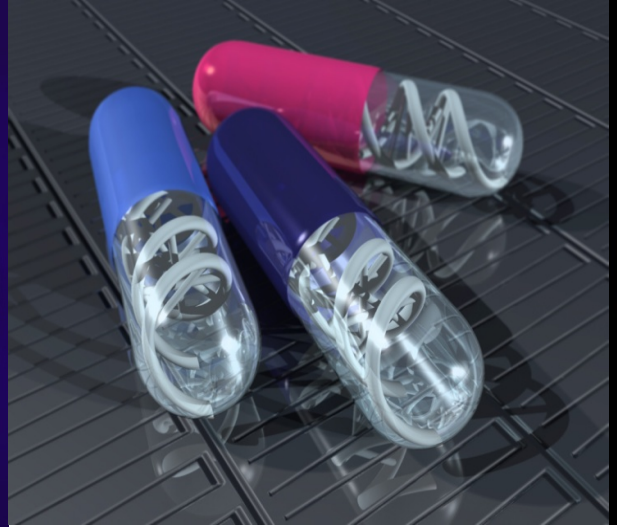
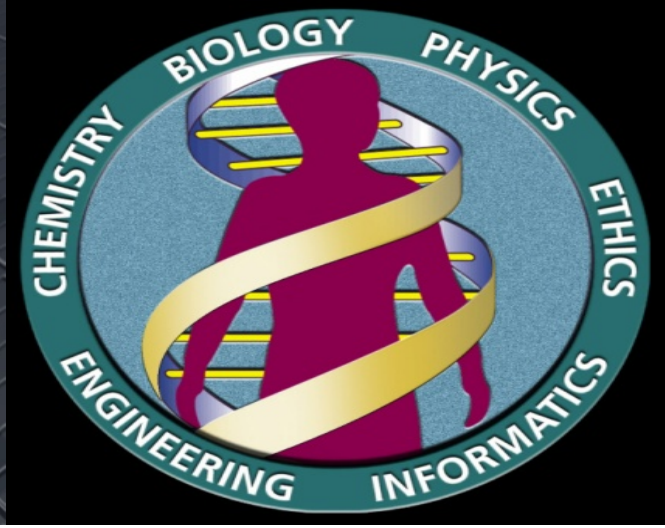
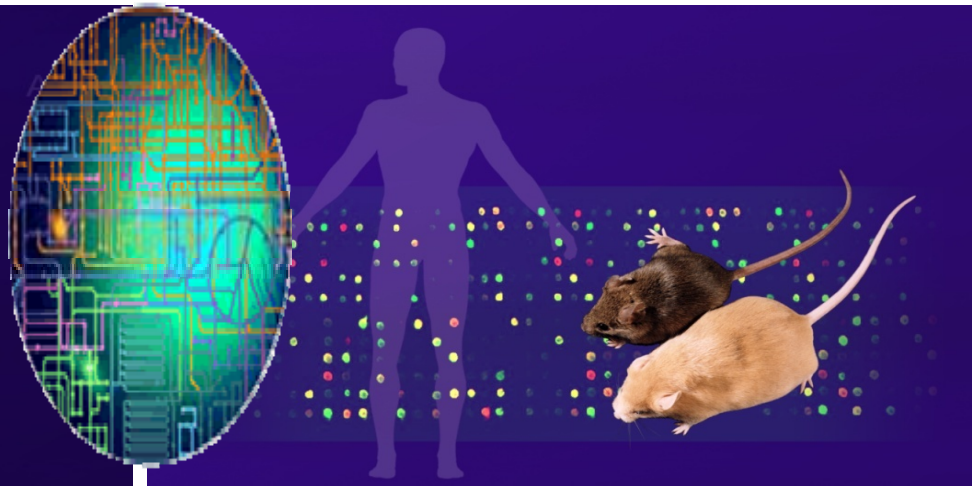




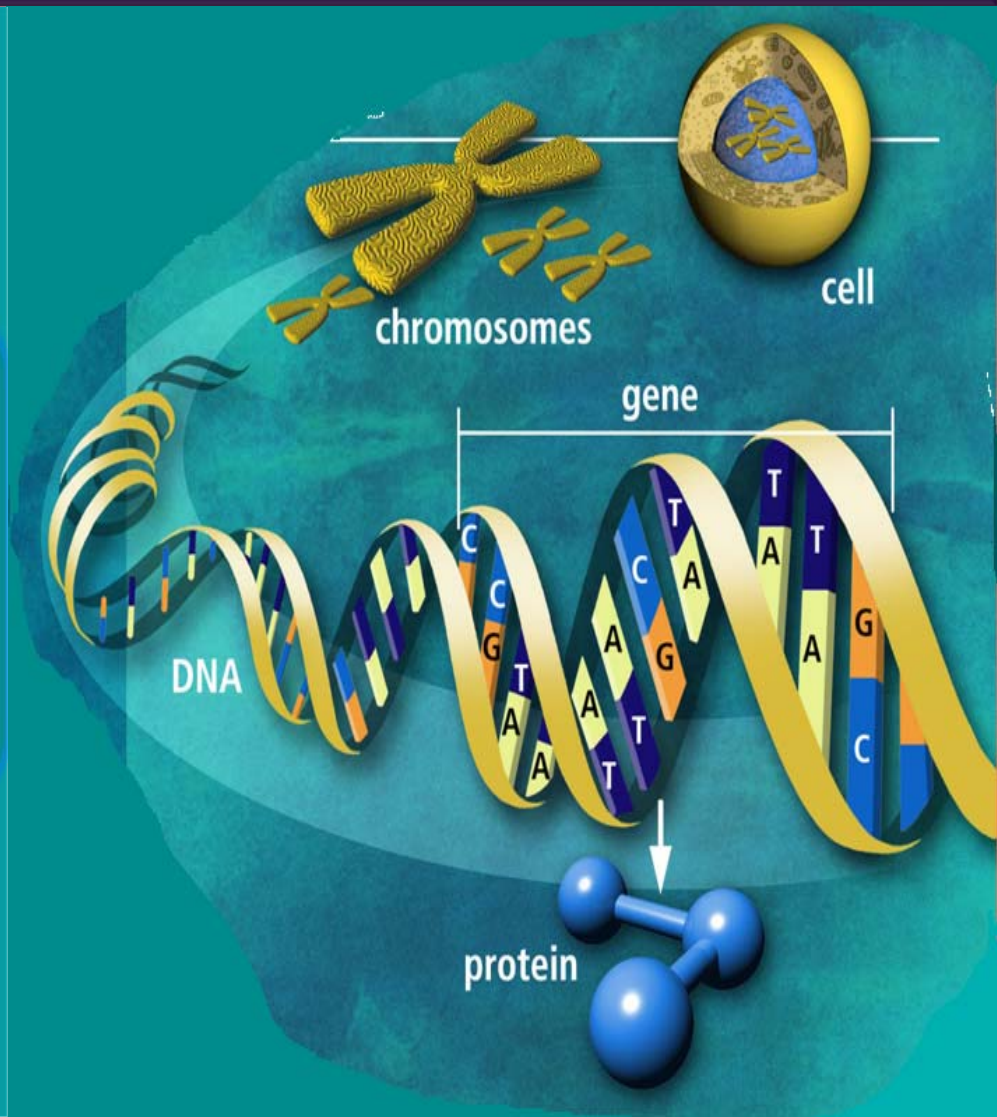
Геномни бази данни за човека

Денис Синеков F55739, НБУ
“ Приложна биология”- II курс



ДНК – МОЛЕКУЛАТА НА ЖИВОТА

Биологичната информация, необходима за възпроизвеждане на един организъм се съдържа в *ДНК*.

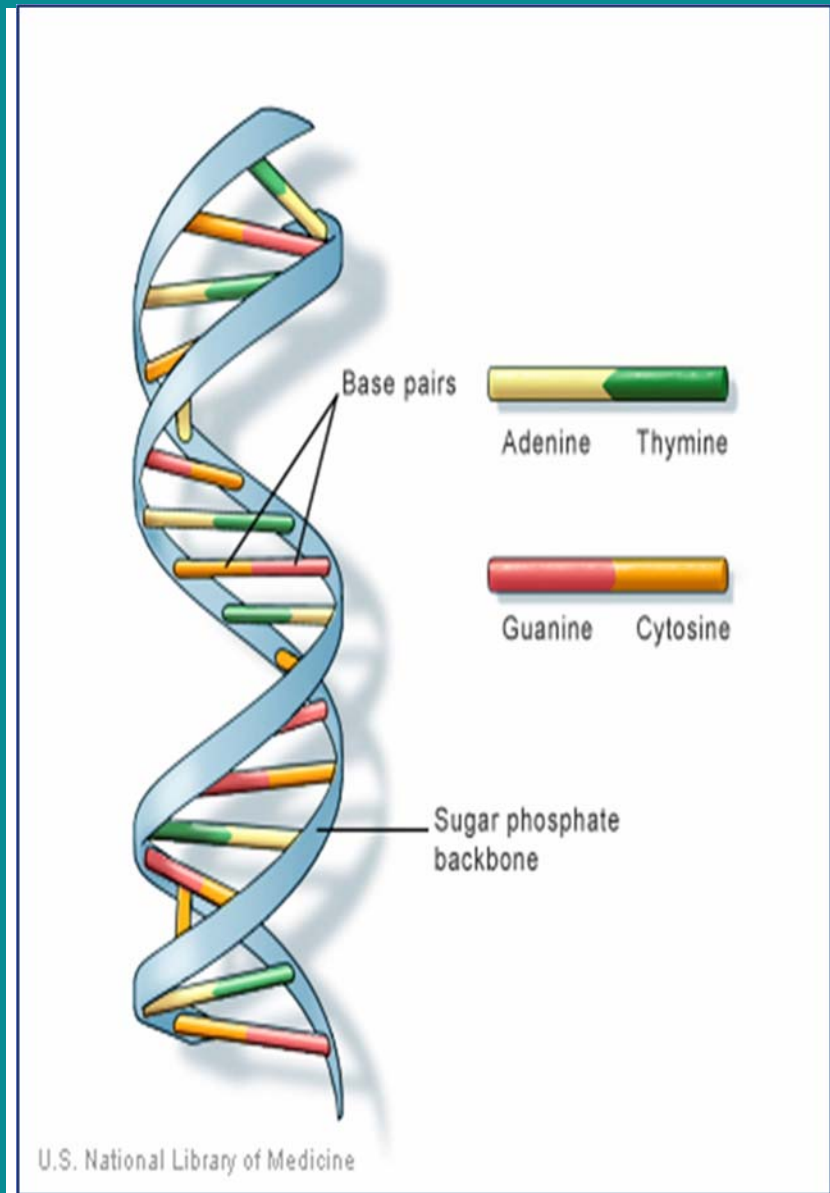


ДНК – МОЛЕКУЛАТА НА ЖИВОТА

ДНК молекулите имат специфична триизмерна структура, известна като двойна спирала. Моделът на двойната спирала е изграден през 1953 г. от Уотсън и Крик.

ДНК е полимер, съставен от дълга верига мономери, наречени нуклеотиди, поради което молекулата на ДНК се означава като полинуклеотид.

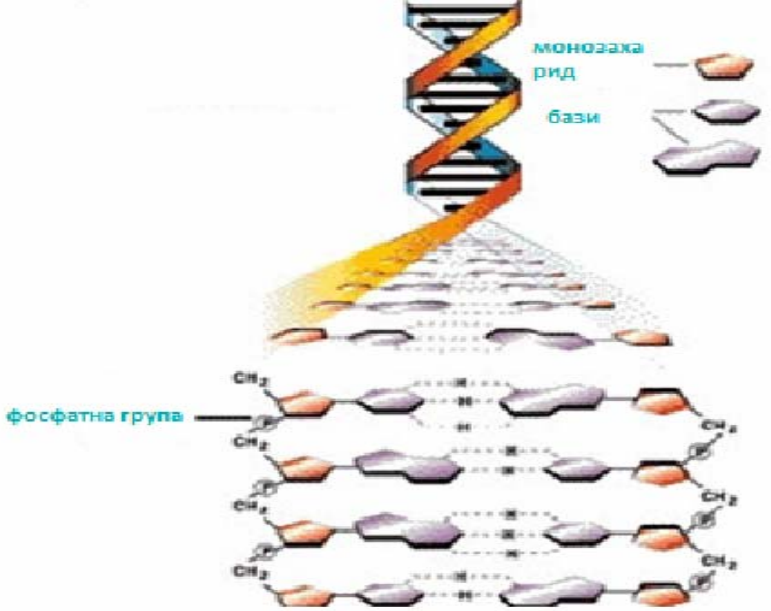
Всеки нуклеотид има три части – един монозахарид, азот-съдържаща пръстеновидна структура, наречена база и фосфатна група.

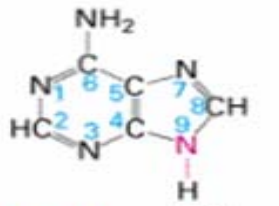


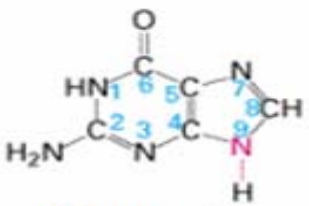
Монозахаридът, представен в ДНК е петвъглеродна пентоза, наречена 2'-дезоксирибоза, в която хидроксилната група (-OH) при втория въглероден атом (2'-C) е заместена с водород.

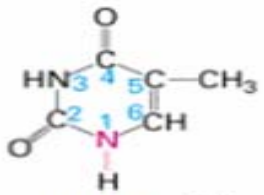
Нуклеотидите съдържат фосфатна група (PO4), свързана с 5'-въглеродния атом на пентозата.

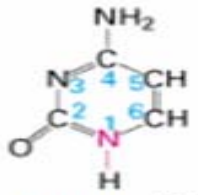
Нуклеотидите съдържат една от четирите бази – аденин, гуанин, тимин и цитозин. Аденинът и гуанинът са изградени от двоен въглеродно-азотен пръстен, и се наричат пурини. Цитозинът и тиминът представляват еднопръстенни съединения – пиримидини. Азотните бази са свързани с пентозата чрез връзка между първия въглероден атом на пентозата (1') и азотът в положение 9 - в пурините и положение 1 - в пиримидините. Тази връзка се означава като N-глюкозидна връзка. Пентозата, свързана с азотната база се нарича нуклеозид.




Аденин (A)

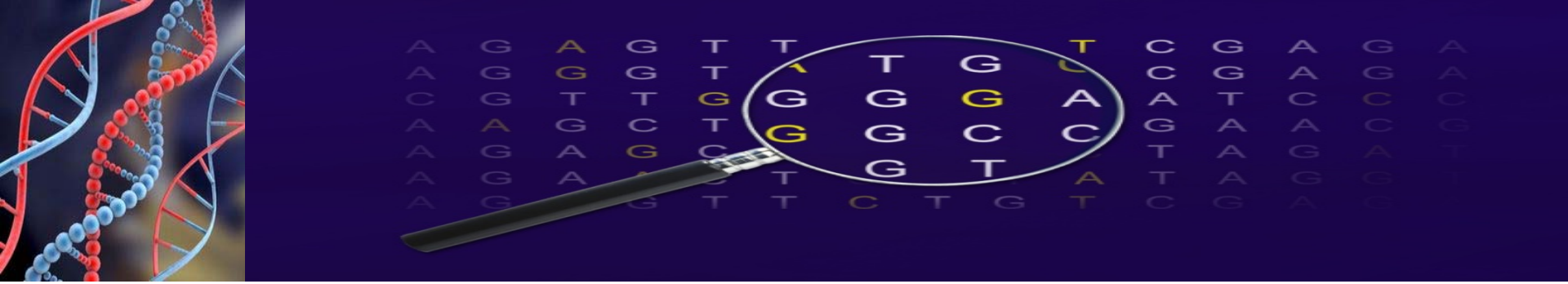

Гуанин (G)


Тимин (T)


Цитозин (C)

agbiosafety.unl.edu/ education/dnax.html

distancelearning.ksi.edu/ demo/bio378/lecture.html

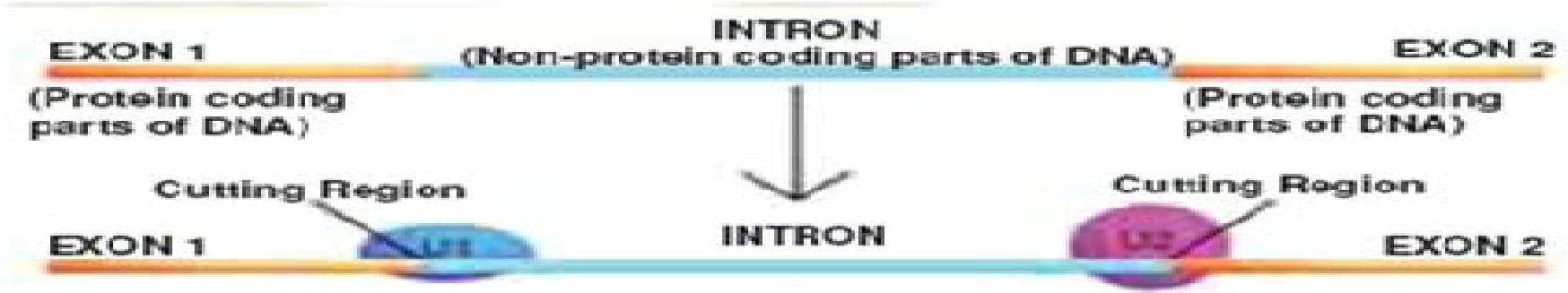


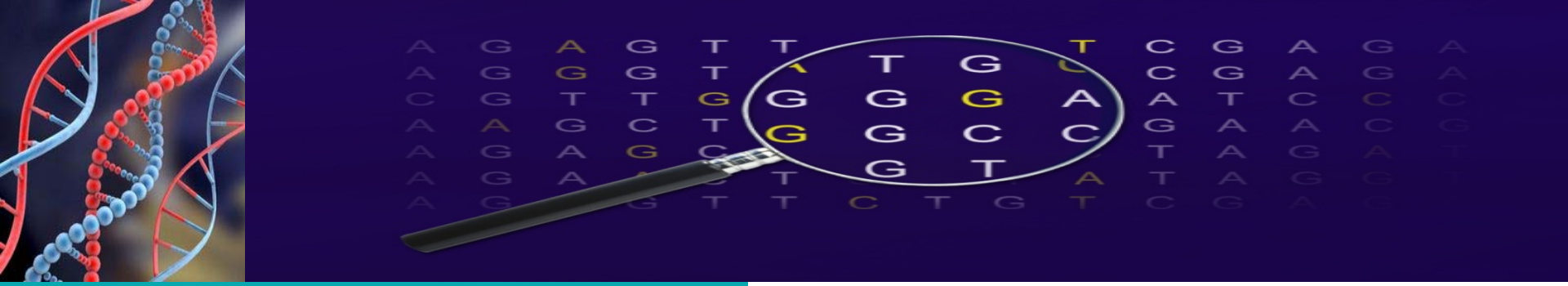
Всяка ДНК молекула е пакетирана отделна хромозома, като цялата генетична информация, съдържаща се в хаплоидния набор хромозоми на един организъм, се нарича *геном*.

Човешкият геном **съдържа около $3x10^9$ нуклеотидни двойки, организирани в 24 хромозоми (22 , Х+У). Всяка хромозома съдържа от $50x10^6$ до $250x10^6$ нуклеотидни двойки. ДНК молекула с тези размери има дължина от 1.7 до 8.5 см., когато е деспирализирана.**

Капацитетът на ДНК да кодира генетична информация е огромен. Разнообразието на една верига ДНК, състояща се от n нуклеотида е 4n. Дори и за късите секвенции ДНК разнообразието много голямо. На практика има известни ограничения в секвенциите, които могат да носят полезна информация, но и тогава комбинациите от секвенции са много.

Кодиращата информация при еукариотните гени е обикновено серия ДНК секвенции, наречени екзони. Те са разделени от серия секвенции, които не носят полезна генетична информация и се означават като интрони. Броят на интроните варира силно от нула до около 50 в някои гени. Дължината на екзоните и интроните също варира, като обикновено интроните са по-дълги и всъщност представляват по-голямата част на гена. Преди генетичната информация да се експресира в белтък, интроните трябва да се премахнат от РНК чрез процес, който се нарича сплайсинг. Този процес превръща РНК в непрекъсната кодираща секвенция. Интроните са характерни за висшите организми и обикновено не се откриват в бактериите.



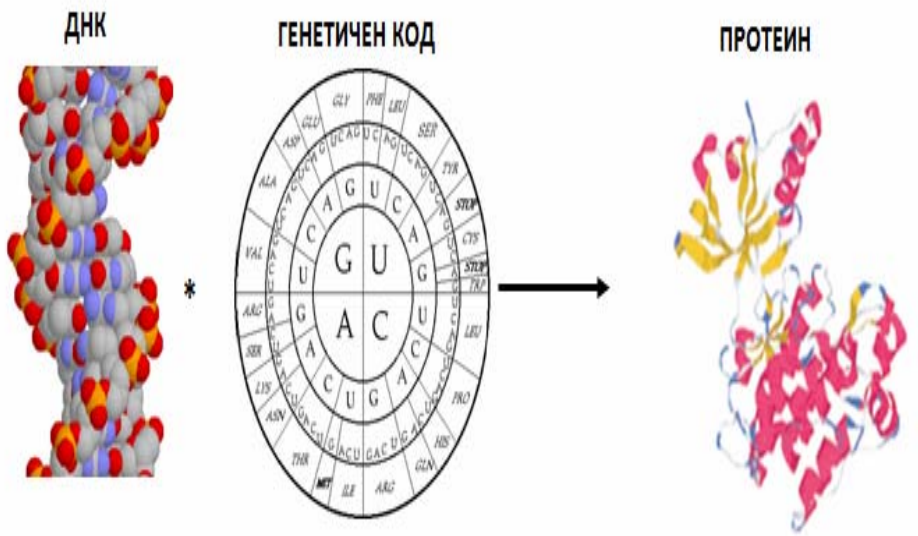


ГЕНЕТИЧЕН КОД

Информацията за първичната структура на белтъците е закодирана в *нуклеотидната последователност* на ДНК. Той е ключът който свързва езика на нуклеотидите с езика на протеините.

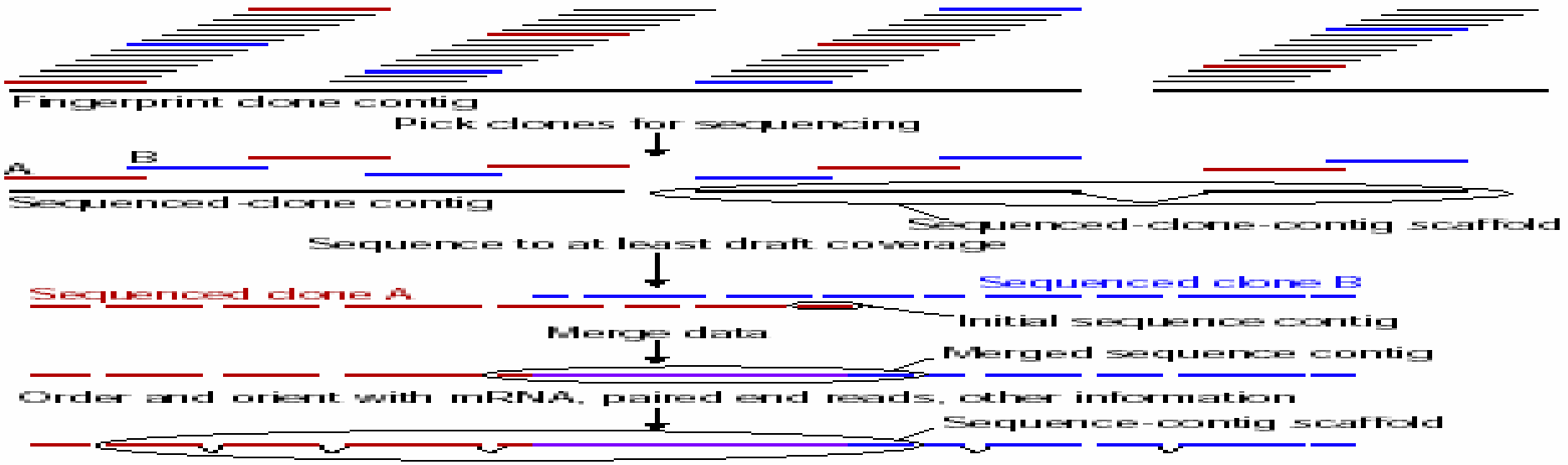
Генетичният код е:

- ***триплетен***
- ***колинеарен***
- ***неприпокриващ се***
- ***с една рамка на четене***
- ***универсален***

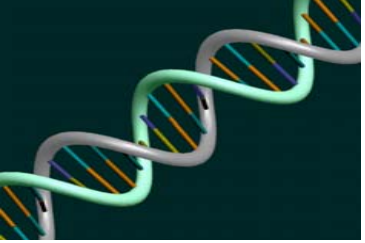


СЕКВЕНИРАНЕ

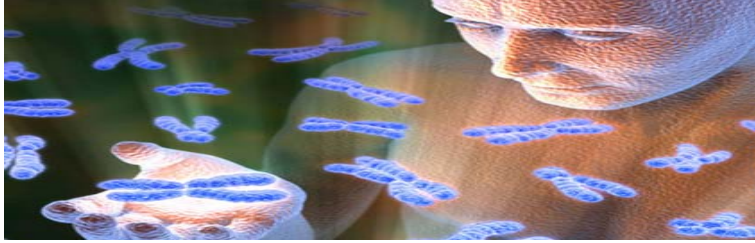
След като генетичната азбука беше открита, оставаше да се прочете написаното в ДНК. Като първа стъпка в тази насока бяха създадени методи за определяне на последователното подреждане на четирите бази по дължината на ДНК, процедура известна като "секвениране на ДНК". Бяха разработени различни методи за секвениране. Понастоящем има автоматични секвенииращи роботи, които работят на този принцип и могат да секвенират ДНК със скорост от една база за секунда. С помощта на този метод беше секвенирано огромно количество ДНК и бяха установени пълните секвенции на човешкия, мишия и други геноми.



Създаване на скелета на една секвенция

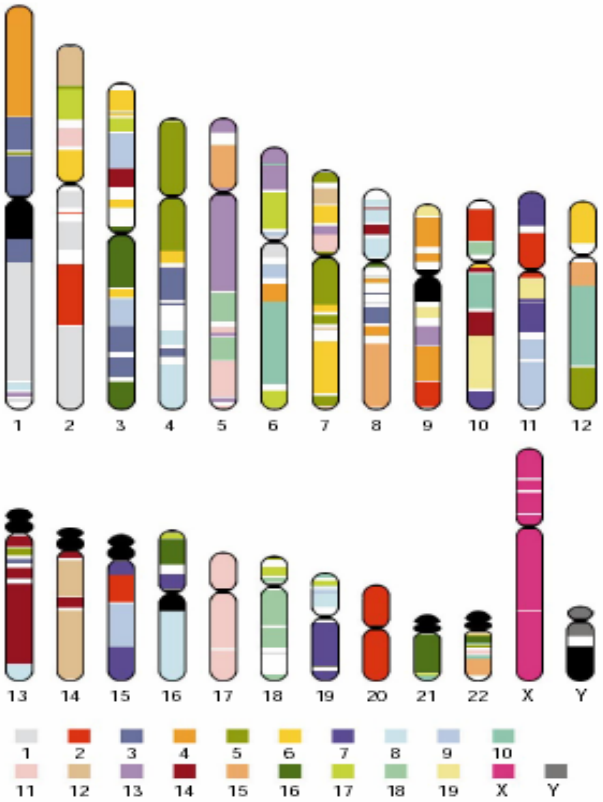


ЧОВЕШКИ ГЕНОМ

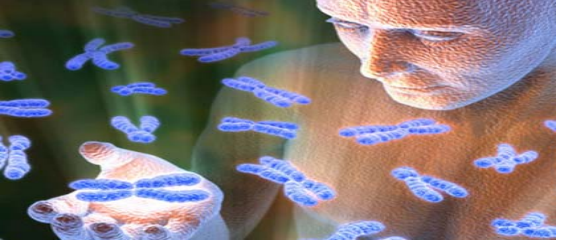


Предпоставки за развитие на проекта по човешкия геном

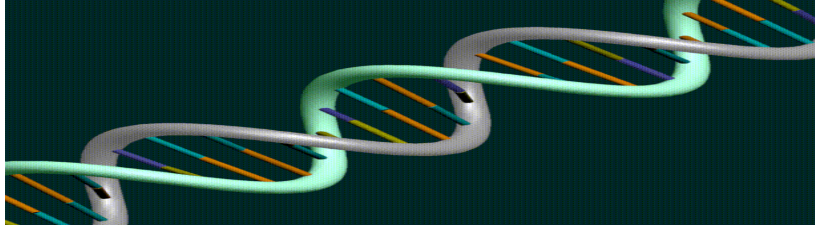
1956:Създаване на първи цитогенетични карти. Чрез светлинна микроскопия на оцветени тъкани се установява, че човешките клетки съдържат 46 хромозоми, или общо 24 различни типа хромозоми



Човешкият кариотип

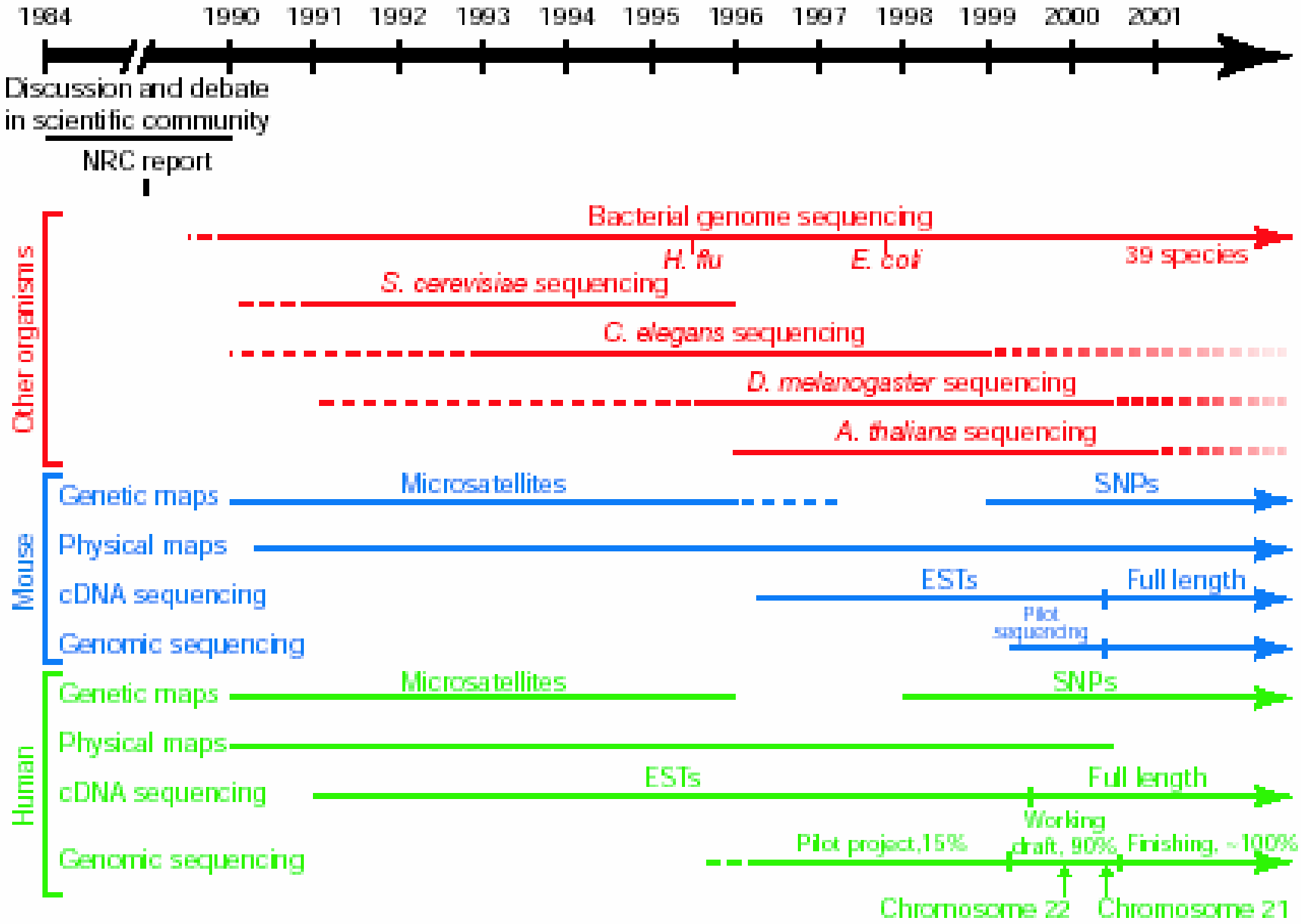


ЧОВЕШКИ ГЕНОМ



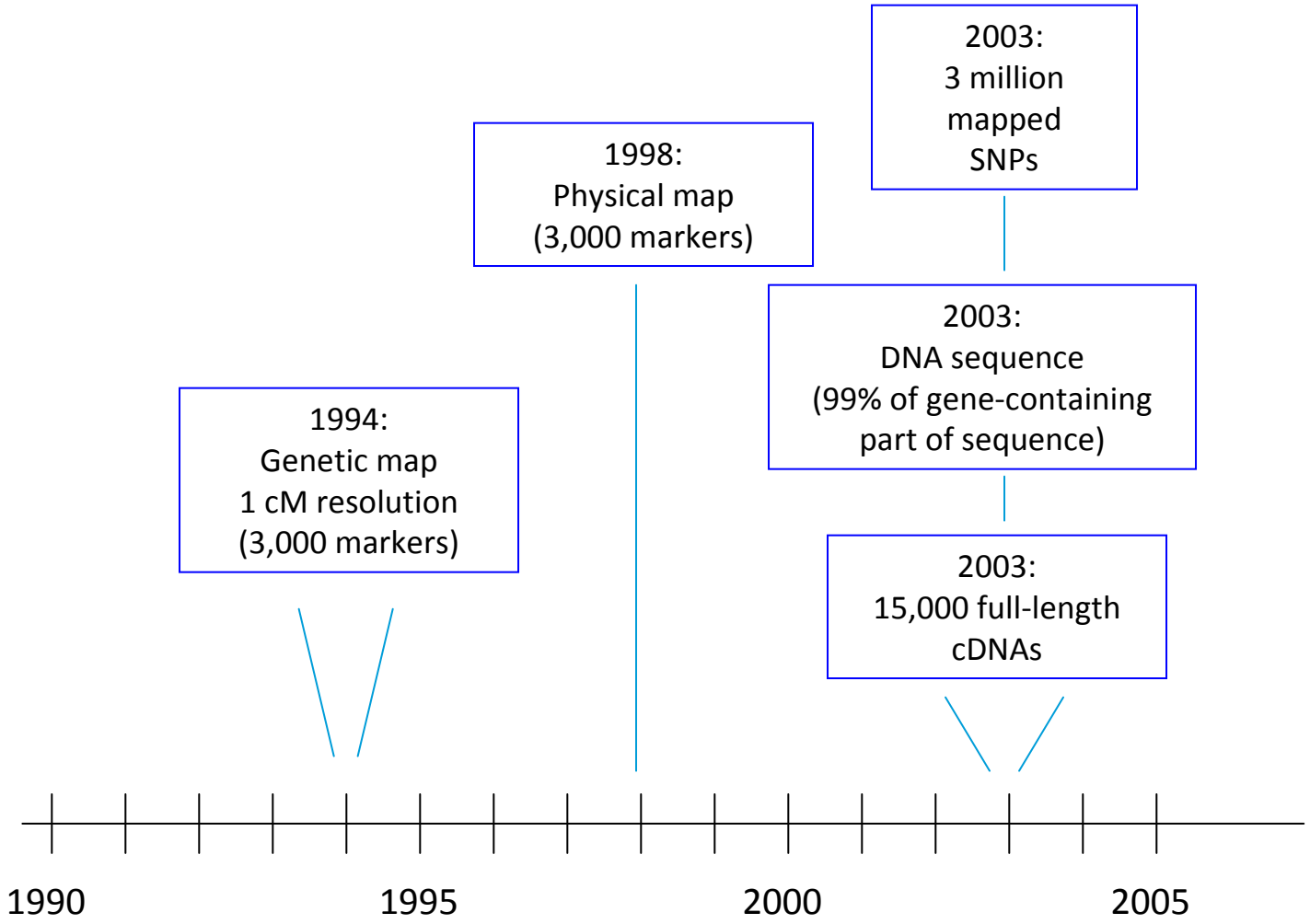
Предпоставки за развитие на проекта по човешкия геном

- **1977: Фред Зангер и колеги от Кембридж, Англия публикуват метода за дидеокси ДНКсеквениране, който и повече от четвърт век по-късно е основа за секвениране на ДНК**
- **1980:Въвежда се използването на случайните ДНК маркери (RFLPs).**
- **1980: Зангер и др. Публикуват пълната секвенция на човешката митохондриална ДНК (16.5 кб)**
- **1987: US Department of Energy предлага инициативата за човешкия геном с три основни направления: детайлно физично картиране на всички хромозоми; развитие на подходящи технологии; подобраване на мрежите за комуникация и капацитетите на база данните.**
- **1989: Създава се и HUGO (Human Genomic Organization) за координиране на международните усилия.**



Разпределение във времето на многомащабните геномни анализи

Ключови достижения в първия етап от проекта за човешкия геном



Ключови достижения в първия етап от проекта за човешкия геном

- 1990: Официално стартира проекта за човешкия геном вUS.
- 1991: Създава се геномната база данни (GBD) като депозитар на данните от картирането на човешката ДНК.
- 1992: Вайсенбах и колеги, Франция, публикуват първата човешка генетична карта, основана на микросателитни маркери
- 1993: Кохен и колеги, Франция, публикуват първа генерация физична карта на човешкия геном, основана на клонове с големи ДНК инсерции
- 1995: Ландер и др., Масачузетски Технологичен Институт публикуват първата детайлна генетична карта на човешкия геном
- 1998: GeneMap'98, първата карта на гените въз основа на маркерите е публикувана от международния консорциум

Ключови достижения в първия етап от проекта за човешкия геном

- **1999: Международния консорциум начело с учени от Зангер Центъра, Англия публикуват първата пълна секвенция на човешка хромозома, хромозома 22**
- **2000-2001: най-интензивно натрупване на резултати - секвенираната част нараства от 15 на 90%**
- **2001: Публикуване в най-груб вид секвенцията на човешкия геном, обхващащ 90% от еухроматина (кодираща част) и 70% от общата секвенция (еухроматин+хетерохроматин) от Международния консорциум и частната компания Celera.**
- **2003: Завършване на 100% секвенирането на еухроматиновата част и на 94% от общата секвенция на човешкия геном. Определената от консорциума дължина на генома е 3.2 гигабази, а от Celera – 2.9 гигабази.**

Ключови достижения в първия етап от проекта за човешкия геном

2003

Human Chromosome 6 Completed, October 2003.

Human Chromosome 7 Completed, July 2003.

Human Chromosome Y Completed, June 2003.

Human Genome Project Completion, April 2003

Human Chromosome 14 Finished - Chromosome 14 is

the fourth chromosome to be completely sequenced.

Ключови достижения в първия етап от проекта за човешкия геном

2004

Human Gene Count Estimates Changed to 20,000 to 25,000, October 2004.

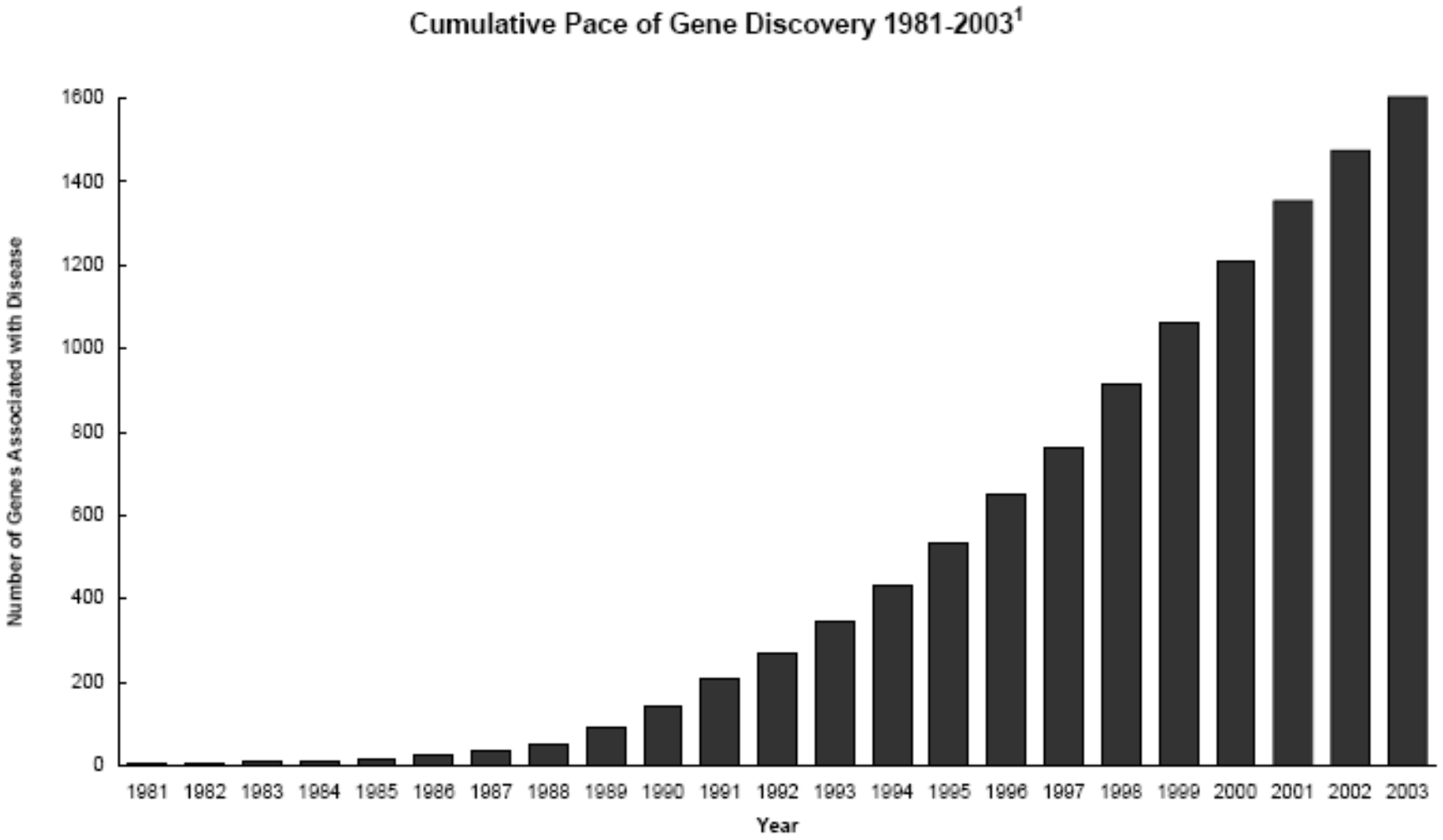
Human Chromosome 5 Completed, September 2004.

Human Chromosome 9 Completed, May 2004.

Human Chromosome 10 Completed, May 2004.

Human Chromosome 19 Completed, March 2004.

Human Chromosome 13 Completed, March 2004.



http://www.genome.gov/Pages/News/PaceofDiseaseGeneDiscovery.pdf

Основни достижения в проекта по човешкия геном

- **Човешкият геном съдържа около 30 000 - 32 000 гена. Макар само 1.5 пъти повече от тези на кръглия червей, те са по-сложно устроени, с повече възможности за допълнителни взаимодействия и нагъвания, водещи до по-голямо разнообразие на кодираните от тях белтъчни продукти.**
- **24% от генома е общ с тези на други еукариотни организми. Допълнително е наблюдавано 22% подобие с другите гръбначни животни.**
- **Само 1% от последователностите са "чисто човешки", т.е. не са открити при други организми. Дори и този % обаче е условен, защото все още геномите на най-близко родствените организми като например шимпанзето не са секвенирани.**
- **Изненадващо, в човешкия геном е открита 1% ДНК, която се съдържа само при прокариотите и не е идентифицирана в никои други еукариоти. Стотици човешки гени вероятно са произлезли чрез хоризонтален трансфер от бактериите в гръбначните. Няколко десетки гени вероятно са произлезли от транспозони. Въпреки че голяма част от човешкия геном се е развил въз основа на транспозоните, в процеса на еволюция те са се инактивирали и в съвременния геном са неактивни.**
- **44% от генома са повторени последователности. При другите организми повторените последователности са много по-малко. Увеличаването броя на повторените последователности предполага еволюционно презастраховане от грешки.**

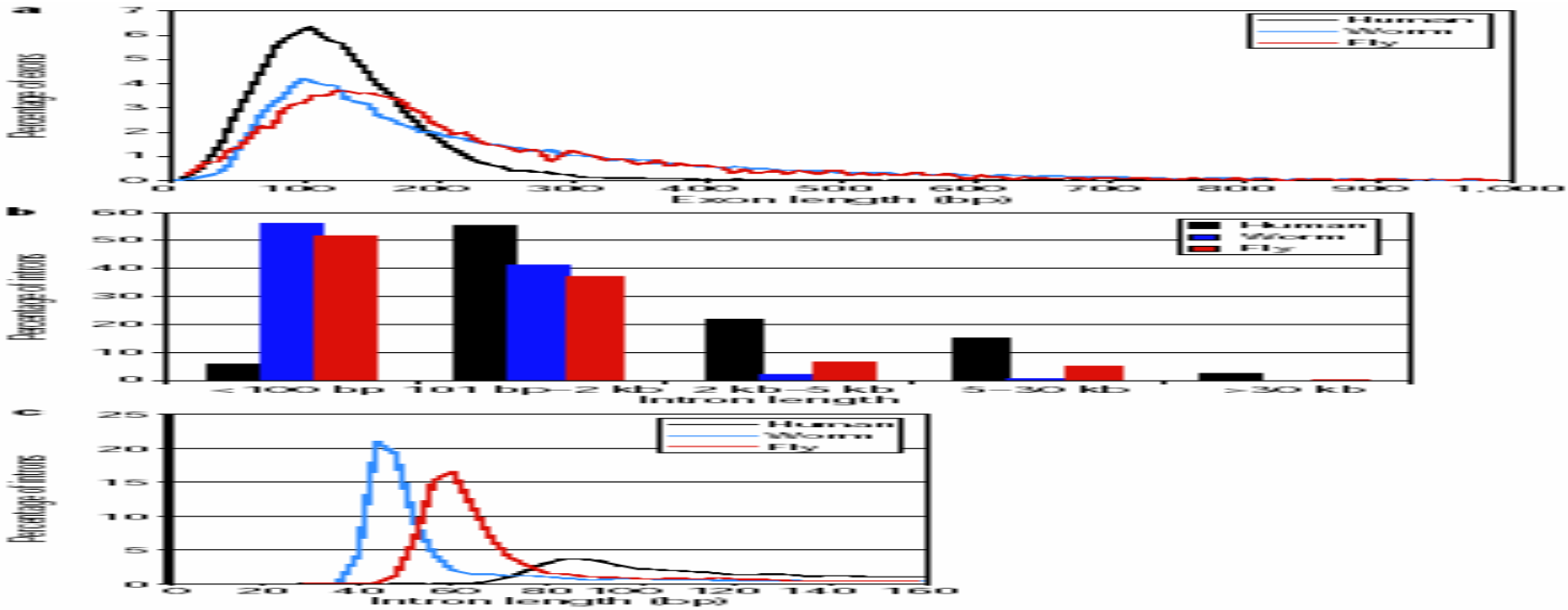
Основни достижения в проекта по човешкия геном

- **Размерът на целия геном беше определен на 3.2 Гб. Беше изчислен размера на всяка хромозома , големината на центромера и хетерохроматина.**

Table 8 Chromosome size estimates

Chromosome*	Sequenced bases† (Mb)	FCC gaps‡		SCC gaps‖		Sequence gaps#		Heterochromatin and short arm adjustments**(Mb)	Total estimated chromosome size (including artefactual duplication in draft genome sequence)†† (Mb)	Previously estimated chromosome sizes‡‡ (Mb)
		Number	Total bases in gaps§ (Mb)	Number	Total bases in gaps¶ (Mb)	Number	Total bases in gaps* (Mb)			
All	2,692.9	897	152.0	4,076	142.7	145,514	80.6	212	3,289	3,286
1	212.2	104	17.7	347	12.1	11,803	6.5	30	279	283
2	221.6	50	8.5	296	10.4	12,880	7.1	3	251	255
3	186.2	71	12.1	336	11.8	14,689	8.1	3	221	214
4	168.1	39	6.6	343	12.0	12,768	7.1	3	197	203
5	169.7	46	7.8	337	11.8	10,304	5.7	3	198	194
6	158.1	15	2.6	275	9.6	5,225	2.9	3	176	183
7	146.2	27	4.6	195	6.8	4,338	2.4	3	163	171
8	124.3	41	7.0	249	8.7	8,692	4.8	3	148	155
9	106.9	19	3.2	122	4.3	6,083	3.4	22	140	145
10	127.1	14	2.4	163	5.7	8,947	5.0	3	143	144
11	128.6	29	4.9	193	6.8	8,279	4.6	3	148	144
12	124.5	26	4.4	168	5.9	8,226	4.6	3	142	143
13	92.9	12	2.0	115	4.0	5,065	2.8	16	118	114
14	86.9	13	2.2	40	1.4	775	0.4	16	107	109
15	73.4	18	3.1	104	3.6	5,717	3.2	17	100	106
16	73.1	55	9.4	102	3.6	4,757	2.6	15	104	98
17	72.8	41	7.0	95	3.3	4,261	2.4	3	88	92
18	72.9	22	3.7	113	4.0	4,324	2.4	3	86	85
19	55.4	49	8.3	108	3.8	2,344	1.3	3	72	67
20	60.5	7	1.2	33	1.2	489	0.3	3	66	72
21	33.8	4	0.1	0	0.0	0	0.0	11	45	50
22	33.8	10	1.0	0	0.0	0	0.0	13	48	56
X	127.7	141	24.0	182	6.4	4,282	2.4	3	163	164
Y	21.8	6	1.0	19	0.7	113	0.1	27	51	59
NA	5.1	0	0	134	0.0	577	0.3	0	0	0
UL	9.3	38	0	7	0.0	566	0.3	0	0	0

- **Кодиращите секвенции са по-малко от 5% от генома.**
- **Геномната карта показва значително разнообразие в разпределението на елементите на човешкия геном, такива като гени, транспозони, ГЦ съдържание, скорост на рекомбинация, което насочва учените към търсене на тяхната функция. Някои участъци съдържат много гени с неголям брой бази, при други е обратно.**
- **Екзоните (кодиращите белтъци области) при човека са по-къси от тези при другите организми, т.е. еволюцията е вървяла в посока на скъсяване на екзоните. Обратно, в човешкия геном се наблюдава удължаване на интроните .**



Разпределение на екзоните и интроните по дължина при човека, кръглия червей и винарката

Основни достижения в проекта по човешкия геном

- Областите, които са богати на гени са с по-високо Г+Ц съдържание (тези двойки образуват 3 връзки помежду си и придават по-голяма ригидност на ДНК молекулата), докато в бедните на гени области преобладават А+Т двойките
- Скоростта на рекомбинации е много по-висока в краищата на хромозомите, което съдейства получаването на поне една рекомбинация за всяка хромозома при всеки мейозис.
- Скоростта на мутации при мъжете в мейозиса е около два пъти по-висока, отколкото при жените, показващо по-голямо натрупване на мутации в мъжете.
- Идентифицирани са повече от 1.4 милиона присъствия на единичен нуклеотиден полиморфизъм (SNP). Това определя уникалността в геномната секвенция на всеки индивид.
- Пълният набор от белтъци (протеома), кодиран от човешкия геном е по-сложен от този при другите организми. Отчасти това се дължи на присъствието на специфични за гръбначните мотиви (около 7% от общата ДНК). Често мотивите са знак за специфична функция. По-често обаче разнообразието е свързано с факта, че при гръбначните съществуващите компоненти в генома се подреждат в по-разнообразни комбинации.

Значение на резултатите от пълното секвениране на човешкия геном

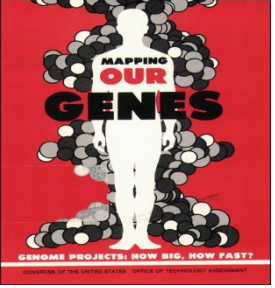
Медицински проучвания. Ключово приложение на изследванията върху човешкия геном е възможността да се открият свързани с болести гени. Досега са описани повече от 30 такива гени. Въз основа на пренатални и пресимптоматични изследвания може родителите да вземат решение за раждането на дете с наследствени разстройства. Очаква се също така този проект да служи като рамка за развитие на нови терапии за различни болести, както и да се разширят съществуващите. Широкото приложение на геномните изследвания се очаква да доведе до радикална промяна в подхода за работещите в здравеопазването – вече не основната цел да е третиране на напредналата болест, а предотвратяване на болестта .

Антропология. Информацията е полезна в антропологическите и исторически проучвания на произхода на човека, движенията на преисторическите популации и социални структури.

Значение на резултатите от пълното секвениране на човешкия геном

Филогенетика. За разлика от бактериите, където съседно разположените гени са свързани функционално, при бозайниците близко разположените гени рядко имат общи функции, но имат обща история и филогенеза. Проучването на геномни сегменти може да хвърли светлина върху генетични преноси станали преди 500 млн години или сравнително скоро, преди около 20 милиона години. Например е установено, че човекът и мишката са имали общ предшественик преди около 100 млн години.

Анализи в следствието и правосъдието. Точността на тестовете, основани на ДНК анализ се използват за идентифициране на заподозрян или да се установи тясно биологично родство. В тези случаи трябва да се има предвид, че диагностичните маркери могат да варират от една популация до друга.



Проблеми, които могат да възникнат от прилагане на знанията, получени в проекта по човешкия геном

- Познаването на човешките гени крие опасността от прилагане на генетични техники за “подобряване” качеството на човешкия род.
- Съществува също и опасност чрез генно инженерство да се получи пренатрупване на индивиди с качества, за които в момента е било решено, че са положителни.
- Огромна е и опасността застрахователните компании да пристъпят към многомащабно генетично тестване на кандидат клиентите си за присъствие на гени, които определят предразположеност към такива болести като диабет, кардиоваскуларни заболявания, ментални увреждания, рак. Тогава на напълно здрави индивиди, които носят такива алели, може да бъде отказана медицинска застраховка. Обезпокоителното е, че много хора носят такива алели, които може никога да не се проявят, а това означава % на хората, които ще бъдат дискриминирани да е много голям.

Информационни геномни ресурси. Бази данни за човешкия геном

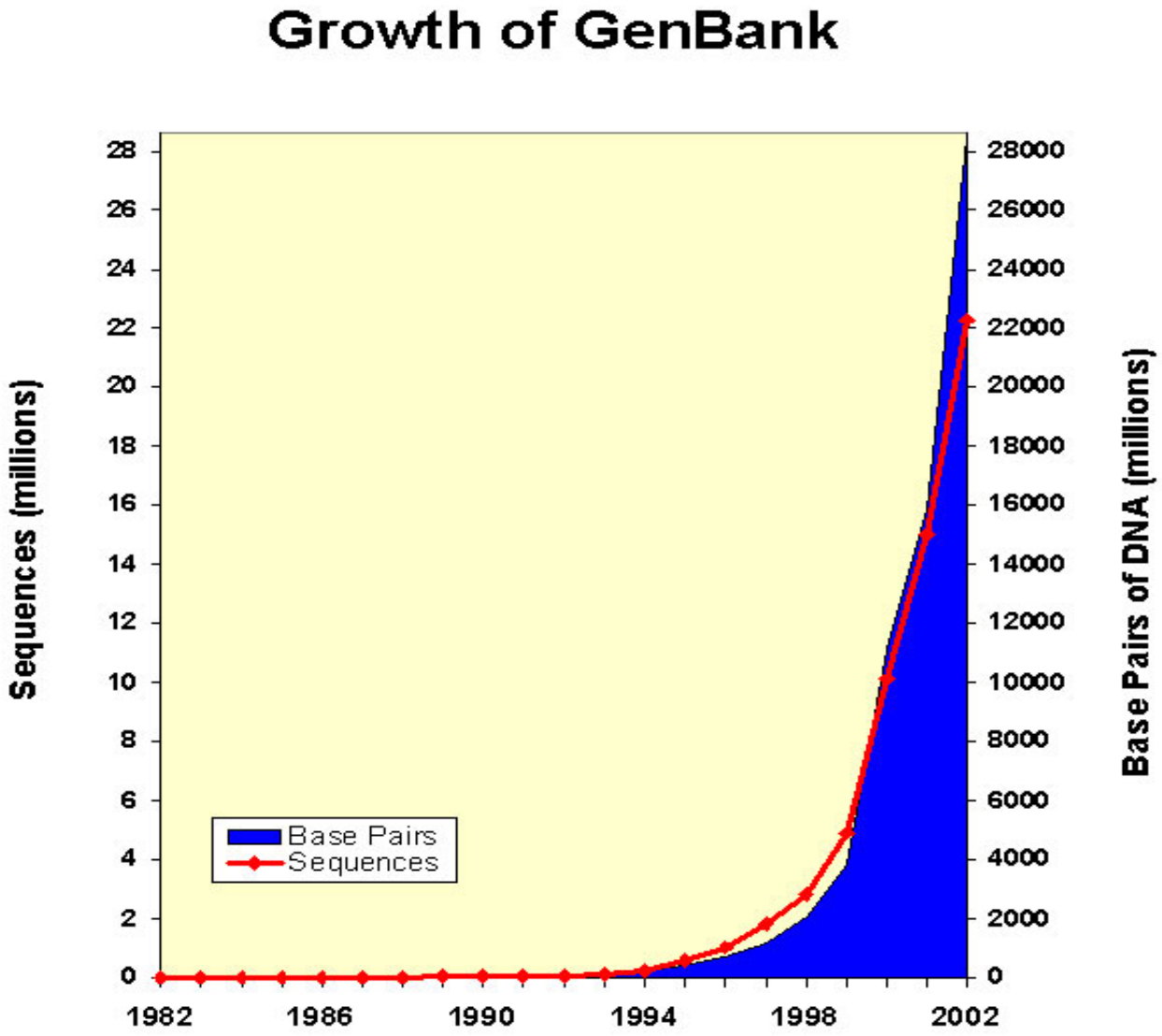
Информационните геномни ресурси са общи и специализирани. Към *общите* спадат базите данни на EMBL, на DDBJ и на GenBank. Последната представлява ДНК- база данни , натрупани от източници на секвенции директно чрез частни малки проекти или чрез широкомащабни проекти. Осигурен е обмен с EMBL и DDBJ. Обемът на данните непрекъснато се изменя. Например през 1997 г. само 6% от информацията в GenBank е за генома, а 74 % - за секвенции от типа EST.

***Самата база данни в GenBank* има 17 подразделения, които са кодирани с три букви. Някои от тях са: PLN (растения, гъби и водорасли),BCT (бактерии) RNA (структурна РНК), SYN (синтетични), EST (секвенции от типа EST), PAT (патенти), STS (месторазположения на секвенции), GSS (геномни обслужващи секвенции) и други.**

DDBJ (ДНК база данни на Япония)
EMBL Nucleotide Sequence DB (Европейска лаборатория по молекулярна биология)
GenBank (Национален център по биотехнологична информация)

1982: 600,000 Bases

2002: 28.5 Billion Bases



източник: www.ncbi.nlm.nih.gov

Информационни геномни ресурси. Бази данни за човешкия геном

Специализираните **геномни ресурси се наричат "бутици" и обикновено представляват разширение на първичните ДНК-бази данни. Целта е да се фокусира вниманието върху специфичните видове геноми и особености на подреждане на информацията. По-известните и по-често употребявани бази данни с Web-сайтове са следните:**

UniGene. **Това е един опит за разработване на описателна картина от мрежа неподредени генно-ориентирани кластери, получени от секвенции на базата данни GenBank. Колекцията представлява гени на множество организми, а всеки кластер се отнася до уникален ген и включва информация за тъканите, за разположението върху картата и други подробности. Допълнително са указани субституционалните номера на новите секвенции от типа EST. Информацията от тази база данни се използува също за селекция на реагенти за проектните карти на екеспериментаторите и шириокомащабен експресен генетичен анализ. Ресурсът използува NCBI-home page.**

Информационни геномни ресурси. Бази данни за човешкия геном

TDB (TIGR-data base) **съдържа ДНК и протеинови секвенции, генни изрази от секвенции, характеристики на клетките, протеинова фамилна информация, таксонометрични данни за микроби, за растения и за човека. Специфична е микробиалната база данни съчетана с TIGR-геномните проекти за секвенции и двоични бази данни, с човешкия идекс-проект за гени и подобни проекти за мишка, плъх, ориз и други биологични видове. Част от данните се обслужва от FTP-сайт, но могат да се намерят и на TIGR - home page.**

GSDB (Genome Sequence Data Base) **е база данни разработена от Националния Център за геномни ресурси, New Mexico. Тя съдържа, актуализира и разпределя една пълна колекция от ДНК-секвенции и съответната информация за геномните лаборатории. Оперира в режим "on-line", като клиент-сървър рационална база данни, предлагаща секвенции за широкомащабни проекти. Данните се проверяват по отношение на коректност на съдържанието и точност на разпределението.**

- The International Human Genome Sequencing Consortium published their results in Nature, 409(6822):860-921, 2001
  - ***Initial Sequencing and Analysis of the Human Genome***

- Celera Genomics published their results in Science, 291(5507), 1304-1351, 2001
  - ***The Sequence of the Human Genome***

DNA sequence databases

GenBank, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bldg. 38A, 8600 Rockville Pike, Bethesda, Maryland 20894, USA

EMBL, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

DNA Data Bank of Japan, Center for Information Biology, National Institute of Genetics, 1111 Yata, Mishima-shi, Shizuoka-ken 411-8540, Japan