последователности са сходни, може да се каже, че "Ако нещо е вярно за вашата секвенция, то вероятно е вярно и за другата!". Само си представете колко време може да спестите, защото изследване на даден ген в лаборатория отнема години, а търсене на подобие в една база данни отнема секунди. И никой не може да ви нарече измамник!

Когато два протеина или гена са много сходни, биолозите ги наричат хомолози, което е “модерна” дума за два протеина или гена, имащи общ произход, сходни функции и подобни структури. Оттук идва проблемът – колко точно подобни са “много” подобни. Правилото гласи, че ако вашите секвенции са дълги повече от 100 аминокиселини (или 100 нуклеотида), могат да се нарекат "хомоложни", ако 25 на сто от аминокиселините са идентични, за ДНК ще е необходимо най-малко 70 на сто идентичност, за да се изкаже същото твърдение. Ако вашите стойности са под посочените границата която сте посочили, то всяко едно предположение е еднакво вероятно. Под 25% е междинна зона, за която важи, че нищо не е сигурно по отношение на наблюдаваните прилики.

- **Нищо не е сигурно по отношение на наблюдавано подобие.**

За някои протеини се знае, че имат под 15 на сто идентичност в аминокиселинната си последователност и имат напълно еднаква 3-D структура, докато други протеини с 20% идентичност имат различни.

- **Хомоложността или липсата на хомоложност никога не е гарантирана.**

25-процентовата гранична стойност определя междинна зона, която е относителен показател. В действителност нещата са много по-сложни. В повечето случаи, за да се уверите, че две секвенции са наистина хомоложни, трябва да използвате и друга информация, получена като резултат от вашето търсене. Тези битове от данни включват:

- **Очакваната стойност (E-value)**, която показва колко вероятно е подобието между вашата секвенция и последователността от базата данни, да се дължи на случайни събития;

- Дължината на подобните фрагменти между двете секвенции;

- Аминокиселинната консервативност;

- Броя на инсерциите и делециите.

Не бъркайте хомоложност и подобие! Хомоложността е бинарно отношение, докато подобието е количествена характеристика. С други думи, две секвенции са, или

хомоложни, или не-хомоложни, но те винаги ще имат измеримо ниво на подобие. Следователно, не може да се каже, че две секвенции са 40% хомоложни, точно както не може да се каже за една жена, че е 40% бременна.

В тази глава ще покажем как се интерпретира резултат при търсене в дадена база данни.

Най-популярният метод за работа с бази данни – BLAST

Само до преди тридесет години биолозите са търсели в базите данни, чрез отпечатване на цялото съдържание на хартиен носител, залепвайки принтираното копие на стената на кабинета, записвайки тяхната секвенция на късче хартия и прекарвайки часове наред в ръчно претърсване на информацията на стената, пиейки кафе. С наличните вече милиони секвенции тези дни са отминали. За щастие сега могат да се ползват компютрите, за да се стартира най-успешния инструмент в областта на биоинформатиката – BLAST („основен инструмент за търсене чрез локален алайнмънт" – **B**asic **L**ocal **A**lignment **S**earch **T**ool), който е предназначен за сканиране на наличните данни по параметрите на вашата заявка.

В този раздел ще покажем два основни начина как да се извърши BLAST (чрез използване на протеинови или ДНК секвенции) и най-важното - ще покажем как да се тълкуват получените резултати. В края на този раздел ще дадем идеи как и кога да се използва BLAST.

"Бластване" на протеинови секвенции

Бластване на протеинови секвенции е това, което искате да направите, ако вече имате протеинова последователност и искате да намерите други подобни на нея в съответната база данни. Съществуват основно две разновидности на протеинов BLAST:

- blastp - сравнява една протеинова последователност с цялата протеинова база данни;
- tblastn – сравнява една протеинова последователност с цялата нуклеотидна база данни.

Ако не сте сигурни как да подходите с вашата протеинова секвенция, тогава е по-добре да се използва blastp. Когато се съмнявате, винаги използвайте blastp! Въпреки това, ако подозирате, че blastp не е най-подходящия избор, по-добре е да се опитате да отговорите на въпроса: „Какво се опитваме да открием?“ и погледнете Таблица 5-1.

Таблица 5-1. Избор на подходящ протеинов BLAST.

Какво искате да направите	**Избор на правилен BLAST**
Искате да намерите нещо характерно, свързано с функцията на даден протеин.	Използвайте blastp, за да сравните този протеин с всички от БД.
Искате да откриете нови гени, кодиращи протеини.	Използвайте tblastn, за да сравните вашия протеин с ДНК секвенциите, транслирани в шестте възможни рамки на четене (три за всяка от веригите).

- blastp сървъра се намира в Националния център по биотехнологична информация (NCBI) в САЩ;
- blastn сървъра се намира в Швейцарския EMB-мрежов сървър.

Тези два сървъра имат малки разлики в интерфейса. Ако знаете как да работите и с двата, трябва да сте сигурни, че може да използвате всеки един от BLAST сървърите, изброени в края на тази глава.

Две от основните причини, за да познавате повече сървъри, извършващи BLAST са:

- **База данни:** Различните BLAST сървъри не ви дават достъп до едни и същи база данни. Ако не намерите базата данни, от която имате нужда на даден сървър, винаги може да я потърсите на друг.
- **Скоростта:** Най-популярните сървъри често са претоварени. Когато това се случи, добрата опция е да проведете търсенето на друг сървър.

Малко вероятно е два различни BLAST сървъра да ви дадат един и същ резултат при определена заявка от ваша страна. Основната причина за това е, че различните бази данни често се различават малко, както и версиите на програмите за BLAST. Резултатите, обаче рядко се различават драстично.

Стартиране на blastp в NCBI

Пример: "Има ли протеини, подобни на желязната супероксид дисмутаза при *Arabidops thaliana,* в протеиновата база данни Swiss-Prot?". Добър въпрос. Ето как да намерите отговора:

1. **Стартирайте вашия браузер и заредете NCBI BLAST сървъра:**
   http://blast.ncbi.nlm.nih.gov/Blast.cgi
   Страницата за BLAST в NCBI изглежда, както е показано на Фигура 5-1.

2. **В раздела за протеини, кликнете на връзката протеин-протеин BLAST (blastp).**
   Появява се прозорец подобен на този показан на Фигура 5-2.

3. **В раздела за протеини, кликнете на връзката протеин-протеин BLAST (blastp).**
   Появява се прозорец подобен на този показан на Фигура 5-2.

4. **Поставете вашата секвенция в прозореца за търсене.**
   Има два начина, по които може да поставите вашата секвенция на blastp сървъра:

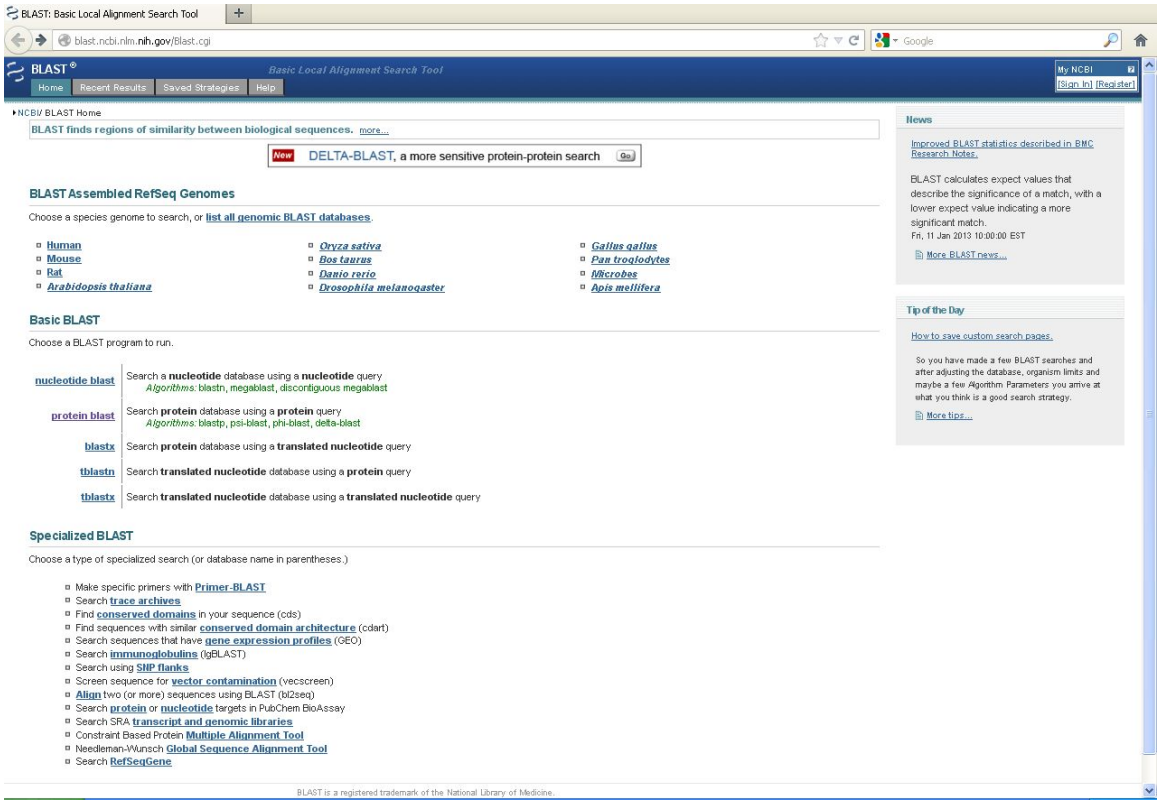
   A. Ако тази секвенция е извлечена от дадената база данни, може да дадете нейния идентификационен номер (номер за достъп). За нашия конкретен пример в Swiss-Prot, идентификационния номер на желязната супероксид дисмутаза при *Arabidopsi thaliana* е P21276. За целта въведете P21276 в прозорчето за търсене, както е показано на Фигура 5-2.

   B. Ако секвенцията не съществува в базата данни, трябва да я въведете във FASTA формат. Фигура 5-3 ви показва как изглежда секвенцията на нуклеолина при хамстер във FASTA формат.

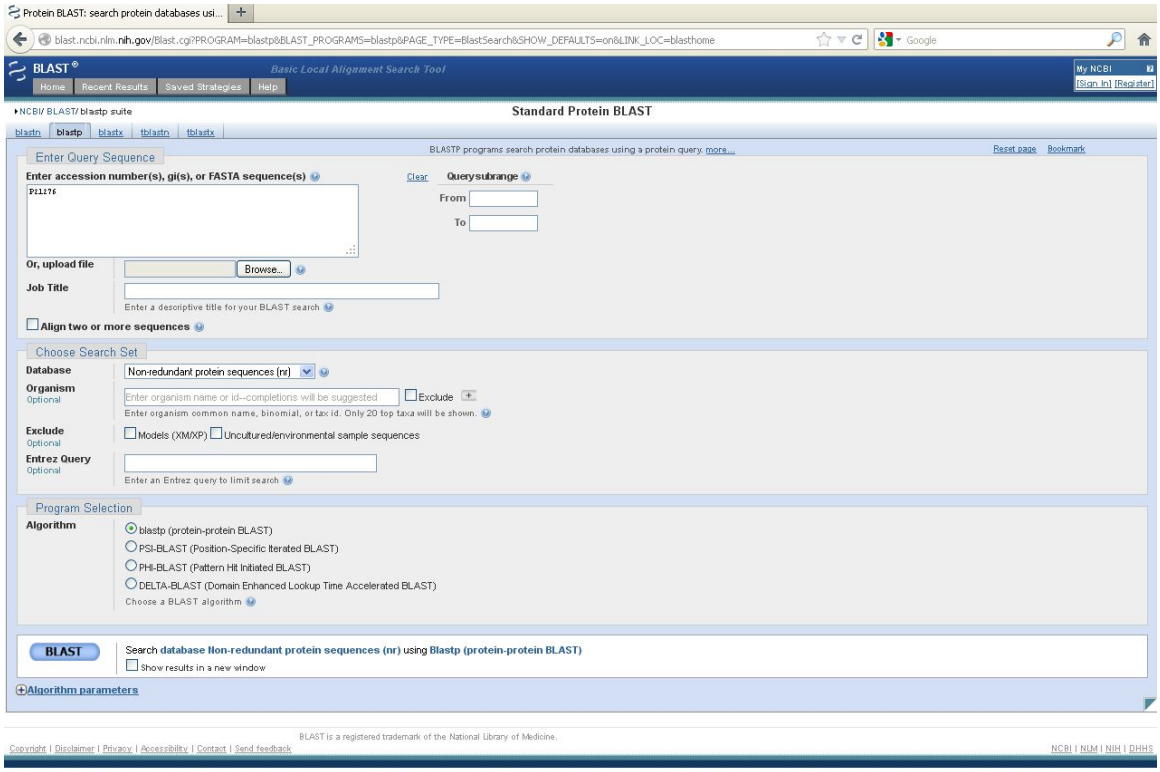
Секвенцията, която поставяте в blastp е вашата заявка (query).

Секвенциите, които ви връща blastp се наричат попадения или съвпадения (hits).

Базата данни, в която търсите е таргетна (прицелна) база данни.



Фигура 5-1. Начална страница на BLAST в NCBI.



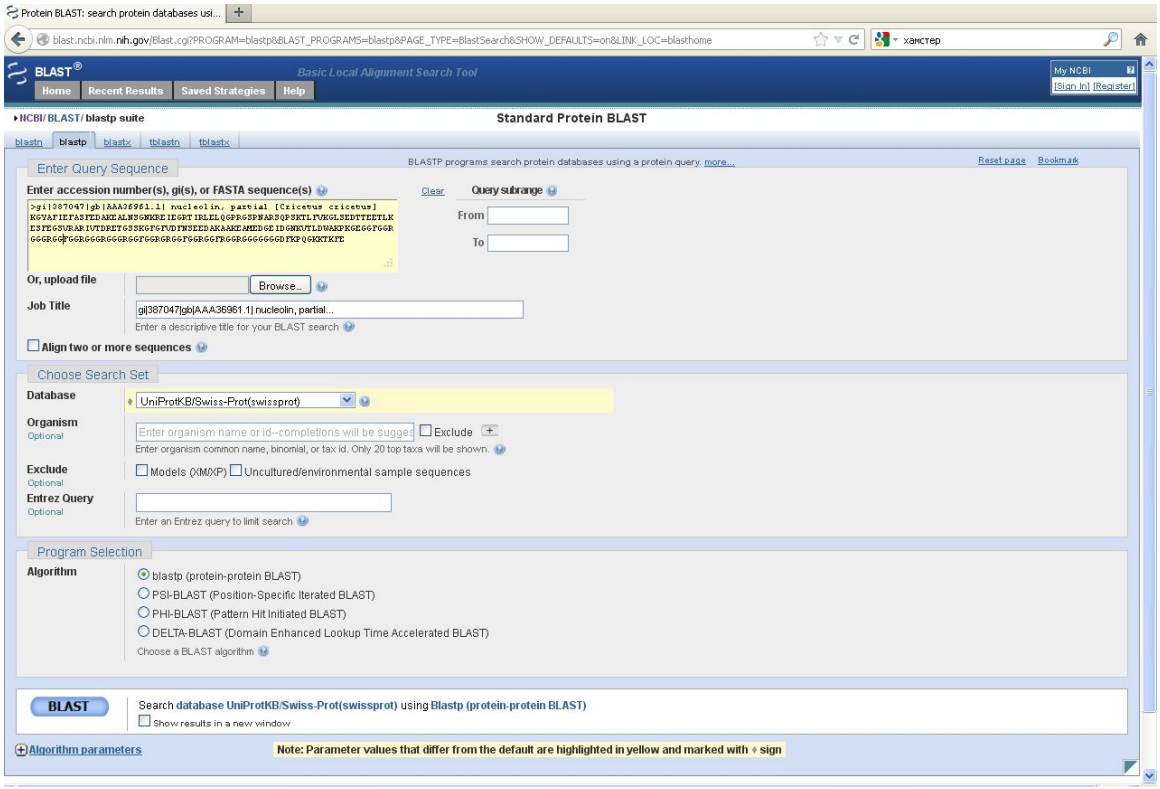
Фигура 5-2. blastp сървъра на NCBI.

5. **Изберете Swiss-Prot от Choose Search Set: Database падащото меню.**

6. **Кликнете на бутона BLAST (и изчакайте).**

Появява се страница подобна на тази показана на Фигура 5-4.

Понякога NCBI BLAST сървъра е много претоварен. Той е известен сред учените с това, че на него могат да стартират голям брой дейности точно към края на работния ден и впоследствие могат да управляват от къщи през нощта. Това е едно от обясненията, защо не е удобно да се ползва BLAST през нощта. За щастие NCBI не е единственият BLAST сървър и винаги може да опитате късмета си на друг сървър (виж Таблица 5-6 в края на тази глава за списък на най-популярните BLAST сървъри).



Фигура 5-3. Осигуряване на секвенция във FASTA формат на NCBI сървъра.

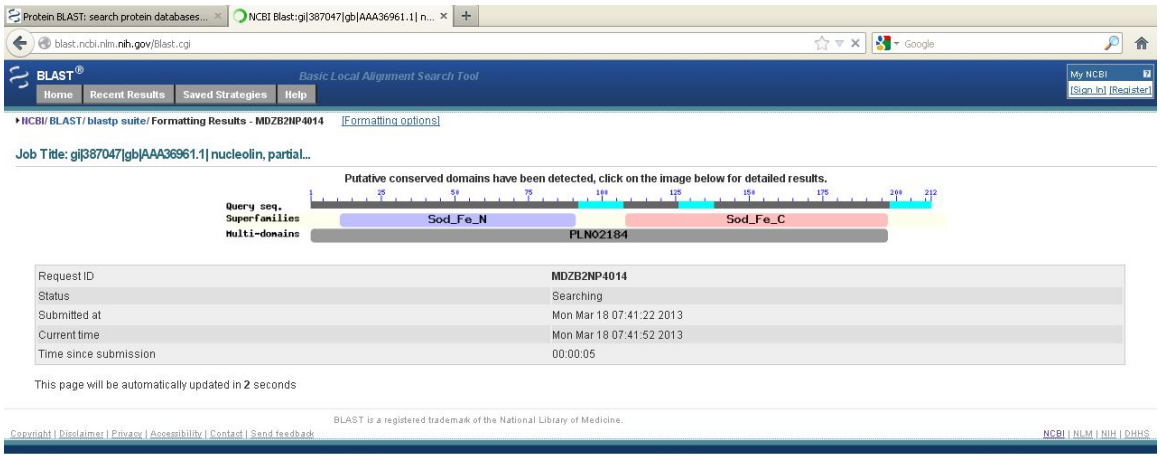
Ако най-близко разположения до вас сървър е прекалено бавен, може да се възползвате от различните часови зони. Таблицата по-долу показва коя част от света спи, докато вие си вършите работата сутрин или в следобедните часове.

Вашето местоположение	**Сутрин**	**Следобед**
USA	Япония	Европа

Европа	USA	Япония
Япония	Европа	USA

7. Щракнете върху Format! бутона от междинната страница и изчакайте резултата.

Когато щракнете върху Format! бутона, се отваря нов прозорец на браузъра. Веднага след като приключи търсенето, BLAST показва резултатите в този нов прозорец. За съжаление, самия процес на търсене може да отнеме повече време, отколкото първоначално формата показва.



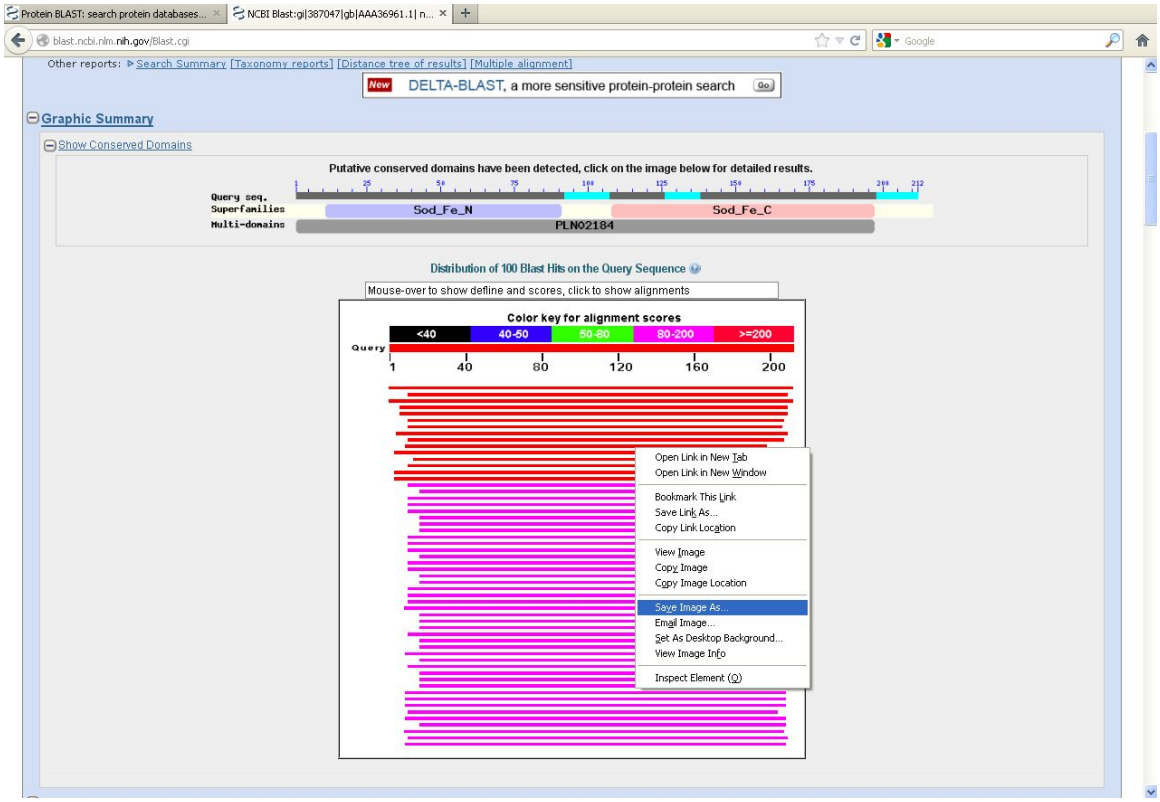
Фигура 5-4. BLAST страница с междинен резултат.

Едно стандартно търсене срещу NR или Swiss-Prot отнема само няколко минути. Ако BLAST показва съобщение, че вашето търсене ще отнеме повече от 20 минути, по-добре е да опитате късмета си по друго време или на друг BLAST сървър. НЕ натискайте никакъв бутон, докато чакате! Ако не получите задоволителен отговор, НЕ отправяйте отново същата заявка, това само ще направи нещата по-лоши за всички (включително и за вас)!

8. Запишете резултатите.

Не се опитвайте да разпечатате резултатите си, защото раздела за алайнмънт може да бъде огромен. Ако искате да запазите конкретно търсене, добре е да го запишете във файл с помощта на командите: File□Save As. За съжаление, цялостното запаметяване на една BLAST NCBI страница не е лесно. Обикновено от запаметените вече страници не може да се визуализират активните графики, когато се презареждат във вашия браузър. За този проблем има две решения (все още твърде незадоволителни):

A. **Съхранение на графични изображения:** Позиционирайте мишката върху графичното изображение, щракнете с десния бутон и изберете *Save Image As* от контекстното меню, което се появява, както е показано на Фигура 5-5. Запаметените изображения са в .gif формат. Въпреки, че тези изображения могат да се включват в електронни документи, трябва да е ясно, че всички активни хипервръзки са загубени.



Фигура 5-5. Съхраняване на графични изображения в NCBI.

B. **Отпечатайте резултата в pdf файл:** Ако имате инсталиран PDF принтер, може да отпечатате резултата в pdf файл. В зависимост от използвания PDF принтер, хипервръзките отново могат да се загубят.

Извършване на blastp в мрежата на EMB

BLAST сървърът работещ в мрежата на EMB е подобен на този в NCBI, но с различен интерфейс, подобен на много BLAST сървъри по света. Ако знаете как да използвате blastp EMBnet, ще знаете и как да използвате почти всички blastp сървъри. Основната разлика между blastp интерфейса на NCBI и интерфейса на EMBnet (показани на Фигура 5-6) е, че blastp EMBnet интерфейса дава възможност за избор измежду много

повече опции. Недостатък е, че трябва да се уверите лично, дали отделните избрани параметри са съвместими. В следващите стъпки ще покажем как да се избегнат всякакви грешки.

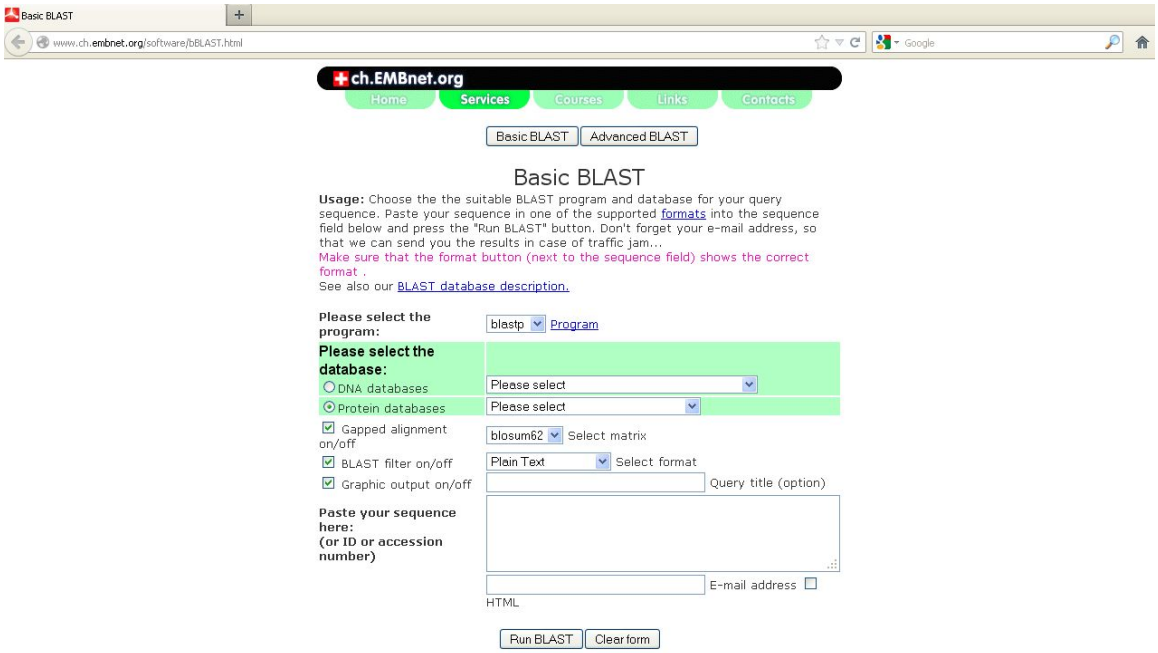
1. **. Заредете на вашия браузър следния адрес:**
   http://www.ch.embnet.org/software/bBLAST.html
   Появява се основната BLAST страница в EMBnet.

2. **Изберете blastp от падащото меню Select the Program.**
   Падащото меню ви позволява да изберете разновидност на BLAST, която искате да използвате.

3. **След като се уверите, че сте избрали опцията Protein Database, изберете протеинова база данни от придружаващото меню.**
   Базата данни, която изберете трябва да бъде съвместима с BLAST програмата, която сте избрали в стъпка 2; EMBnet сървърът няма да извърши тази проверка вместо вас. Видът на базата данни, в която може да търсите с конкретна заявена протеинова секвенция, зависи от това кой BLAST сте избрали в стъпка 2.

Вид BLAST	**Тип база данни**
blastp, blastx	протеини
blastn, tblastn, tblastx	ДНК

Правилният избор на геномна база данни може да направи по-лесен анализа на вашите резултати. Геномните бази данни се специфицират с име на съответния вид или дума от името (като *C. elegans*). Това може да е особено полезно, ако вашата заявка принадлежи към много голямо семейство. Ако знаете името на вида, който ви интересува, това е мястото, където да го изберете.



Фигура 5-6. Интерфейс на BLAST EMBnet.

Ако не може да намерите базата данни, която ви интересува, тогава я потърсете на други BLAST сървъри, като сървъра поддържан от Европейския Биоинформатичен Институт:

http://www.ebi.ac.uk/Tools/services/web/toolform.ebi?tool=ncbiblast&context=protein

4. **Имате възможност да поставите вашата секвенция в един от следните формати:** plain_text; SwissProt_ID or AC; TrEMBL_ID or AC; EMBL_ID or AC; EST_ID or AC; GenPept_gi; Yeast_ORF (може да ги видите като кликнете на бутона formats). Изберете от падащото меню: SwissProt_ID or AC.
   Plain text е FASTA формат без първия ред (този, който започва с >). Всички други формати се отнасят за идентификационни номера, а не за секвенции. Най-добре е да се запазят маркираните по подразбиране полета в различните селектори от дясната страна на формата:

   A. **Gapped Alignments On/Off:** дава възможност да се стартира BLAST в по-стара версия, която не дава възможност за отчитане на дупки. За да имате същия режим на работа, както NCBI Blast, запазете тази отметка.

   B. **BLAST Filter On/Off:** изключва филтър. Целта на този филтър е да предотврати отклонение от достоверността, при наличие на региони с повтаряне на една аминокиселина. Тази опция по подразбиране е включена и е добре да остане така, освен ако няма по-добра причина за изключването и (вижте раздел "Контрол на маскиране на последователности", по-надолу в тази глава).

C. **Graphic Output On/Off:** Запазете тази отметка включена, ако искате същия графичен изглед като в NCBI.

5. **Въведете идентификационния номер (или самата последователност) в полето за секвенции.**
   Нашия ID номер е P21276. Ако имате самата секвенция я поставете от паметта на компютъра в същия прозорец.

6. **Щракнете върху бутона Run BLAST.**

7. **Запишете резултатите.**
   BLAST резултатите могат да бъдат огромни. Никога не ги разпечатвайте (освен, ако не планирате да се отървете от целия си запас от хартия). Когато запазвате BLAST резултати, съхраняването на графичните изображения е изключително трудно. Ако все пак искате да ги запазите, съхранете ги поотделно. Поставете показалеца на мишката върху изображението, щракнете с десния бутон и изберете *Save Image As* от контекстното меню, което се появява. Ако имате достъп до програмата PDF принтер, отпечатайте резултата в PDF файл, това е най-добрият начин за поддържане на електронни записи за печат.

Анализиране на BLAST резултат

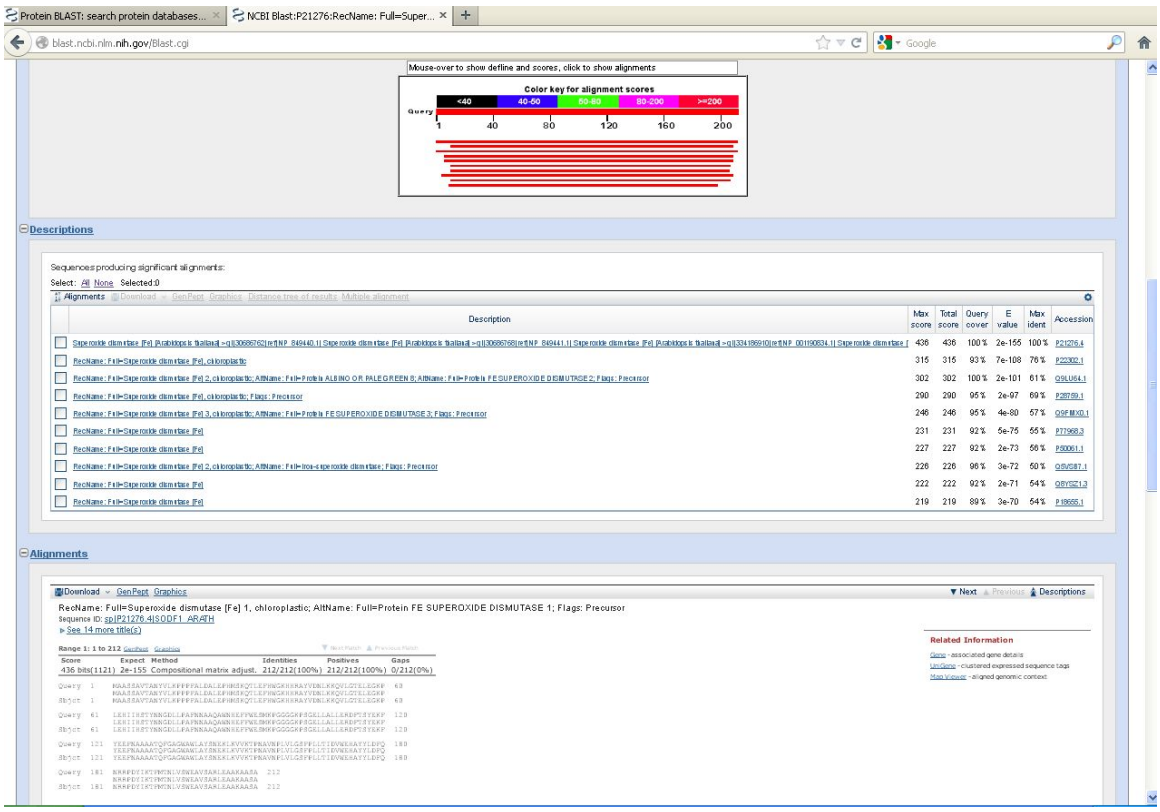
Всички видове BLAST дават подобни резултати. Те са доста сложни и често са объркващи дори и за опитни потребители. Фигура 5-7 показва цялостен изглед на BLAST резултат с неговите четири характерни области. Ако сте разбрали всичко от този резултат може да се наречете биоинформатик и другите основателно могат да заподозрат във вас наличие на магически сили!

Детайлен преглед на един BLAST резултат

При всеки BLAST резултат има четири области. Тези секции се показват винаги в една и съща последователност, както е показано на Фигура 5-7. Те включват:

1. **Графично изображение:** показва в кои области вашата заявка е подобна на други секвенции. В зависимост от сървъра, който използвате този дисплей може да е доста различен при някои сървъри дори може да липсва .
2. **Списък със съвпадения:** името на секвенциите подобни на вашата заявка, подредени по подобие.
3. **Алайнмънти:** Отделни алайнмънти между вашата заявка и всяко едно от намерените съвпадения.
4. **Параметри:** Списък на всички параметри, използвани при търсенето.

Всеки един от тези елементи съдържа много информация. Много полезно за вашите научни изследвания би било да се познава значението на всяко едно от показаните на екрана елементи.

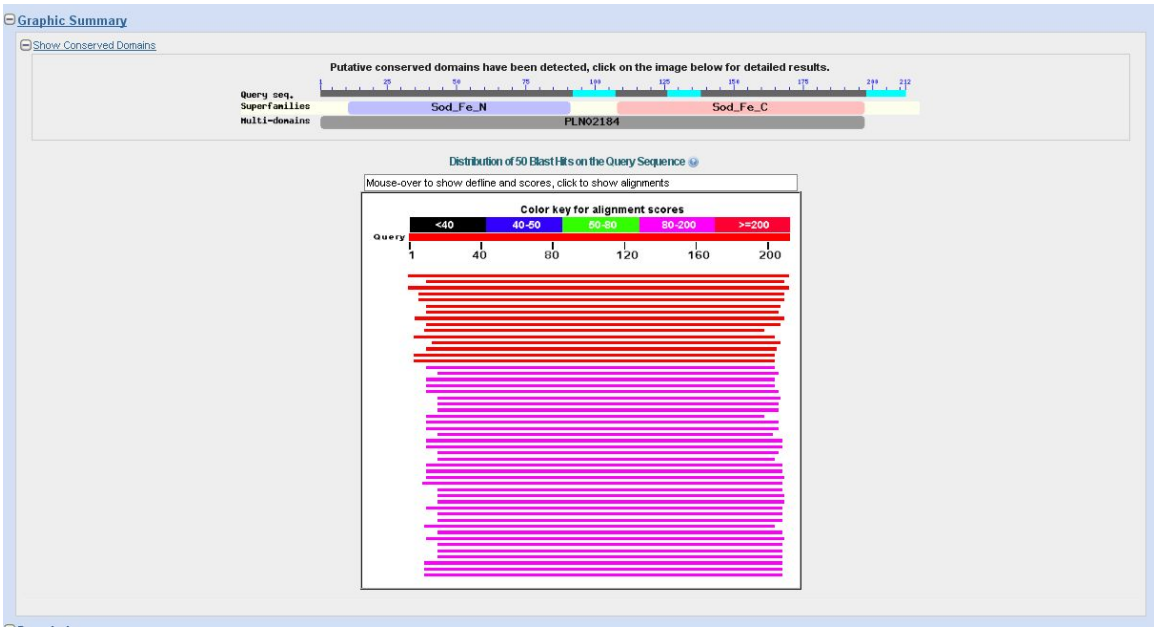


Фигура 5-7. Схематично представяне на основните компоненти на BLAST резултат (не е включена областта с параметри, разположени най-долу).

Графичното изображение

Екранът на Фигура 5-8 показва графично визуализиране на резултатите. Вашата заявена секвенция се показва най-отгоре. Всяка следваща ивица представлява част или цялата секвенция, подобна на вашата заявка, както и региона във вашата секвенция, където се наблюдава това подобие. Червените ивици показват секвенциите, които са най-подобни, розовите ивици показват малко по-слабо подобие, а зелените ивици показват подобие, което не е значимо. Ивиците в синьо и черно показват съвпадения с най-ниски стойности. Обикновено червените, розовите и зелените съвпадения са подходящи. Черните хитове са много слаби: протеини, които са толкова слабо подобни на заявката, че техния алайнмънт най-вероятно няма смисъл от биологична гледна точка (те се намират на граничната зона).

Това графично представяне ви показва и съвпадения, които не се простират по цялата дължина на вашата заявена секвенция. Това е полезно за откриване на домени. Например, на Фигура 5-8 най-горе се намират протеини, които са хомоложни на супероксид дисмутазата (СОД). Близо до най-долните резултати има по-къси, съответстващи на домена на СОД, към който се свързва РНК. Тези хитове показват протеини, които също съдържат този РНК свързващ домен, но са по друг начин свързани със СОД.



Фигура 5-8. Графичен дисплей на NCBI BLAST.

Това изображение в NCBI е активно. Ако позиционирате показалеца на мишката върху една от ивиците, то името на съответната последователност се появява в малък прозорец в горния край. Ако кликнете върху тази ивица, браузърът ще ви отведе към съответния алайнмънт, намиращ се в рамките на същата страница (Имайте предвид, че това графичното изображение в EMBnet не е активно).

Когато две цветни ивици са свързани с тънка черна линия, това означава, че един и същи протеин съответства на вашата заявка в две различни области (Имайте предвид, че тънка черна линия се вижда в NCBI; на EMBNet сървъра ще видите прекъсната черна линия). Тези съвпадения са независими, но те биха могли да бъдат обединени в по-дълго съвпадение.

Сектор със списък на съвпаденията (hit list)

Повечето потребители приемат областта със списъка на съвпаденията, като любимата част от BLAST резултата (подобен на показания на Фигура 5-9). Тази част ви показва веднага, дали вашата заявена секвенция изглежда като нещо, което вече се намира в базата данни, и дали може да му се доверите като добро попадение. Всяка ивица съдържа четири важни характеристики:

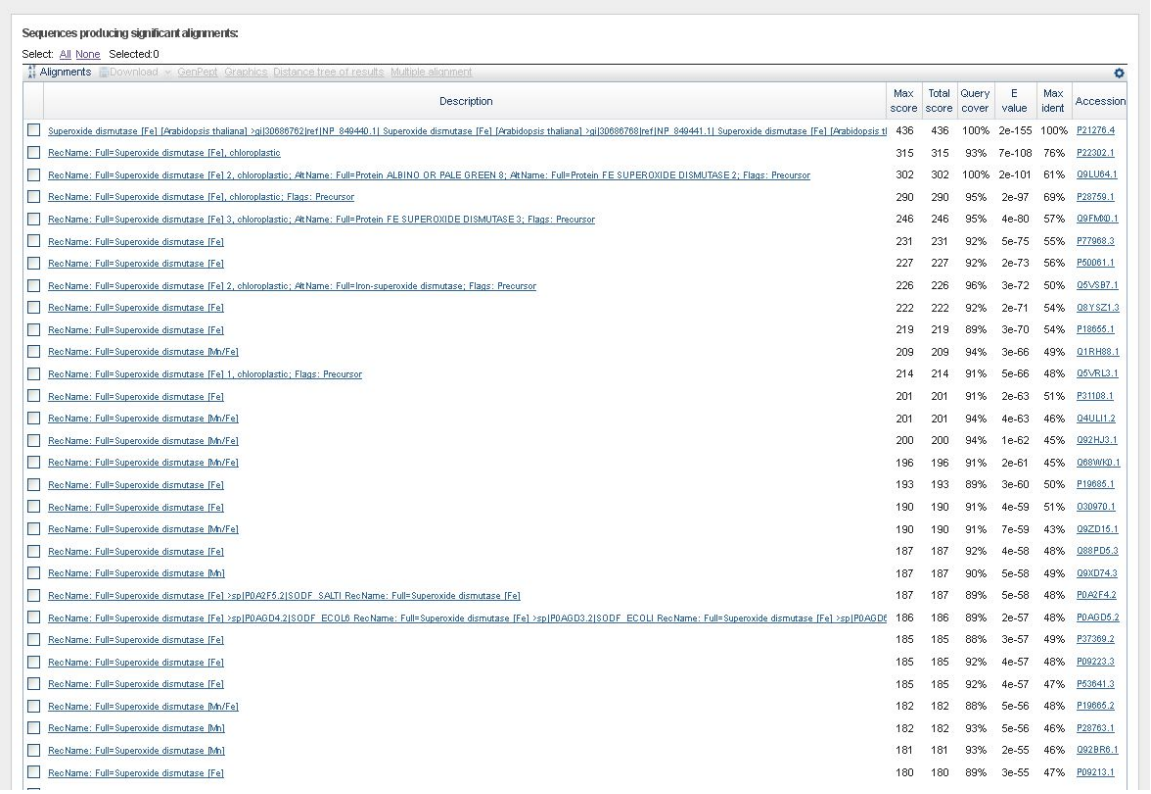
- **Номер за достъп на секвенцията и нейното име:** Тази хипервръзка може да ви отведе до записа в базата данни, съдържащ тази секвенция. В тази част може да намерите много полезна информация, описваща секвенцията. Swiss-Prot връзката (|SP| на Фигура 5-9) може да ви предостави допълнителна информация.

- **Описание:** Описанието на анотацията позволява да разберете веднага, дали намерените резултати биха били интересни за вас. Разбира се, никога няма гаранция, че дадената анотация е коректна и затова ние препоръчваме внимателно да проверите цялата анотация преди да се радвате напразно.

- **Bit score:** Мярка за статистическа значимост на алайнмънта. По-високата стойност показва по-високо подобие на двете секвенции. Попадения със стойност на Bit score под 50 са ненадеждни.

- **E-стойност (очаквана стойност):** Оценка на вероятността колко пъти да очаквате добро попадение на случаен принцип (за конкретна база данни). E-стойността ви предоставя най-важната мярка за статистическа значимост (вижте следващия раздел - "E-стойности, подобие и хомоложност"). По-ниската E-стойност предполага повече подобие между последователностите и по-голяма сигурност, че съответното попадение наистина е хомоложно на вашата заявка. Така например, секвенции идентични на вашата заявка имат E-стойности много близки до 0. **Попаденията със стойности над 0,001 са на граничната линия.**

- **Геномна връзка:** Вече са известни и на разположение много секвенирани геноми, което ви дава възможността да се запознаете с геномната локализация на потенциалното съвпадение. Когато тази информация е достъпна се обозначава с лилаво G. Кликнете на G и ще намерите всичко, което трябва да знаете за този ген.



Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Max ident	Accession
Superoxide dismutase [Fe] [Arabidopsis thaliana] >gi|30686762|ref|NP_849440.1| Superoxide dismutase [Fe] [Arabidopsis thaliana] >gi|30686768|ref|NP_849441.1| Superoxide dismutase [Fe] [Arabidopsis th	436	436	100%	2e-155	100%	P21276.4
RecName: Full=Superoxide dismutase [Fe], chloroplastic	315	315	93%	7e-108	76%	P22302.1
RecName: Full=Superoxide dismutase [Fe] 2, chloroplastic; AltName: Full=Protein ALBINO OR PALE GREEN 8; AltName: Full=Protein FE SUPEROXIDE DISMUTASE 2; Flags: Precursor	302	302	100%	2e-101	61%	Q9LU64.1
RecName: Full=Superoxide dismutase [Fe], chloroplastic; Flags: Precursor	290	290	95%	2e-97	69%	P28759.1
RecName: Full=Superoxide dismutase [Fe] 3, chloroplastic; AltName: Full=Protein FE SUPEROXIDE DISMUTASE 3; Flags: Precursor	246	246	95%	4e-80	57%	Q9FM90.1
RecName: Full=Superoxide dismutase [Fe]	231	231	92%	5e-75	55%	P77968.3
RecName: Full=Superoxide dismutase [Fe]	227	227	92%	2e-73	56%	P50061.1
RecName: Full=Superoxide dismutase [Fe] 2, chloroplastic; AltName: Full=Iron-superoxide dismutase; Flags: Precursor	226	226	96%	3e-72	50%	Q5VSB7.1
RecName: Full=Superoxide dismutase [Fe]	222	222	92%	2e-71	54%	Q8YSZ1.3
RecName: Full=Superoxide dismutase [Fe]	219	219	89%	3e-70	54%	P18655.1
RecName: Full=Superoxide dismutase [Mn/Fe]	209	209	94%	3e-66	49%	Q1RH88.1
RecName: Full=Superoxide dismutase [Fe] 1, chloroplastic; Flags: Precursor	214	214	91%	5e-66	48%	Q5VRL3.1
RecName: Full=Superoxide dismutase [Fe]	201	201	91%	2e-63	51%	P31108.1
RecName: Full=Superoxide dismutase [Mn/Fe]	201	201	94%	4e-63	46%	Q4ULI1.2
RecName: Full=Superoxide dismutase [Mn/Fe]	200	200	94%	1e-62	45%	Q92HJ3.1
RecName: Full=Superoxide dismutase [Mn/Fe]	196	196	91%	2e-61	45%	Q88WK0.1
RecName: Full=Superoxide dismutase [Fe]	193	193	89%	3e-60	50%	P19685.1
RecName: Full=Superoxide dismutase [Fe]	190	190	91%	4e-59	51%	O30970.1
RecName: Full=Superoxide dismutase [Mn/Fe]	190	190	91%	7e-59	43%	Q9ZD15.1
RecName: Full=Superoxide dismutase [Fe]	187	187	92%	4e-58	48%	Q88PD5.3
RecName: Full=Superoxide dismutase [Mn]	187	187	90%	5e-58	49%	Q9XD74.3
RecName: Full=Superoxide dismutase [Fe] >sp|P0A2F5.2|SODF_SALTI RecName: Full=Superoxide dismutase [Fe]	187	187	89%	5e-58	48%	P0A2F4.2
RecName: Full=Superoxide dismutase [Fe] >sp|P0AGD4.2|SODF_ECOL6 RecName: Full=Superoxide dismutase [Fe] >sp|P0AGD3.2|SODF_ECOLI RecName: Full=Superoxide dismutase [Fe] >sp|P0AGD6	186	186	89%	2e-57	48%	P0AGD5.2
RecName: Full=Superoxide dismutase [Fe]	185	185	88%	3e-57	49%	P37369.2
RecName: Full=Superoxide dismutase [Fe]	185	185	92%	4e-57	48%	P09223.3
RecName: Full=Superoxide dismutase [Fe]	185	185	92%	4e-57	47%	P53641.3
RecName: Full=Superoxide dismutase [Mn/Fe]	182	182	88%	5e-56	48%	P19665.2
RecName: Full=Superoxide dismutase [Mn]	182	182	93%	5e-56	46%	P28763.1
RecName: Full=Superoxide dismutase [Mn]	181	181	93%	2e-55	46%	Q92BR6.1
RecName: Full=Superoxide dismutase [Fe]	180	180	89%	3e-55	47%	P09213.1

Фигура 5-9. Списък с попадения на NCBI BLAST.

Областта с алайнмънти

Без значение какво казваме за E-стойностите и статистическата значимост, цифрата си е само една цифра. Когато говорим за реални неща, биолозите се доверяват само на алайнмънта. Те са убедени, че алайнмънта не може да лъже, и това е така, ако знаете как точно да го анализирате.

BLAST показва алайнмънтите точно под списъка, както е показано на Фигура 5-10. Във всеки алайнмънт на BLAST може да намерите следните характеристики.

- Име (имена): Първата характеристика е името. Понякога има повече имена на секвенции, това се случва всеки път, когато има няколко секвенции, идентични на вашата заявка. Например, ако базата данни съдържа няколко секвенции идентични на

вашата заявка, всички тези съвпадения ще бъдат съобщени като един резултат в раздела с алайнмънтите. Това дори може да стане и с така наречените Non-Redundant бази данни!

- Процент на идентичност: Тази характеристика ви дава конкретен заместител на E-стойността. Запомнете магическата цифра: идентичност над 25% е добра стойност (идентичност е броят на еднаквите остатъци, разделени на броя на сравнените остатъци – разкъсванията обикновено се игнорират). Положителните полета дават мярка за остатъците, които са идентични или подобни - представени с + в истинския алайнмънт. Gapsfield показва остатъците, които не са включени в алайнмънта.

E-стойност, подобие и хомоложност

Често високо ниво на подобие между две секвенции се свързва с общ прародител и с еднаква 3-D структура. Биолозите наричат такива секвенции хомоложни.

Често хомоложните секвенции притежават сходни биохимични функции. Ако изучавате един протеин, най-добрия възможен вариант е наличие на добре характеризирана протеинова последователност, която несъмнено е хомоложна на вашия протеин. Всички търсят в базите данни с надеждата да намерят точно тази секвенция. Остава открит въпроса как да се покаже, че двете секвенции са хомоложни? Гледайте на хомоложните секвенции като на роднини в едно семейство. Всички знаем, че роднините са склонни да си приличат, но също така знаем, че двама души с еднакъв цвят на очите, не са непременно брат и сестра. От друга страна, ако те имат един и същ тип коса, същите черти на лицето и т.н., ние може да се изкушим да предположим, че те са наистина роднини. По същия начин се подхожда и при сравняване на секвенции. Колко подобни трябва да са секвенциите, за да се приемат за хомоложни? Отговорът е ясен: Трябва да са подобни повече от 25 на сто от аминокиселините в протеина и над 70 на сто от нуклеотидите в ДНК. Над тази граница може да бъдете почти сигурни, че двата протеина имат подобна структура и общ прародител. Под тази граница никой не може да бъде сигурен, дали наистина наблюдаваното сходство означава нещо.

Предупреждение:

Бъдете внимателни - 25% и 70% са гранични стойности за секвенции, които съдържат повече от 100 аминокиселини и нуклеотиди. Възможно е да се получи почти перфектна идентичност между къси сегменти (например 10 остатъка) на напълно несвързани протеини или ДНК секвенции. Всички харесват визуализация

чрез процентна идентичност, но за съжаление това не винаги е добър индикатор. Например, резултат от конкретен алайнмънт е незадоволителен, когато аминокиселините са много подобни, но не са идентични. Освен това, как да се покаже разликата между 60 съвпадащи остатъка по дължината на секвенция с дължина 100 остатъка и 120 съвпадащи остатъка в секвенция с дължина 200? По-дългата, вероятно е по-значима, но процента на идентичност не споменава нищо за това. За това E-стойностите са мощен инструмент за сравняване на двойки секвенции с различно подобие и различни дължини. Те също ви помагат да решите доколко може да се доверите на заключението за хомоложност. E-стойностите показват, дали да се радвате или да поизчакате, за да видите какво наистина ще се получи. Те ви показват възможността алайнмънта ви да е станал случайно.

E-стойността има конкретен смисъл: Това е броят на случаите, в които базата данни дава наличие на съвпадение в резултат на случайни събития. Ние приемаме, че дадено съвпадение е добро, ако е малко вероятно да се случи в резултат на случайни събития. Резултати, свързани с ниски E-стойности са най-надеждни. Казваме, че те са най-сигурни, защото можем да им се доверим и да твърдим възможна хомоложност. На теория, един алайнмънт може да се приеме за достоверен при E-стойност по-ниска от единица. На практика обаче, това не е вярно, защото BLAST използва приблизителна формула за изчисляване на E-стойности и силно ги подценява. В света на секвенциите, подобие с E-стойност над 10^{-4} (0.0001) не е задължително да е значима. Ако искате да сте сигурни в наличието на хомоложност, то E-стойността трябва да бъде по-ниска от 10^{-4}. BLAST не е единствената програма, която използва E-стойности. Принципът е винаги един и същ.

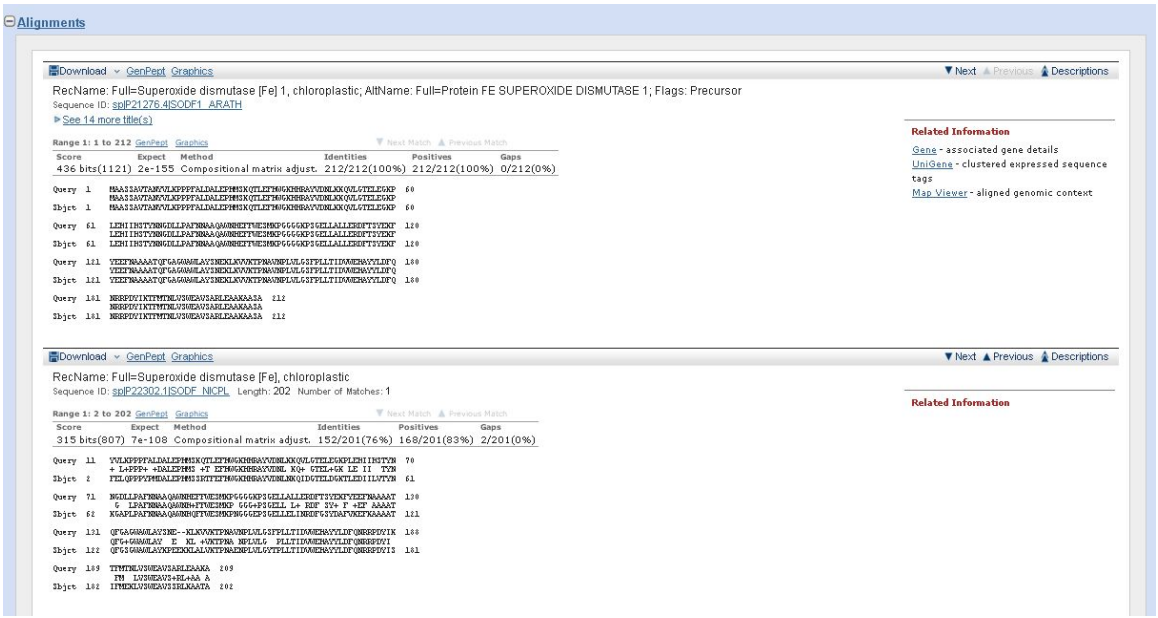
- **Дължина:** Това е дължината на алайнмънта, която показва колко дълги фрагменти от вашите секвенции BLAST е сравнил. Това е най-добрият начин да се види, дали дадено съвпадение има смисъл. Понякога, много къси алайнмънти може да имат високи E-стойности, но те не са много смислени.

- **Горната секвенция:** Вашето запитване/заявка. **Долната секвенция:** Съвпадението или „обекта" от базата данни.

- **Лентата между секвенциите:** Между двете секвенции е линията, която има (+) знак за подобни аминокиселини, буква за еднакви остатъци или празно място за несъответствията.

- **XXXXX региони:** Във вашата заявка, BLAST въвежда X-сове автоматично, с цел да се маскират региони, съдържащи многократно повтарящи се остатъци. Специалистите наричат тези региони сегменти с ниска сложност. Тези области могат да причинят проблеми при търсене. X-совете оказват автоматично на BLAST да игнорира съответните сегменти. Маскирането става само в заявената секвенция.

- **Числата:** числата от дясната страна на секвенциите посочват координатите на съвпаденията върху заявената секвенция, както и върху откритата в базата данни. BLAST прави локален алайнмънт, който може да съдържа само част от заявената секвенция и съответното попадение.



Фигура 5-10. Алайнмънти по двойки, установени от BLAST.

Един добър алайнмънт не трябва да съдържа твърде много дупки/разкъсвания и по-скоро трябва да има няколко области на високо сходство, отколкото единични еднакви остатъци разпределени по дължина на секвенцията. Този критерий е особено важен при алайнмънти, разположени в близост до приетата гранична стойност. Имайте предвид, че понякога се съобщава за повече от един алайнмънт. Това се случва, когато между вашата заявка и съответното попадение може да се намери подобие на повече от едно място. На графичното изображение, тънка черна (NCBI) или пунктирани линия (EMBnet) свързва тези независими алайнмънти.

Параметрите на BLAST

Не е нужно да се тревожите много за смисъла на информацията, която се появява в долната част на страницата с резултата. Ако промените параметрите по подразбиране на BLAST, тази част пази следа от това, давайки сигурността, че може, ако се наложи да възпроизведете това търсене. Никой не може да гарантира, че може да се

възпроизведе даден резултат абсолютно точно, дори ако използвате един и същ BLAST сървър. Обновяването на всеки един от компонентите на сървъра, може да доведе до промени в резултатите от дадено търсене. Компонентите, които могат да бъдат обновени включват:

- Базите данни;
- Самата BLAST програма;
- Параметрите по подразбиране на сървъра.

Реално вие нямате контрол над тези параметри. Те се променят непрекъснато с течение на времето. Информацията, която се появява в долната част на търсенето, може да предложи причина за възможна появила се грешка.

BLAST на ДНК секвенции

Ако мислите, че BLAST на ДНК последователности е като BLAST на протеинови секвенции, то вие сте едновременно прави и грешите. Макар да е вярно, че принципът е един и същ, и че на практика BLAST на ДНК изисква операции, подобни на тези при протеини, при ДНК BLAST не винаги работи толкова добре. По-бързо и по-точно е да се бластват протеини (blastp), отколкото нуклеотиди.

Ако знаете отворената рамка на четене на вашата секвенция, по-добре е първо да транслирате ДНК секвенцията в аминокиселинна последователност и да проведете BLAST с нея (вижте съответния раздел, "BLAST на протеинови секвенции", по-горе в тази глава). Ако не знаете рамката на четене, тогава вие трябва да изберете една от следните BLAST програми (виж Таблица 5-2), които работят с нуклеотиди.

Таблица 5-2. Налични BLAST програми за секвенции.

Програма	Заявка	База данни	Използване
blastn	ДНК	ДНК	много сходни ДНК секвенции
tblastx	TrDNA	TrDNA	изучаване на протеини и EST-та
blastx	TrDNA	протеин	анализ на заявената ДНК секвенция

„Tr" идва от транслиране. Това означава, че ДНК секвенцията се задава в BLAST и BLAST превежда тази последователност в нейните шест възможни рамки на четене (три на правата верига и три на комплементарната). Когато BLAST попадне на стоп кодон го заменя с X. Шестте рамки на четене означават, че не е необходимо да се тревожите за ДНК веригата, която заявявате в BLAST. Програмата пробва всяка една от тези шест възможни рамки на четене.

Избор на подходящ BLAST за дадена ДНК секвенция

Списъка с програми, показан в Таблица 5-2 с всички възможни акроними може да ви изглежда малко сложен, но не се оставяйте да ви объркa. Просто си задайте въпросите в Таблица 5-3 в посочения ред, за да определите коя програма е най-подходящата за вашия конкретен случай.

Таблица 5-3. Избор на подходящ BLAST за ДНК.

Въпрос	**Отговор**
Интересувам ли се от некодираща ДНК?	Да: Използвайте blastn. Никога не забравяйте, че blastn е само за тясно свързани ДНК секвенции (с повече от 70% идентичност).
Искам ли да открия нови протеини?	Да: Използвайте tblastx.
Искам ли да открия протеини, кодирани в заявената ДНК секвенция?	Да: Използвайте blastx.
Имам ли съмнения относно качеството на моята ДНК секвенция?	Да: Използвайте blastx, ако подозирате, че вашата ДНК секвенция кодира протеин, но че тя може да съдържа грешки при секвенирането

Blastx може да коригират определени грешки. Ако не сте сигурни за вашата протеинова секвенция, то blastx може разкрие подобието му с най-добре секвенираните ДНК фрагменти. Ако установите, че вашата секвенция съдържа неочаквана, отместена рамка на четене, не се вълнувайте веднага. Грешки при

секвенирането са най-вероятните причини за наблюдаваната отместена рамка на четене.

Подбиране на параметри при BLAST на ДНК

От гледна точка на сървъра, няма реална разлика между BLAST на протеини и BLAST на ДНК последователности. Механизмът е един и същ и може да следвате сляпо всяка една от стъпките при BLAST на протеинови секвенции, както е описано по-горе в този раздел. Запомнете добре следните стъпки:

- **Избор на база данни:** Изберете база данни съвместима с BLAST програмата, която сте решили да използвате. Освен ако не използвате blastx, базата данни трябва да бъде изградена на базата на ДНК секвенции. Повечето BLAST сървъри не проверяват, дали сте направили това правилно. Затова ще очаквате да получите резултат, но това няма да се случи.
- **Ограничаване на вашето търсене:** ДНК базите данни са по-бавни. При възможност е добре да се ограничи търсенето до подмножество на базата данни. Например, ако се интересувате само от *Drosophila* (плодова мушица), е по-добре да търсите само в генома на *Drosophila*.
- **„Пазарувайте на повече места”:** Не се колебайте да „пазарувате на повече места”, за да намерите BLAST сървъра с база данни, която ви интересува.
- **Използвайте филтриране:** Геномните последователности са пълни с повтори. Не забравяйте да филтрирате поне някои от тях.

Как нещата работят с помощта на BLAST

Хората, които извършват биоинформатичен анализ знаят, че почти нищо не може да се направи без BLAST! Какъвто и биологичен въпрос да се зададе, голяма е вероятността BLAST да ви покаже някои идеи, преди да сте започнали да използвате по-сложни програми или да се занимавате с проектиране на скъпо струващи експерименти. В Таблица 5-4 показваме четири конкретни примера за това. За всяко от тези приложения има изчерпателно и подходящо решение. BLAST е бърз и ефикасен метод и може да бъде напълно достатъчен в повече от случаите. Разбира се, ако имате твърде много време на разположение, винаги може да подходите към по-сложен и труден начин.

Таблица 5-4. Отговор на биологични въпроси с помощта на BLAST.

Какво трябва да направите	**Как да го направите с BLAST**	**Сложната алтернатива**
Намиране на гени в генома	Накъсайте вашата геномна секвенция на малки (2 до 5 килобайта) припокриващи се последователности. Използвайте blastx, за да бластнете всяко парче от генома срещу мРНК секвенции. Това става по-добре, ако нямате интрони (бактерии).	Стартиране на софтуер за предсказване на гени
Предсказване на функция на протеини	Използвайте blastp, за да бластнете вашата протеинова секвенция срещу Swiss-Prot. Ако получите добри попадения (с повече от 25 процента идентичност) по цялата дължина на протеина, то тогава вие сте решили проблема си, защото вашия протеин има същата функция, като протеина от Swiss-Prot.	Проведете доменен анализ или експеримент в лабораторни условия
Предсказване на 3-D структура на протеин	Използвайте blastp, за да бластнете вашия протеин срещу PDB (база данни за протеинова структура). Ако получите добро попадение (идентичност повече от 25 процента), може да приемете, че този вашия протеин имат сходна 3-D структура.	Провеждане на хомоложно моделиране, рентгенови лъчи или ядрено-магнитно резонансен анализ на вашите протеини.
Намиране на членовете на протеиново семейство	Използвайте blastp (или по-мощен вариант PSI-BLAST) и го стартирайте срещу NR. Така вие имате всички членове на даденото протеиново семейство и може да проведете множествен алайнмънт (виж глава 9) и да съставите филогенетично дърво.	Клониране на нови членове на семейството, използвайки PCR техники.

Контролиране на BLAST: избор на подходящи параметри

Казват: „Властта е нищо без контрол”. Това изказване обобщава доста добре как може да получите най-добри резултати от BLAST- като го контролирате. В този раздел ще ви покажем основните параметри на BLAST - какво определят те и как може да ги промените според вашите нужди.

При използване на BLAST, първо правило е да определите максимално точно и ясно какво искате да правите. Отнасяйте се към BLAST, като към древен оракул. За да получите отговор на вашите въпроси трябва да бъдете кристално ясни. Таблици 5-3 и 5-4 могат да ви помогнат да прецизирате въпросите си така, че те да са възможно най-конкретни и ясни.

Параметрите по подразбиране на BLAST са напълно оптимизирани и добре проверени. Ако не получавате нищо смислено с тях, не очаквайте чудеса, ако ги промените. Таблица 5-5 изброява основните причини, поради които бихте искали да се променят параметрите по подразбиране. Тези причини са валидни, както за ДНК, така и за протеини.

Повечето BLAST сървъри използват малко различаващи се начини за дефиниране на параметрите. В този раздел ще се съсредоточим върху сървъра NCBI. Ще откриете, че е лесно да адаптирате параметрите според вашите конкретни нужди въпреки, че могат да се окажат на пръв поглед малко скрити. Всеки път, когато искате да персонализирате търсенето си, потърсете разширен режим на сървъра, който дава повече възможности.

Таблица 5-5. Някои от причините, за да промените BLAST параметрите по подразбиране са следните:

Причина	**Параметър, който да промените**
Секвенцията, от която се интересувате съдържа много еднакви остатъци.	Филтър на секвенцията (автоматично маскиране)
BLAST не отчита никакви резултати.	Матрицата на заместванията или gap penalties.
Вашето попадение е с гранична E-стойност.	Матрицата на заместванията или gap penalties, да се провери стабилността на

	попадението
BLAST докладва твърде много попадения.	Базата данни, в която търсите или филтрирате докладваните попадения по ключова дума; да се повиши E-праговата стойност или да се отхвърлят секвенции, твърде подобни на заявката (тези с много ниски E-стойности)

Контрол върху маскирането на секвенциите

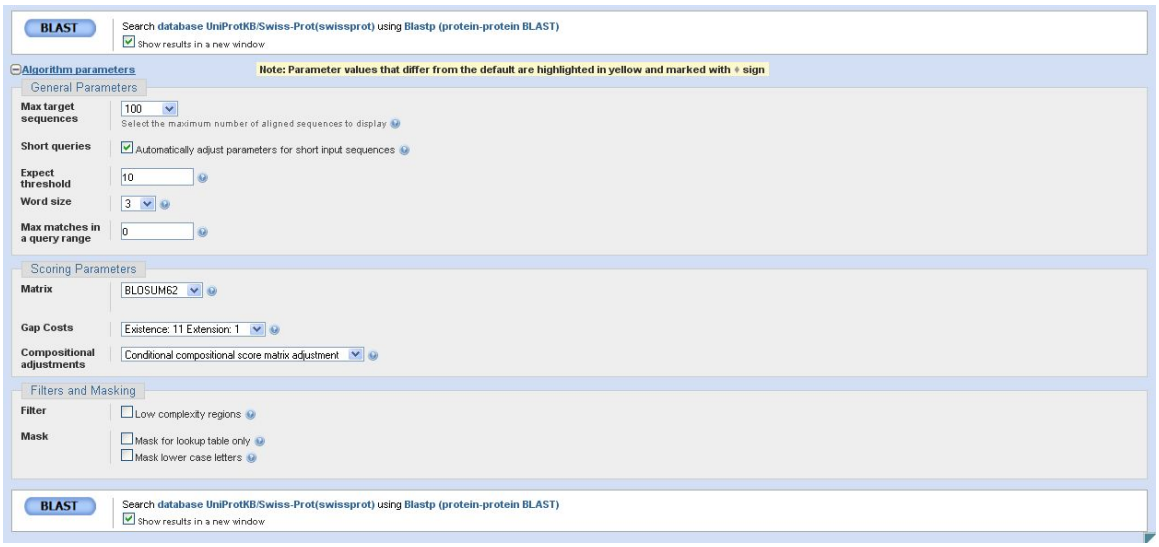
Когато BLAST търси в дадена база данни, има едно основно важно предположение, а то е, че BLAST приема всички ваши секвенции за усреднени последователности. Например, ако търсите протеинови последователности, BLAST предполага, че средния протеинов състав на всеки един протеин е като средния протеинов състав на цялата база данни. На практика, това предположение е твърде далеч от реалността, защото не е валидно винаги.

Маскиране на протеинови последователности

Много протеини съдържат области, известни като региони с ниска сложност (или с ниска ентропия). Например, тези региони могат да са сегменти, които съдържат много пролини или много остатъци от глутаминова киселина. Ако BLAST сравнява две богати на пролин области, то този алайнмънт ще има много ниска E-стойност, поради големия брой еднакви аминокиселини. За съжаление, има голяма вероятност тези две богати на пролин области да не са свързани по никакъв начин. В действителност, тези области богати на една и съща аминокиселина са известни с това, че могат да доведат до объркващ BLAST резултат. За да се избегне този проблем при анализ на протеините, BLAST филтрира районите с ниска сложност. За да направи това, той замества тези региони с последователност от X. Ако се интересувате специално от тези области с ниска сложност и не искате да бъдат игнорирани при търсенето, е необходимо да премахнете отметката от квадратчето Low Complexity, намиращо се в съседство до опциите Mask for lookup table only и Mask lower case letters, както е показано на Фигура 5-11.

Представете си, че току що сте клонирали и секвенирали даден протеин. В секвенцията на този протеин намирате 10 пролинови остатъка (**PPPPPPPPPP**). На този етап вероятно се чудите, дали това е грешка на секвенирането или друг вид грешка. Ще сте много по-уверени, ако има реален протеин, който съдържа същия участък от аминокиселини. За да намерите този протеин, напишете своята секвенция **PPPPPPPPPPPPPPPPPP** и я задайте като заявка на blastp. Не забравяйте да премахнете отметката от квадратчето Low Complexity. Някои домени са много често срещани в протеиновите секвенции, като Zn пръсти в тандеми или фибронектинови домени. Ако вашият протеин съдържа този вид домен, BLAST докладва за много съвпадения с протеини, които съдържат същите области, но не са свързани. За да направите търсенето по-интересно е добра идея филтрирането на тези домени. Ето как:

1. **Използвайте CD търсене, InterProScan или Pfscan, за да намерите домени във вашия протеин.**
2. **Прочетете информацията дадена за домените, за да разберете колко широко разпространени са домените, които сте намерили.**
3. **Заменете секвенциите на ниско информационните домени с X-ове - или ги пренапишете с малки букви - и след това поставете отметка на квадратчето Mask lower case letters, разположено до Filter опцията във BLAST формата (виж Фигура 5-11).**
4. **Стартирайте стандартен blastp, както е показано в началото на тази глава.** Получените резултати трябва да ги тълкувате, както е показано в раздел "Анализ на BLAST резултат", по-горе в тази глава.

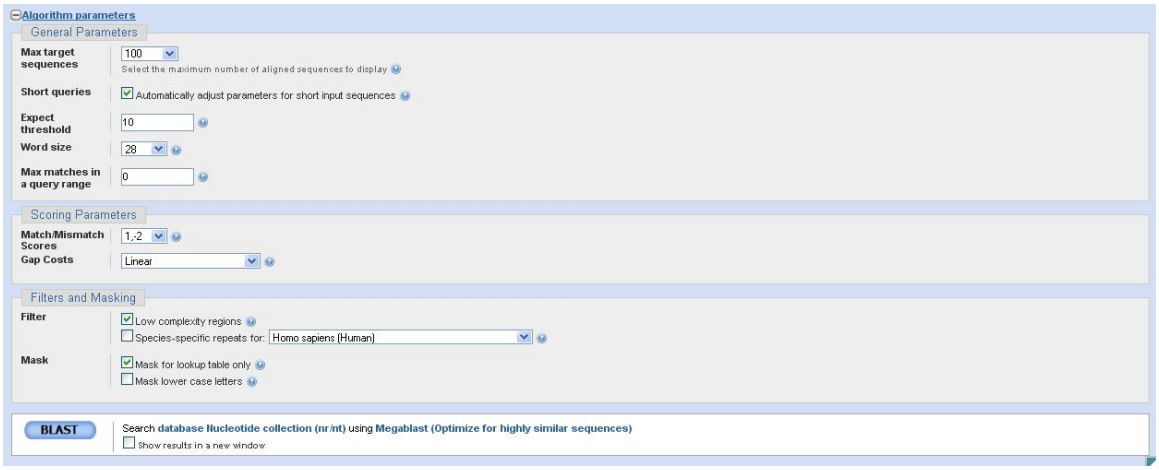


Фигура 5-11. Разширени параметри на blastp в NCBI.

Ако мислите, че е проблем запазването на резултатите при работа с протеини, нещата изглеждат още по-сериозни, когато имате работа с ДНК секвенции. В ДНК има много повтарящи се последователности, които в действителност нямат голямо значение. Повечето геноми съдържат такива повторения. 60% от човешкият геном е съставен от такива повтарящи се секвенции! Ако искате да избегнете преминаването през тях се налага да маркирате Species-specific repeats for: Human квадратчето, което се появява на страницата blastn, както е показано на Фигура 5-12. Избирането на тази опция води до филтриране на търсенето, което изключва тези повтори при човека. Не винаги е безпроблемно мащабно секвениране на даден геном, както се опитват да ни убедят. Например в много случаи, краищата на генома съдържат малки парченца и части от използвания за клониране организъм (което е причината, когато търсите в човешкия геном да откривате части и парченца от генома на дрожди или *E. coli*), или пък други помощни секвенции. Това означава, че преди да се зарадвате от намирането на ген в човешкия геном (или който и да било друг геном), трябва да се убедите, че този ген не е замърсяване, донесено от организма, използван за клониране. Представете си срама от откриването на това, че раковия ген, който сте изучавали в продължение на три години е в действителност протеин от *E. coli*!

Промяна на параметрите за BLAST алайнмънт

Фигура 5-11 ви показва всички параметри, които може да промените на BLAST NCBI сървъра. Те включват: E-стойност, дължина на заявката, както и средствата за филтриране. Сред тях има два важни параметъра, които са свързани с начина, по който BLAST прави алайнмънт. Това са gap penalties (наречен Gap Costs на страницата на BLAST NCBI сървъра) и substitution matrix (наречен Match/Mismatch Scores на страницата на BLAST NCBI сървъра). Ако промените някой от тези параметри, BLAST може да даде в резултата различни хитове. Попаденията, които най-вероятно се променят при едно или друго стартиране на BLAST са резултати, които са на границата (тези с E-стойност по-висока от 0,0001).



Фигура 5-12. Филтриране на ДНК секвенциите, докато използвате blastn.

Дължината на думата (word) е тайната рецепта на BLAST. В BLAST, дължината на т.н. дума е минималния размер на елемента, който BLAST се опитва да сравни между вашата секвенция и секвенцията в базата данни. Например, ако думата е определена с размер 6, BLAST ще прегледа всички възможни под-секвенции, съдържащи се с размер 6 бази във вашата заявка и ще ги сравни с всички секвенции в базата данни с размер 6 подобни на тези във вашата секвенция. Дългите думи правят BLAST по-бърз и по-малко чувствителен. Кратките думи правят точно обратното. Добра причина да се варира с тези параметри е да се провери стабилността на граничните попадения. Ако даденото попадение не изчезва, когато промените substitution matrix или gap penalties, то има по-големи шансове да бъде биологично значимо.

Контрол върху BLAST резултата

Ако вашата заявка принадлежи към голямо протеиново семейство, BLAST резултата може да създаде проблеми, тъй като базата данни съдържа твърде много секвенции почти идентични с вашата.

Понякога това изобилие на хомоложни секвенции може да попречи да видите хомоложна секвенция, която е по-слабо свързана, но все пак свързана с експерименталните ви данни. Ако подобно нещо се случи, вижте Таблица 5-5, където има пет решения за коригиране на този проблем. В следващите раздели ще покажем как се прилагат тези решения.

Избор на правилна база данни

Изберете база данни, която е подходяща за нуждите ви. Ако BLAST докладва за твърде много съвпадения, може да се намери изход от това затруднение, като търсите в БД Swiss-Prot вместо NP (Swiss-Prot е 100 пъти по-малка от NP) и търсите само в един геном или само с PDB, ако сте заинтересовани основно от структурни прилики.

Ограничаване чрез Entrez заявка

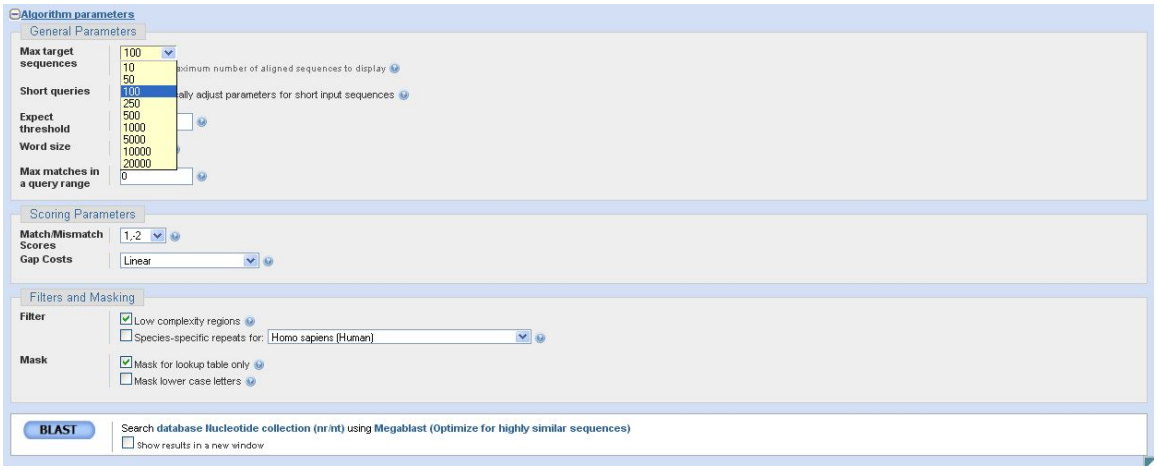
Ако се интересувате само от един конкретен организъм и имате твърде много попадения, използвайте ограничаване на резултатите чрез отметка в Entrez Query квадратчето (виж Фигура 5-12) или асоциираното с него падащо меню. Може да използвате това поле за комбиниране на ключови думи с операторите AND и OR. Например, ако искате BLAST да изведе резултати само за протеази и да се игнорират протеазите от генома на човека, напишете **“protease NOT human [Organism]”** (не забравяйте кавичките). Ако искате да научите повече по тази тема, вижте онлайн информацията за отправяне на заявки в Entrez на адрес:

http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helpentrez&part=EntrezHelp

Очаквана E-стойност

Това е ключов момент при отчитане на съвпадения. BLAST не отчита попадения с E-стойност, по-висока от посочената в съответното поле. Ниска стойност на този параметър кара BLAST да отчита само добри попадения. Може да филтрирате резултатите и от двете страни на границите и да предотвратите доклад на полуидентични секвенции. Например на Фигура 5-13 ние поискахме най-добрите секвенции (E-стойност по-ниска от $10e^{-100}$) да не се показват.

Трябва да сте наясно, че ако търсенето дава повече от 100 попадения, трябва да повишите броя на отчетените описания в раздел параметри, както е показано на Фигура 5-13.



Фигура 5-13. Преформатиращи възможности на BLAST.

Направете така, че BLAST да може да бъде повторен с помощта на PSI-BLAST

BLAST понякога не е достатъчен. Представете си, че искате да откриете всички членове на много голямо протеиново семейство, като се започне със секвенцията, която имате. При стартиране на BLAST може да откриете само най-близко свързаните последователности. Какво ще кажете за възможността да намерите и други далечни братовчеди, които вашата заявена последователност не може да разпознае? PSI-BLAST прави точно това - първо търси секвенции, които са тясно свързани с вашата и след това постепенно и внимателно разширява кръга за да се включат секвенции, които не изглеждат толкова интересни в началото, но при по-задълбочен анализ се оказва, че са свързани по някакъв начин с вашата секвенция. PSI-BLAST прави опит всеки път за привличане на нови членове от дадено семейство с помощта на най-обикновени свойства сред вече избраните членове. Всеки нов цикъл е едно повторение.

PSI-BLAST на протеинови секвенции

В следният пример използваме PSI-BLAST за да си отговорим на въпроса, дали легхемоглобина при растенията е свързан с човешкия хемоглобин.

1. **Заредете във вашия браузър основната страница на BLAST NCBI:**
   http://blast.ncbi.nlm.nih.gov/Blast.cgi

2. **Кликнете върху blastp връзката (включваща използването на алгоритмите blastp, psi-blast, phi-blast).**
   PSI-BLAST страницата се появява, както е показано на Фигура 5-14.

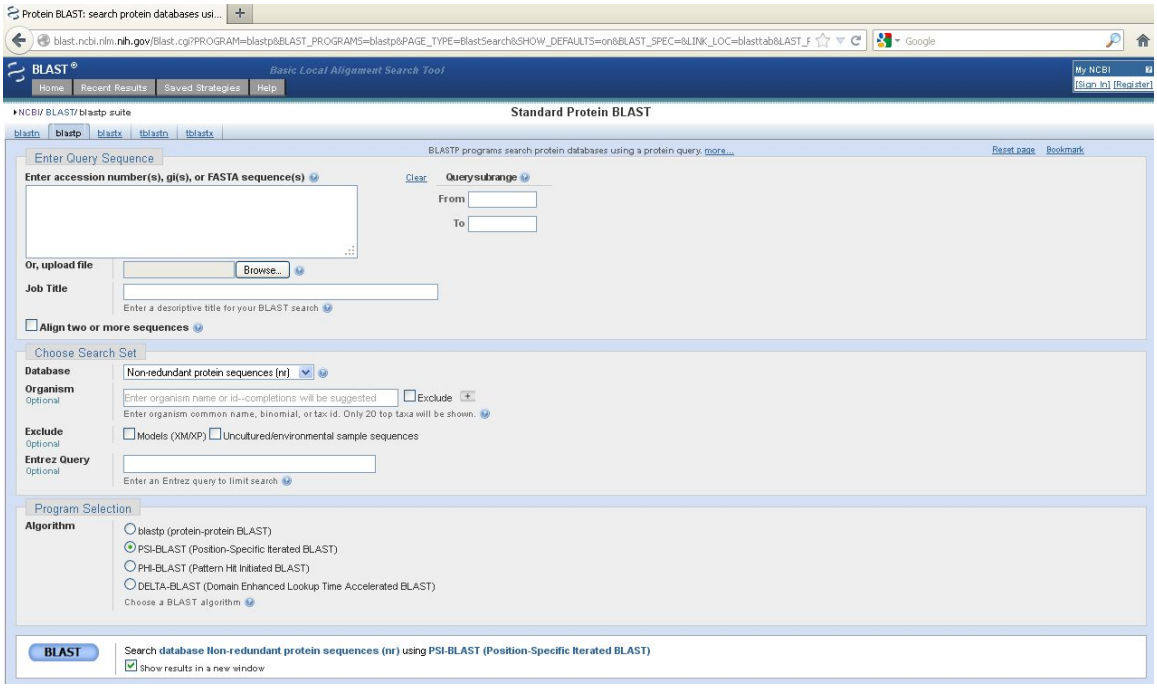
3. **Въведете номера за достъп на вашата секвенция или по-добре поставете пълната секвенция в прозореца.**
   Процедурата, по която задавате вашата секвенция в PSI-BLAST е точно същата като процедурата в blastp (виж Фигура 5-2 и 5-3).

   A. Ако вашата последователност съществува в базата данни, може да въведете само номера и за достъп.
   За този пример ние използваме човешки хемоглобин, протеин от Swiss-Prot с номер за достъп P69905.2.

   B. Ако вашата секвенция не съществува в базата данни, трябва да я поставете във FASTA формат.

4. **Изберете PSI-BLAST (Position-Specific Iterated BLAST) от опциите за алгоритъм (виж Фигура 5-14).**
   PSI е протеинова база данни с известна 3-D структура. Избираме я за нашия пример, защото е малка и подходяща база данни.
   Целта е да се попадне на секвенция в PSI, зададена като легхемоглобин и така да се убедим, че хемоглобина и легхемоглобина са структурно свързани.
   Ако при първоначално стартиране на PSI-BLAST не се открива хомология, може вместо PDB да използвате NR (цялата база данни). NR може да съдържа междинна секвенция - между вашата и други PDB секвенция, които PSI-BLAST може да открие едва след няколко повторения.

5. **Кликнете на BLAST бутона.**
   Времето, което отнема процеса на BLAST може да е по-дълго от това, което се показва на екрана. Бъдете търпеливи! Появява се междинна страница, в която трябва да изчакате!
   Появява се нова страница в нов прозорец, озаглавена BLAST резултати. Това е мястото, където резултатите се показват, когато са готови. Не забравяйте да запазите отворен прозореца „преформатиране на BLAST" - ще се нуждаете от него!



Фигура 5-14. Задаване на секвенция в PSI-BLAST.

6. **Проверете резултатите.**
   Има голямо количество много близки секвенции и само няколко родствено отдалечени. Не бива да се вълнуваме от такъв резултат. Както можеше да се очаква, има много хемоглобинови протеини и няколко миоглобина.
   Обърнете внимание, че на Фигура 5-15 списъка с попаденията на PSI-BLAST е различен от този, който виждаме при нормален BLAST. Има три нови особености, свързани с всяка секвенция:

   A. **A check box:** Ако има отметка в това квадратче, PSI-BLAST ще използва съответната секвенция, за да се получи позиционно специфична матрица при следващото повторение на PSI-BLAST.

   B. **The New symbol:** Показва секвенциите, които PSI-BLAST съобщава като първи попадения.

   C. **The green pill:** Показва секвенциите, които вече PSI-BLAST е използвал за получаване на текущия резултат.
      Ако от описанието прецените, че някоя от секвенциите не е свързана с вашата заявка, не се колебайте да я премахнете. По този начин, дадената несвързана секвенция няма да бъде използвана в следващата итерация.

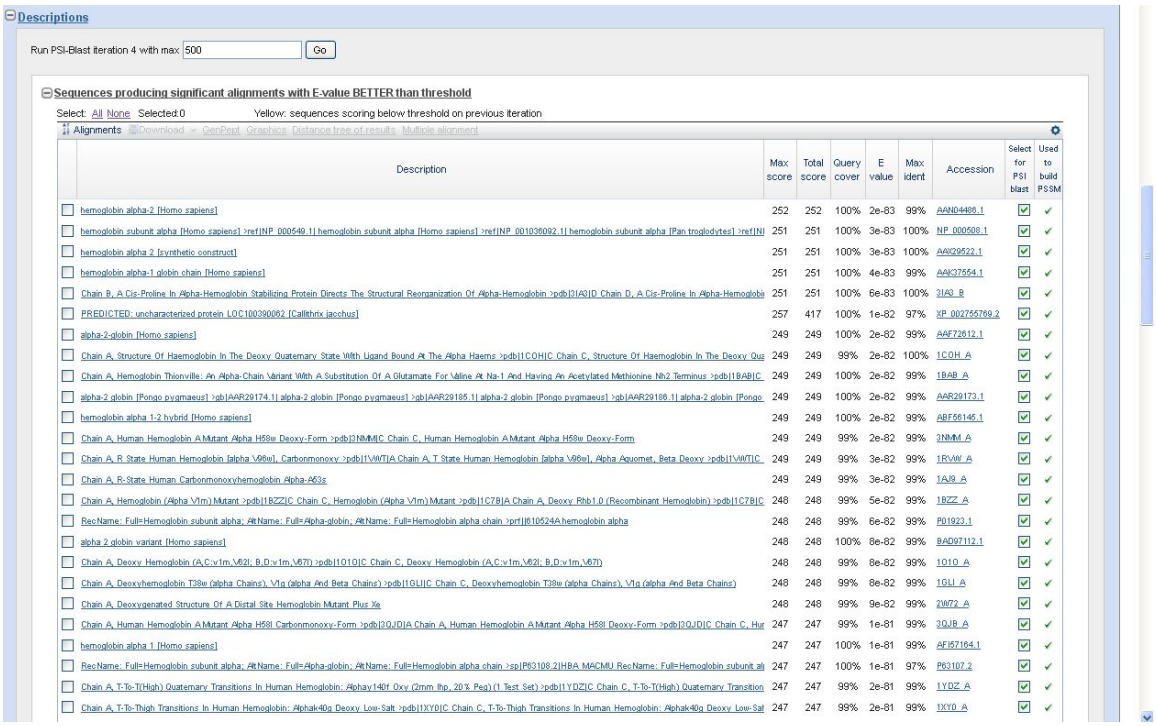
7. **Кликнете върху Run PSI-BLAST Iteration 2 бутона (разположен в близост до горната част на страницата).**
   Появява се преформатиращият BLAST прозорец.

8. **Кликнете върху Format! бутона, намиращ се на преформатиращият BLAST прозорец.**
   Резултатите се появяват в Result of BLAST прозореца.

9. **Продължете с повтаряне на стъпки 7 и 8.**
   Легхемоглобинът се появява по време на вторият цикъл, но не притежава много добра E-стойност. Въпреки това след третата итерация (Фигура 5-15), неговата E-стойност се подобрява значително, което предполага, че това е истински родственик на семейството на хемоглобина.

Избягване на грешки при стартиране на PSI-BLAST

В примера за легхемоглобина (малко по-горе) се придвижваме от една итерация към следващата, защото очевидно анотираните секвенции, които се добавят при всяка итерация са свързани с хемоглобина. Това е знак, че вървим в правилната посока.



Фигура 5-15. Списък с попадения в PSI-BLAST.

От друга страна, понякога това е път без изход. Ако след втората итерация всеки нов хит е алкохолдехидрогеназа, значи е време да се откажете! За съжаление няма просто правило, според което да се каже кога трябва да спрете или кога да продължите.

Единственият полезен източник на информация за това е анотацията. Обърнете внимание, че имате възможността (и дори задължението!) да премахнете очевидно неверни съвпадения между две итерации.

Избор на правилната секвенция

Основна трудност в PSI-BLAST е да се реши кои секвенции може да запазите от една итерация към следваща. PSI-BLAST контролира това решение автоматично с праговите E-стойности (включване на прагова E-стойност е разяснено в раздел форматиране), но може да използвате квадратчето свързано с всяко съвпадение за да промените процедурата от автоматична на ръчна.

Включването на прагови стойности е нож с две остриета, защото:

- Ако сте задали прага твърде ниско, PSI-BLAST няма да попадне на никаква секвенция.
- Ако сте задали прага твърде високо, PSI-BLAST ще улови несвързани секвенции.

Намирането на несвързани секвенции е много подобно на улавянето на бълхи - става все по-зле и по-зле. Първата неподходяща секвенция на която попаднете, води до други такива, несвързани с вашата заявка и нещата бързо могат да излязат от контрол. Единствената причина за понижаване на прага е пълна липса на резултати след първата итерация. В противен случай е добре да се запазят настройките по подразбиране и да се разгледат много внимателно попаденията с E-стойност под този праг. След всяка итерация може да се добавят хитовете, които изглеждат подходящи (в зависимост от техните анотации) и да се премахват тези, които са избрани, но изглеждат без значение. Всичко това отнема време и прилича малко на изкуство, но обикновено се отплаща добре.

Промяна на заявката е средство, с което да проверите своите резултати. Ако използвате някое от съвпаденията като първоначална заявка се очаква, че PSI-BLAST ще намери оригиналния въпрос след няколко повторения.

Избягване на грешки при работа с мулти-доменни протеини

Ако смятате, че домен е нещо като малка автономна част от протеина, който може самостоятелно да се нагъва, не е изненада, че повечето протеини са мулти-доменни. Един и същи домен може да се открива в много различни протеини.

Ако вашият протеин е мулти-доменен, при стартиране на PSI-BLAST може да имате неприятности, тъй като всеки домен се конкурира с другите и всички те могат да привлекат различни последователности от несвързани протеини. Резултатът при такава заявка е много труден за тълкуване.

Когато това се случи, най-добрата стратегия протеинът да се нареже на отделните му домени и след това да проверите всяка една от секвенциите независимо.

Може да намерите къде са разположени отделните домени с помощта на NCBI CD сървъра или pfscan EMBnet.

Има правило, според което, ако не намерите никакви домени във вашия протеин и той е по-голям от 200 аминокиселини, е добре да го нарежете на малки фрагменти с дължина от по 200 аминокиселини и след това да анализирате тези сегменти един по един. Крайъгълния камък в това правило е, че остава скучната задача да се наредят след това всички парчета от пъзела. Но понякога това е забавно.

Откриване и използване на домени в протеини с BLAST и PSI-BLAST

Може да използвате BLAST, за да откриете свои собствени домени и да ги използвате за сканиране на протеинови бази данни. В този BLAST доменът се нарича PSSM (Position Specific Substitution Matrix).

Когато стартирате PSI-BLAST, може да зададете опцията резултатите да са под формата на PSSM. За да направите това, следвайте следните лесни стъпки описани в документацията на BLAST онлайн на адрес:

www.ncbi.nlm.nih.gov/blast/blastcgihelp.shtml#pssm

След като вашият PSSM е готов всичко, което трябва да направите е cut и paste на параметрите във формуляра в съответния раздел на BLAST за напреднали (виж Фигура 5-11).

Когато предоставите PSSM, BLAST го използва на мястото на единична заявена секвенция и го сравнява с всяка секвенция в базата данни, която сте избрали.

Търсенето на хомология в Интернет е навсякъде

Онлайн документацията за BLAST е огромна и трябва да я използвате колкото може повече! Най-полезна информация може да намерите на страниците:

При избор на BLAST и база данни тази страница е много добра:

www.ncbi.nlm.nih.gov/blast/producttable.shtml

А тази ви казва всичко, което трябва да знаете за всеки един параметър:

www.ncbi.nlm.nih.gov/blast/blastcgihelp.shtml

Ако се опитате да стартирате примерите, дадени в тази глава на NCBI сървъра, трябва да знаете, че понякога има твърде много трафик, за да успеете да направите каквото и да било полезно. За щастие, по целия свят се появяват непрекъснато нови BLAST сървъри. В Таблица 5-6 ви представяме списък на едни от най-стабилните и актуални сървъри.

Таблица 5-6. BLAST и PSI-BLAST сървъри по света.

Държава/Континент	**Програма**	**URL**
САЩ	Blast/PSI-BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi
Европа	BLAST	www.expasy.ch/tools/blast/
Европа	BLAST	http://www.ch.embnet.org/software/bBLAST.html
Европа	BLAST	http://www.ebi.ac.uk/Tools/sss/ncbiblast/nucleotide.html
Япония	Blast/PSI-BLAST	http://blast.ddbj.nig.ac.jp/top-e.html

NCBI-BLAST има малка сестра, известна като WU-BLAST (където WU абревиатурата идва от Вашингтонски Университет). Специалистите могат да прекарват часове аргументирайки се за това, защо WU-BLAST е по-чувствителен и по-продуктивен във вмъкване на разкъсвания от NCBI-BLAST. Трябва да пробвате за себе си (виж Таблица 5-7) и след това да се включите в дебатите...

Таблица 5-7. WU-Blast.

Адрес	**Описание**
http://www.advbiocomp.com/blast.html	AB-BLAST (не е онлайн сървър)
http://www.genome.wustl.edu/tools/blast/	blastn и tblastn, за анализ с цели геноми
http://www.ebi.ac.uk/Tools/sss/ncbiblast/nucleotide.html http://www.ebi.ac.uk/blast/uk/BrassicaDB/blast_form.html	Всички видове BLAST и WU-Blast
http://brassica.bbsrc.ac.uk/BrassicaDB/blast_form.html	Друг WU- BLAST сървър

BLAST е един от най-популярните методи за търсене в бази данни, но не е единствения. Три от основните алтернативи на BLAST са:

- **Smith и Waterman (SSEARCH):** Това е по-бавен, но може би по-точен метод от BLAST.
- **FASTA:** Това е метод малко по-бавен от BLAST, но се твърди, че е по-точен при сравняване на ДНК секвенции.
- **BLAT:** Използвайте този метод за бързо локализиране на къси ДНК секвенции в генома или намиране на близки (бозайници сравнени с бозайници) протеини в генома.

На практика, интерфейсът за използване на тези методи е много подобен на този използван за BLAST. Не трябва да имате проблем да адаптирате съдържанието на тази глава към всяка базирана на сходство база данни. В Таблица 5-8 ви даваме информация за най-стабилните и актуални сървъри, които извършват тези услуги. Нито една от тези програми няма BLAST-подобни способности за филтриране на региони с ниска сложност!

Таблица 5-8. Алтернативни методи за хомоложно търсене.

Страна/ Континент	**Програма**	**Адрес**
USA	FASTA	http://www.xmarks.com/site/fasta.bioch.virginia.edu/fasta/lalign.htm
Европа	FASTA	http://www.ebi.ac.uk/Tools/sss/ncbiblast/nucleotide.html
Европа	SSEARCH	www.ch.embnet.org/software/GMFDF_form.html
Япония	SSEARCH/ FASTA	http://www.ddbj.nig.ac.jp/searches-e.html
USA	BLAT	www.genome.ucsc.edu

Упражнение 1

BLAST

1. **1. Характеризиране на непозната секвенция:**
   В папка sequences в (**http://web.uni-plovdiv.bg/vebaev/Bioinformatics/**) имате файл seq.txt, анотирайте секвенцията и определете от кой ген/организъм е тя.
   В NCBI, изберете BLAST → nucleotide blast и от опцията за организми, изберете Others (nr etc.): _____

2. **Търсене на идентичност.**
   Има ли идентична секвенция на тази от първата задача в мишка и плъх? _____

3. **Търсене на ортолози.**
   Потърсете с BLAST, дали генът ERBB2 от човек има ортоложна мРНК секвенция при шимпанзето (*Pan troglodytes*)? _____

4. **Предвиждане на протеинова функция.**
   В папка sequences имате файл protein.txt, анотирайте секвенцията и определете каква функция има тя (изберете опцията protein blast).
   От кое супер семейство е протеинът: _____
   Какви домени има протеинът: _____

5. **Намиране на семейства протеини и сходна доменна структура (намиране на хомология).**
   Намерете колкото може повече членове на семейството протеини AGO от *Drosophila* като използвате BLAST и секвенцията на AGO1 гена. Колко Argonaute протеина има? _____
   От кое суперсемейство е протеинът? _____
   Селектирайте това суперсемейство и от там потърсете протеини със сходна архитектура (similar domain architecture).

6. **Определяне на геномна/генна локализация с BLAST.**
   Използвайте BLAST от сайта http://www.arabidopsis.org/ в Tools (Datasets: TAIR10 intergenic) и определете геномната локализация на секвенцията от файл gene.txt:
   Хромозома _____, Позиции _____-_____
   Между кои два гена е разположена дадената секвенция?
   _____ и _____

Упражнение 2

BLAST

1. Определете от кой организъм е секвенцията и какво кодира:
   ATCTCTGTGCTTCTTTGTCTACACTTTTGGAAAAGGTGATGATATCATTGCTTTTCCCCA
   AATGTAGACAAAGCAATACCGTGAT

2. Потърсете гена от предната задача в *www.arabidopsis.org* и посочете:
   геномната локализация: _____
   дължината на секвенцията: _____
   има ли други гени около този: _____

3. Извлечете секвенцията на първия интрон на гена (FASTA формат, *SequenceViewer*), който е на ~1777 бази downstream спрямо гена от задача 2 и следната информация за него (*Sequence ruler*):
   Номер за достъп: _____
   Описание: _____
   Геномна локализация: _____

Упражнение 1

Сравняване на секвенции

(алайнмънт)

1. **Задача 1. Алайнмънт на 2 нуклеотидни секвенции. Сравняване на мРНКи на гена DLL1.**

   A. Влезте в **NCBI.** От базата данни HomoloGene (за хомоложни гени) намерете гена **DLL1** при човек и мишка. От страницата на генния запис намерете номера за достъп до информационните РНКи. Изберете да виждате секвенцията във ФАСТА формат.

   B. Отворете програмата за биоинформатичен анализ **Geneious.** Създайте нова папка и в нея файл, с копираната от вас секвенция на човека. Съхранете секвенцията, като обърнете специално внимание на типа на съхраняваната от вас молекула! В нов файл направете същото с мишата секвенция. Изберете секвенциите, на които искате да направите алайнмънт, след което изберете менюто **Alignment**.

2. **Алайнмънт на 2 протеинови секвенции. Сравняване на RRAS2 протеини.**

   A. Влезте в **NCBI.** От базата данни HomoloGene (за хомоложни гени) намерете гена **RRAS2** при човек и плъх. От страницата на генния запис намерете номерата за достъп до протеините. Използвайте програма **Geneious**, за да ги сравните!

3. **Задача 3. Множествен алайнмънт. Сравняване на повече от 2 секвенции.**

   A. **Сравнете секвенциите:**
   NM_024674; XM_513232; XM_849892; NM_145833; NM_001109269; NM_001031774

   B. Каква секвенция и от кой организъм анализирате?

Ген (име, АС)	Организъм

4. .Намиране на разлика в транскрипционни варианти.

A. Намерете разликата в транскриптите на ген *Clk* в *Drosophyla melanogaster.* Къде е разликата? _____

Изграждане на филогенетични дървета

В тази глава ще научите:

- Какво могат да направят за вас филогенетичните дървета;
- Как да направите правилния множествен алайнмънт за правилното дърво;
- Как да изчислявате разстоянието между вашите секвенции;
- Как да построите филогенетично дърво по метода Neighbor Joining;
- Как да оценявате качеството на вашето филогенетично дърво чрез bootstrapping;
- Къде да намерите ресурси за филогенетика в Интернет.

В биологията нищо няма смисъл, освен в светлината на еволюцията.

— Theodosius Dobzhansky (1973)

С помощта на компютър и няколко онлайн ресурси можете да правите т.н. **филогенетичен анализ**. Това е научен метод, който позволява да реконструирате еволюционната история на група организми или секвенции. Ще ви покажем как да си подбирате секвенции, да ги анализирате с множествен алайнмънт, а него да превърнете във филогенетично дърво (дендрограма). Ще разгледаме два метода:

- **ClustalW:** Това е лесен метод, който обаче не можете да контролирате – той действа като „черна кутия”. Използва т.н. Neighbor Joining, надежден и доста използван метод за филогенетика.

- **Phylip:** По-сложен метод, който ви позволява да контролирате всички параметри при построяване на филогенетичното дърво.

Какъвто и метод да изберете, не забравяйте, че висококачествения множествен алайнмънт е основния фактор за добро филогенетично дърво.

Ползата от филогенетичните дървета

Целта на **филогенията** е да възстанови историята на живота и да обясни съвременното многообразие на живите същества. То може да се представи като голямо генеалогично дърво (дървото на живота). Основният принцип на филогенетиката е да групира живите същества според степента им на подобие едни с други. Съответно, колкото са

по-близки два вида (например човека и шимпанзето), толкова по-близки са с техния общ предшественик. Систематиката не е нова наука, напротив, тя е едно от първите неща които са започнали да правят биолозите – да класифицират.

Филогенетиката (молекулярната филогения) е специален раздел от систематиката, който се занимава със сравняване на хомоложни гени и други секвенции от различни видове, за да възстанови тяхната филогения на молекулно ниво. Този метод може да се приложи и за различните гени-членове на едно генно семейство, за да се възстанови по подобен начин историята на семейството (Имайте предвид, че тези дървета имат смисъл само ако вярвате в еволюцията!).

Молекулярните истории, които могат да ни разкажат филогенетичните дървета, са много разнообразни. Разбираме дали (и кога) са се случвали мутации, делеции, дупликации или видообразуване, загуба на функции и поява на нови и всякакви други събития, оформили живите същества в сегашния им вид и отношения.

Напоследък филогенетиката вече не е само метод или направление, а се оформя като цяла отделна наука.

В контекста на биоинформатиката има три основни причини да използвате филогенетични методи.

- **Определяне на най-близкородствените видове на вида, от който се интересувате.** Например, ако изследвате нов вид бактерия, можете да секвенирате нейната рибозомна РНК и да потърсите нейното място във филогенетичото дърво, съставено от всички известни бактериални рРНК. Това ще ви даде много добра представа с каква бактерия работите.
- **Откриване на функцията на ген.** Можете да използвате филогенетичните дървета, за да се уверите, че гена, който изследвате е ортоложен на друг добре известен ген от друг вид.
- **Проследяване на произхода на даден ген.** Повечето гени в един геном пътуват заедно през еволюционното време. Понякога обаче отделни гени могат да прескочат от един вид на друг – например чрез вирусна инфекция – т.н. хоризонтален генен трансфер.

Филогенетиката използва гени, за да възстанови еволюционните истории на видовете. Гените (секвенциите) са много по-лесни и точни за интерпретиране от морфологичните данни, използвани в класическата систематика. Секвенции може да се сравняват по много по-обективен начин, отколкото морфологични характеристики.

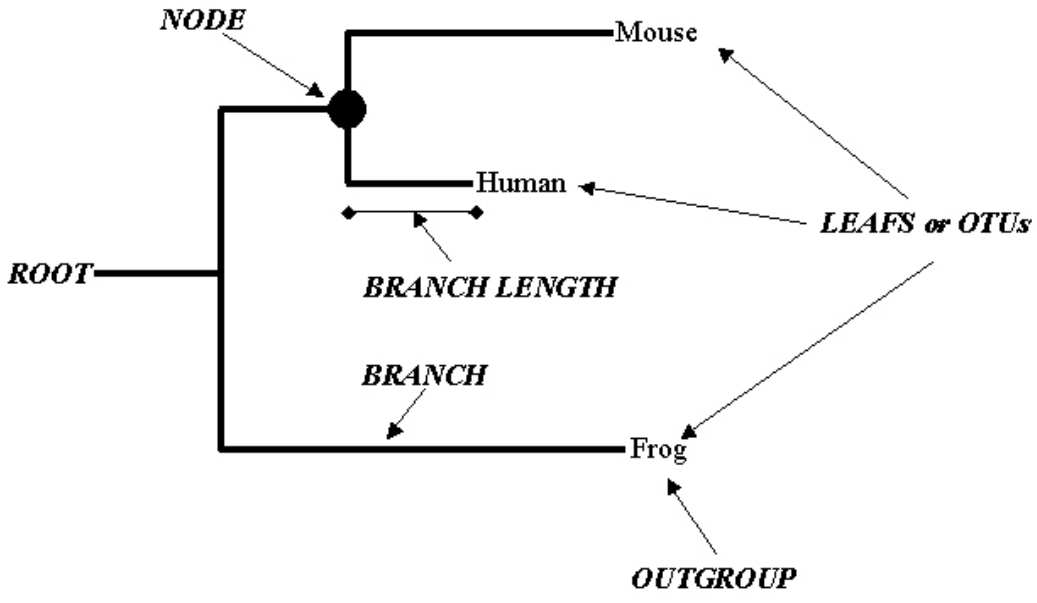
ЗАПОМНЕТЕ! Филогенетиката разполага с голямо разнообразие от методи и често е трудно да се спрете на един от тях. Това, което е ключово за един добър филогенетичен анализ обаче, е качеството на данните. Ако имате много добри данни,

всеки разумен филогенетичен метод ще ви даде смислено филогенетично дърво. От друга страна, ако данните ви не са добри, нито един метод няма да ви помогне. *„Garbage in, garbage out”* – това може да се каже за почти всички методи за анализ.

По конкретно, добри данни във филогенетиката означава: висококачествен и точен множествен алайнмънт, направен от правилно подбрани секвенции. Като начало, ще ви покажем как да си подготвите достатъчно добри първични данни за филогенетичен анализ.

Кое какво е във вашето филогенетично дърво

Когато работите с филогенетични дървета, неизбежно ще срещнете специфична терминология. Филогенетичния жаргон става все по-често срещан в биологичните публикации. Някои от тях са показани на Фигура 9-1 и са обяснени по-долу в текста.



Фигура 9-1. Основните компоненти на едно филогенетично дърво.

В едно филогенетично дърво различаваме **крайни или терминални възли (terminal nodes)** и **междинни или вътрешни възли (internal nodes**), както и **крайни или външни разклонения (external branches**) и **вътрешни разклонения (internal branches).** Разклонението дефинира връзката между дадена група и останалата част от дървото.

Терминалните възли на едно филогенетично дърво представят съществуващи в момента **таксономични единици**. Таксономична единица по принцип може да бъде ген или част от ген, друга секвенция, геном, индивид, вид, или по-голям таксон. Крайните възли или „листата" на филогенетичното дърво са всички съвременни (**extant**) таксономични единици. Те се наричат **оперативни таксономични единици**, или съкратено **OTE (operational taxonomic units, OTUs**). Вътрешните възли представят **предполагаеми** таксономични единици – предшественици на реално съществуващите, и затова понякога се наричат **хипотетични таксономични единици, XTE (hypothetical taxonomic units, HTUs)**.

Когато дължината на разклоненията не е пропорционална на броя изменения, случили се през съответния период от време, тези разклонения са **без оразмеряване (unscaled**). При филогенетично дърво с оразмеряване разклоненията са **оразмерени (scaled)**, т.е. **дължината на всяко разклонение (branch length)** е пропорционална на броя изменения (например нуклеотидни замени), които са се случили по дължината му.

Клада (clade) се нарича групата от всички таксони, произлезли от даден общ предшественик, плюс самия общ предшественик. В молекулярната филогенетика терминът "клада" се използва обикновено за всяка изследвана група таксони, която има общ предшественик, не споделян от всеки друг таксон извън групата.

Филогенетичните дървета могат да имат или да нямат корен (Фигура 9-1а). При филогенетичното **дърво с корен (rooted tree)** съществува специален възел, наречен **корен (root),** от който уникален път води до всеки друг възел от дървото.

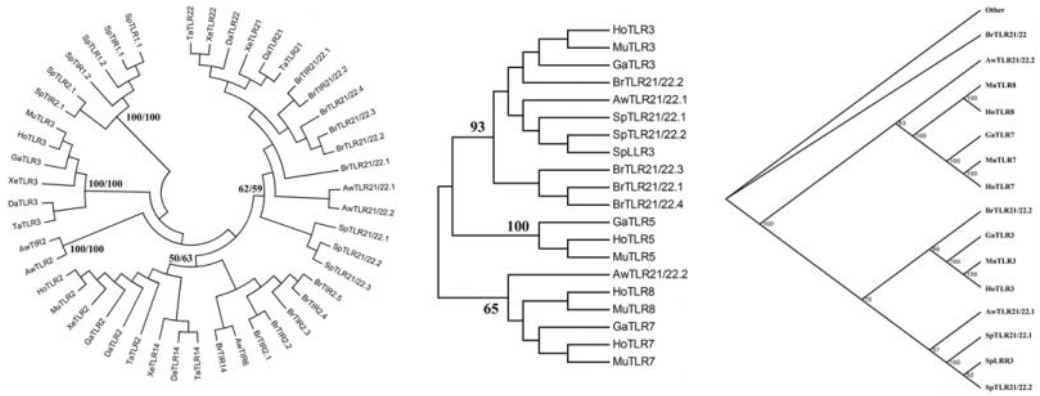
Коренът символизира най-близкия във времето **общ предшественик (common ancestor**) на всички таксономични единици, представени в даденото филогенетично дърво. Дървото с корен е ориентирано по стрелата на времето, т.е. то дава представа за посоката на еволюционния път. Чрез него таксономичните единици се дефинират като предшественици и потомци.

Филогенетичното **дърво без корен (unrooted tree)** показва само степента на родство или подобие между таксономичните единици, без да дефинира еволюционния път между тях.

Основния проблем с филогенетичните дървета е, че никой от начините за построяване реално няма методи, с които да покаже направлението на еволюцията. Те не могат да кажат кой наистина е най-древният възел. Коренът, който виждате на

филогенетичните дървета обикновено е произволен, и следователно няма биологичен смисъл. Затова биолозите често предпочитат да използват дървета без корен, за да избегнат объркванията.

Единственият подходящ начин да вкорените вашето дърво е, да използвате **външна група** (outgroup). Например, ако правите филогенетично дърво на приматите, можете да използвате мишката като външна група; ако правите филогенетично дърво на бозайниците, можете да използвате пилето като външна група и т.н. Коренът трябва да бъде поставен между външната група и останалите. Но все пак проблема с произволността на групата и съответно – корена – си остава.



Фигура 9-1а. Различни типове филогенетични дървета.

Подготовка на данните: как да изберете правилните секвенции за правилното дърво

Когато построявате филогенетично дърво, се основавате на предположението, че всички секвенции, които използвате имат общ предшественик. Това предположение е разумно, ако имате достатъчно сходни секвенции. Тогава ще искате да използвате филогенетичото дърво, за да разберете, кой на кого е най-близък роднина.

ЗАПОМНЕТЕ! Това виждане подразбира, че еволюцията винаги е дивергентна – т.е. ако две секвенции са подобни, предварително изключваме възможността това подобие да е възникнало чрез конвергентна еволюция. В повечето случаи това има смисъл – на молекулно ниво конвергентна еволюция наистина съществува, но е повече изключение, отколкото правило.

Какво да използваме: ДНК или протеини?

За да установите връзката между две секвенции, ще трябва да измерите тяхното време на дивергенция от общия им предшественик, или еволюционното разстояние (дистанция) между тях. Дали да използваме ДНК или протеинови секвенции? Отговорът на този прост въпрос е доста сложен.

- **Ако вашите ДНК секвенции са с повече от 70% идентичност:**

Можете да направите множествен алайнмънт на ДНК секвенциите. Ако обаче секвенциите ви са протеин-кодиращи, това не е препоръчително.

ВНИМАНИЕ! Проблема с множествените алайнмънти на ДНК секвенции е, че матриците на заместване за ДНК не са много добри, освен това ДНК секвенциите винаги са по-дълги и понякога алайнмънтът се изчислява доста бавно.

- **Ако вашите ДНК секвенции са с по-малко от 70% идентичност:**

Ако те кодират протеини, е по-сигурно да транслирате секвенциите си в протеинови и да направите множествен алайнмънт с протеиновите секвенции.

ПОЛЕЗЕН СЪВЕТ: Ако между секвенциите ви почти няма разлика на ниво протеин, можете да насложите ДНК секвенциите обратно върху протеиновия алайнмънт. Използвайте за тази цел **pal2nal** на адрес:

http://www.bork.embl.de/pal2nal/

Ако не разполагате с ДНК секвенцията, кореспондираща на вашите протеини, може да използвате **Protogene**, за да „заловите" нужните ви ДНК секвенции. Protogene е инструмент към Tcoffee, намира се на www.tcoffee.org

След като вече имате хомоложни ДНК и протеинови секвенции, имате две възможности:

- Ако повечето синонимни сайтове имат различия (мутации) - това е типично „мутационно насищане". В този случай, измерването на разстоянията направено на ниво протеини ще е по-точно.
- Ако синонимните сайтове не са мутационно наситени – тогава измерването на дистанциите ще е по-точно на ниво ДНК (поне на теория). Това е така, защото в ДНК са записани и тихите мутации, които нямат фенотипна изява.

ПОЛЕЗЕН СЪВЕТ: На практика, ако секвенциите ви са близки, винаги е по-лесно да работите на ниво протеин. Това може да не е толкова точно, както ако работите на ниво ДНК, но в повечето случаи ще получите смислени резултати.

Неутралните мутации като молекулен часовник

*В мисленето ни за еволюцията отдавна са се заселили идеи като: естествен отбор, оцеляване на най-приспособените, борба за съществуване и т.н. Всичко това е известно като **нео-Дарвинизъм** (или още Синтетична теория на еволюцията). Тази популярна теория твърди, че единствената причина мутациите да се запазват в популацията е, защото те са полезни и увеличават приспособеността на индивидите, които ги носят. Но тази теория няма добро обяснение за синонимните мутации и другите безвредни изменения в ДНК. През 60-те години, увеличаващото се количество молекулярни данни позволява на Motoo Кимура да развие т.н. **Неутрална теория** (неутрализъм). Тя отправя предизвикателство към нео-дарвинизма с твърдението, че повечето мутации са или неутрални, или летални. Леталните рано или късно се изчистват от популацията чрез естествения отбор, но съдбата на неутралните мутации се определя предимно от случайни събития (генен дрейф).*

Неутралните мутации не винаги са невидими на фенотипно ниво – някои от тях едва изменят протеините, без да се засяга тяхната функция. След като бъдат фиксирани в популацията, някои неутрални мутации може да се окажат полезни –

благодарение на тях индивидите, които ги носят може да се окажат приспособени към условия, които още не са настъпили (пре-адаптация).

Основната сила на Неутралната теория обаче е, че дава основа за изграждането на филогенетиката, на базата на идеята за ***молекулния часовник****. Ако повечето мутации са неутрални по отношение на ефекта си върху приспособеността, тогава съдбата им ще се определя предимно от генния дрейф и фиксирането на нова мутация ще става през повече или по-малко равномерни интервали, т.е. ще имаме молекулен часовник. Така че ако се занимавате с филогенетика, точно тези неутрални мутации са това, което ви интересува. Най-добрата мярка за еволюционно разстояние е да преброите неутралните (или почти неутрални) мутационни събития, които са се случили във вашите секвенции, след като те са се отделили от общия предшественик. Можете да мислите за всяка от тези мутации като за тиктакане на работещ часовник, натрупващи се равномерно, една по една, в течение на милионите години.*

Избор на секвенции за генно или видово дърво

Хомоложните гени са гени, имащи общ предшественик. Те могат да имат три типа отношения (съществуват три типа хомолози):

- **Ортолози:** Те са разделени само от събитие на **видообразуване (*speciation)*** — феномен, при който от общия предшественик бавно се отделят две подгрупи, които постепенно се превръщат в различни видове. Два гена са ортоложни, ако съответстват на един и същ прародителски ген. Биолозите очакват ортолозите да имат еднаква или сходна функция и структура. На Фигура 9-2 A1 и A2, както и B1 и B2 са ортолози.
- **Паралози:** Това са хомолози, разделени от събитие на дупликация – удвояване на гена-предшественик в генома. Едното копие може да е запазило старата функция, а другото да е добило нова. Очаква се паралозите да имат различни, но свързани функции. На Фигура 9-2 A1 и B1 са паралози.
- **Ксенолози:** *Xeno* е гръцка дума, означаваща „чужд”. Ксенолозите са резултат на хоризонтален генен трансфер (латерален трансфер) на ДНК между два различни вида. Типичен пример за латерален трансфер е присъствието в някои бактерии на изолевцил-тРНК синтетаза, която те са взели от техните гостоприемници. Този протеин изглежда помага на бактериите да придобият устойчивост срещу

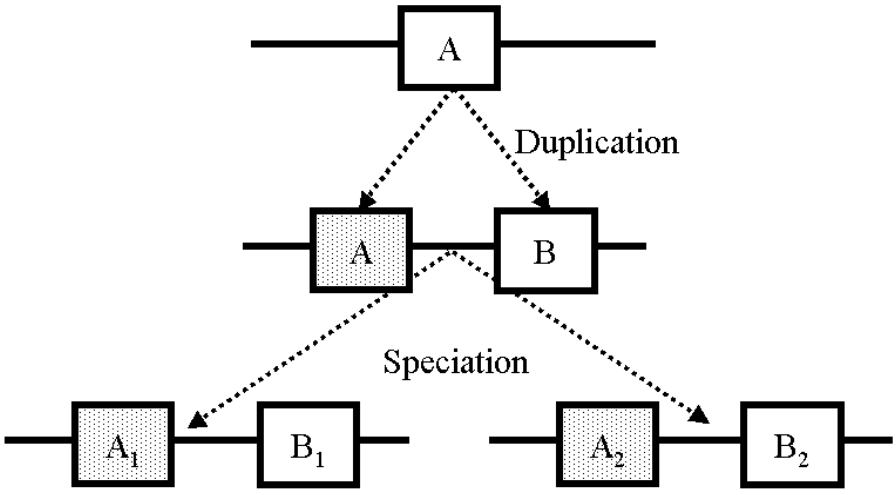
антибиотици. Новопридобитата изолевцил-тРНК синтетаза е ксенолог на „оригиналните" бактериални аминоацил-тРНК-синтетази.

ЗАПОМНЕТЕ! Два гена са **ортоложни** (**ортолози**) ако се намират в два различни организъма, произлизат от общ ген-предшественик и са разделени от събитие на видообразуване (а не генна дупликация). На теория, ортоложните гени имат еднаква (или подобна) функция в двата сравнявани вида. Два гена са **параложни** (**паралози**) ако произлизат от събитие на генна дупликация. За паралозите е по-вероятно да имат различна функция, отколкото за ортолозите.

ПОЛЕЗЕН СЪВЕТ: Когато си избирате група хомоложни гени, за да правите филогенетично дърво, това което създавате е всъщност **генно дърво**.

Ако си изберете **паралози** – членове на едно генно семейство, вашето генно дърво ще разказва историята само на това семейство гени. Ще можете да ги използвате само за да проследите последователността от дупликации, чрез които членовете на семейството са възникнали от гена предшественик.

Ако за вашето филогенетично дърво си изберете **ортолози** от различни видове, тогава вашето дърво пак ще е генно дърво, но ще прилича до голяма степен на **видово дърво** и ще ви позволи да възстановите събитията на видообразуване, които са се случили по време на дивергенцията на видовете, които изследвате. Най-добрия пример за такова генно дърво е филогенетичното дърво на 16S рРНК, което биолозите са използвали за реконструкцията на голямото дърво на Живота. Рибозомните РНК съществуват във всички видове и са типични ортолози.



Фигура 9-2. Ортолози и паралози (обяснението е в текста).

Разбира се, можете и да правите нещата „наобратно" и да използвате филогенетичното дърво, за да разберете, дали секвенциите ви са ортолози или паралози. Тази стратегия работи само ако ви е известно истинското видово дърво.

ВНИМАНИЕ! Два хомоложни гена в различни видове не са задължително ортолози! Те могат да бъдат и паралози – като A1 и B2 на Фигура 9-2. Объркването на ортолози и паралози в едно филогенетично дърво е неизчерпаем източник на проблеми. За съжаление, те често е трудно да се избегнат.

ПОЛЕЗЕН СЪВЕТ: На теория, няма просто решение на проблема с установяване на ортология, понякога това дори е невъзможно. На практика обаче, учените са изобретили емпиричен критерий, който работи доста добре – **реципрочно най-добро BLAST съвпадение** (reciprocal BLAST best match). Ето как се прилага:

1. **Изберете секвенция A от геном A.**
2. **Пуснете BLAST търсене със секвенция A срещу целия геном B.**
   Най-добрият резултат (top hit) е секвенция B.
3. **Вземете секвенция B и пуснете BLAST търсене с нея срещу целия геном A.**
   Ако при това обратно търсене секвенция A се появи като най-добро съвпадение (top hit), можете с голяма сигурност да считате секвенция A и секвенция B за ортолози от геном A и геном B.
   Този резултат не ви гарантира 100%, че секвенция A и секвенция B са наистина ортолози, но е най-доброто, което можете да направите, за да се уверите в това.

ВНИМАНИЕ! Тази стратегия има смисъл само когато и двата генома са завършени и имат приблизително сходен размер на протеома.

Eugene Koonin е съставил висококачествена колекция от ортолози към NCBI. За да го направи, той е използвал подобни, но малко по-сложни критерии. Колекцията е известна като COG (Clusters of Orthologous Groups) и можете да я намерите онлайн на http://www.ncbi.nlm.nih.gov/COG/

Други колекции на ортолози са HOGENOM (бивша HOBACGEN) и HOVERGEN, разработени към Университета в Лион, Франция. Можете да ги намерите на http://pbil.univ-lyon1.fr/.

Създаване на перфектния набор от секвенции

Тук “перфектен” е малко иронично казано. Все пак, в Таблица 9-1 сме изброили някои важни съвети, с които е добре да се съобразите, когато събирате секвенции за филогенетичен анализ.

Таблица 9-1. Подготовка на набор секвенции за създаване на филогенетично дърво.

Съвет	**Обяснение**
Избягвайте фрагментирани секвенции	Непълните секвенции създават много проблеми при множествения алайнмънт и филогенетичната реконструкция. Избягвайте ги на всяка цена, в краен случай поне се опитайте да използвате един и същ (хомоложен) фрагмент от всички секвенции.
Избягвайте ксенолози	Освен ако специално не ги изследвате, нямате друга причина да ги включвате.
Избягвайте рекомбинантни (химерни) секвенции	Някои протеини се състоят от части от няколко различни протеина. Като филогения, един такъв протеин ще има повече от един предшественик. Стандартните методи за изграждане на дървета не са екипирани да разплитат такива взаимовръзки.
Избягвайте големите и сложни генни семейства	Те може да са много трудни за анализ. Една такава проблемна фамилия е тази на ABC транспортерите. Ако можете или изобщо избягвайте големи семейства, или работете с по-малки и ясно обособени подсемейства в рамките на едно голямо.
Опитайте се да направите малък набор	Анализирането на големи набори от данни отнема много време. Може да ви се наложи да си инсталирате софтуера локално.
Добавете външна група (outgroup) към вашия набор от	Външната група е секвенция, за която знаете, че се е отделила от общия предшественик много преди останалите в групата. Например, конският хемоглобин ще ви бъде външна група, ако изследвате хемоглобините на

секвенции	приматите. Външната група позволява да сложите „корен” на вашето дърво, който символизира първия общ предшественик на всичките ви секвенции. Изборът на подходяща външна група е сложен въпрос, но със сигурност трябва да се уверите поне, че секвенцията е достатъчно близка до другите, за да могат да се включат в общ множествен алайнмънт.

ЗАПОМНЕТЕ: Живи вкаменелости не съществуват! Освен ако не използвате ДНК от фосилни останки, всички секвенции, които сравнявате са на една и съща възраст. Дори ако някой от видовете е запазил по-голямо подобие с общия предшественик, неговите секвенции са акумулирали също толкова много неутрални мутации, както и най-„прогресивния” член на групата.

Подготовка за филогенетичното дърво - множествения алайнмънт

Много методи за построяване на филогенетични дървета изискват специална Таблица, наречена матрица на разстоянията (distance matrix). Тя съдържа разстоянията (броя еволюционни събития), които разделят всяка двойка секвенции във вашия набор за анализ. На теория, можете да създадете тази матрица, като сравните всички секвенции две по две, но на практика това не работи достатъчно добре – методите за сравняване по двойки не са достатъчно точни. Единственият начин да направите такава матрица на разстоянията включва две стъпки:

1. Направете множествен алайнмънт, който включва двете секвенции.

2. Измерете разстоянието между тях на двойната проекция (pairwise projection) на секвенциите. Това е алайнмънта на двойката секвенции, който вие извличате от множествения алайнмънт. Той е обогатен с информацията, съдържаща се във **всички** сравнявани секвенции.

ЗАПОМНЕТЕ! Качеството на вашия множествен алайнмънт е истинския лимитиращ фактор за качеството на вашето филогенетично дърво – няма начин да направите хубаво дърво от лош алайнмънт!

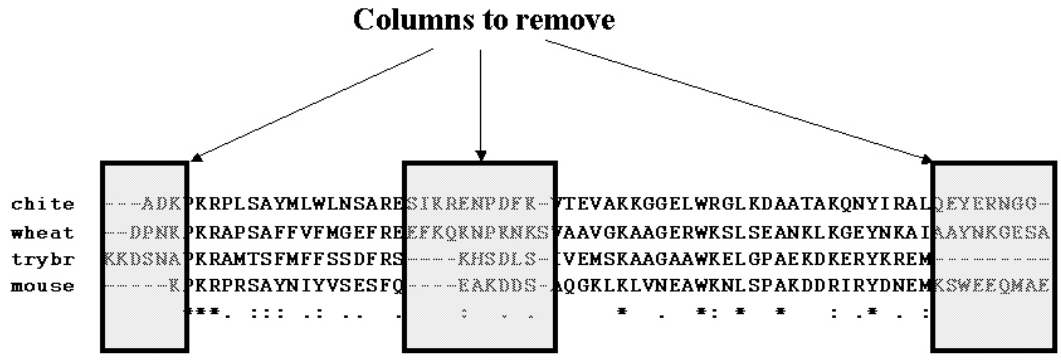
Направете множественият алайнмънт

За целта можете да използвате всеки от методите, представени в главата за множествен алайнмънт. Ето отново най-използваните сървъри за множествен алайнмънт в Интернет:

- **ClustalW**: www.ebi.ac.uk/clustalw
- **MUSCLE**: phylogenomics.berkeley.edu/cgi-bin/muscle/input_muscle.py
- **Tcoffee**: www.tcoffee.org

Уверете се, че имате правилния множествен алайнмънт

Преди да използвате вашия множествен алайнмънт за построяване на филогенетично дърво, убедете се че той е възможно най-верния. Тук ще ви покажем няколко критерии и стъпки, които може да направите, за да подобрите съвместимостта на вашия множествен алайнмънт. На Фигура 9-3 са показани типовете региони, които можете да премахнете (вижте текста по-долу).



Фигура 9-3. Подготовка на множественият алайнмънт.

За да се уверите, че вашият множествен алайнмънт е точен и подходящ, следвайте тези стъпки:

1. **Уверете се, че имате максимален брой колони без дупки.**

   Дупките причиняват проблеми в повечето методи за построяване на филогенетични дървета. Някои като ClustalW, могат да изтриват автоматично всяка колона, съдържаща дупки.

2. **Премахнете крайните части на вашия множествен алайнмънт.**

   N- и C-краищата на протеините обикновено са слабо консервативни и не се сравняват добре. Можете спокойно да ги премахнете (вижте Фигура 9-3).

3. **Премахнете регионите с много дупки от вашия множествен алайнмънт (вижте Фигура 9-3)**.

   Вътрешните региони с много дупки често съответстват на бримки и други слабо консервативни участъци.

4. **Запазете най-информативните блокове.**

ЗАПОМНЕТЕ! Идеалният множествен алайнмънт за изграждане на филогенетично дърво трябва да е висококачествен, да е изграден от секвенции с ниско ниво на идентичност, така че всяка позиция да носи белег от историята на семейството (групата организми).

Ето няколко насоки:

- Добрият блок в един множествен алайнмънт типично е с дължина около 20-30 АК и съдържа малко консервативни позиции. Такива блокове са идеални за реконструкция на висококачествени филогенетични дървета.
- Може да използвате сървъра на Tcoffee за да оцените вашият множествен алайнмънт и да премахнете колоните, за които е вероятно да не са подравнени правилно.
- Най-добре можете да редактирате вашия множествен алайнмънт с помощта на Jalview, редактор на множествени алайнмънти.

ПОЛЕЗЕН СЪВЕТ: По-бърз (но и по-съмнителен) начин е да използвате директно текстов редактор като Notepad++, за да премахнете ненужните ви блокове. Ето как:

- Дръжте натиснат клавиша Alt и използвайте мишката, за да изберете цели колони във вашия алайнмънт.
- След като изберете всичко, което искате да премахнете, натиснете Delete, за да изтриете избраните блокове.

ВНИМАНИЕ! Не се опитвайте да правите нищо друго със секвенциите и алайнмънтите с текстовия редактор! За да редактирате секвенциите си, използвайте подходящ алайнмънт редактор като Jalview.

Изграждане на точния тип филогенетично дърво

Има три основни групи методи, които можете да използвате за изграждане на филогенетично дърво – **методи на разстоянията (distance methods), методи на икономията (parsimony methods) и методи на вероятността (likelihood methods).** Никой от тях не е всемогъщ, всички си имат плюсове и минуси. В тази част ще разберете как да построите филогенетични дървета, базирани на разстояния и на максимална вероятност. Дърветата, базирани на разстояния не винаги са най-прецизните, но са приложими за много цели и са лесни за изграждане. В повечето възможни ситуации можете да прилагате методи за базирано на разстояния изграждане на дървета. Методите на максимална икономия и максимална вероятност (Maximum Parsimony, Maximum Likelihood) на теория са най-прецизни, но са по-бавни. От двете групи, експертите донякъде са склонни да отдават леко предпочитание на вероятностните методи.

Когато решите да използвате базиран на разстояния метод за вашето филогенетично дърво, трябва да вземете предвид четири основни неща:

- **Вашият множествен алайнмънт:** По кой метод? С какви секвенции?
- **Измерването на разстоянията:** По коя матрица на заместване?
- **Алгоритъма за изграждане на дървото:** UPGMA, NJ, Fitch или Kitsch?
- **Какъв тип филогенетично дърво искате:** Дърво без разстояния или с показване на дистанциите? Дърво с корен или без корен?

В този раздел ще разберете как да си построите филогенетично дърво и да го покажете в желания от вас формат.

Изчисляване на вашето филогенетично дърво

Няма най-добра рецепта за филогенетично дърво – това е донякъде „въпрос на вкус”.

Има два много различни начина за правене на дървета. Първия използва ClustalW. Той е бърз, безпроблемен и без много внимание към детайлите. Втория е по-прецизен и

бавен, и ви позволява повече контрол върху параметрите. Той се прави с пакета Phylip, с който можете да правите наистина сериозна филогения и се използва от най-добрите специалисти в областта. Въпреки това, Phylip е изненадващо лесен за употреба.

Бързо и лесно създаване на филогенетично дърво с ClustalW

Ако просто ви трябва филогенетично дърво и сте сигурни, че не искате да знаете нищо за детайлите по създаването му, ClustalW е програмата за вас.

ClustalW е преди всичко програма за множествен алайнмънт, но някои от ClustalW сървърите дават възможност да го използвате и за производство на филогенетични дървета.

ВНИМАНИЕ! Докато правите множествен алайнмънт, ClustalW генерира водещо дърво, което му помага да изгради алайнмънта, но това **не е** филогенетично дърво! Всяко дърво с окончание на файла .dnd произведено от ClustalW **не е** филогенетично дърво! Не можете да го използвате вместо филогенетично дърво – ще получите неверни резултати.

Тук ще разберете как да направите истинско филогенетично дърво с EBI сървъра на ClustalW. Предполага се, че вече сте направили и редактирали своя множествен алайнмънт. За примера тук сме използвали готов „dummy" алайнмънт, изтеглен от www.tcoffee.org/dummy_aln2.html.

Той съдържа хомоложни фрагменти от белтъка гама-фибриноген, който е секвениран при много видове. Преименували сме фрагментите за яснота с видовите имена:

O12957 (овца), O02672 (мишка), O02683 (жираф)

O02690 (chevrotain), O02681 (beluga) O02687 (кашалот)

O02673 (Rorqual, кит от род Balaenoptera), O02688 (прасе), O12959 (пекари)

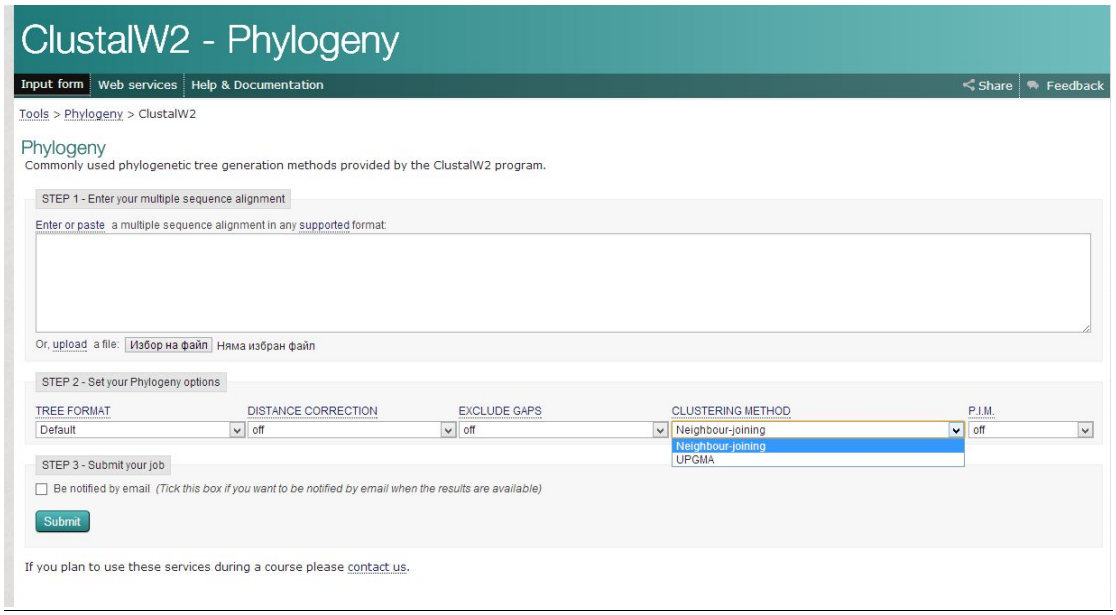
O02677 (едногърба камила), O02689 (тапир), O02682 (кон)

O02676 (хиена), O02680 (койот), O12954 (хипопотам)

1. **Отидете на http://www.ebi.ac.uk/Tools/phylogeny/clustalw2_phylogeny/**

2. **Поставете вашия множествен алайнмънт в прозореца, както е показано на Фигура 9-4.**
   Можете да въведете вашия алайнмънт във всякакъв формат, включително ALN (форматът на ClustalW), MSF или FASTA. Ако използвате алайнмънт с ALN формат, задължително въведете и първия (заглавен ред “CLUSTAL W ...(и т.н.)”

3. **От менюто за метод Type изберете, както е показано на Фигура 9-4.**
   Това меню ви позволява да избирате от няколко различни метода за построяване на филогенетични дървета. Смята се че NJ (Neighbor Joining method) е най-добрия за такъв тип задачи.



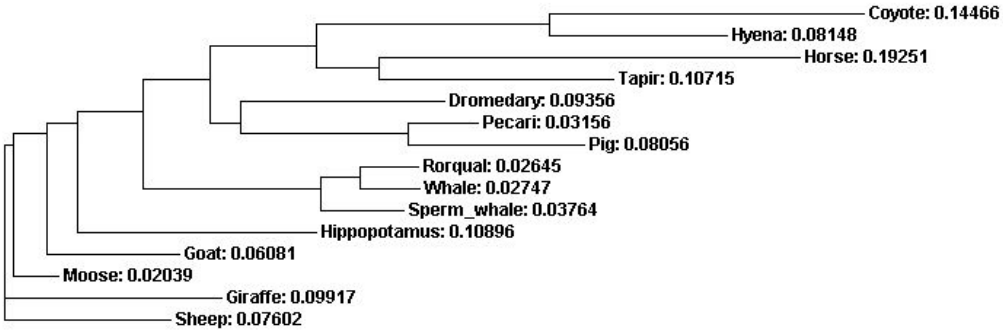
Фигура 9-4. Създаване на филогенетично дърво с ClustalW.

4. **От менюто Correct Dist. (вдясно, точно до менюто Tree Type) изберете On**
   Тази опция позволява да се внасят корекции за множествени замени. Има свойството да удължава клонките на филогенетичното дърво. Най-полезна е за далекородствени видове.

5. **От менюто Exclude Gaps изберете On.**
   С това казвате на ClustalW да игнорира всяка колона, в която има дупка. Специалистите обикновено предпочитат да игнорират позициите с дупки. Включването на тази опция ви гарантира, че ще бъде сравнен еднакъв брой остатъци от всички секвенции.

6. **Натиснете бутона Submit.**
   Почакайте, докато ClustalW покаже вашето NJ дърво, което можете да видите на Фигура 9-5. Намира се в долния край на страницата с резултати. Това дърво е много по-точно от водещото и ясно показва генетичната връзка между хипопотама и кита, която е потвърдена в едно изследване наскоро.

7. **Запазете графичното изображение на вашето дърво.**
   Най-лесният начин да запазите вашето дърво е да направите екранно копие (print screen) с клавиша (Prt Sc). След това може да го поставите и редактирате във всеки редактор на изображения.



Фигура 9-5. NJ филогенетично дърво, съставено с ClustalW.

От алтернативните можете да изберете дали дървото ви изглежда като кладограма (неоразмерено, с еднаква дължина на разклоненията) или като филограма (дължината на разклоненията е пропорционална на разстоянието).

От алтернативните бутони Show distances/Hide distances можете да изберете, дали разстоянията да бъдат показани до името на всяко разклонение, или да не се виждат.

9. Натиснете бутона View PH File.

Появява се вашето филогенетично дърво в Newick формат (Вижте Фигура 9-6)

10. Използвайте опцията File□Save As на браузъра, за да запазите текстовата версия на вашето дърво.

ЗАПОМНЕТЕ! ClustalW представя филогенетичното дърво в т.н. Newick текстов формат, показан на Фигура 9-6. Когато запазвате файла, вие всъщност запазвате Newick-версията на вашето дърво. Този формат е най-стандартният начин да запазите едно филогенетично дърво, повечето филогенетични програми го разпознават и използват. Ако искате да запазите само един файл за вашето филогенетично дърво, използвайте този!

```
(
(
(
Coyote:0.14466,
Hyena:0.08148)
:0.10650,
(
Horse:0.19251,
Tapir:0.10715)
:0.02904)
:0.04845,
(
Dromedary:0.09356,
(
Pecari:0.03156,
Pig:0.08056)
:0.07672)
:0.01404)
:0.03101,
(
(
Rorqual:0.02645,
Whale:0.02747)
:0.01784,
Sperm_whale:0.03764)
:0.08182)
:0.02954,
Hippopotamus:0.10896)
:0.01439,
Goat:0.06081)
:0.01489,
Moose:0.02039)
:0.00428,
Giraffe:0.09917,
Sheep:0.07602);
```

Фигура 9-6. Филогенетично дърво в Newick формат.

Създаване на филогенетично дърво с Phylip

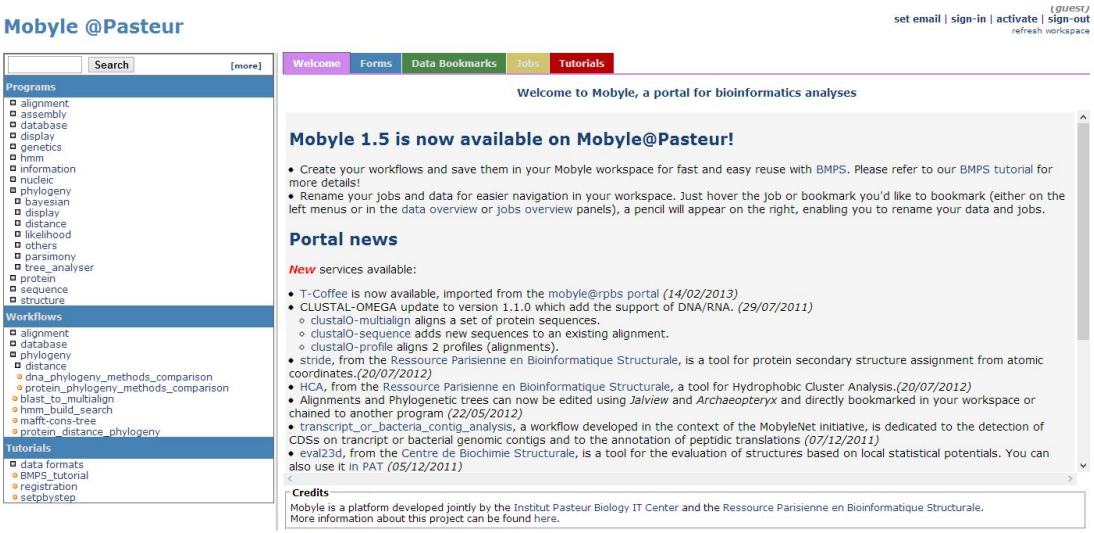
Когато стане въпрос за филогенетични дървета, Phylip е най-известната марка. Заедно с PAUP, това е един от най-използваните ресурси за създаване на висококачествени филогенетични дървета.

Едно от предимствата на Phylip е, че ви улеснява да направите т.н. bootstrap на вашето дърво (да проверите неговата надеждност; по-долу ще разберете как става това).

Ако решите да ставате специалист по филогенетика, Phylip има всичко, което ви е нужно. Може да го използвате да построите всякакъв вид филогенетично дърво, с всеки метод, който си изберете. По-долу ще ви демонстрираме стъпките, необходими за да произведете филогенетично дърво с bootstrap информация.

ВНИМАНИЕ! Преди да отидете на този сървър, трябва да имате готов висококачествен множествен алайнмънт. Един такъв, създаден с учебна цел, можете да изтеглите от www.tcoffee.org/dummy_aln2.html.

1. **Отидете на адрес http://mobyle.pasteur.fr/cgi-bin/portal.py#welcome**



Фигура 9-7. Онлайн програми на сървъра на института Пастьор.

Появява се страницата **с програми**, както е показана на Фигура 9-7.

Тази страница съдържа всички програми от пакета Phylip, които са инсталирани на сървъра. Можете да създадете филогенетично дърво, използвайки ги една след друга – трябва само да ги използвате в правилния ред.

2. **Потърсете подпрограмата protdist чрез менюто в ляво**
   Отваря се страницата Phylip: Protdist към портала Mobyle@Pasteur. Ето какво трябва да знаете тук:

   A. За да направите дърво базирано на разстояния, първо трябва да генерирате матрица на разстоянията (distance matrix), съдържаща разстоянията между двойките секвенции, изчислени на базата на множественния алайнмънт.

B. Някои методи ви позволяват да изчислите разстоянието между двойка ДНК или протеинови секвенции. Има и различни програми за близки и далечни секвенции.

C. **Protdist** е метод, използващ матрица на заместване (substitution matrix) за измерване на разстоянието между секвенциите във вашия алайнмънт и да генерира матрица на разстоянията (distance matrix). С тази матрица след това можете да захраните всяка от програмите на Phylip, които работят с разстояния.

D. **Protpars** прави същото, но се опитва да изчисли броя мутации на ниво ДНК, необходими за появата на съществуващите различия на ниво протеин. Protpars работи по-добре с близкородствени секвенции.

3. **Ако използвате портала Mobyle@Pasteur за пръв път от вашия компютър, се появява форма, в която трябва да въведете вашия *e-mail* адрес и за потвърждение да повторите знаците от изображение.**
   Програмата запомня вашия IP адрес и не ви изисква тези данни при повторно използване.

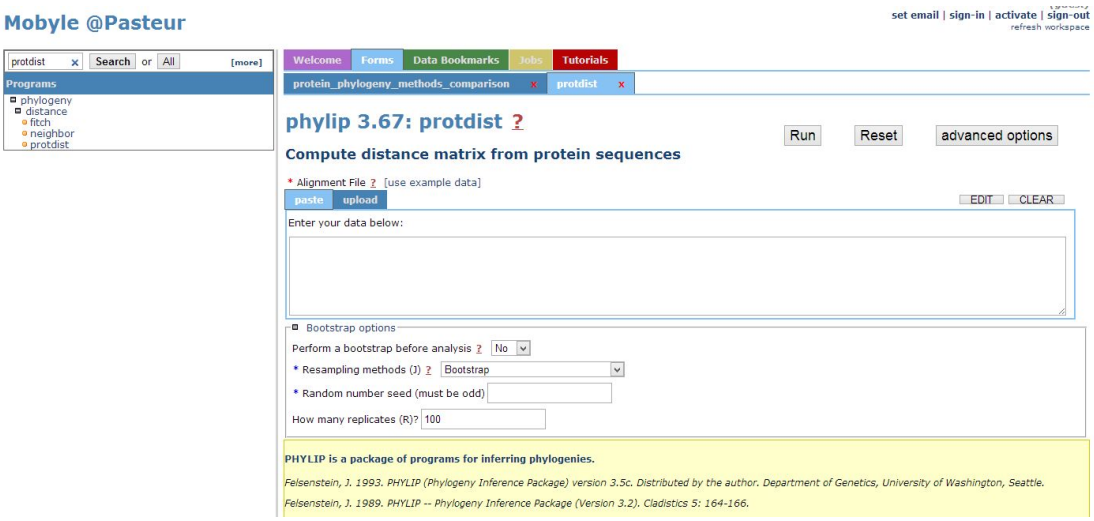
   A. Ако сървърът е претоварен, или ако алайнмънта ви е голям, Phylip ви изпраща резултатите на e-mail, а не ви ги показва онлайн.

   B. Не давайте фалшив e-mail! Почти винаги щом сървърът се претовари, той решава да ви прати резултатите по e-mail и ако той е фалшив, никога няма да ги получите! Обикновено те идват по е-мейла за няколко минути.

4. **Поставете вашия множествен алайнмънт в прозореца Sequence, както е показано на Фигура 9-8.**
   Не забравяйте да копирате заедно с алайнмънта и заглавния ред, започващ с CLUSTALW.

5. **Маркирайте Perform a Bootstrap Before Analysis, както е показано на Фигура 9-8.**
   Това е начин за оценка на качеството на вашето филогенетично дърво.

6. **Въведете за „seed" нечетно число, например 1 или 3.**
   „Seed" контролира генерирането на случайни числа по време на bootstrap анализ. Различни seeds водят до различни резултати.



Фигура 9-8. Protdist сървърът.

7. **Оставете bootstrap за метод.**

8. **Оставете броя на копията (replicates) на *100.***
   Броят копия (replicates) е броят на bootstrap циклите, които правите. Обикновено, този брой трябва да е поне 100. Запомнете колко копия сте поискали, защото ще трябва да въведете тази цифра по-късно!
   За истински анализ ви трябват поне 100 цикли, но ако просто искате да разберете как работи сървъра, можете да поставите и по-малки числа, дори 2. В този случай изчислението ще стане бързо, но не очаквайте вашия bootstrap да има биологичен смисъл! Ако изберете 2, обикновено ще получите резултатите онлайн, но при по-големи стойности, обикновено ще ви се изпратят по e-mail.

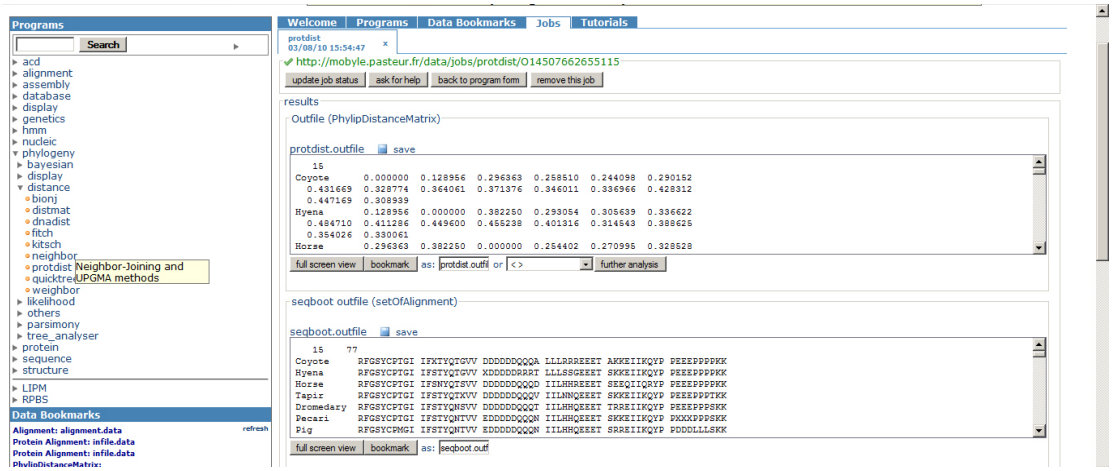
9. **Върнете се в горната част на страницата и натиснете бутона Run Protdist.**
   Ще видите резултати като показаните на Фигура 9-9. (Алтернативно, ще получите по е-мейла линк към URL и след като го отворите, ще видите пак подобна страница.) Страницата с резултатите ви съдържа три полета:

   A. **Outfile (PhylipDistanceMatrix)** – това е най-важната част от вашия резултат – самата матрица на разстоянията, която програмата Protdist е генерирала на базата на вашия множествен алайнмънт.

   B. **seqboot outfile (setOfAlignment)** – това са всички 100 (или толкова, колкото сте задали) bootstrap алайнмънти, използвани за да се провери здравината на вашето дърво.

   C. **standard output (Text)** – текстово описание на Bootstrapping алгоритъма и всички използвани параметри.

Можете да видите на цяла страница (като натиснете бутона full screen view) или да запишете като отделен файл всяко поле с резултати.



Фигура 9-9. Част от страницата с резултати на Protdist.

10. **Копирайте в Clipboard матрицата на разстоянията от полето Outfile (PhylipDistanceMatrix).**

11. **От списъка с програми (намира се вдясно на всички страници на портала Mobyle@Pasteur, включително и на вашата страница с резултати) изберете Phylogeny -> Distance ->. Neighbor.**
    Neighbor е програма, която превръща вашата матрица на разстоянията във филогенетично дърво, създадено по метода Neighbor Joining.

12. **Изберете Distance method.**
    Ако искате вашето дърво да е с корен, изберете UPGMA (Advanced options). Ако искате да е без корен, изберете Neighbor-joining. Neighbor Joining и UPGMA са свързани методи, но от двата Neighbor Joining дава по-точни дървета. UPGMA дава дървета с корен – корена е опит да се определи позицията на общия предшественик.
    Neighbor-Joining не се опитва да определи общия предшественик и произвежда дървета без корен, които можете да „вкорените" по-късно.
    Можете да използвате и други методи, например Fitch, но изглежда най-много използвани и най-добри по всеобщо съгласие са двата NJ метода.

13. **Поставете матрицата на разстоянията, генерирана с Protdist в полето Distances matrix File (PhylipDistanceMatrix).**

14. **Придвижете се надолу по страницата на Neighbor до секцията Bootstrap options.**

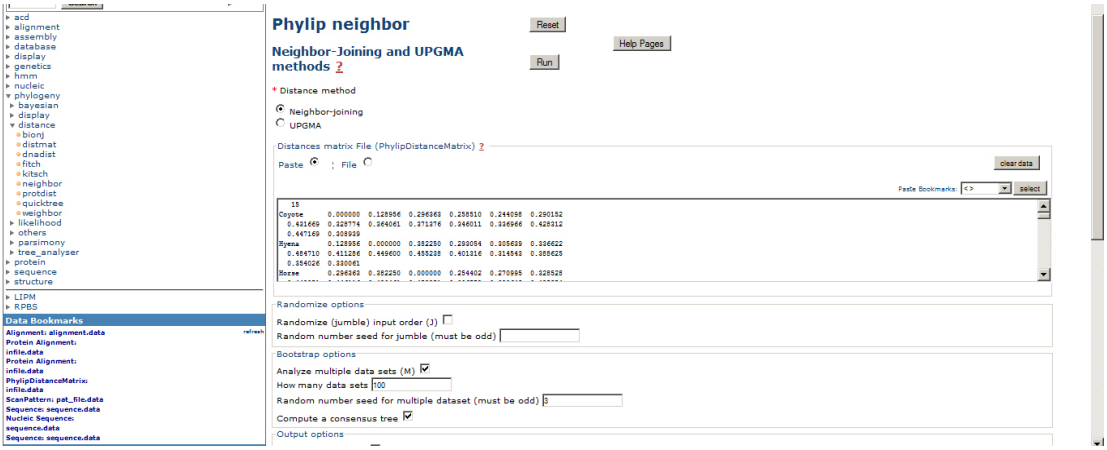
    A. **Изберете опцията Analyze Multiple Data Sets.**

    B. **Въведете същия брой за „number of data sets", който сте въвели на стъпка 8, както е показано на Фигура 9-10.**
    Тази секция НЕ Е идентична с това, което въведохте в Protdist. На тези стъпки

решавате колко Bootstrap цикъла искате да направите. Сега, на стъпка 14 е време да напомните на програмата колко цикъла е направила!

C. **Настройте** Random number seed for multiple dataset **на нечетно число (най-добре 1 или 3) и маркирайте бокса Randomize.**
Тук няма нужда да използвате същите числа като в стъпка 7.

D. **Маркирайте бокса Compute a Consensus Tree.**
Тази опция казва на Neighbor да подаде дървото с най-добро съответствие с данните и броя копия, които сте въвели. Има смисъл само ако използвате bootstrap и сте избрали число по-голямо от 1 на стъпки 8 и 14 (multiple datasets).



Фигура 9-10. Входната страница на Neighbor.

15. Отидете горе на страницата и натиснете бутона Run.

Резултатите ви се появяват на нова страница с няколко полета.

A. consense.outfile – съдържа графична версия в текстов формат на вашето консенсусно дърво. Числото над всяко разклонение показва неговата стабилност (например 50 означава по-стабилно, 20-по-малко стабилно). Изглежда като показаното на Фигура 9-11.

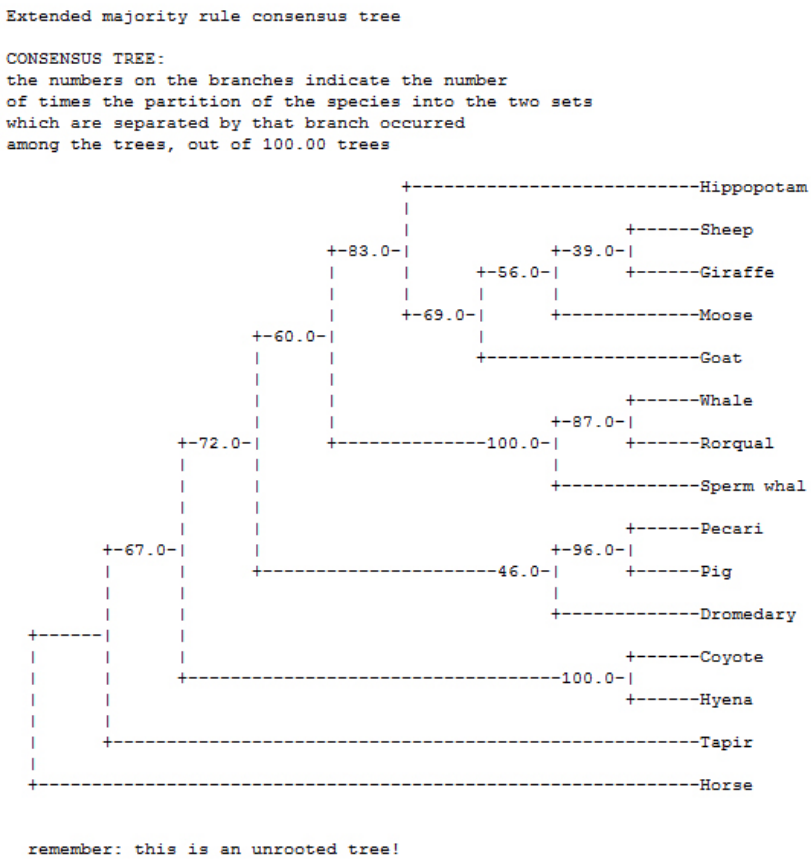
B. consense.outtree - вашето консенсусно дърво в Newick формат.

C. neighbor.outfile - програмата Neighbor дава като резултати и дърветата, които е използвала за да изгради консенсуса в outfile и outtree. Съдържа толкова дървета, колкото копия (replicates) сте въвели в стъпки 8 и 16.

16. Запишете резултатите си.

С онлайн инструменти е трудно да се покаже филогенетично дърво с bootstrap стойности. Ако искате да запазите тези стойности, най-добре е да запазите файла outfile.consense като дърво в Newick формат. (Това е стандартен текстов файл, който можете да запазите чрез опцията File Save As на браузъра си).

Ако искате да покажете отново резултатите си в друг графичен формат, най-удобно е да запазите файла outtree.consense и да го преобразувате в GIF или PDF файл с помощта на сървъра Phylodendron (Вижте по-долу как да направите това).



Фигура 9-11. Консенсусното дърво, получено като резултат на Neighbor.

Какво е bootstrapping?

Това е начин да проверим, дали филогенетичното дърво, което сме генерирали, има биологичен смисъл като проверим неговата здравина. Това означава да проверим, дали всяка част от алайнмънта еднакво подкрепя нашето дърво.

Проверява се здравината на всяко разклонение в дървото и се преценява, дали можем да му се доверим или не, когато правим крайните изводи.

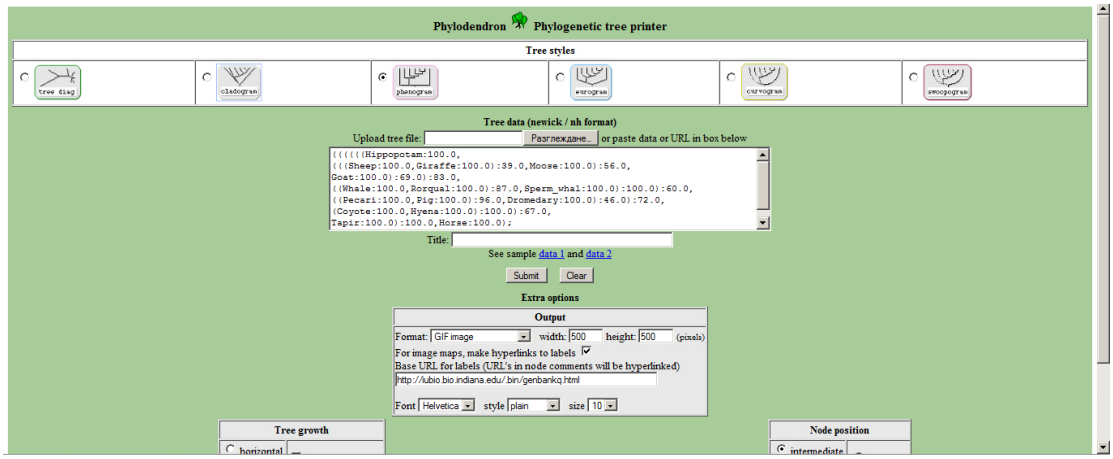
Bootstrapping е процес от много стъпки. Започва се със случаен избор на колони от първоначалния алайнмънт, като целта е да се генерира bootstrap алайнмънт, в който някои колони липсват, а други са повторени два или повече пъти. Това се нарича random sampling. Процедурата генерира множество такива bootstrap алайнмънти, които приличат повече или по-малко на оригиналния. Целта на всичко това е да се провери, дали всички колони във вашия първоначален алайнмънт разказват подобна еволюционна история, както се очаква. На следващата стъпка, всеки случаен алайнмънт се превръща в матрица на разстоянията (distance matrix) и всяка матрица се превръща във филогенетично дърво. За да изгради консенсусното дърво, Phylip осреднява всички дървета, които е генерирал от

bootstrap алайнмънтите. За да оцени качеството на всяко разклонение от консенсусното дърво, Phylip преброява колко bootstrap дървета съдържат това конкретно разклонение. Добрите разклонения са тези, които се появяват във всяко bootstrap дърво. На тях можете да се доверите. В практиката, ще имате нужда от поне 100 bootstrap цикъла, за да можете да използвате истински този начин на проверка на вашето филогенетично дърво.

Визуализация на вашето филогенетично дърво

Phylodendron е мощен уеб сървър, който ви позволява да превърнете текстов файл в Newick формат в графично изображение на вашето дърво. Можете да контролирате всеки аспект от изображението и да го запишете във всеки формат, който си изберете. По-долу ще ви покажем как да използвате Phylodendron, за да превърнете текстов файл на филогенетично дърво в GIF изображение, което може да бъде включено във всеки документ.

1. **Отидете на адрес http://iubio.bio.indiana.edu/treeapp/**
   Оттук можете да се запознаете с помощната документация на Phylodendron.
2. **Кликнете върху линка Phylodendron Web Form.**
   Появява се входната страница на Phylodendron (Фигура 9-12).
3. **Поставете вашето дърво в Newick format в прозореца Tree data, както е показано на Фигура 9-12.**
4. **На горния ред под Tree Styles, изберете радио бутона Phenogram.**
   Ако изберете Tree Diag, трябва да сте сигурни, че дървото ви е без корен, иначе програмата ще блокира. Можете да изпробвате и други, по-рядко използвани изгледи за вашето дърво (eurogram, curvogram и т.н.).
5. **В секцията Extra Options, изберете GIF image от менюто Format.**
6. **Ако искате разстоянията да имат значение, изберете бокса Use Node Lengths в секцията Tree Growth.**
7. **Натиснете бутона Submit.**
8. **Изберете File Save As от браузъра, за да запазите резултата като GIF файл, който можете да включвате в други документи**.



Фигура 9-12. Входната страница на Phylodendron.

Свободни ресурси за филогенетика в Интернет

Филогенетичните анализи в Интернет не са чак толкова разпространени, както търсенето в бази данни и някои други методи за анализ на секвенции. Все пак има достъпни няколко висококачествени ресурси, оборудвани с всичко, което ви трябва, за да построите вашето филогенетично дърво.

Повечето програми за филогенетика са достъпни за изтегляне от Интернет, но трябва да си ги инсталирате локално на вашия компютър. Има все пак и такива, които можете да ползвате онлайн. В Таблица 9-2 са изброени основните от тях.

Таблица 9-2. Онлайн ресурси за филогенетика.

Адрес	Описание
www.ebi.ac.uk/	Можете да използвате ClustalW за множествени алайнмънти ИЛИ за генериране на NJ филогенетични дървета, но НЕ и за двете неща наведнъж!
www.genebee.msu.ru/clustal/basic.html	Сървърът на Genebee прави истински филогенетични дървета само в една стъпка
www.tcoffee.org	Tcoffee също прави истински филогенетични дървета само в една стъпка

http://www.jalview.org/	Можете да използвате Jalview за създаване на NJ филогенетични дървета. Това е мощен ресурс, комбиниращ редактиране на алайнмънти със създаване на дървета
http://www.atgc-montpellier.fr/phyml/	Мощен метод за създаване на дърветата с максимална вероятност
http://mobyle.pasteur.fr/cgi-bin/portal.py?form=bionj	Интерфейс за BioNJ, нов NJ метод.
http://sites.univ-provence.fr/evol/figenix/	Мощен Java инструмент за събиране на членове на протеинови семейства и свързването им във филогенетично дърво
http://bioweb.pasteur.fr/phylogeny/intro-en.html	Web интерфейс за филогенетични ресурси на Пастьоровия Институт.

Общи ресурси

В Таблица 9-3 са изброени някои важни ресурси за филогенетика. Някои от тях имат доста добри обучителни материали, от които бързо можете да научите всичко за филогенетичните дървета и други свързани теми.

Таблица 9-3. Общи ресурси за филогенетика в Интернет.

Адрес	**Описание**
http://evolution.genetics.washington.edu/phylip/software.html	„Домашната" страница на Phylip; а също и много богата колекция от филогенетични ресурси.
http://www.ucmp.berkeley.edu/subway/phylo/phylosoft.html	Много пълен списък с филогенетични ресурси
http://paup.csit.fsu.edu/index.html	„Домашната" страница на PAUP, филогенетичен ресурс за създаване на дърветата с максимална икономия. Това е комерсиален пакет, но е на разумна цена.

http://www.ncbi.nlm.nih.gov/About/primer/phylo.html	Въвеждащ курс по филогения към NCBI.
http://www.techfak.uni-bielefeld.de/bcd/Curric/MathAn/mathan.html	Отличен курс по създаване на филогенетични дървета

Колекции с ортоложни бази данни

Такива колекции от ортолози и други хомолози са много полезни за функционални и филогенетични анализи.

Някои бази данни за хомолози са изброени в Таблица 9-4.

Таблица 9-4. Колекции от ортоложни гени.

Адрес	**Описание**
www.ncbi.nlm.nih.gov/COG/	Клъстерите с ортоложни секвенции, поддържани към NCBI. Всеки клъстер съдържа протеини от бактериални геноми.
http://pbil.univ-lyon1.fr/databases/hovergen.php	Колекция от ортоложни гени на гръбначни.
http://pbil.univ-lyon1.fr/databases/hogenom.html	Колекция от ортоложни бактериални гени.
http://systers.molgen.mpg.de/	Друга колекция от ортоложни секвенции.
http://rdp.cme.msu.edu/ http://bioinformatics.psb.ugent.be/webtools/rRNA/ssu/ http://bioinformatics.psb.ugent.be/webtools/rRNA/lsu/	Три обширни колекции от рРНК секвенции, много полезни за класифициране на нови организми. Към тях има и подходящи филогенетични програми.

Упражнение 5

Филогенетични дървета

1. Направете филогенетично дърво от 5 организма по иРНК на гена за porin и сравнете секвенциите с Geneious, като добавите опция за дължина на клоните.

2. Направете филогенетично дърво на 5 организма (като включите човек) по иРНК на гена MYCBP (MYC binding protein).
   Кой е най-близко родственият еволюционно вид до човека?
   _____

3. На адреса с файловете по биоинформатика – http://web.uni-plovdiv.bg/vebaev/Bioinformatics/sequences, имате секвенции на ген лин-28 от 5 организма във файл seqs.fa. Ако знаете че еволюционно най-близка до човешката секвенция е мишата, то намерете секвенцията на мишката.

   A. Проверете дали намерената от вас секвенция е наистина миша.

   B. Мишата секвенция е номер _____

   C. Ако знаете че секвенцията на плъха е най-отдалечена еволюционно и не формира клъстер/група, можете ли да я намерите? Да/Не? _____

Приложение (списък с био-бази данни)

Първични бази данни

The International Nucleotide Sequence Database (INSD) (http://www.insdc.org/) състои се от:

1. DDBJ (DNA Data Bank of Japan)
2. European Nucleotide Archive (European Bioinformatics Institute)
3. GenBank [1] (National Center for Biotechnology Information)

Трите бази данни са хранилища на нуклеотидни секвенции за всички организми. Те са унифицирани бази данни съдържащи всякакъв тип информация. Поддържат синхронизация и информацията между тях се обновява всекидневно.

Мета-бази данни

Те представляват бази данни от бази данни и в повечето случай информацията в тях е организирана и оптимизирана за даден проект, организъм или тип информация.

1. BioGraph (University of Antwerp, Vlaams Instituut voor Biotechnologie).
2. Bioinformatic Harvester[2] (Karlsruhe Institute of Technology) – Интегрира 26 вида бази данни с генни ресурси.
3. ConsensusPathDB – Молекулярно-функционална база данни интегрираща 12 бази данни.
4. Entrez[3] (National Center for Biotechnology Information).
5. Enzyme Portal – интегрира информация за ензими. (European Bioinformatics Institute).
6. euGenes (Indiana University).
7. MetaBase[4] (KOBIC).
8. mGen - съдържа 4 от най-големите бази данни GenBank, Refseq, EMBL и DDBJ
9. PathogenPortal
10. SOURCE (Stanford University)

Геномни бази данни

Тази бази данни съдържат информация за определен организъм.

1. Bioinformatic Harvester.
2. SNPedia.
3. CAMERA Ресурс за микроорганизми и метагеномика.
4. Corn, геномна база данни за царевица.
5. EcoCyc ресурс за а *E. coli K-12.*
6. Ensembl автоматична анотация за човек, мишка и др. гръбначни организми.
7. Ensembl Genomes.
8. PATRIC, the PathoSystems Resource Integration Center.
9. Flybase, геномна информация за Drosophila melanogaster.
10. MGI Mouse Genome (Jackson Lab).
11. JGI Genomes of the DOE-Joint Genome Institute – съдържа информация за много еукариотни и микробиални геноми.
12. National Microbial Pathogen Data Resource. Ресурс за патогени - Campylobacter, Chlamydia, Chlamydophila, Haemophilus, Listeria, Mycoplasma, Neisseria, Staphylococcus, Streptococcus, Treponema, Ureaplasma, и Vibrio.
13. Saccharomyces Genome Database, геномен ресурс за дрожди.
14. Xenbase, геномен ресурс за Xenopus tropicalis и Xenopus laevis.
15. Wormbase, геномен ресурс за Caenorhabditis elegans.
16. Zebrafish Information Network, геномен ресурс на Danio rerio.
17. TAIR, The Arabidopsis Information Resource.
18. UCSC Malaria Genome Browser, геномен ресурс за малария (*Plasmodium falciparumata*).
19. RGD Rat Genome Database: геномен ресурс за Rattus norvegicus

Бази данни за протеини

1. UniProt[6] универсална база данни за протеини (UniProt Consortium: EBI, Expasy, PIR).
2. PIR Protein Information Resource (Georgetown University Medical Center (GUMC)).
3. Swiss-Prot[7] Protein Knowledgebase (Swiss Institute of Bioinformatics).
4. PROSITE база данни за протеинови семейства и домени.
5. DIP база данни за взаимодействия на протеини (Univ. of California).
6. Pfam база данни за протеинови семейства(Sanger Institute).
7. ProDom база данни за протеинови семейства (INRA/CNRS).
8. SignalP 3.0 сървър за предсказване на локализационен сигнал на протеини.
9. SUPERFAMILY протеинни суперфамилии.
10. InterPro[8] протеинови семейства, домени, класификация

Структурни бази данни за протеини

1. Protein Data Bank (PDB):
   A. Protein DataBank in Europe (PDBe).
   B. ProteinDatabank in Japan (PDBj).
   C. Research Collaboratory for Structural Bioinformatics (RCSB).
2. SCOP Structural Classification of Proteins.
3. CATH база данни за класификация на протеини.
4. PDBsum.

Бази данни за РНК

1. Rfam, РНК семейства.
2. mirBase, база данни за миРНК.
3. snoRNAdb, база данни за сноРНК.

Бази данни за протеин-протеин взаимодействия

1. BIND Biomolecular Interaction Network Database.
2. BioGRID CCSB Interactome.
3. DIP Database of Interacting Proteins.
4. IntAct molecular interaction database: NetPro.
5. STRING: STRING is a database of known and predicted protein-protein interactions. (EMBL).
6. The Cell Collective.
7. MINT: Molecular INTeraction database.

Специализирани бази данни

1. CGAP Cancer Genes (National Cancer Institute).
2. Clone Registry Clone Collections (National Center for Biotechnology Information).
3. DBGET H.sapiens (Univ. of Kyoto).
4. Edinburgh Mouse Atlas.
5. GreenPhylDB – филогенетична база данни.
6. GDB Hum. Genome Db (Human Genome Organisation).
7. HGMD disease-causing mutations (HGMD Human Gene Mutation Database).
8. HUGO (Official Human Genome Database: HUGO Gene Nomenclature Committee).

9. HvrBase++ Human and primate mitochondrial DNA.
10. INTERFEROME The Database of Interferon Regulated Genes.
11. List with SNP-Databases.
12. NCBI-UniGene (National Center for Biotechnology Information).
13. Oncogenomic databases – компилация от бази данни с онкогени.
14. OMIM Inherited Diseases (Online Mendelian Inheritance in Man).
15. OrthoMaM (A database of Orthologous Mammalian Markers).
16. SNPSTR database – база данни за SNP полиморфизми.
17. TRANSFAC – база данни за мотиви на транскрипционни фактори.
18. TreeBASE – филогенетична база данни.
19. Treefam TreeFam (Tree families database) – филогенетична база данни.