КОМПЮТЪРНИ МОДЕЛИ

ПРОФ. ПЛАМЕНКА БОРОВСКА

КАТЕДРА ИНФОРМАТИКА

ТЕХНИЧЕСКИ УНИВЕРСИТЕТ – СОФИЯ

Протеини

- Протеините изобилстват в месото, рибата и зеленчуците
- Всички протеини са изградени от едни и същи основни градивни елементи, наречени *аминокиселини*.
- Аминокиселините са вече доста сложни органични молекули, съставени от *въглерод, водород, кислород, азот, и серни атоми*.
- Така цялостната рецепта за един протеин е от вида

$$C_{1200}H_{2400}O_{600}N_{300}S_{100}$$

Протеини

- Ранните години на биохимията са посветени на намирането на по-добър начин за представяне на протеините, за предпочитане чрез формула, която да обясни техните биологични (или дори хранителни) свойства.
- С времето биохимиците са открили, че *протеините са големи молекули (макромолекули), изградени от голям брой аминокиселини (обикновено от 100 до 500),* подбрани от селекция от 20 "flavors" с имена като аланин, глицин, тирозин, лутамин и т. н.
- В Таблица 1-1 е даден списък на тези 20 градивни елемента, с пълните им имена, трибуквени кодове, както и еднобуквени кодове (*код на IUPAC, създаден от Международният съюз по теоретична и приложна химия*).

Table 1-1	The 20 Amino Acids and Their Official Codes		
#	*1-Letter Code*	*3-Letter Code*	*Name*
1	A	Ala	Alanine
2	R	Arg	Arginine
3	N	Asn	Asparagine
4	D	Asp	Aspartic acid
5	C	Cys	Cysteine
6	Q	Gln	Glutamine
7	E	Glu	Glutamic acid
8	G	Gly	Glycine
9	H	His	Histidine
10	I	Ile	Iscleucine
11	L	Leu	Leucine
12	K	Lys	Lysine
13	M	Met	Methionine
14	F	Phe	Phenylalanine
15	P	Pro	Proline
16	S	Ser	Serine
17	T	Thr	Threonine
18	W	Trp	Tryptophan
19	Y	Tyr	Tyrosine
20	V	Val	Valine

Протеинови секвенции

- Даден вид протеин (като инсулина, напр.) винаги съдържа точно един и същи брой от всички аминокиселини (наричани *остатъци*), винаги в едно и също съотношение.
- По този начин, формулата за протеин изглежда така:

  insulin = (30 glycines + 44 alanines + 5 tyrosines + 14 glutamines + . . .).
- Аминокиселините са свързани заедно в една верига и функционалността на един протеин се определя не само от състава му, но също и от структурата му (реда на изграждащите го аминокиселини).

Протеинови секвенции

- Първата секвенция на аминокиселините в протеина на инсулина е определена през 1951 г.
- Действителната рецепта за човешки инсулин, от който произтичат всичките му биологични свойства, е следната верига:

  insulin =
  MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLV EALYLVCGERGFYTPKTRREAEDLQVGQVELGGGPG AGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN

  - *Анализът на протеинови секвенции (последователности) остава централна тема на биоинформатиката.*

История на анализа на секвенциите

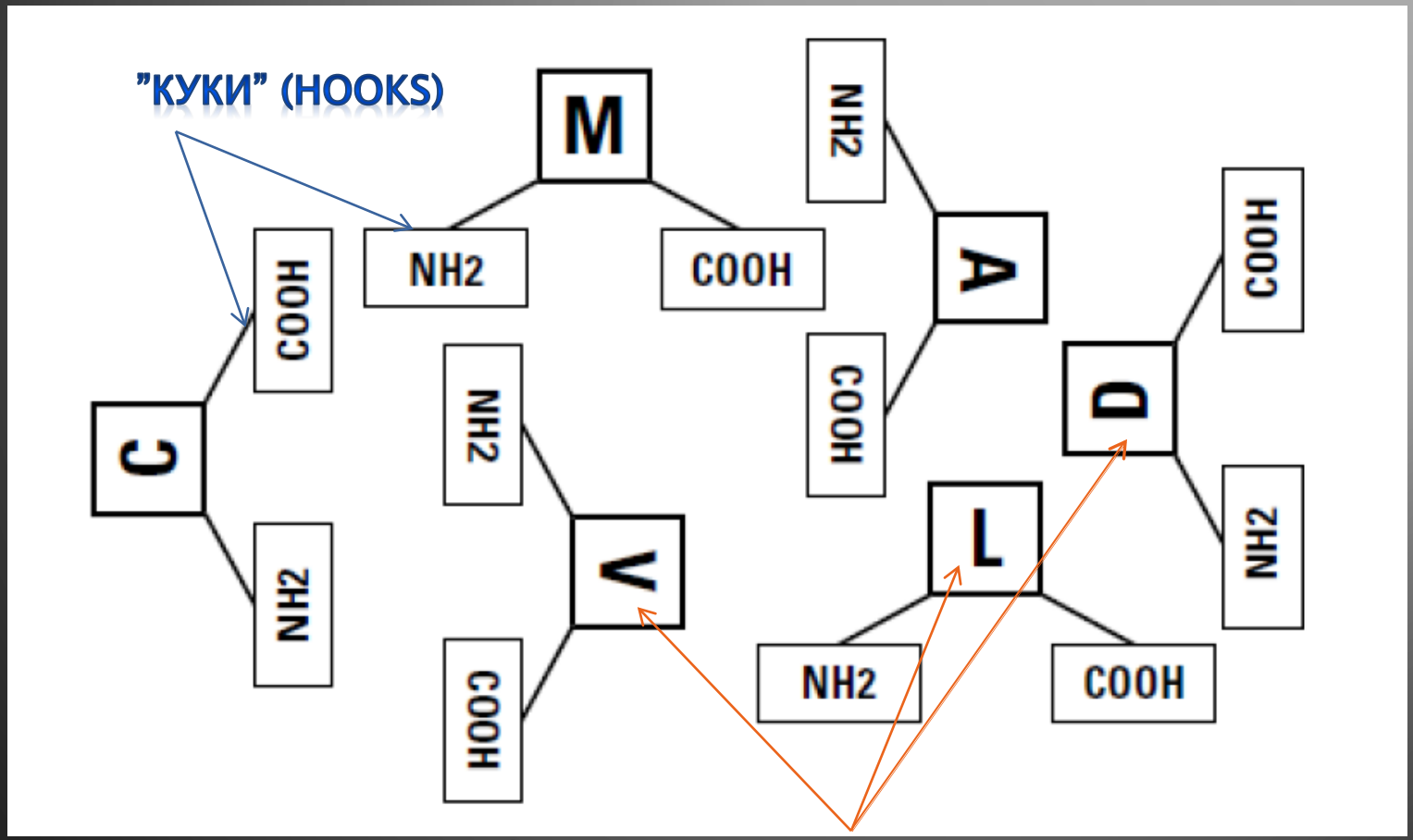
- Алфред Сангър печели първата си Нобелова награда за секвениране на инсулина и открива модерната епоха на молекулярната и структурната биология.
- Молекулните последователности са първите *основни набори от данни за биологията*.
- В началото на 1960-те години, протеиновите секвенции се натрупват бавно, те се събират, анализират и сравняват ръчно.

Разчитане на протеинови секвенции от N до C

- Двадесетте аминокиселини молекули, открити в протеините имат различни структури (*bodies*), но всички те имат една и съща двойка - NH2 и COOH.
- Тези групи от атоми се използват, за да формират т. нар. *пептидни връзки* между следващите остатъци в последователността.
- Протеиновата молекула сама по себе си се изгражда на основата на пептидни връзки
- Пептидна връзка CO-NH се формира, когато една свободна група NH2 се свърже химически с една група COOH

пептидни връзки
CO–NH

Свободни плуващи аминокиселини и техните "куки" за изграждане на пептидни връзки



Свободни плуващи аминокиселини

Разчитане на протеинови секвенции от N до C

- В резултат на този верижен процес (захващане на плуващите аминокиселини една с друга посредством техните куки) изградената протеинова молекула ще остане с неизползван NH2 в единия край и неизползван COOH в другия край.
- *Тези краища се наричат (съответно) N-край (N-terminus) и C-край (C-terminus) на протеиновата верига.*
- Протеиновата секвенция или протеиновият фрагмент се определят като *поредица от съставни аминокиселини, изброени подред от N-края до C-края.*
- Примерна протеинова последователност

MAVLD= Met-Ala-Val-Leu-Asp= Methionine–Alanine-Valine–Leucine-Aspartic

3-D структура на протеини

- Точната последователност на съставните а минокиселини на протеина определят протеиновата молекула.
- Свойствата на протеина (например, способността му за смилане на захар или да стане част от мускулните влакна) се определят не само от неговия състав, а съществено зависят от неговата пространствена (3-D) структура
  - Създадената протеинова молекула не е просто една верига от аминокиселини, а изключително гъвкав обект, компактна, добре пакетирана топка от нагънат стринг (well-bundled ball of string).

3-D СТРУКТУРА НА ПРОТЕИНИ

- *Крайната 3-D форма на протеиновата молекула е уникална и продиктувана от неговата последователност*,
- Съществуват видове аминокиселини, които са *хидрофобни* (например остатъци L, V, I) - не са на повърхността и не взаимодействат със заобикалящата ги вода, докато други са *хидрофилни* (например остатъци D, S, K), които активно търсят такава възможност.
  - Протеиновата верига отразява също така и други влияния, като например електрическите заряди, носени от някои от аминокиселините, или на тяхната възможност да се “поберат” с техните непосредствени съседи.

Логическата връзка Секвенция ⇨ Структура ⇨ Функция

- *Първата 3-D структура на протеин е определена през 1958 г. от Kendrew и Perutz, използвайки сложна техника на рентгеновата кристалография.*
- Освен, че са спечелили още една Нобелова награда в зараждащата се област на молекулярната биология, с това постижение лекарите разбират, че протеините са точни и специфични форми, кодирани в последователността на аминокиселините.
- Следователно, те *прогнозират, че протеини с подобни секвенции ще се нагънат в подобни пространствени структури и, обратно, че протеини с подобни структури ще бъдат кодирани с подобни секвенции от аминокиселини*.
- *Функцията на един протеин се оказва пряка последица от неговата 3-D структура*.
  - *Логическата връзка Секвенция ⇨ Структура ⇨ Функция понастоящем е централна концепция на молекулярната биология и биоинформатиката.*

Структурна биоинформатика

- Проиграването на компютърни модели на протеиновата структура и визуализацията им на монитора на компютъра е, разбира се, много по-лесно от манипулирането на 3-D пъзел от около хиляда части.
- В резултат на това все по-голям *дял от биоинформатиката е посветен на развитието на кибер-инструменти за навигация между секвенции и 3-D структури* - *структурна биоинформатика*.
- Визуализация на протеини в Интернет

Типична протеинова 3-D (схематична) структура от 400 аминокиселини

Въпреки тяхната голяма сложност, протеиновите молекули са доста малки. Една единствена бактерия е изградена от хиляди различни протеини, всеки от тях в хиляди копия - повече от достатъчно доказателство, че живите организми не са прости!

Анализ на ДНК секвенции

- През 1950-те години, докато учени като Kendrew и Perutz все още се борят за определяне на първите 3-D структури на протеини, други биолози вече получават много косвени доказателства (чрез генетични експерименти), че *дезоксирибонуклеиновата киселина (ДНК) - вещество, което създава нашите гени - е също голяма макромолекула.*
- *Тази молекула е като дълга верига, усукана в двойна спирала и всяка връзка във веригата е сдвояване на две от четирите съставни части, наречени нуклеотиди.*
  - *Нуклеотидът* се състои от фосфатна група, свързана с монозахариден остатък, които се свързани с една от *4 вида азотсъдържащи органични основи, символизирана от четирите букви A, C, G, и T.*

Анализ на ДНК секвенции

- Молекулярните биолози, обаче, трябва да изчакат до 1970-те години, за да може да се определят секвенциите на ДНК молекулите и да получат непосредствен достъп до секвенциите от нуклеотидите на гена
- *Това е революция (А. Сангър печели втората си Нобелова награда!), защото малката азбука на ДНК секвенциите (4 нуклеотида, в сравнение с 20-те аминокиселини) позволява много по-просто и по-бързо разчитане и бързо се достига до пълна автоматизация.*
- В момента в световен мащаб определянето на ДНК секвенции е по-бързо (с порядъци) в сравнение с темпа на секвенирането на протеини.

Четене на ДНК секвенции по правилния начин

- Аналогично на 20-те аминокиселини в протеините, *4-те нуклеотида на ДНК имат различни “тела”, но всички те имат един и същ чифт “куки” : 5‘ фосфорилна и 3’ хидроксилна (произнася се пет-прим и три-прим)*, в зависимост от тяхната позиция в дезоксирибозната захарна молекула, която е част от механизма за изграждане на веригата нуклеотиди.
- Молекулата на ДНК се изгражда на базата на формирането на връзки между позициите 5’ и 3’ на съставните нуклеотиди.
- След като се свържат нуклеотидите, в резултатната ДНК има неизползвана фосфорилна група (PO4) в 5‘ края и неизползвана хидроксилна група (OH) в 3‘ края .
  - Тези краища са съответно наречени *5'-край (5'–terminus) и 3'-край (3'–terminus) на веригата ДНК*.

ДНК секвенции

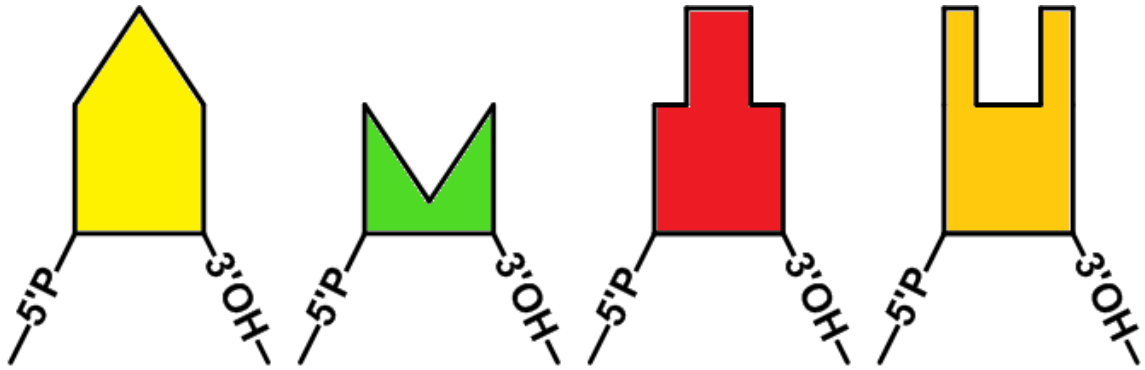
The IUPAC code for DNA sequences

The following table lists the one-letter codes (IUPAC codes) used to work with DNA sequences. Official IUPAC codes, from the International Union of Pure and Applied Chemistry, are defined for all possible two- and three-way ambiguities. The table shows only the ones most frequently used.

Most Common Letters Used for DNA Nucleotide Sequences

1-Letter Code	*Nucleotide Name*	*Category*
A	Adenine	Purine
C	Cytosine	Pyrimidine
G	Guanine	Purine
T	Thymine	Pyrimidine
N	Any nucleotide (any base)	(n/a)
R	A or G	Purine
Y	C or T	Pyrimidine
--	-----	None (gap)

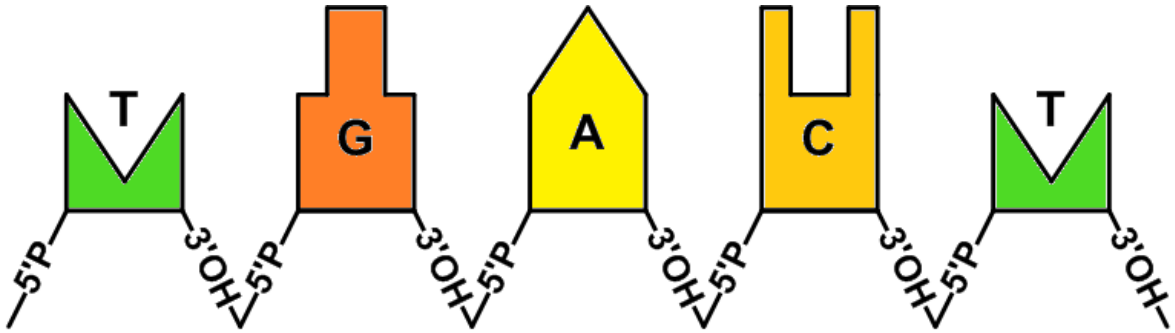
4 нуклеотида, изграждащи ДНК



4 вида азотсъдържащи органични бази, означавани с четирите букви A, C, G, и T

Adenine-Cytosine-Guanine-Thymine

Верига нуклеотиди, изграждащи нишките ДНК



Последователността на показаната (къса!) ДНК е **TGACT = Thymine-Guanine-Adenine-Cytosine-Thymine**

Анализ на биологични секвенции

- *ДНК и РНК са изградени от 4 нуклеотидни бази.*
- Три от тези бази са еднакви при ДНК и РНК: *гуанин (G), аденин (A) и цитозин (C)*.
- Четвъртата база е различна - за ДНК е *тимин (T)*, докато при РНК в четвъртата база липсва метилова група и се нарича *урацил (U)*.
- Всяка база има две точки, които могат да се присъединят ковалентно към две други бази в двата края, образувайки линейна верига от мономери.
- Тези вериги могат да бъдат доста дълги, с много милиони бази, често срещани при повечето организми.

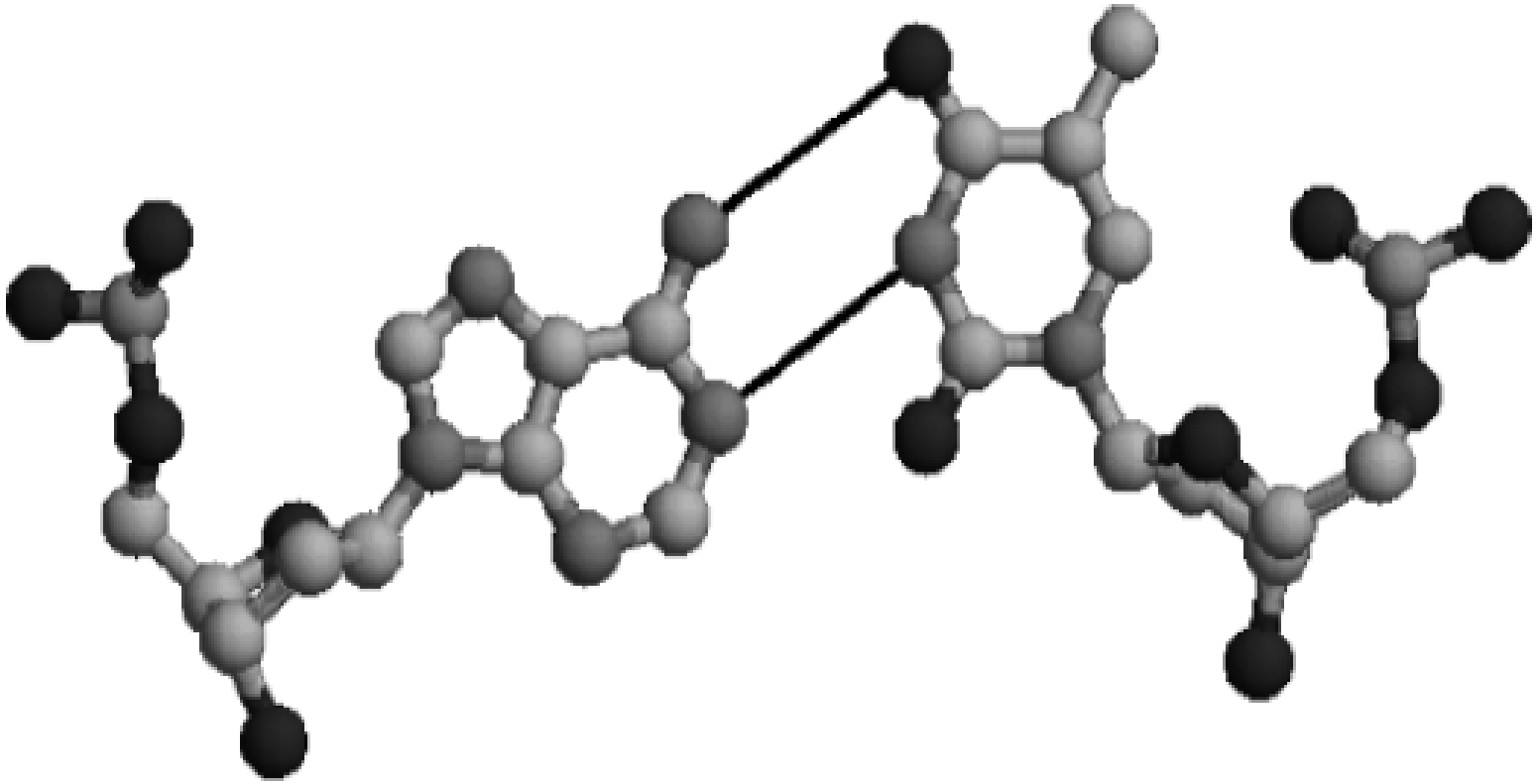
Анализ на биологични секвенции

- Друга интересна особеност на нуклеотидните бази е, че четирите бази *се свързват в две ексклузивни двойки поради наличието на заредени атоми в краищата им.*
- Три от тези връзки се формират между C и G, докато две се формират между A и T (или A и U за РНК).
- Тези връзки, които са значително по-слаби от ковалентните връзки между атомите, са достатъчни за стабилизиране на структура като известната *двойна спирала* (double helix), при която базите се подреждат в линия почти перпендикулярна на оста на спиралата.

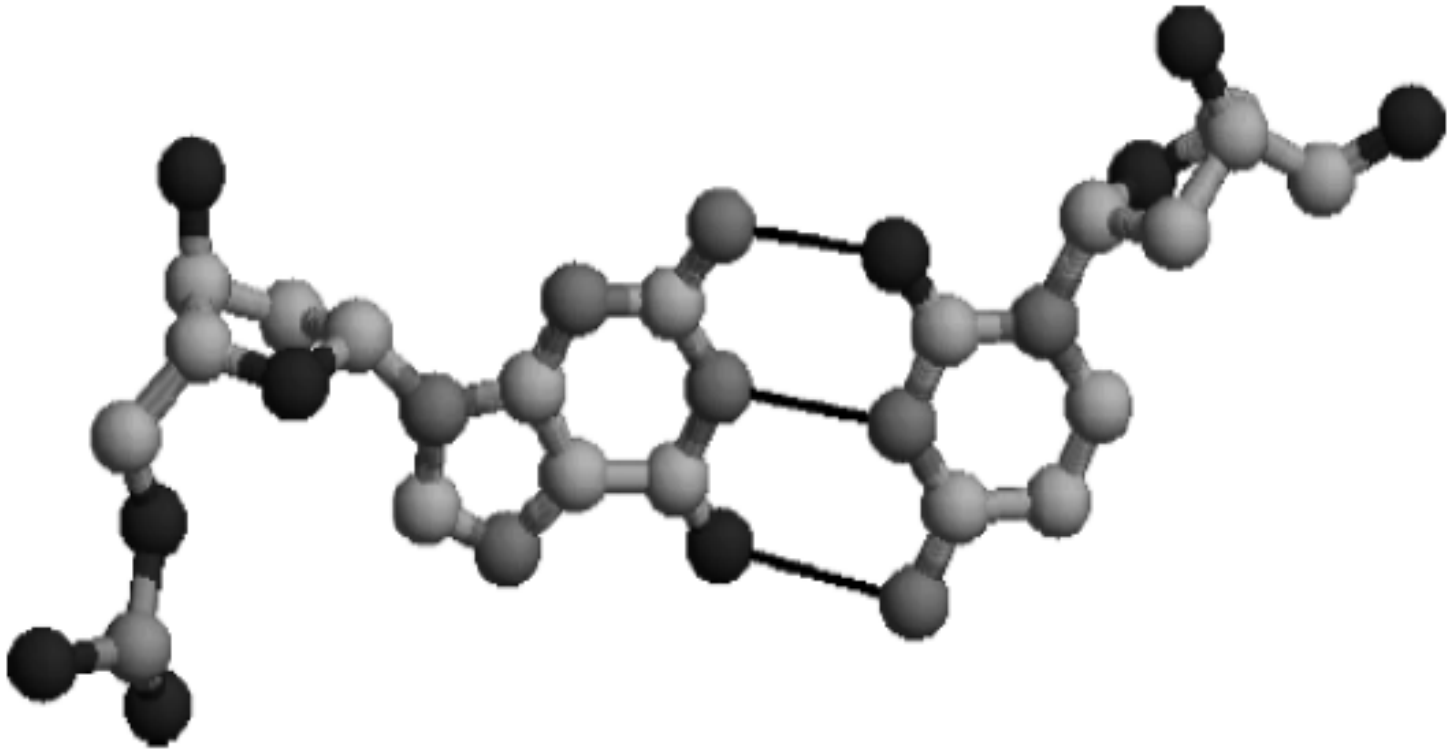
Важни последствия от двойната спирала

- Когато е налице базата G в едната верига, на съответното място в другата верига на двойната спиралата е базата C и за двете вериги се казва, че се *допълват една друга*. Веригите често се наричат нишки.
- Това допълване означава, че информацията е дублирана, т.е. съхранява се в двете вериги; следователно, е необходимо само една верига да се съхранява цялата информация.
- Поради спецификата на структурата на нуклеотидните бази, ДНК молекулите имат *посока*. *Фосфатният "гръбнак" на двойната спирала е прикачен към захарните пръстени на различни места: хидроксилните групи 3'и 5'.*

Нуклеотидните бази Аденин (А) и тимин (Т) (тънките черни линии показват трите водородни връзки между двете бази)



Нуклеотидните бази Гуанин (G) и цитозин (C) (тънките черни линии показват трите водородни връзки между двете бази)



Анализ на биологични секвенции

- ДНК веригите на двойната спирала са ориентирани в противоположни посоки, съответните краища на спиралата са *3’ terminus* на едната верига и *5’ terminus* на другата верига.
- При записване на секвенцията се започва с нуклеотидите от *5’ terminus* ("най-левия" нуклеотид) на молекулата на ДНК до *3’ terminus* ("най-дясния" нуклеотид).
- *При вертикално ориентираната двойна спирала веригата, започваща от 3‘ -края се нарича възходяща, а веригата, започваща от 5' -края веригата — низходяща.*

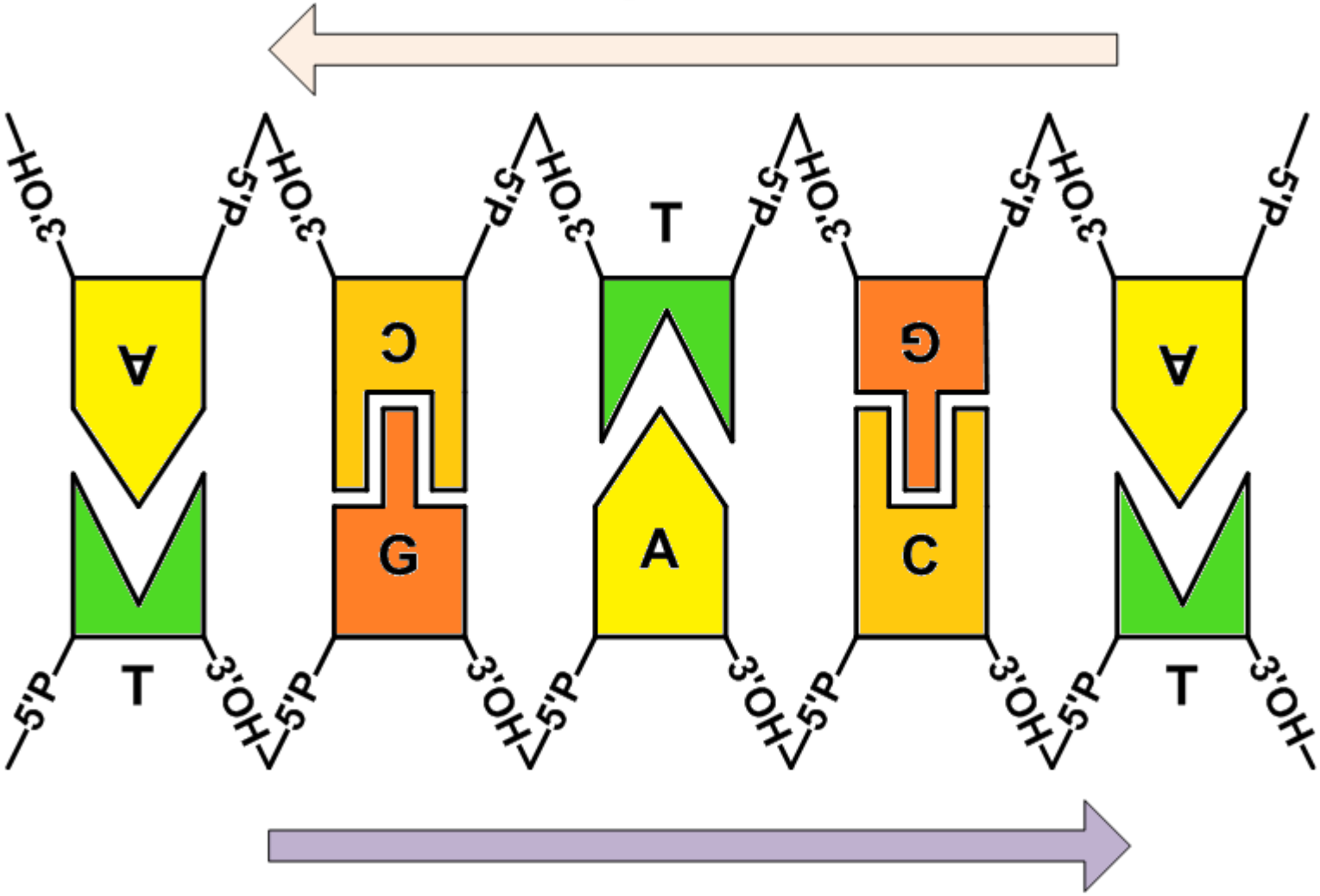
ДВЕТЕ СТРАНИ НА ДНК СЕКВЕНЦИИТЕ

- В същата лаборатория, в която Kendrew и Perutz се опитват да разберат първата 3-D структурата на един протеин, *Watson и Crick изясняват през 1953г. известната двойна спирална структура на молекулата на ДНК.*
- Но това, което прави това откритие толкова важно, спечелило Нобелова награда за молекулярна биология, не е спиралната форма, *а откритието, че молекулата на ДНК се състои от две допълващи се нишки.*

ДВЕТЕ СТРАНИ НА ДНК СЕКВЕНЦИИТЕ

- *Под взаимно допълване се има предвид, че тимин (T) на едната нишка е винаги свързан с аденин (A) (и обратното) и гуанин (G) винаги се свързва с цитозин (C).*
- *Тези двойки A-T и G-C, въпреки че не са свързани чрез химична връзка, имат стриктна едно-към-едно реципрочна връзка.*
- Когато се знае последователността на нуклеотидите в рамките на една нишка, може автоматично да се изведе последователността на другата.
- Това невероятно свойство, както и спиралната структура са крайъгълните камъни, за разкритието на всичко за ДНК секвенциите.
- Например, когато живите организми се възпроизвеждат, всеки от техните гени трябва да бъде дублиран (точно копие).
- За тази цел, природата разделя веригите на ДНК и прави *две взаимно допълващи се нишки, благодарение на магическата двустранна структура на ДНК молекулата.*

Двете взаимно допълващи се нишки на молекулата на ДНК



Двете страни на ДНК секвенциите

- Двойната структура на ДНК прави неясно дефинирането на секвенцията: дори при спазване на правилото за четене на нуклеотидите от края 5’ до края 3’, има възможност за избор на едната или другата нишка.
- Така, за всяка позиция,една и съща молекула ДНК съответства на две — тотално различни— секвенции, свързани с комплементарна релация. Но, този проблем не е толкова сложен, просто трябва да се има предвид всеки път, когато се анализират ДНК сеявенции.
- Повечето програми за работа с биологични бази данни като BLAST, вземат предвид това свойство и анализират и двете секвенции при формиране на резултата от търсенето.
- В случаите, когато и двете нишки имат значение, винаги трябва да се прави пълен анализ.

Палиндроми в ДНК секвенциите

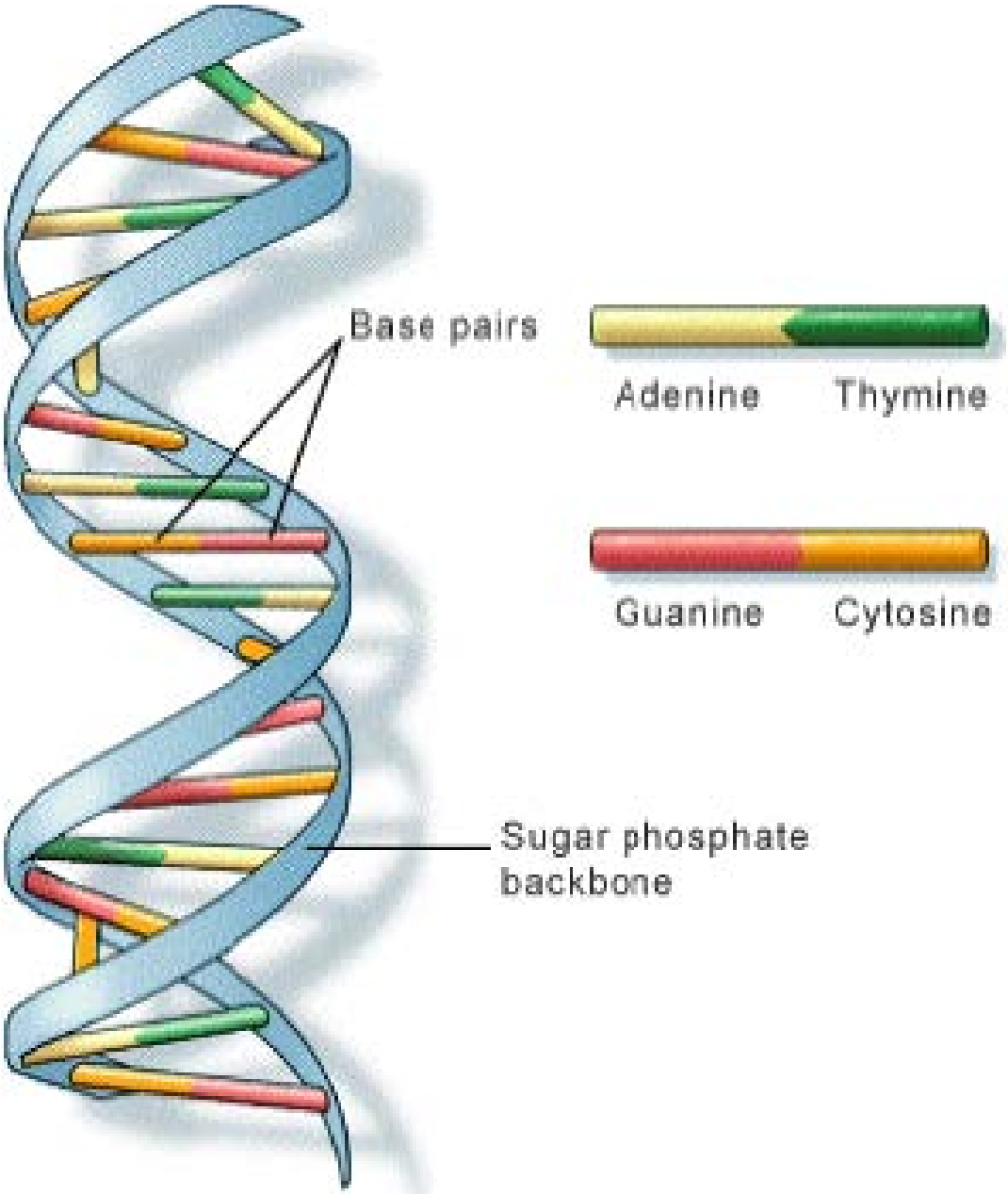
- Начинаещите в анализа на ДНК последователности обикновено са объркани от идеята на обратните допълващи се последователности.
- Двете поредици

ATGCTGATCTTGGCCATCAATG и

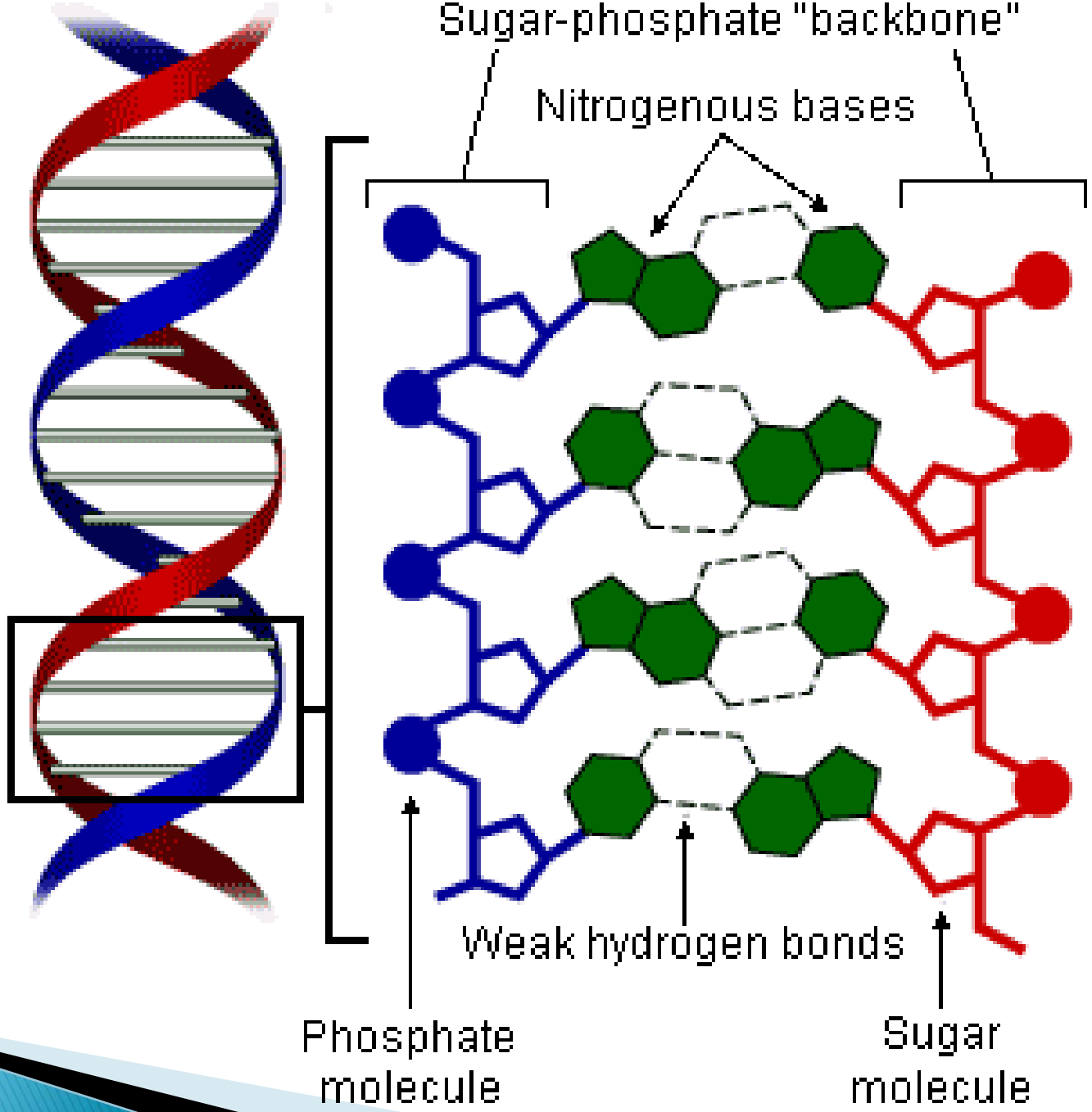
CATTGATGGCCAAGATCAGCAT

съответстват на една и съща ДНК молекула.

- Едно важно свойство на взаимното допълване на ДНК е фактът, че областите на ДНК могат да съответстват на последователности, които са идентични, когато се четат от две допълващи се нишки.



U.S. National Library of Medicine


Sugar-phosphate "backbone"
Nitrogenous bases
Weak hydrogen bonds
Phosphate molecule
Sugar molecule

The DNA structure is a double helix looking something like a twisted ladder.

The two cross-linked helices are always right-handed, twisting in the same direction as a normal screw.

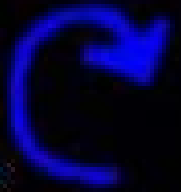
LEFT HANDED

RIGHT HANDED

Names for clockwise and anticlockwise helices:

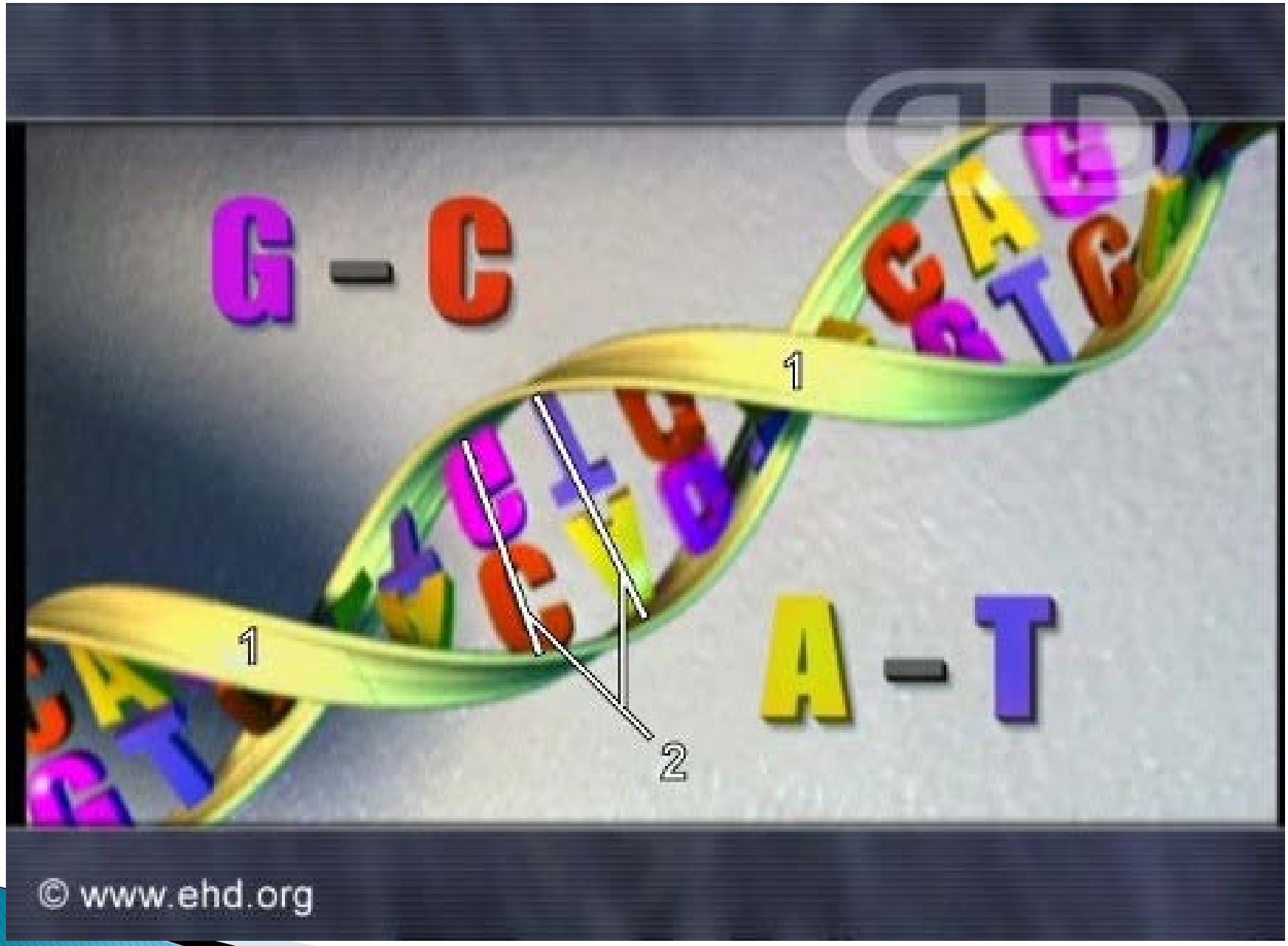
ANTICLOCKWISE

Away from your eye

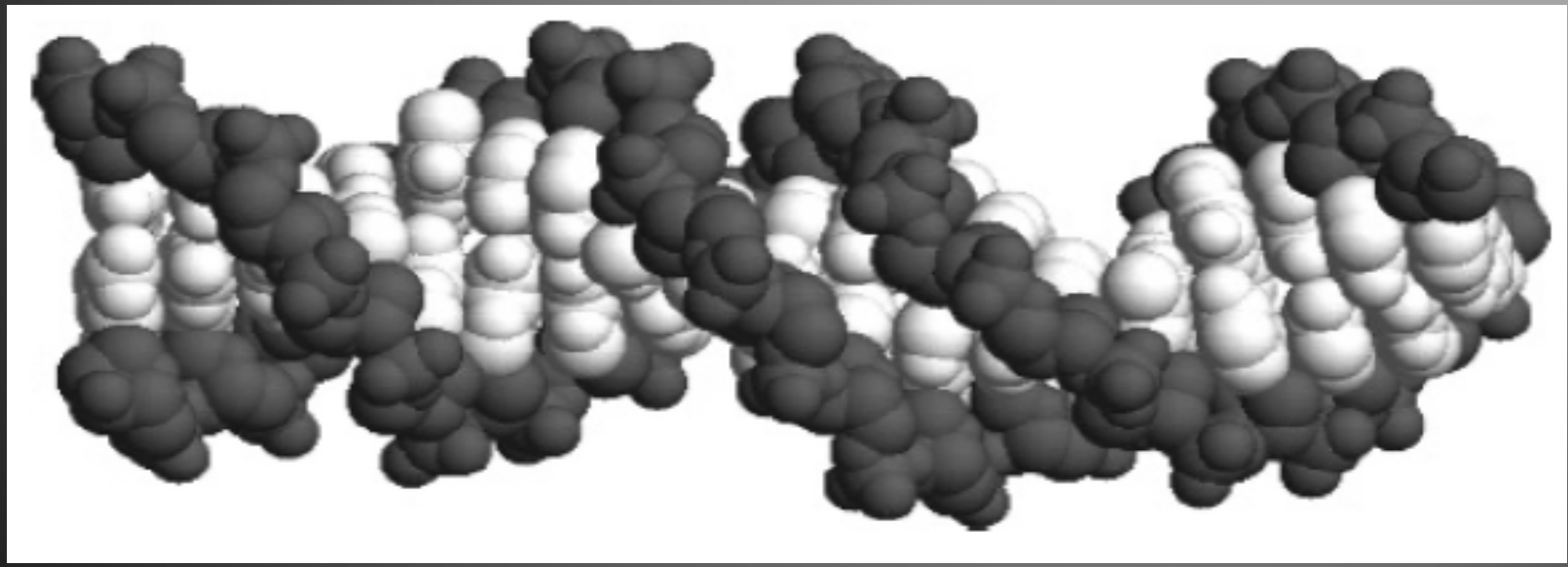


CLOCKWISE

Away from your eye



Двойната спирала на ДНК



Обобщени модели и тяхното използване

- Отношенията между ДНК, РНК, протеини, структура и функция следват един обобщен модел. За съжаление, както повечето обобщения, той е крайно опростен за много ситуации.
- Биоинформатиците по принцип се занимават с информацията на по-абстрактно ниво: ДНК, РНК и аминокиселинните секвенции са "само" низове от букви.
- Понякога е лесно да се забрави, че това са реалните изображения на молекулите, които съществуват в клетъчния свят и следователно, трябва да си *взаимодействат с физическата среда като цяло, да не говорим за съществуването им в клетъчна среда.*
- *Колко трябва да знаят биоинформатиците за контекста на реалния свят на данните*, които анализират, зависи от анализа, който се извършва.
- В някои случаи са достатъчни доста повърхностни знания, докато други изискват по-дълбоко разбиране на основните физични и биологични процеси по време на работа.

РАМКИ ЗА ЧЕТЕНЕ

- Ефектът от избора на коректна и некоректна рамка за четене може да бъде изследван с използване на средството *Transeq, включено в пакета от програми EMBOSS* .
- Интерфейс на *Transeq, предоставен от EBI*

  http://www.ebi.ac.uk/emboss/transeq/