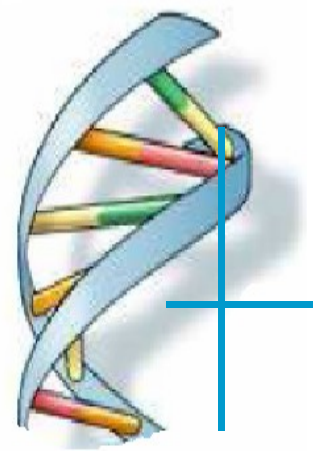


Бази данни в биоинформатика

гл. ас. д-р Веска Ганчева
vgan@tu-sofia.bg

Human Genome Project (HGP) 2003

~25, 000 genes in human DNA

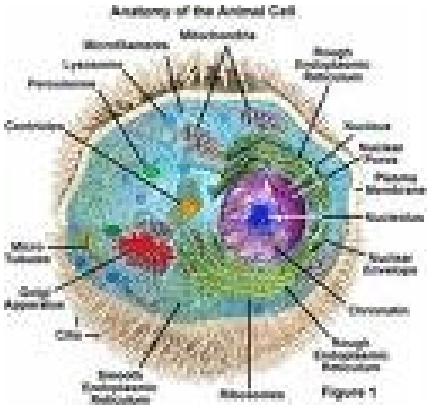
~3 billion base pairs

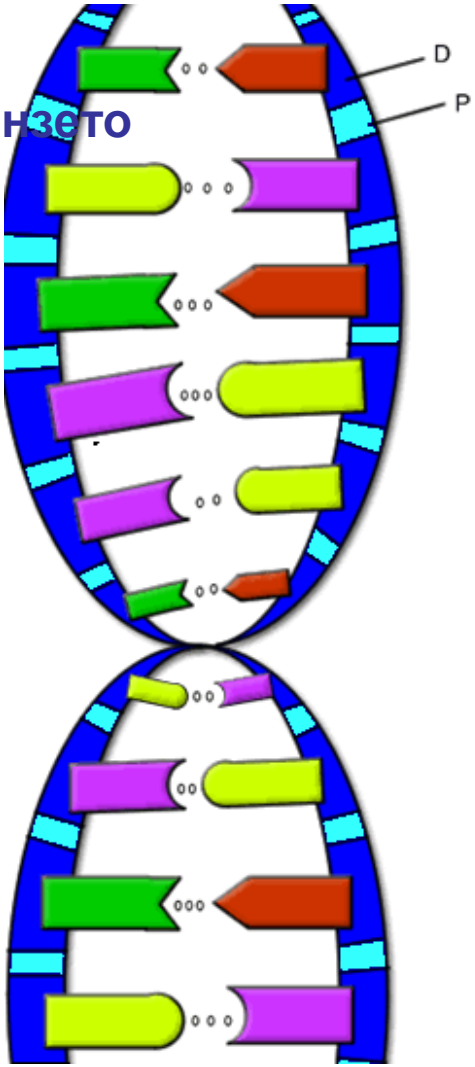
~9 GB

3% са кодиращи, 97% - празна?

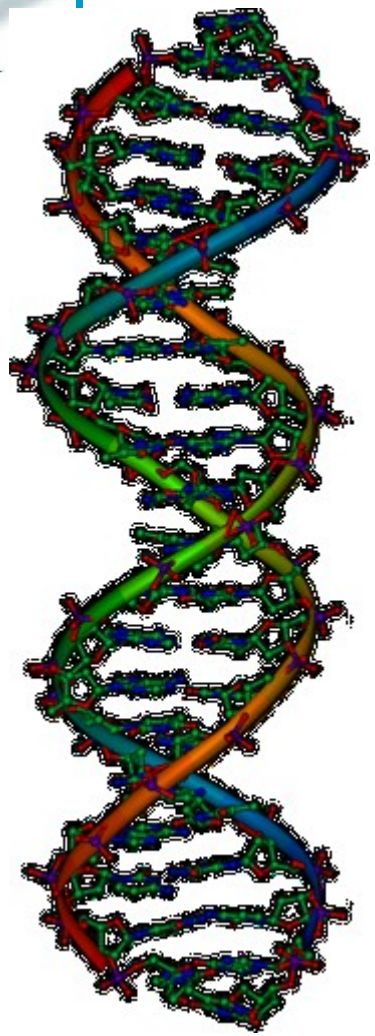
98% съвпадение с геномите на шимпанзето

28 алгоритъма

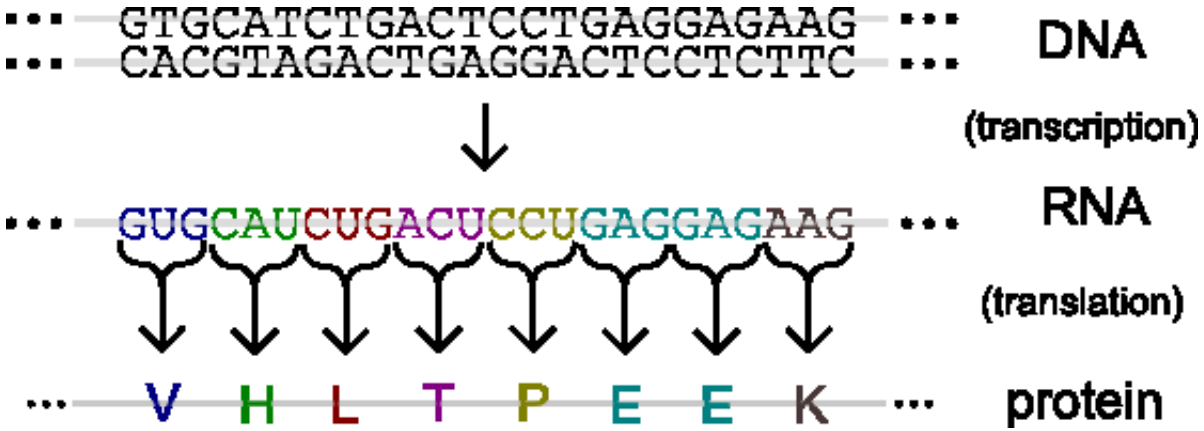




ДНК - носител на генетичната информация



Двойноспирална верига на ДНК. Съставена е от нуклеотидни бази, които се свързват комплементарно една с друга. Последователността при нуклеотидите определя последователността при аминокиселините изграждащи белтъците.



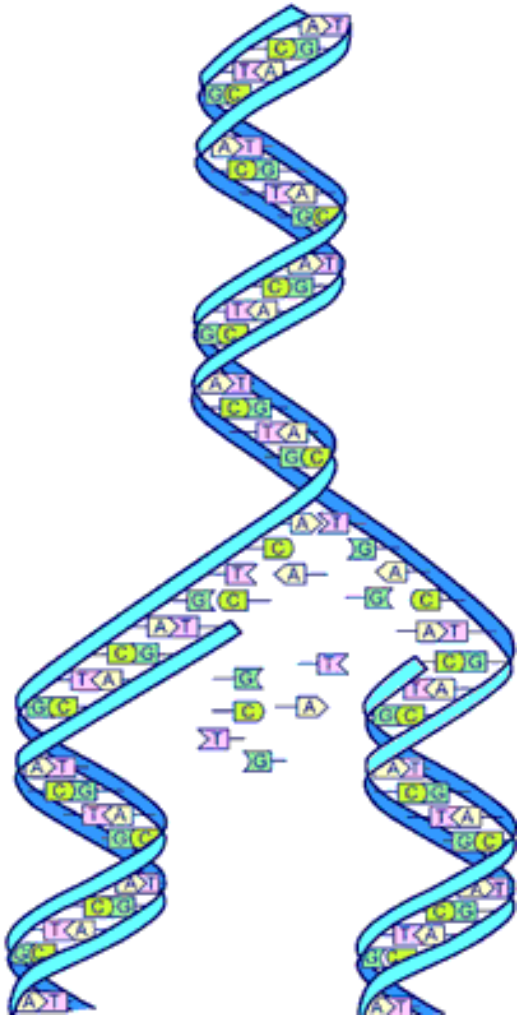
ДНК

- Една верига ДНК съдържа гени - области, които **регулират гените** – около 3% и други области, които нямат функции или функциите им все още не са познати – 97%;
- ДНК се състои от две нишки с връзки помежду им, които могат да се разделят;
- ДНК кодира генетичната информация благодарение на четири „строителни елемента“, наречени бази: аденин, тимин, гуанин, цитозин. Те се обозначават съкратено като A, T, G и C и имат свойството всяка да „се чифтосва“ само с една от останалите три бази: A+T, T+A, G+C, C+G;
- Редът е от значение: A+T не е еквивалентно на T+A, както и C+G не е едно и също с G+C.
- Тъй като са възможни само 4 комбинации, базите на едната от нишките са достатъчни, за да се опише последователността.
- Редът, в който са разположени базите по дължината на ДНК, е важен — ДНК-последователността (или секвенцията) е описанието на гените.

РНК - **посредник между ДНК и протеините**

- Т (тимин) е заменен от U (урацил)
- РНК транскрипция – синтез на РНК
- Една от главните функции на RNK е копирането на генетична информация от ДНК чрез транскрипция и превеждането и в белтъци чрез транслация
- 1.1 Рибозомна rRNK - синтеза на полипептидните вериги на белтъците
- 1.2 Транспортна tRNK - пренася аминокиселините до рибозомите, където се свързват чрез пептидни връзки в последователност, определена от mRNK
- 1.3 Информационна (матрична) mRNK - матрица (шаблон), по който се осъществява белтъчният синтез - NRA-транслация

Репликация



- Репликацията или синтез на ДНК се осъществява чрез разделяне на двете нишки и създаване на „втората половина“ на така получената единична верига чрез потапяне в „супа“, която съдържа всичките четири бази. Тъй като всяка база може да се комбинира само с една от останалите три бази, подредбата на базите в съществуващата верига определят еднозначно какви бази ще има в новообразуваната верига и как ще са подредени. По този начин, всяка единична верига образува точно копие на оригиналната ДНК, като събира необходимите бази в „супата“, освен ако не настъпи мутация.
- Мутациите са химически грешки в този процес: една база може случайно да бъде пропусната, вмъкната или грешно копирана или пък веригата може да бъде скъсена или удължена; всички други основни мутации могат да се опишат като комбинация от тези случайни „операции“.

Генетичен код

- Последователността от нуклеотиди по една ДНК-верига в един ген дефинира един белтък, който организмът трябва да произведе, като използва информацията, съдържаща се в последователността.
- Отношението между нуклеотидната последователност и тази от **аминокиселини** в белтъка се определя от прости клетъчни правила на транслация, познати под името генетичен код.
- Генетичният код се състои от трибуквени думи (***кодони***), образувани от последователност от три нуклеотида (напр. ACT, CAG, TTT). Тези кодони след това могат да се транслират посредством информационна РНК и впоследствие транспортна РНК до кодон, съответстващ на определена аминокиселина.

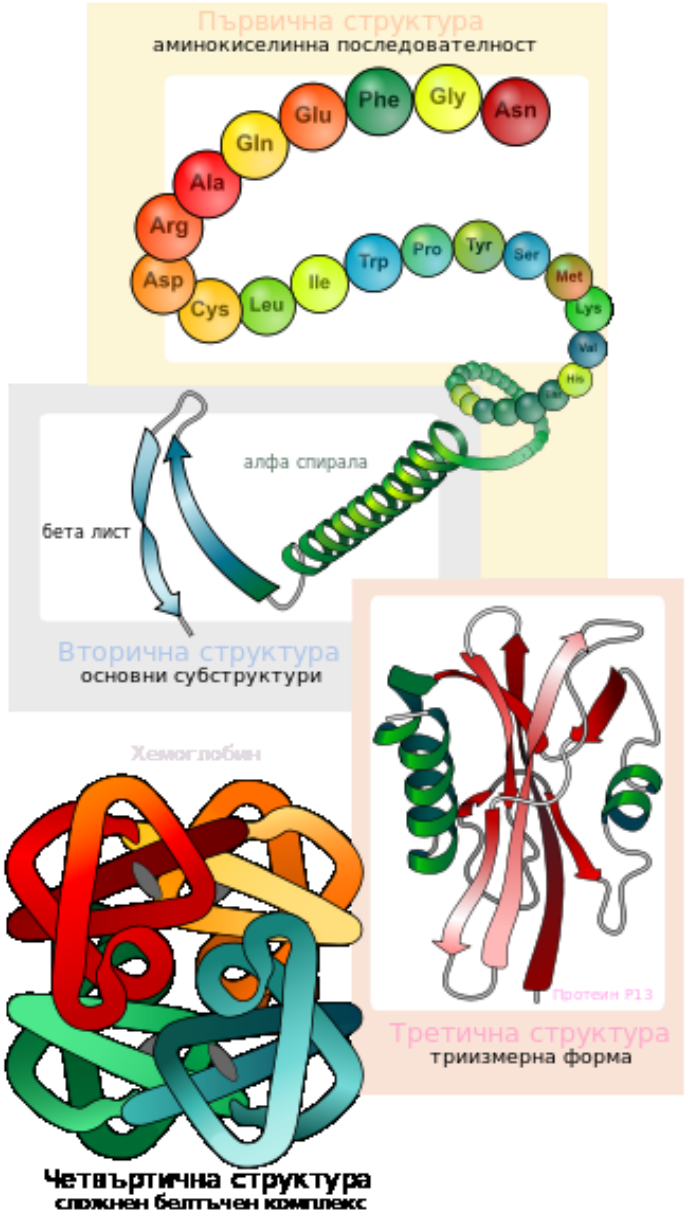
Аминокиселини

- 20 познати аминокиселини, описани с 20 букви
- Поредицата от аминокиселини изграждаща един протеин наричаме протеинова (полипептидна) секвенция.

Протеини

- Показателно за протеините е, че те имат добре оформена пространствена структура.
- Разпънатата полипептидна група няма биологична активност и функцията на протеина идва от формата му, която е пространственото разположение на молекулите в протеина. Формите и структурите на протеините се разделят в четири основни нива.
  - Първична (primary) структура - последователността от аминокиселини представена от секвенцията
  - Вторична (secondary) структура - количеството от последователности, които полипептидната верига приема. Има два основни типа вторична структура в протеините: **a-helix** и **ß-sheet**.
  - Третична (tertiary) структура - основните третични форми - кълбовидни форми, дълги влакна
  - Квадратична (quaternary) структура - връзките между 2 и повече полипептидни вериги, които се свързват в една обща молекула

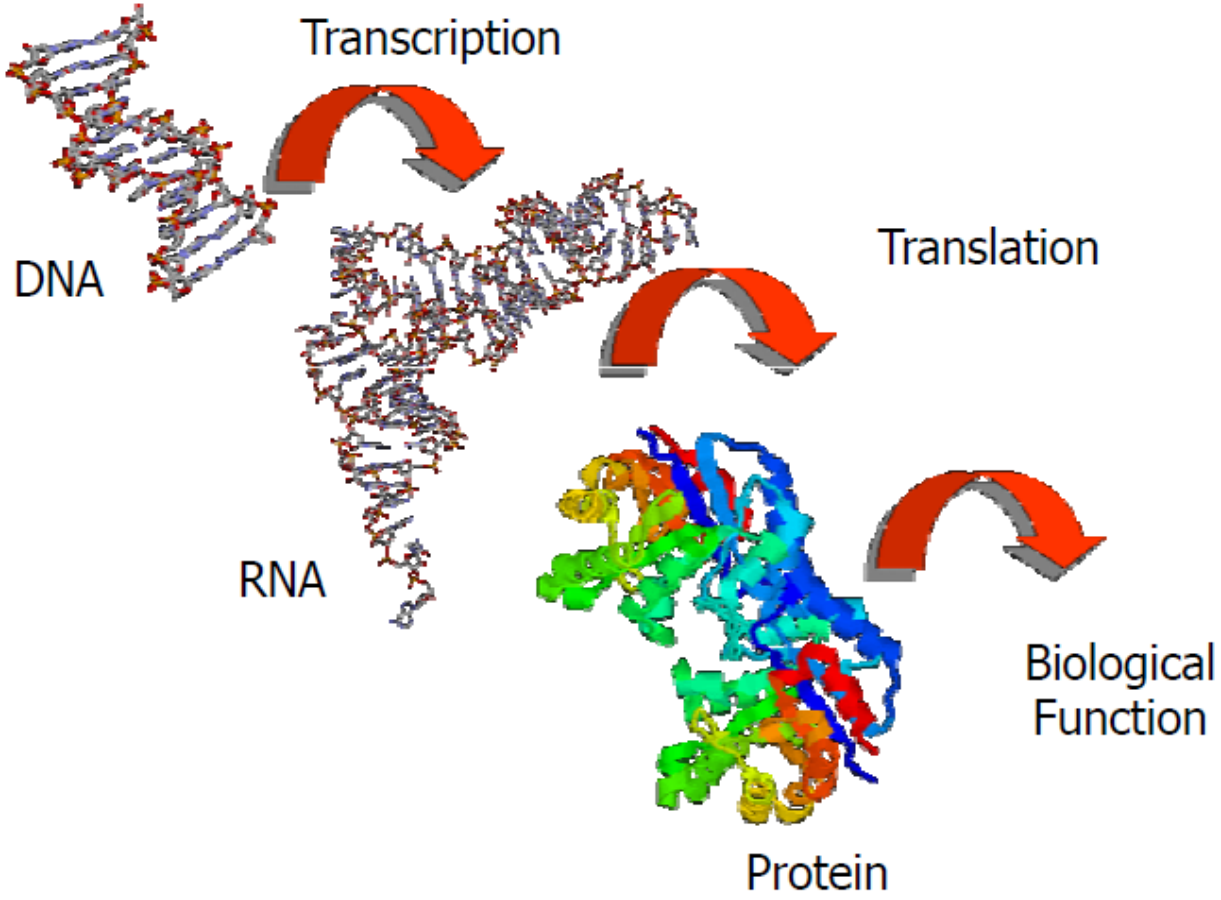
Протеини


Първична структура
аминокиселинна последователност
Phe
Gly
Asn
Glu
Gln
Ala
Arg
Asp
Cys
Leu
Ile
Trp
Pro
Tyr
Ser
Met
Lys
Val
His
алфа спирала
бета лист
Вторична структура
основни субструктури
Протеин P13
Третична структура
триизмерна форма
Хемоглобин
Четвъртична структура
сложнен белтъчен комплекс

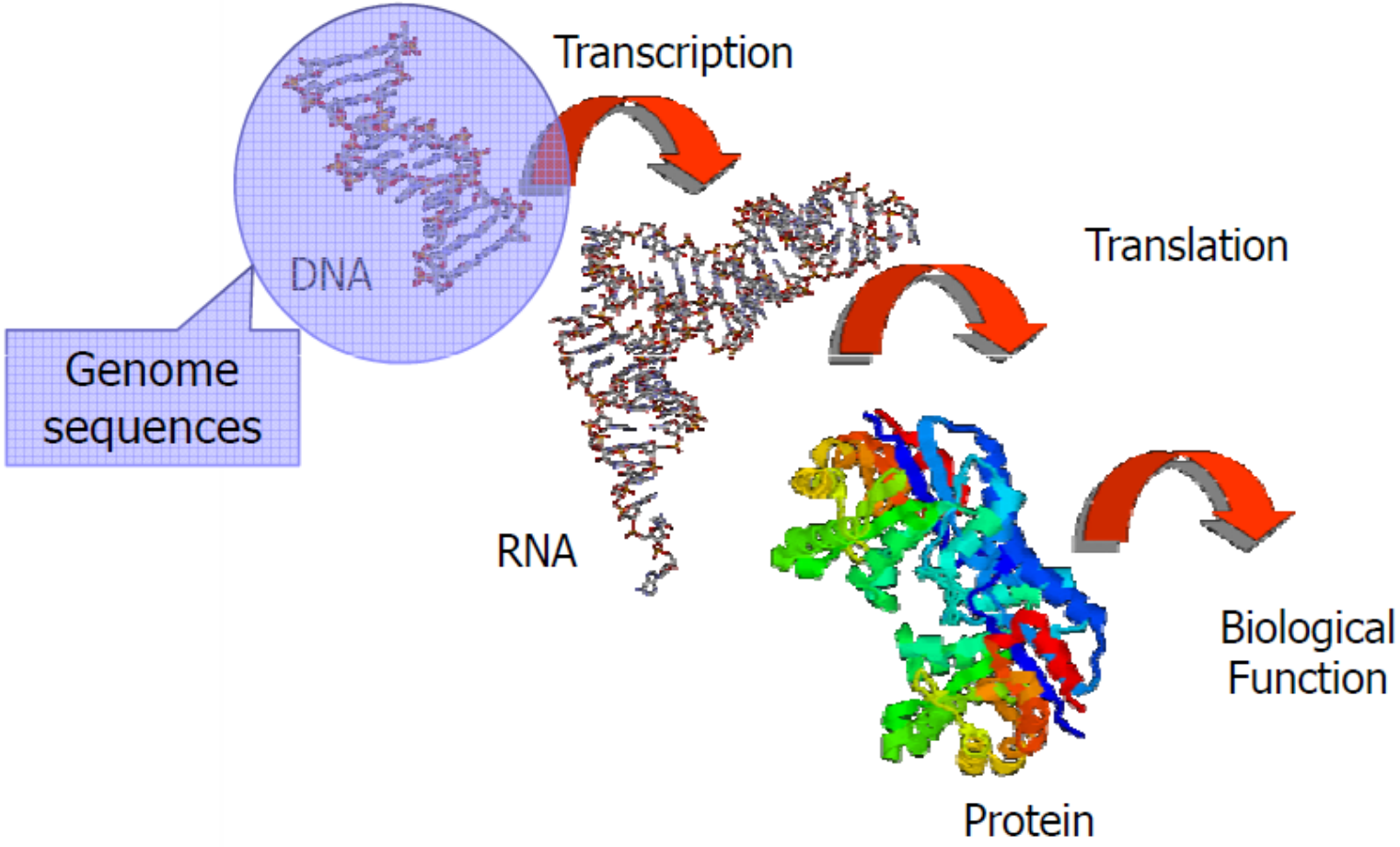
Протеини

- Една от основните причини за изследване структурата на протеините е изясняването на молекулярните патологии т.е. дефектни протеини които причиняват болести. Например Sickle Cell Anemia (анемия) се причинява от смяната само на една аминокиселина с друга в кръвните клетки.

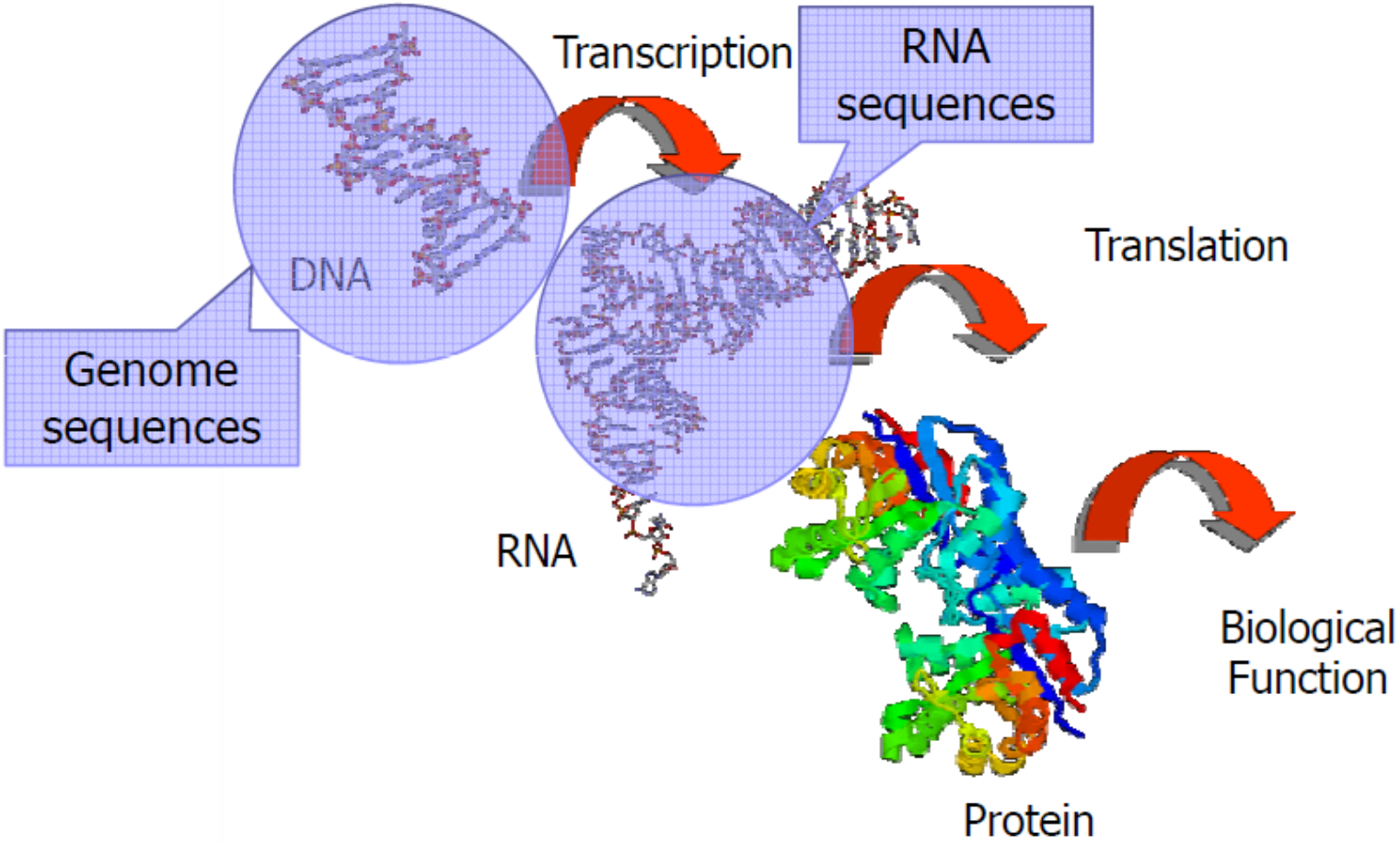
Biological Data



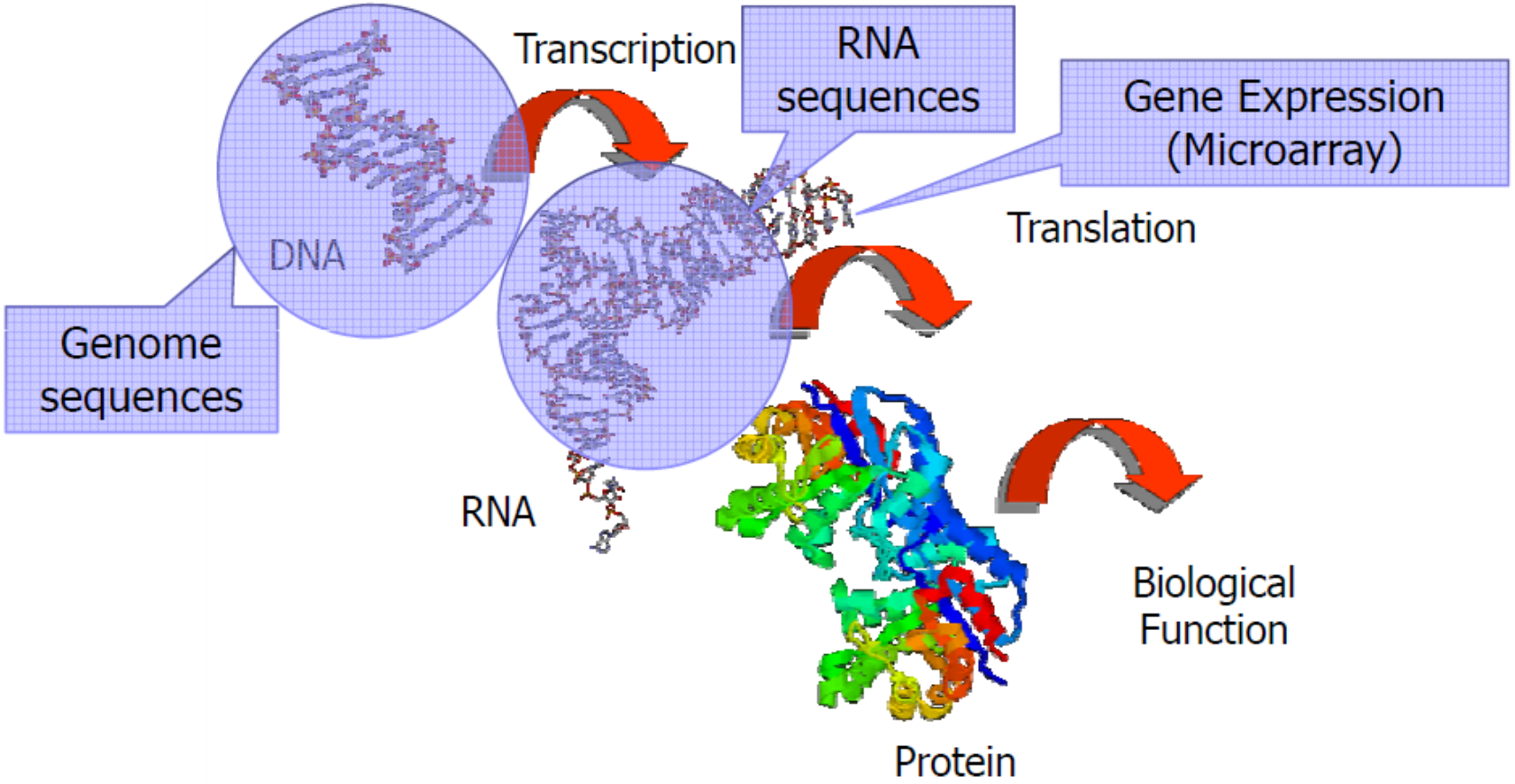
Biological Data



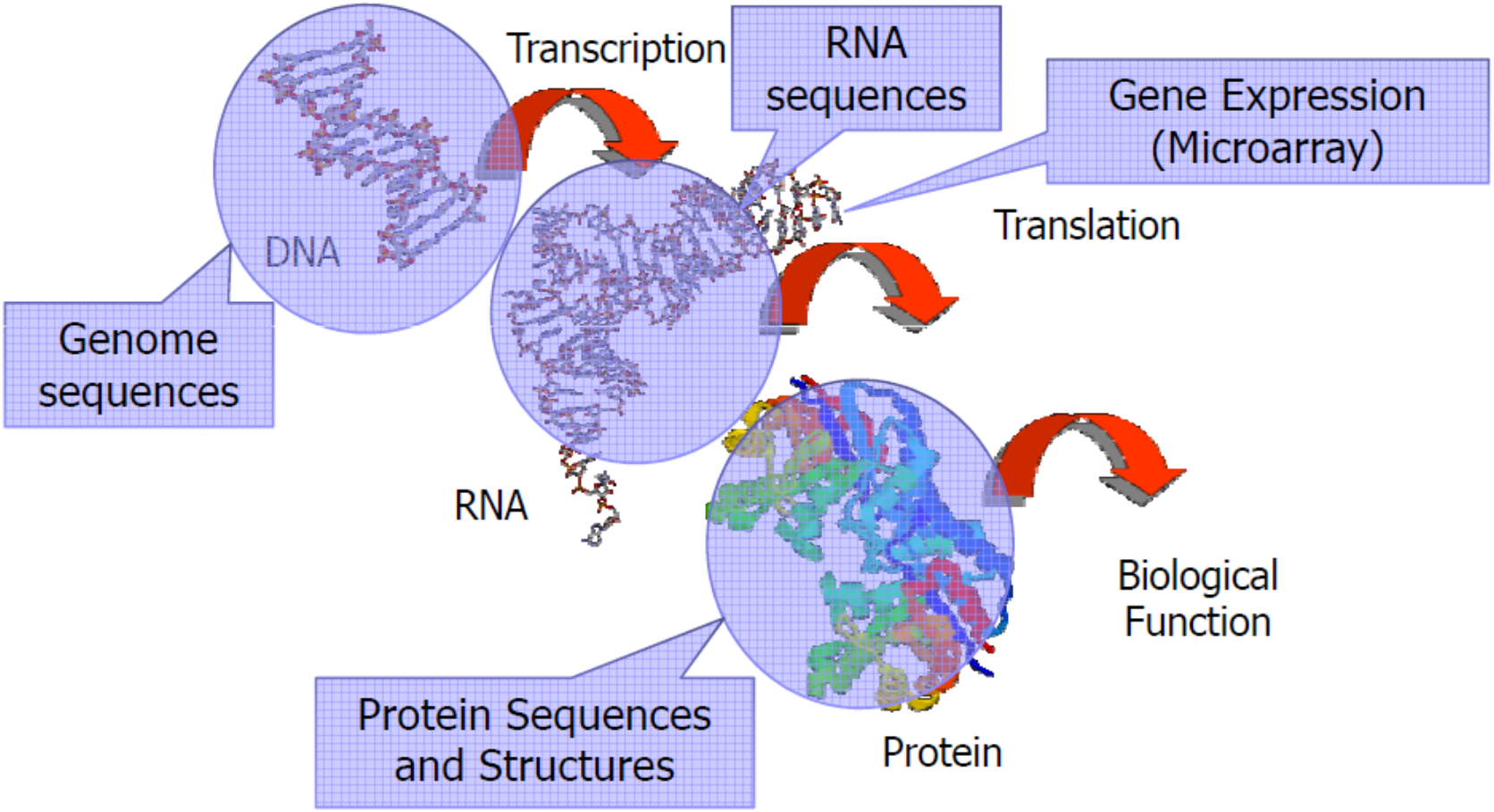
Biological Data



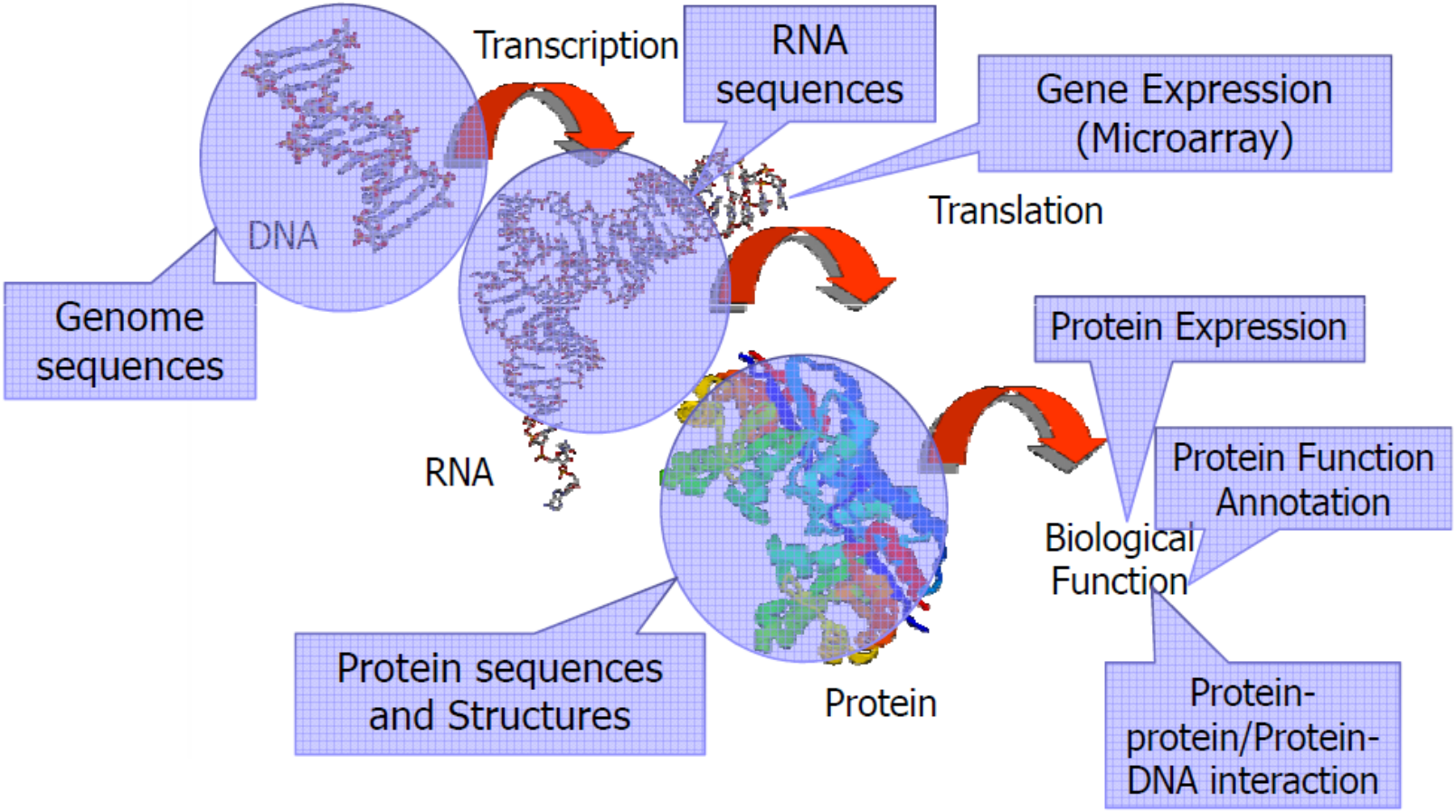
Biological Data



Biological Data



Biological Data



Еволюцията информира за всичко в биологията

- Големи проекти за секвениране на генома предоставят данни - какво от това?
- Еволюционната история свързва всички организми и гени и ни помага да ги разберем и прогнозираме
  - взаимодействия между гените (генетични мрежи)
  - дизайн на лекарства
  - предсказване функциите на протеините
  - разработка на противогрипна ваксина
  - произход и разпространение на болести
  - произход и миграции на хора

Какво е Биоинформатика?

Интегрира развитието на информационните и компютърни технологии, които се прилагат към биотехнологиите и биологичните науки.

Използва инструменти за създаване на биологична база данни, управление, съхранение, извличане на данни в глобални комуникационни мрежи.

Запис, анотация, съхранение, анализ и търсене / извличане на нуклеотидни последователности (гени и RNAs), протеинови последователност и структурна информация.

Базите данни на последователности и структурна информация, както и методи за достъп, търсене, визуализиране и извличане на информацията.

Под-дисциплини в биоинформатиката

Три важни под-дисциплини в рамките на биоинформатиката, включващи изчислителна биология:

- Разработване на нови алгоритми и статистически данни, с които да се оценят отношенията между биологичните данни.
- Анализ и интерпретация на различни типове данни, включително и нуклеотидни последователности, аминокиселини, протеини и структури на протеини.
- Разработване и прилагане на инструменти, които позволяват ефективен достъп и управление на различни видове информация.

Дейности в биоинформатиката

Две области: организация и анализ на биологичните данни

Организация в Биоинформатика

- Създаване на базите данни от биологична информация.
- Поддържането на тези бази данни.

Анализ на биологичните данни

- Разработване на методи за прогнозиране на структурата и / или функцията на новооткритите протеини и структурни последователности на РНК.
- Клъстериране на протеинови последователности и развитие на протеинови модели.
- Сравняване на подобни протеини и генериране на филогенетични дървета с цел проучване на еволюционни отношения.

Цели на Биоинформатика

Целите на биоинформатиката са основно три:

- Организация на данни по такъв начин, че да позволява на изследователите да получат достъп до съществуващи информация и да представят нови вписвания.
- Да се разработят инструменти и ресурси, които помагат за анализа на данните. Развитие на тези ресурси и обширни познания на компютърната теория, както и задълбоченото познаване на биологията.
- Използването на тези инструменти, за да се анализират отделните системи в подробности и сравнение с някои, които са свързани.

Три нива на биоинформатиката

- Анализ на един ген (протеин). Например:
  - Сходство с други известни гени
  - Филогенетични дървета; еволюционни отношения
  - Идентификация на точно определени области в редицата
  - Поредни особености (физически свойства, свързване, промяна сайтове)
  - Прогнози на вторична и третична структура
- Анализ на пълни геноми. Например:
  - Кои генни семейства са налице, кое липсва?
  - Местоположение на гените в хромозомите, корелация с функция или еволюция
  - Разширяване / дублиране на генни семейства
  - Наличието или липсата на биохимични пътища
  - Идентификация на "липсващите" ензими
  - Големи събития в еволюцията на организмите
- Анализ на гени и геноми, по отношение на функционалните данни.
  - Expression Analysis; Microarray данни; mRNA
  - Протеомика; протеинови измервания, ковалентни промени
  - Сравнение и анализ на биохимичните пътища
  - Идентифициране на основните гени или гените, свързани с конкретни процеси

Компютърна биология

- Прилагането на основните технологии на компютърните науки (напр. алгоритми, изкуствения интелект, бази данни и т.н.) за проблемите на биологията. Компютърната биология е особено вълнуваща днес, защото проблемите са достатъчно големи, за да се мотивират ефективни алгоритми и търсенето на постижения на биологията и на компютърната наука се увеличава.
- Компютърната биология включва:
  - Намирането на гени в ДНК последователности на различни организми.
  - Разработване на методи за прогнозиране на структурата и / или функцията на новооткритите протеини и структурни последователност на РНК.
  - Клъстеризация на протеинови последователности в семействата, свързване на последователности и развитие на протеинови модели.

Сравнение

- ***Биоинформатика:*** Проучване, разработване или прилагане на компютърни подходи и средства за разширяване използването на биологични, медицински, поведенчески или здравни данни, включително и тези, които се придобиват, съхраняват, организират, архивират, анализират или визуализират.
- ***Компютърна биология:*** Разработване и прилагане на данни и теоретични аналитични методи, математическо моделиране и компютърна симулация за изследването на биологични, поведенчески и социални системи.

Биологични данни

- Учените в днешно време са зависими от базите данни и достъпа до лавина от информация, която произвеждат. Доставчиците на данни полагат огромни усилия в предоставянето на ресурси данни, които са изчерпателни, лесни за ползване, както и свързани към други бази данни, но различните доставчици на данни използват различни методи. Това означава, че един изследовател може трябва да търси в 10 или повече различни бази данни, за да намери цялата информация, свързана с определен набор от кандидат гени. Ако те правят търсене на регулярна основа, ще искат локални копия на всички тези бази данни.
- Поддържането на актуална база данни и напълно функциониращи версии на всички тези бази данни и инструментите за търсене в тях е огромна и сложна задача.

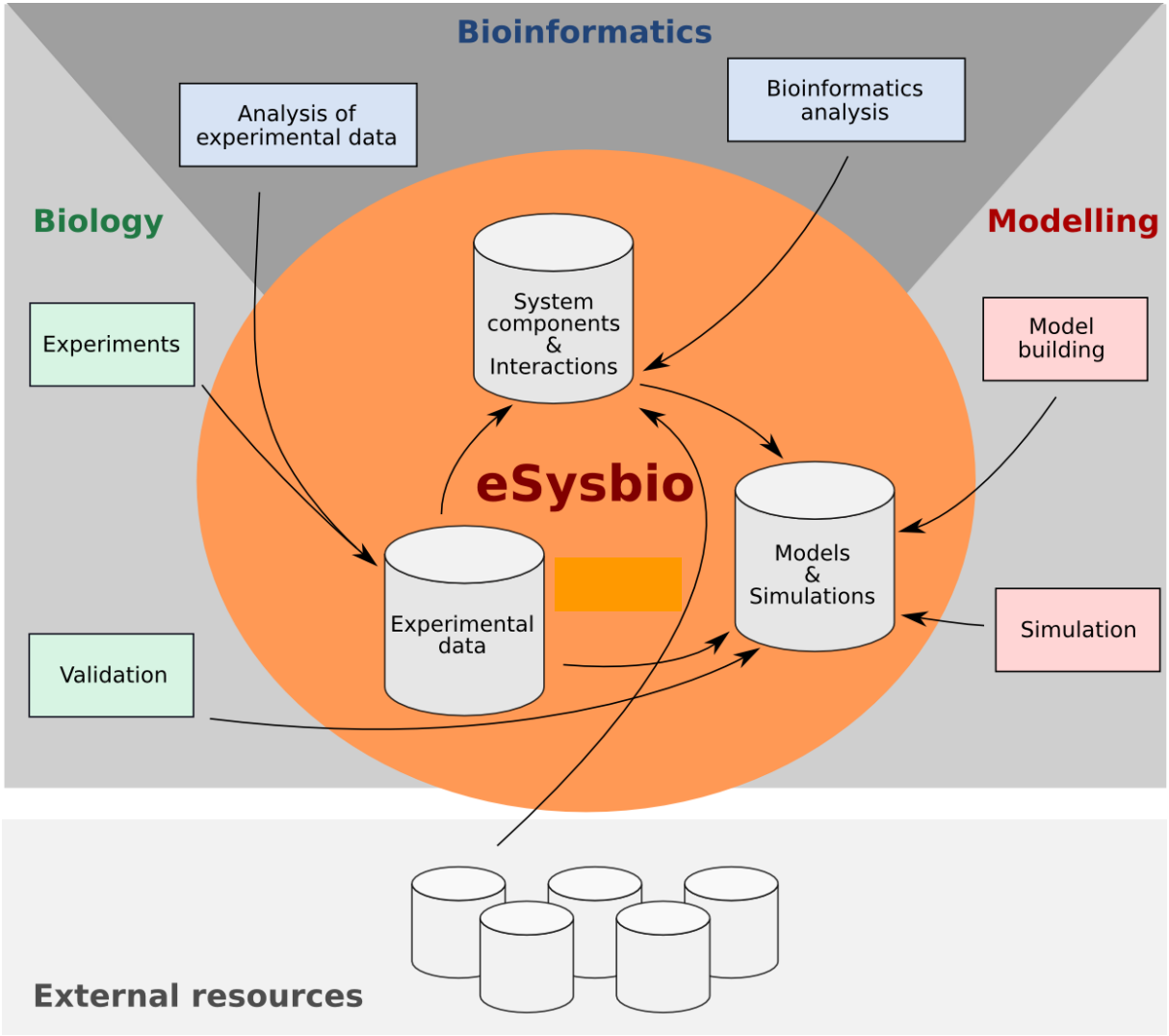
Биологични данни

- Следователно е необходимо да се предложи на биолози достъп до актуална данни на молекулярната биология, свързани с техните изследвания. Това изисква интегриране на големи различни бази данни, пълни с информация, свързани с различните нива на изразяване на живите системи. В условията на висока производителност при производство на системи данни, базите данни са прогресивно нарастващи.
- В обобщение, основното предизвикателство за анализ на данни в науките за живота е да предложи на молекулярните биолози интегрирани и up-to-date view на прогресивно нарастващият обем данни в множество формати.

Билогични данни

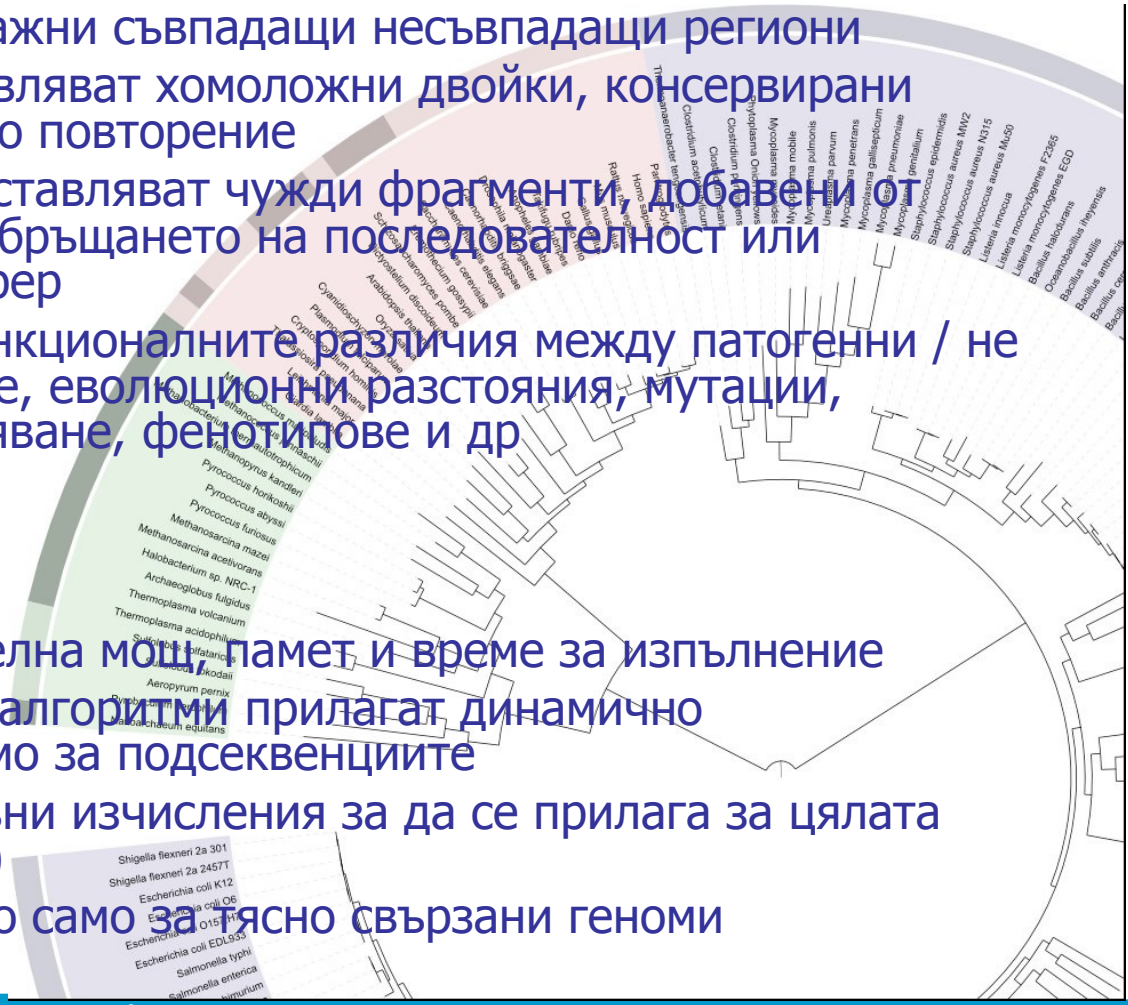
- Биоинформатиката цели разработване на програмни инструменти за подпомагане на биолозите в анализа на техните данни.
- Изследванията в биоинформатиката са свързани с нови алгоритми и нови услуги за повишаване на производителността.
- За 20 години изследвания в биоинформатиката се създадоха нови инструменти, софтуер, бази данни и портали.
- Голямото предизвикателство, специфично за молекулярна биология е, че изследователите трябва да имат постоянен достъп до състоянието на изследванията, за да сравнят резултатите си с наличната информация.
- Молекулярните биолозите се нуждаят от достъп до постоянно актуализирано представителство на всички натрупани знания в своята област. Това изискване не се отнася за други области от науките като физика или химия, където развитието на знанията е много по-бавно.

Биоинформатика



Мотивация

- Значението на сравнение на геноми:
  - Определяне на важни съвпадащи несъвпадащи региони
  - "matches" представляват хомоложни двойки, консервирани региони или дълго повторение
  - "mismatches" представляват чужди фрагменти, добавени от транспониране, обръщането на последователност или страничен трансфер
  - Откриване на функционалните различия между патогенни / не патогенни щамове, еволюционни разстояния, мутации, водещи до заболяване, фенотипове и др
- Проблеми
  - Голяма изчислителна мощ, памет и време за изпълнение
  - Съществуващите алгоритми прилагат динамично програмиране само за подсеквенциите
  - Изисква интензивни изчисления за да се прилага за цялата поредица (O (N^2))
  - Така е приложимо само за тясно свързани геноми

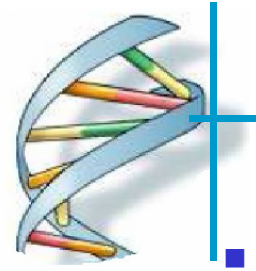


Търсене на секвенции

- Сравнението на секвенции е най-добрият метод за изучаване на еволюционното взаимодействие между гените
- Търсенето е базирано на подравняване - процес на подреждане на две или повече последователности за постигане на максимално ниво на идентичност за целите на оценка на степента на сходство и възможността за хомология.

Търсене на секвенции

- **Паралелно търсене на секвенции**
  - Защо паралелно търсене**?**
  - Търсене: Все по-голяма интензивност на изчисленията
  - Бази данни: Експоненциално нарастващи
  - Нови компютри: Multiple cores/processors
- **Търсенето е независимо**
  - Използване на паралелни библиотеки и езици за разделяне на търсенето в независими задачи
  - Изпълнение на всички задачи в паралел на големи суперкомпютри
- **Време за търсене**
  - 2 дни (serial search) -> 8 минути (parallel search)
- **Повече алгоритми: *Scatter-Search-Gather***



Структура на протеин

- Класическия проблем за разкриване на огъването на протеините може да се дефинира така: ***да се намери структурата на молекулата на протеина в пространството, като е известна само неговата секвенция***.
- Изчислено е, че може да достигне до 8100 различни състояния за произволен средно голям протеин. При наличието на такива огромни по обем данни, тяхното изследване става невъзможно без използване на компютри и евристични алгоритми.
- Сравнително (хомоложно) моделиране предполага, че протеини, които имат сходни секвенции имат и сходни структури.
- Предполага се, че много различни секвенции се извиват по сходни начини и има относително голяма вероятност нова секвенция да следва вече наблюдаван начин на извиване.
- За определяне на 3D структура на протеин често се използва и неговата вторична структура, която дава информация за общото пространствено разположение на протеина.

Real-World приложения

- Био-онтология и извличане на данни
- Визуализация на данните
- ДНК асемблиране, клъстеризация и картиране
- Молекулярна еволюция и филогенетика
- Генна експресия и micro-arrays
- Молекулно моделиране и симулация
- Търсене и подравняване на секвенции
- Предсказване на структурата на протеини

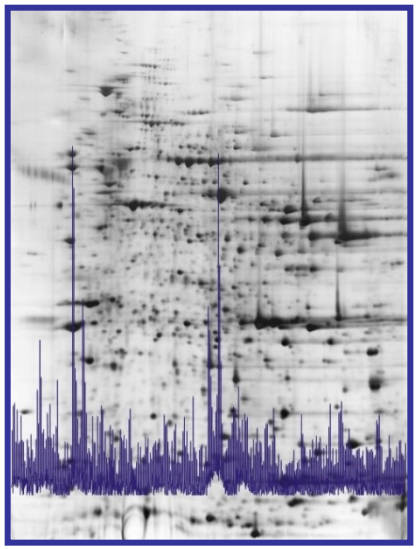
Входни данни за биоинформатика

- Бързо развиващите се биологични изследвания доведоха до огромно количество данни през последните години. Разпределен, динамичен и разнороден характер на тези данни прави обединението им труден проблем. Технологии, които се използват за обединение на данни за биоинформатиката:
  (1) Контролирани речници, които регулират анотациите
  (2) RDF подобни стандарти, които могат еднозначно да идентифицират един обект в уеб
  (4) Децентрализирани технологии, които се занимават с разпределени данни и техники обработка на хетерогенни структури метаданни в децентрализирана среда.

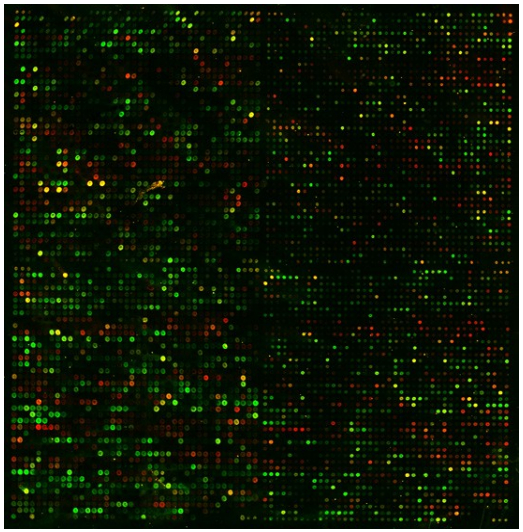
Входни данни за биоинформатика

- Приложения за Биоинформатика:
- Най-големият проблем при изпълнението на приложенията за биоинформатика е достъпа до данни
  - Плоски файлове DB
  - RDBMS
- Проблеми с плоските файлове:
  - Обикновено приложението е написано с предположение за локален достъп до данни
  - Често не е лесно да се промени код или не е наличен
  - Някои сървъри не разполагат с достатъчно дисково пространство за файлове за вход / изход на някои приложения
  - Често решението чрез споделена мрежова файлова система не е възможно
  - Може да има проблеми с производителността или локалната конфигурация

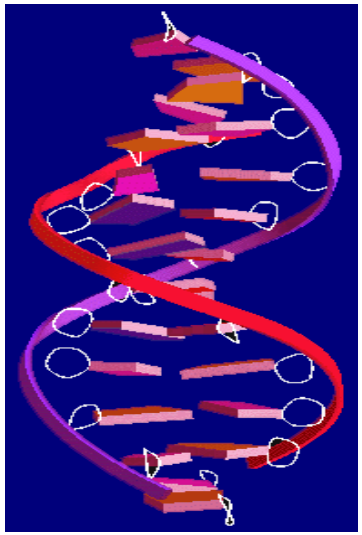
Интеграция на данните



Proteins
(Proteomics)



Microarray
(Trascriptomics)

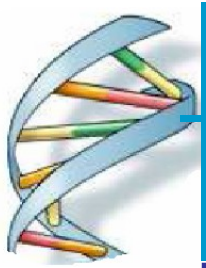


Gene & SNPs
(Genomics)

Инструменти за Биоинформатика

Стандартни и потребителски продукти, които отговарят на изискванията на конкретни проекти. Това са data-mining, които извличат данни от бази данни геномни последователности и визуализиращи средства за анализ и изтегляне на информация от протеомни бази данни. Те могат да бъдат класифицирани като инструменти за търсене на homology и прилика, инструменти за функционален анализ на протеин, инструменти за анализ на последователности и други инструменти.

Основни инструменти: програми като BLAST за търсене на последователности, програми за анализ на последователности като EMBOSS и Staden пакети, предсказване структурата на протеин като THREADER или PHD или програми за молекулярно моделиране като RasMol и WHATIF .



Инструменти за Биоинформатика

- **Homology Инструменти:**

Хомоложни последователности са последователности, които са свързани с отклонения от един общ прародител. Така степента на сходство между две поредици може да се измери. Този набор от инструменти могат да бъдат използвани за идентифициране на прилики между нова заявка за последователности от неизвестна структура и функция и поредици база данни, чиято структура и функция са изяснени.

- **Анализ на функцията на протеин:**

Тази група от програми позволяват да се сравнят протеинови последователности с вторична (или производна) протеинова бази данни, която съдържа информация за мотивите, подписи и протеинови домейни.

Инструменти за Биоинформатика

- **Структурен анализ:**

Този набор от инструменти позволяват да се сравни структура с известна структура от бази данни. Функцията на протеин е по-пряко следствие от неговата структура. Определянето на 2D/3D структура протеин е от решаващо значение при изучаването на неговата функция.

- **Анализ на последователности (секвенции):**

Този набор от инструменти позволява да се проведат нови, по-подробни анализи на последователност, включително еволюционен анализ, идентификация на мутации, региони и отклонения от състава.

Инструменти за Биоинформатика

- **Blast:**
  Blast **(Б** ASIC **М** ЕСТНИ **А** lignment **S** earch OOL **Т)** попада в категорията на инструментите за homology и прилика. Това е набор от програми, предназначени за търсене и се използва за бързо търсене на прилика, независимо от това дали заявката е за протеин или ДНК. Може да се извърши сравнение на нуклеотидните последователности в база данни.

- **FASTA:**
  FAST homology **търсене в** последователности. Използва евристичен алгоритъм за увеличаване на скоростта на сравненията. Основната идея е добавяне на бърза prescreen стъпка, за намиране високо съвпадение на сегменти между две поредици, а след това съвпадение се разширява обхвата на тези сегменти на местните проекция с използване на по-строг алгоритъм като Смит-Уотърман.

- **EMBOSS:**
  EMBOSS **(E**uropean **M**olecular **B**iology **O**pen **S**oftware **S**uite) е софтуерен пакет за анализ. Може да работи с данни в набор от различни формати, а също и с прозрачни данни от Интернет.

Инструменти за Биоинформатика

- **Clustalw:**
  Напълно автоматизиран инструмент за привеждане в съответствие на последователност на ДНК секвенции и протеини.

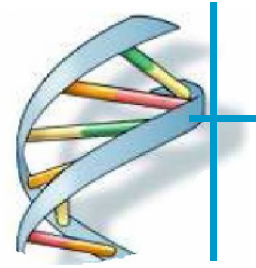
- **RasMol:**
  Мощен инструмент за търсене за показване на структурата на ДНК, протеини, и по-малки молекули. Protein Explorer, производна на RasMol, е лесна за използване.

- **PROSPECT:**
  PROtein Structure Prediction and Evaluation Computer ToolKit е система за предсказване на протеиновата структура, използва изчислителна техника - protein threading за изграждане на 3-D протеинов модел.

Инструменти за Биоинформатика

- **PatternHunter:**
  PatternHunter, базиран на Java, може да идентифицира всички приблизителни повторения в пълния геном в кратко време, използвайки малко памет на настолен компютър. Патентован алгоритъм и структури от данни, написан на Java. Java версията на PatternHunter е само 40 KB, само 1% от размера на Blast, като същевременно предлага голяма част от неговата функционалност.

- **COPIA:**
  COPIA (COnsensus Pattern Identification and Analysis) е инструмент за структурен анализ на протеин за откриване на мотиви (консервирани региони) в семейство на протеинови последователности. Такива мотиви могат да се използват след това за да се определи членството на семейството за нови поредици протеини, прогнозиране вторична и третична структура и функция на протеини и история изследване еволюцията на поредицата.



Средства за паралелна обработка

- ClustalW-MPI
- BLAST
- Smith-Waterman
- FASTA
- HMMER
- SSEARCH
- fastDNAml
- RAxML
- Parallel PhyloBuilder
- PhylTree
- PHYLIP
- PHYML
- MrBayes
- REMD

Езици за програмиране

- **JAVA в Биоинформатика:**
  Тъй като изследователските центрове са разпръснати по целия свят, вариращи от частни към академични, както и се използват набор от хардуерни и операционни системи, Java се очертава като основен език в биоинформатиката. Physiome Sciences' и PatternHunter са два примера на нарастващото приемане на Java в биоинформатиката.

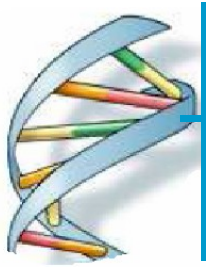
- **Perl в Биоинформатика:**
  Обработка на стрингове, съвпадение на регулярни изрази, анализирането на файл, данни и др. са в общия случай задачите, изпълнявани в биоинформатиката. Perl блесна с такива задачи и се използва от много фирми. И все пак, няма стандартни модули в Perl, предназначени специално за областта на биоинформатиката. Въпреки това, разработчиците са проектирали няколко отделни модули за целите, които са станали доста популярни и се координират от проекта BioPerl.

Езици за програмиране

- **BioJava:**
  В BioJava предоставя Java инструменти за обработка на биологични данни, включва обекти за манипулиране на последователности, динамично програмиране, файлови парсери, прости статистически обработки и др.
- **BioPerl:**
  BioPerl е международна асоциация на разработчиците на Perl инструменти за биоинформатиката и предоставя онлайн ресурс за модули, скриптове и интернет връзки за разработчиците на Perl-базиран софтуер.
- **BioXML:**
  Една част от проекта BioPerl;, това е ресурс, за да се съберат XML документи, DTDs и XML инструменти за биологията на едно място.
- **Biocorba:**
  Интерфейс CORBA за bioperl. С biocorba обекти, написани на bioperl ще могат да комуникират с обекти, написани на biopython и biojava

Езици за програмиране

- **Ensembl:**
  Ensembl е автоматизирана-геном-анотация на EBI. Голяма част от Ensembl код е базиран на bioperl.
- **bioperl-DB:**
  Bioperl-DB е сравнително нов проект, предназначен да прехвърли част от способността Ensembl за интегриране bioperl синтаксис със самостоятелна база данни MySQL към bioperl код базирана.
- **Biopython и biojava:**
  Biopython и biojava са проекти с отворен код с много сходни цели на bioperl. Въпреки това, техният код се изпълнява в Python и Java. С развитието на интерфейса обекти и biocorba, е възможно да се пишат Java или Python обекти, които могат да бъдат достъпени от bioperl скрипт, или да се извикат bioperl обекти от Java или Python код.



Приложение на биоинформатиката

- Интерфейси за бази данни
  - **Genbank/EMBL/DDBJ, Medline, SwissProt, PDB, …**
- Подравняване на секвенции
  - **BLAST, FASTA**
- Множествено подравняване на секвенции
  - **Clustal, MultAlin, DiAlign**
- Откриване на гени
  - **Genscan, GenomeScan, GeneMark, GRAIL**
- Онализ и идентификация на протеини
  - **pfam, BLOCKS, ProDom**
- Идентификация на мотиви (шаблони)
  - **Gibbs Sampler, AlignACE, MEME**
- Предсказване стуктурата на протеини
  - **PredictProtein, SwissModeler**

Основни website

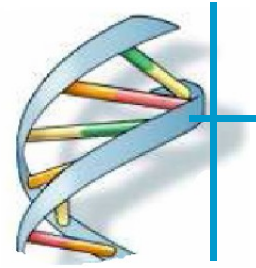
- NCBI (The National Center for Biotechnology Information;
  - **http://www.ncbi.nlm.nih.gov/**
- EBI (The European Bioinformatics Institute)
  - **http://www.ebi.ac.uk/**
- The Canadian Bioinformatics Resource
  - **http://www.cbr.nrc.ca/**
- SwissProt/ExPASy (Swiss Bioinformatics Resource)
  - **http://expasy.cbr.nrc.ca/sprot/**
- PDB (The Protein Databank)
  - **http://www.rcsb.org/PDB/**

NCBI (http://www.ncbi.nlm.nih.gov/)

- Entrez интерфейс за бази данни
  - Medline/OMIM
  - Genbank/Genpept/Structures
- BLAST сървър
- Draft Human Genome
- Much, much more…

EBI (http://www.ebi.ac.uk/)

- SRS интерфейс за бази данни
  - EMBL, SwissProt и др.
- Много сървърно базирани средства
  - ClustalW, DALI, …

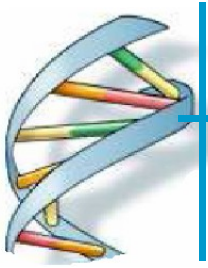


SwissProt (http://expasy.cbr.nrc.ca/sprot/)

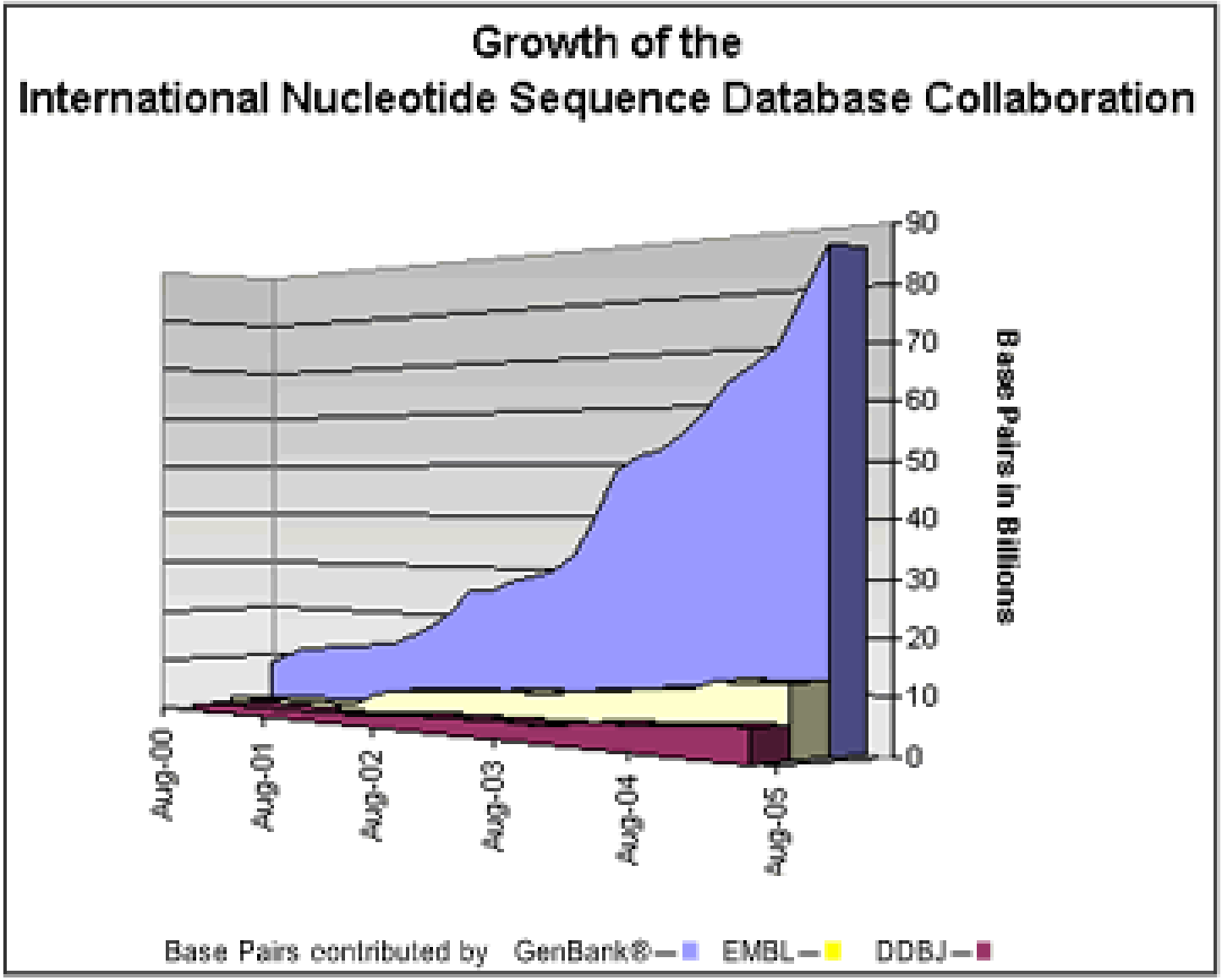
- Процентът на грешки в информацията е значително по-малък в сравнение с повечето други бази данни.
- Обширно кръстосано свързване към други източници на данни.
- SwissProt е "златен стандарт", спрямо който могат да бъдат измерении други бази данни и е най-доброто място, от което да се започне за изследване на специфичен протеин.

Други ресурси

- Human Genome Working Draft
  - **http://genome.ucsc.edu/**
- TIGR (The Institute for Genomics Research)
  - **http://www.tigr.org/**
- Celera
  - **http://www.celera.com/**
- (Model) Organism specific information:
  - **Yeast**: **http://genome-www.stanford.edu/Saccharomyces/**
  - **Arabidopis**: **http://www.tair.org/**
  - **Mouse**: **http://www.jax.org/**
  - **Fruitfly**: **http://www.fruitfly.org/**
  - **Nematode: http://www.wormbase.org/**
- Nucleic Acids Research Database Issue
  - **http://nar.oupjournals.org/ (First issue every year)**



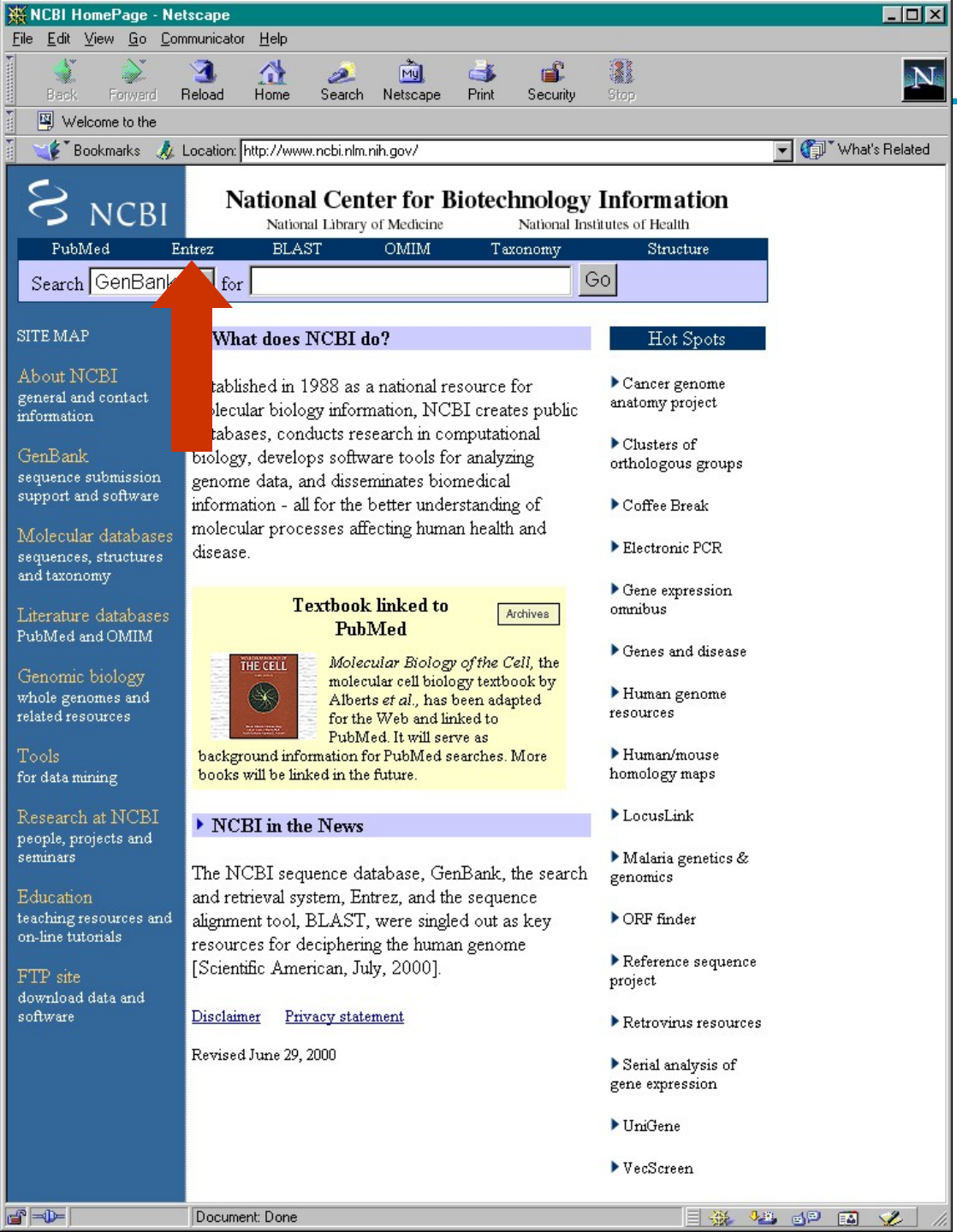
Growth of the International Nucleotide Sequence Database Collaboration



Какво е accession number?

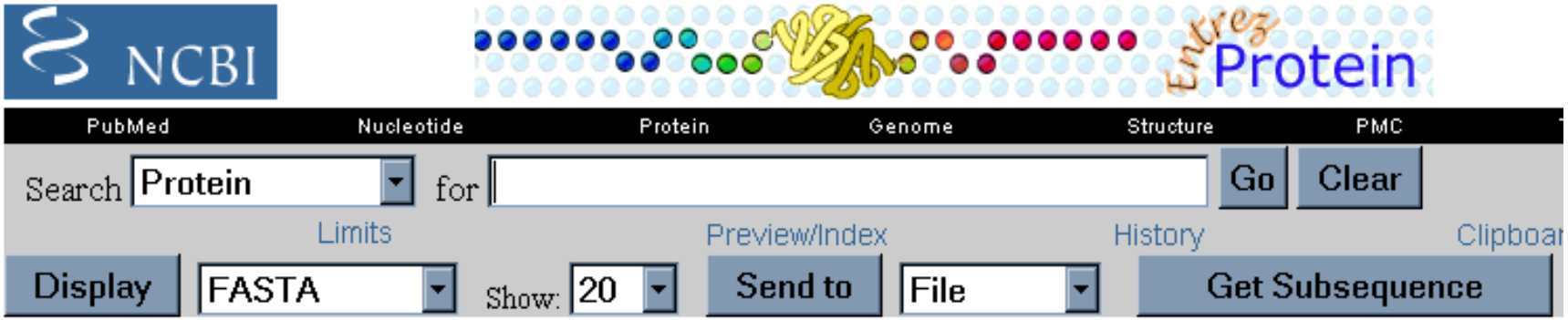
Етикет за идентифициране на секвенциите. Представлява стринг от букви и/или цифри, свързани с дадена молекулярна секвенция.
Пример (all for retinol-binding protein, RBP4):

X02775	**GenBank genomic DNA sequence**	
NT_030059	**Genomic contig**	
Rs7079946	**dbSNP (single nucleotide polymorphism)**	
N91759.1	**An expressed sequence tag (1 of 170)**	**DNA**
NM_006744	**RefSeq DNA sequence (from a transcript)**	
NP_007635	**RefSeq protein**	**RNA**
AAC02945	**GenBank protein**	
Q28369	**SwissProt protein**	
1KT7	**Protein Data Bank structure record**	**protein**



Използване на NCBI за търсене на информация за протеин или ген

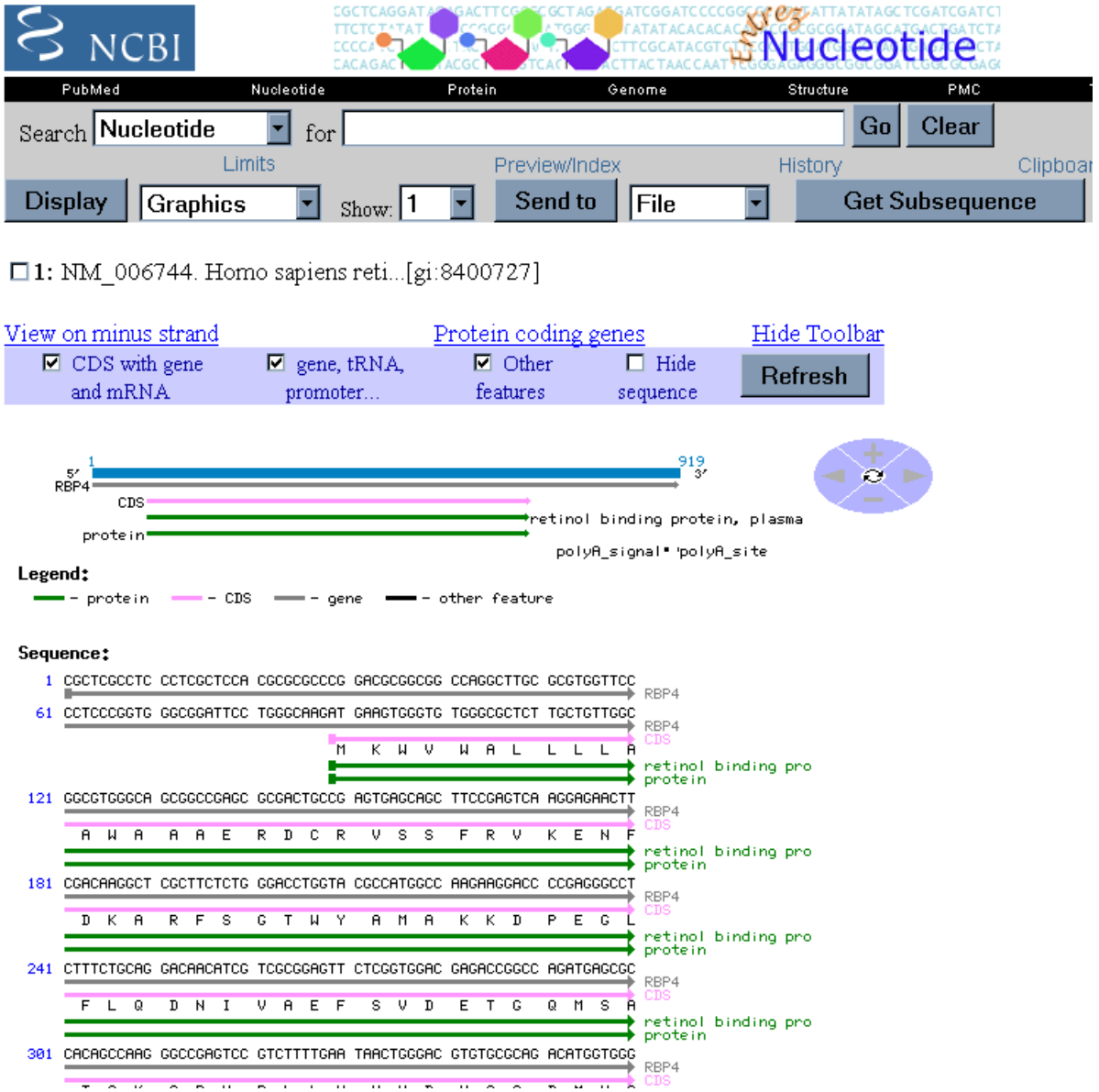
FASTA формат

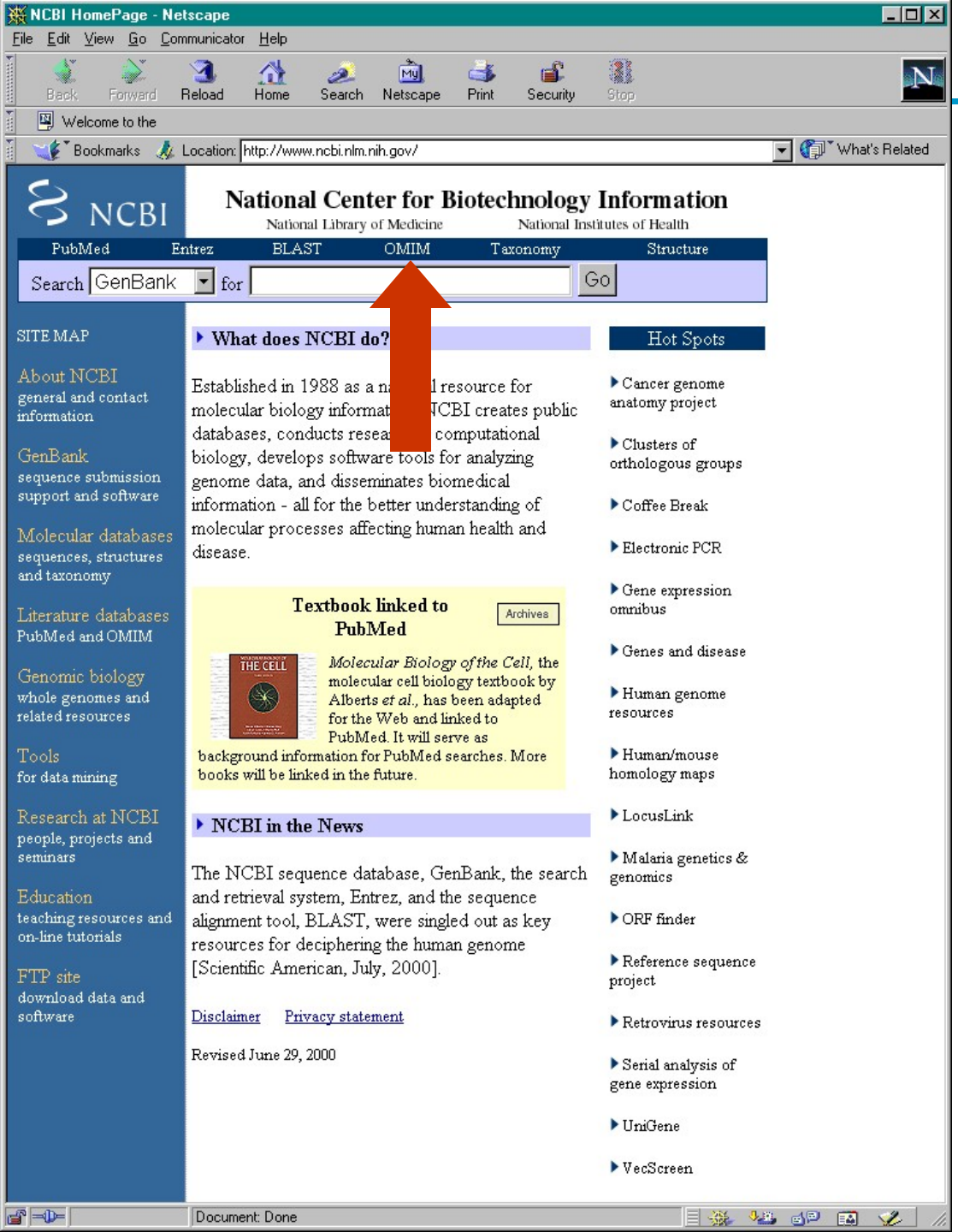


1: NP_006735. RBP4 gene product...[gi:5803139]

```
>gi|5803139|ref|NP_006735.1| RBP4 gene product [Homo sapiens]
MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMS
ATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLN
LDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL
```

Графичен format





Търсене на информация за заболявания: OMIM

NCBI

OMIM
Online Mendelian Inheritance in Man

Johns Hopkins University

PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | OMIM

Search OMIM for rbp Go Clear

Limits | Preview/Index | History | Clipboard | Details

Entrez

OMIM
Search OMIM
Search Gene Map
Search Morbid Map

Help
OMIM Help
How to Link

FAQ
Numbering System Symbols
How to Print
Citing OMIM
Download

OMIM Facts
Statistics
Update Log
Restrictions on Use

Allied Resources
Genetic Alliance
Databases
HGMD
Locus-Specific
Model Organisms
MitoMap
Phenotype

Display Titles Show: 20 Send to Text

Items 1-10 of 10 One page.

1: *180250 Links
RETINOL-BINDING PROTEIN 4; RBP4
RETINOL-BINDING PROTEIN, PLASMA
Gene map locus 10q24

2: *176300 GeneTests, Links
TRANSTHYRETIN; TTR
AMYLOIDOSIS I, INCLUDED
Gene map locus 18q11.2-q12.1

3: *607201 Links
HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN R; HNRPR
Gene map locus 1p36.11

4: *600354 GeneTests, Links
SURVIVAL OF MOTOR NEURON 1, TELOMERIC; SMN1
Gene map locus 5q12.2-q13.3

5: *147183 Links
RECOMBINATION SIGNAL-BINDING PROTEIN SUPPRESSOR OF HAIRLESS, DROSOPHILA, HOMOLOG OF; RBPSUH
Gene map locus 9p13-p12

*180250
RETINOL-BINDING PROTEIN 4; RBP4

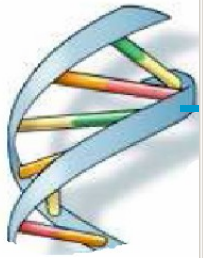
Alternative titles; symbols

RETINOL-BINDING PROTEIN, PLASMA
RETINOL-BINDING PROTEIN DEFICIENCY, INCLUDED
FAMILIAL HYPO-RBP, INCLUDED

Gene map locus 10q24

TEXT

This protein is the specific carrier for retinol (vitamin A alcohol) in the blood. Rask et al. (1987) reported the complete amino acid sequence of serum retinol-binding protein. Pervaiz and Brew (1985) found homology of human serum RBP to bovine beta-lactoglobulin and to protein HCP. By means of a cDNA probe for RBP4, Rocchi et al. (1989) did in situ hybridization and Southern blot analysis of genomic DNA from somatic cell hybrids to map the RBP4 gene to 10q23-q24. Gray et al. (1995) found that the RBP4 gene resides just centromeric of the cluster of CYP2C genes (124020) on 10q24. By the study of recombinant inbred strains, Chainani et al. (1991) showed that the mouse Rbp4 locus is closely linked and just proximal to the locus for phenobarbital-inducible cytochrome P450-2c (Cyp-2c) at the distal end of chromosome 19.



Entrez-PubMed - Netscape

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Shop Stop

Bookmarks Location: http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?SUBMIT=y What's Related

NCBI

PubMed

National Library of Medicine NLM

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books

Search PubMed for retinol-binding protein fish Go Clear

Limits Preview/Index History Clipboard Details

About Entrez

Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities

PubMed Services
Journal Browser
MeSH Browser
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
Cubby

Related Resources
Order Documents
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central

Privacy Policy

Display Summary Sort Save Text Clip Add Order

Show: 20 Items 1-20 of 29 Page 1 of 2 Select page: 1 2

1: Sammar M, Babin PJ, Durliat M, Meiri I, Zchori I, Elizur A, Lubzens E. Related Articles, Nucleotide, Protein

Retinol binding protein in rainbow trout: molecular properties and mRNA expression in tissues.
Gen Comp Endocrinol. 2001 Jul;123(1):51-61.
PMID: 11551117 [PubMed - indexed for MEDLINE]

2: Funkenstein B. Related Articles, Nucleotide, Protein

Developmental expression, tissue distribution and hormonal regulation of fish (Sparus aurata) serum retinol-binding protein.
Comp Biochem Physiol B Biochem Mol Biol. 2001 Jun;129(2-3):613-22.
PMID: 11399497 [PubMed - indexed for MEDLINE]

3: Bellovino D, Morimoto T, Mengheri E, Perozzi G, Garaguso I, Nobili F, Gaetani S. Related Articles

Unique biochemical nature of carp retinol-binding protein. N-linked glycosylation and uncleavable NH2-terminal signal peptide.
J Biol Chem. 2001 Apr 27;276(17):13949-56.
PMID: 11278316 [PubMed - indexed for MEDLINE]

4: Power DM, Elias NP, Richardson SJ, Mendes J, Soares CM, Santos CR. Related Articles

Evolution of the thyroid hormone-binding protein, transthyretin.
Gen Comp Endocrinol. 2000 Sep;119(3):241-55. Review.
PMID: 11017772 [PubMed - indexed for MEDLINE]

5: Cunningham LL, Gonzalez-Fernandez F. Related Articles

Coordination between production and turnover of interphotoreceptor retinoid-binding protein in zebrafish.
Invest Ophthalmol Vis Sci. 2000 Oct;41(11):3590-9.
PMID: 11006257 [PubMed - indexed for MEDLINE]

6: Stenkamp DL, Cunningham LL, Raymond PA, Gonzalez-Fernandez F. Related Articles

Novel expression pattern of interphotoreceptor retinoid-binding protein (IRBP) in the adult and developing zebrafish retina and RPE.

Document: Done

Бази данни за ДНК секвенции

- Основни хранилища:
  - GenBank (US)
    - (http://www.ncbi.nlm.nih.gov/Genbank/index.html)
  - EMBL (Europe)
    - (http://www.ebi.ac.uk/embl/)
  - DDBJ (Japan)
    - (http://www.ddbj.nig.ac.jp/)
- Първични бази данни
  - ДНК последователностите са идентични

Бази данни за секвенции

- Бази данни с анотирани секвенции
  - SWISS-PROT, GenBank etc…
  - за идентифициране на функции и извличане на информация
- Всеки запис в тези бази даннис е анотиран и включва допълнителна справочна информация за поредицата.
- Тази справочна информация позволява да се търси помощта на ключови думи.
- Анотирани бази данни за секвенции обикновено се използват за определяне на функцията на непознати секвенции чрез търсене на сходство в базата данни.
- Ако непозната секвенция съвпада с една от секвенциите от тези бази данни, функцията му често може да се изведе от анотациите, свързани със секвенцията с която има съвпадение.

Бази данни за секвенции

- Слабо анотирани бази данни
  Тези бази данни съдържат предимно секвенции с минимални анотации, защото много от секвенциине нямат характеристики
- Тези бази данни могат да бъдат полезни като източник на нови генни последователности.
- Специализирани бази данни
- Съдържат подмножество на последователности, обикновено на определен вид, форматирани или с обяснителни бележки по специфичен начин.

Основни протеинови бази данни

- SWISS-PROT
  - За разлика от основните бази данни за нуклеотидни секвенции, които съдържат едни и същи данни последователности, основните публични протеинови бази данни имат различна «персонализация».
  - Swiss-PROT е ръчно разработена база данни. Като такава има много високо качество, съобразени анотации, които я правят много подходяща за търсене по ключови думи. Въпреки това, тъй като процесите на валидиране и анотиране отнемат време, не съдържа най-много данни както други протеинови бази данни.

Основни протеинови бази данни

- GenPept/TREMBL
  - GenPept и TrEMBL са бази данни, генерирани автоматично от транслиране на кодиращи секвенции (CDS) от Genbank и EMBL.
  - Качеството на анотации в тези бази данни не е по-високо от Swiss-PROT, но тези бази данни са много по-актуални.
- PIR
  - PIR е с филогенетично базирана анотации, но неговите анотации са различни от тези на Swiss-PROT. Форматът е често по-удобен за търсене на текст, но записите съдържат информация за superfamily на записа, която често е трудно да се намери в други бази данни.
- Трите бази данни са комбинирани в UniProt (http://www.uniprot.org)

Структурни бази данни

- PDB (Protein Databank)
  - Съхранява 3 измерни координати на атомите в биологичната молекула
  - Данните са получени от X-ray кристалография, NMR или моделиране
  - http://www.rcsb.org/pdb/
- Повечето методи за предсказване на третичната структура на протеини са силно зависими от съществуващите протеинови структури. Тези структури са получени чрез експериментални методи или компютърно моделиране.
- Протеиновите бази данни са основни хранилища на протеинови структури, които се съхраняват като координати на атомите.

Структурни бази данни

- Съществуват и други структурни бази данни които добавят стойности към суровите данни, съхранявани в PDB. Например базата данни SCOP класифицира протеините съгласно структурното подобие и еволюционните връзки. Предоставя йерархична класификация на протеините от фамилията и суперфамилията, които са свързани с дадената структура.
- MMDB (Molecular Modeling database)
  - Над 28,000 3D макромолекулни структури
  - (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Structure)
- SCOP (Structural Classification of Proteins)
  - Класификация на протеините според структурните и еволюционни връзки

Genome Databases

- Фокусирана върху един или няколко организма. Събрана на едно място цялата налична информация.
- Примери:
  - Colibase (E. coli and related species)
  - http://colibase.bham.ac.uk/
  - GDB (human)
    - http://www.gdb.org/
  - Flybase (Drosophila)
    - http://flybase.bio.indiana.edu/
  - WormBase (C. elegans)
    - http://wormbase.org
  - AtDB (Arabidopsis)
    - http://www.arabidopsis.org
  - SGD (S. cerevisiae)
    - http://genome-www.stanford.edu/Saccharomyces/