Follow-up of survey of lecture 1

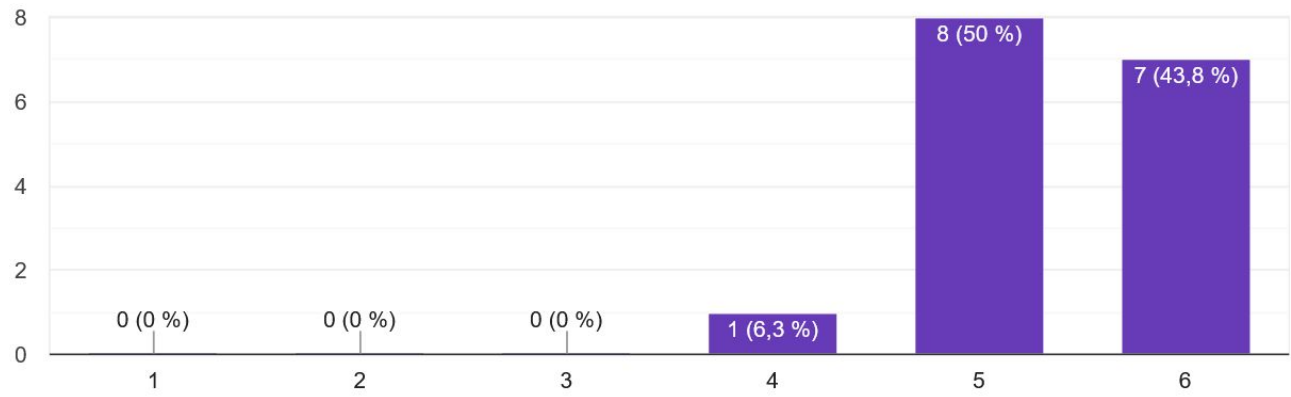
We have 16 and 13 replies to the survey and to the offline activity form by Oct 1st 2021, respectively.
Thank you! Thanks to your feedback, I will

- Leave more time for discussions
- Keep the evaluation transparent: both asking questions and joining discussions count.

Please keep giving feedbacks!

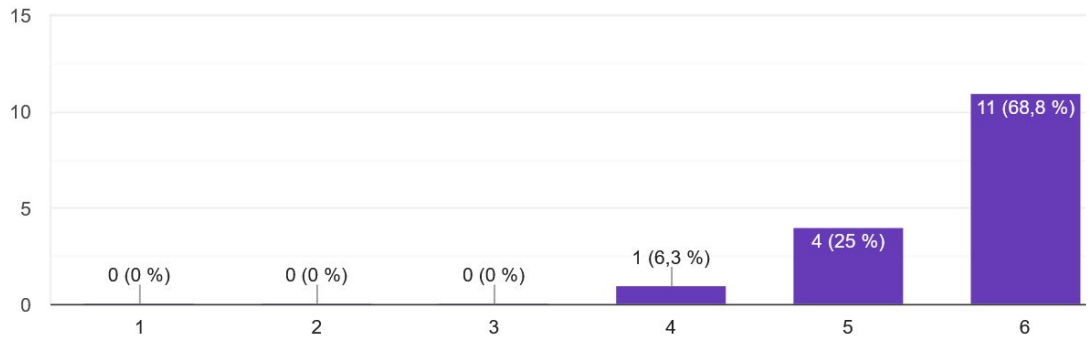
How was your overall impression of today's lecture?

16 Antworten



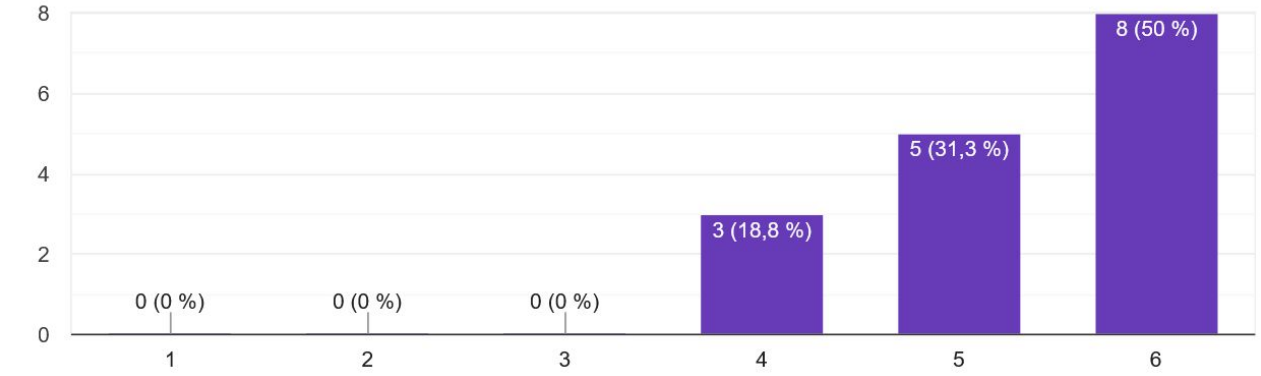
How did you experience the interactions between your peers and David, and among the peers?

16 Antworten



How well could you understand and follow David (the lecturer)?

16 Antworten



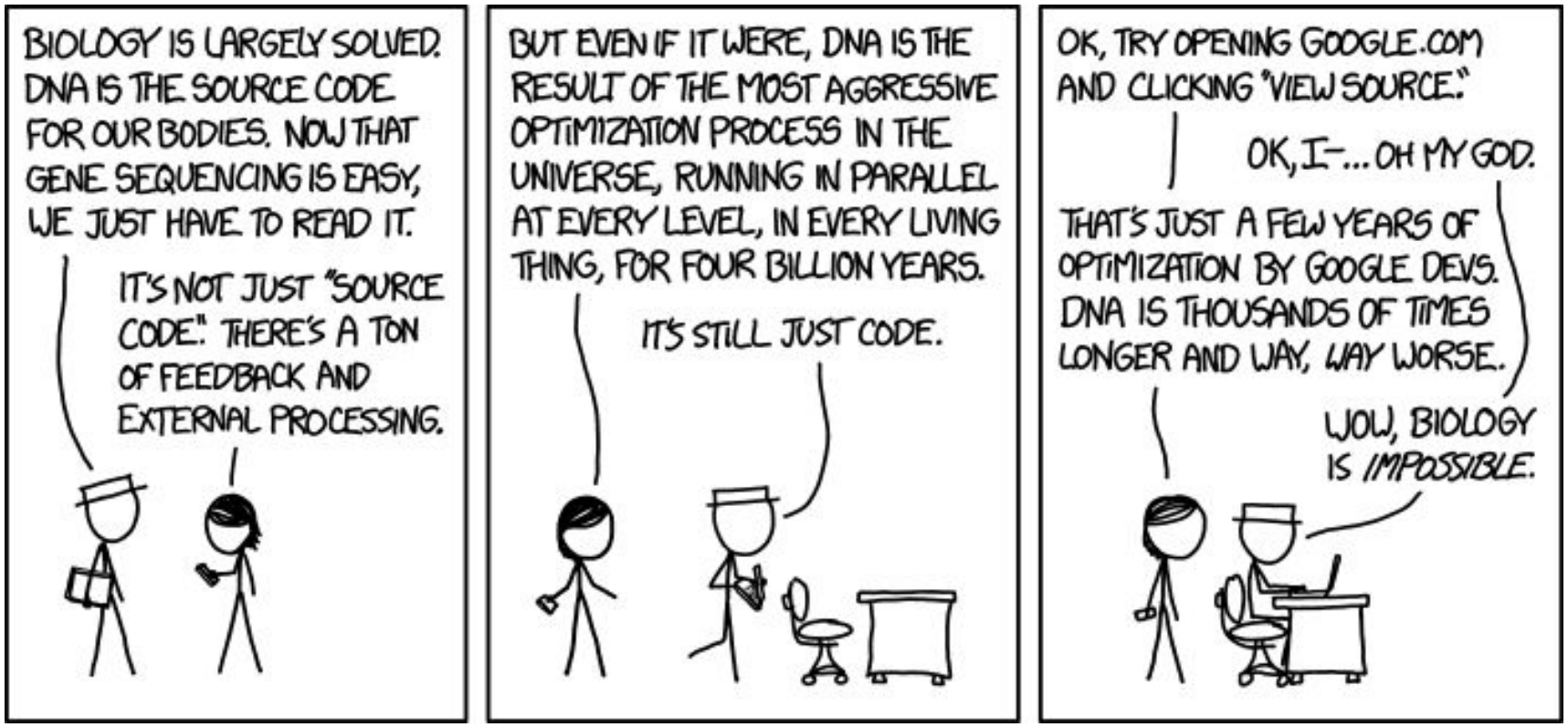
Follow up of questions on the video on Herceptin by Susan Desmond-Hellmann

Link to the video

Questions for the video

1. What is the **indication** of *Herceptin*? (Her2 positive breast cancer) What is its generic (USAN, or United States Adopted Name) name? (Trastuzumab)
2. What is the **gene target** of Herceptin? (Her2, ERBB2)
3. In which year was the **target** of Herceptin described? When was Herceptin **approved**? (1987; 1998 in metastatic cancer and 2005 in the adjuvant setting)
4. What was the **improvement** of Herceptin compared with earlier antibodies? (humanized)
5. Why does a **biomarker** matter besides developing drugs? (diagnostic, higher chance of success due to patient stratification)
6. In the clinical trial of *Herceptin* for **metastatic breast cancer**, how much improvement in the **median survival** did Herceptin achieve? And how much improvement is in the **adjuvant setting** (Herceptin applied directly after operation)? (5.1 months improvement in median survival for metastatic breast cancer. Time to remission doubled in the adjuvant setting)

AMIDD Lecture 2: The Central Dogma and Drug Discovery



DNA by Randall Munroe, https://xkcd.com/1605/

Dr. Jitao David Zhang, Computational Biologist

[1] Pharmaceutical Sciences, Pharma Research and Early Development, Roche Innovation Center Basel, F. Hoffmann-La Roche

[2] Department of Mathematics and Informatics, University of Basel

Google.com, visited by ~200 million people every day

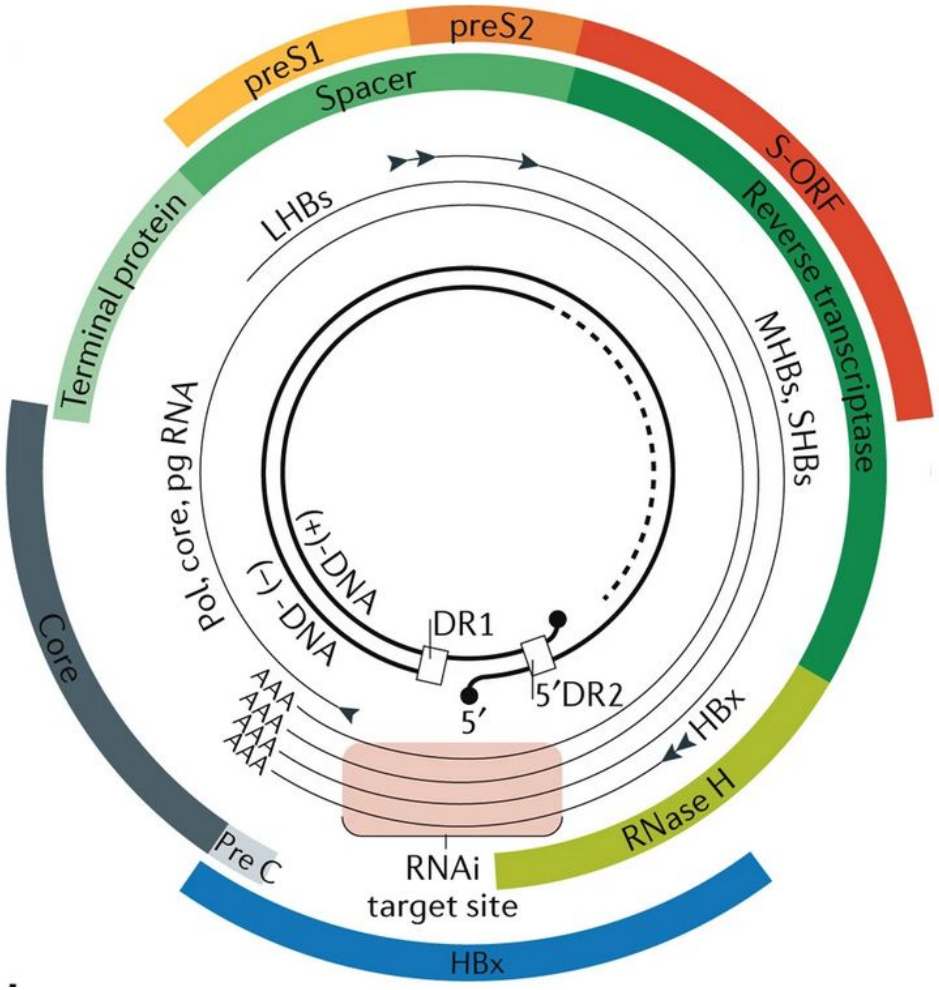
As of 28.09.2021


```
<!doctype html><html itemscope="" itemtype="http://schema.org/WebPage" lang="de-CH"><head><meta charset="U
var f=this||self;var h,k=[];function l(a){for(var b;a&&(!a.getAttribute||!(b=a.getAttribute("eid")));)a=a.
function n(a,b,c,d,g){var e="";c||-1!==b.search("&ei=")||(e="&ei="+l(d),-1===b.search("&lei=")&&(d=m(d))&&
google.y={};google.sy=[];google.x=function(a,b){if(a)var c=a.id;else{do c=Math.random();while(google.y[c])
document.documentElement.addEventListener("submit",function(b){var a;if(a=b.target){var c=a.getAttribute("
var e=this||self;var g=window.performance;google.timers={};google.startTick=function(a){google.timers[a]={
google.rll=function(a,b,c){function d(f){c(f);k(a,"load",d);k(a,"error",d)}h(a,"load",d);b&&h(a,"error",d)
function t(a){r(a.timeStamp)&&k(document,"visibilitychange",t,!0)}google.c.wve&&(google.c.fh=Infinity,h(do
function l(){return window.performance&&window.performance.navigation&&window.performance.navigation.type}
function r(a){return"none"===a.style.display?!0:document.defaultView&&document.defaultView.getComputedStyl
function u(a,b){var c=b(a);a=c.left+window.pageXOffset;b=c.top+window.pageYOffset;var d=c.width;c=c.height
function O(){if(!J){var a=F===E,b=D===C,c=I===H;c=google.c.nli?c:a;if(a&&b){google.c.e("load","ima",String
google.aftq)||void 0===B?void 0:B[b++];)try{c()}catch(R){google.ml(R,!1)}google.aftq=null}}}var Q="src bsr
function U(a){var b=a.parentElement;if(google.c.gip&&b&&"G-IMG"===b.tagName&&(b.style.height||b.style.widt
var b=[function(){google.tick&&google.tick("load","dcl")}];google.dclc=function(a){b.length?b.push(a):a()}
var b=[];google.jsc={xx:b,x:function(a){b.push(a)},mm:[],m:function(a){google.jsc.mm.length||(google.jsc.m
var e=this||self;

var g={};function u(a,b){if(null===b)return!1;if("contains"in a&&1==b.nodeType)return a.contains(b);if("co
var H=e._jsa||{};e._jsa=H;H._cfc=void 0;H._aeh=void 0;var I=function(){this.h=this.g=null},K=function(a,b)
var Q=function(){this.s=[];this.g=[];this.h=[];this.l={};this.i=null;this.j=[];N(this,"_custom")},R=functi
"click"!=h.eventType&&"clickmod"!=h.eventType||(c.preventDefault?c.preventDefault():c.returnValue=!1),(c=a
function _F_installCss(c){}
</script>  <script defer="" src="/xjs/_/js/k=xjs.s.de_CH.Wq5KIF2zgVQ.O/am=QIACAAQAAAAAAAAACAAoAEgEBgIAAADM
window.rwt=function(){return!0};}).call(this);(function(){
window.jsarwt=function(){return!1};}).call(this);(function(){window.google.erd={sp:'hp',jsr:0,bv:1449,sd:t
try{
/*

 Copyright The Closure Library Authors.
 SPDX-License-Identifier: Apache-2.0
*/
var ja,ma,oa,pa,qa,ra,sa,ta,va,wa,Aa,Ba,Ka,La,Na,Oa,Pa;_.aa=function(a){if(Error.captureStackTrace)Error.c
ja=function(a,b,c){return"object"===typeof a?_.ha&&!Array.isArray(a)&&a instanceof Uint8Array?c(a):_.ia(a,
_.n=function(a,b){return null!=a?!!a:!!b};_.p=function(a,b){void 0==b&&(b="");return null!=a?a:b};_.na=fun
qa=function(a){a=["object"==typeof globalThis&&globalThis,a,"object"==typeof window&&window,"object"==type
sa("Symbol",function(a){if(a)return a;var b=function(f,g){this.j=f;pa(this,"description",{configurable:!0,
sa("Symbol.iterator",function(a){if(a)return a;a=Symbol("c");for(var b="Array Int8Array Uint8Array Uint8Cl
_.ua=function(a){var b="undefined"!=typeof Symbol&&Symbol.iterator&&a[Symbol.iterator];return b?b.call(a):
_.q=function(a,b){a.prototype=va(b.prototype);a.prototype.constructor=a;if(Aa)Aa(a,b);else for(var c in b)
sa("WeakMap",function(a){function b(){}function c(l){var m=typeof l;return"object"===m&&null!==l||"functio
var f="$jscomp_hidden_"+Math.random();e("freeze");e("preventExtensions");e("seal");var g=0,k=function(l){t
Ba(l,f)&&Ba(l[f],this.j)?delete l[f][this.j]:!1};return k});
```

~15k characters

Hepatitis B virus, affetcing ~290 million people every day

Genotype B, 3.2kb



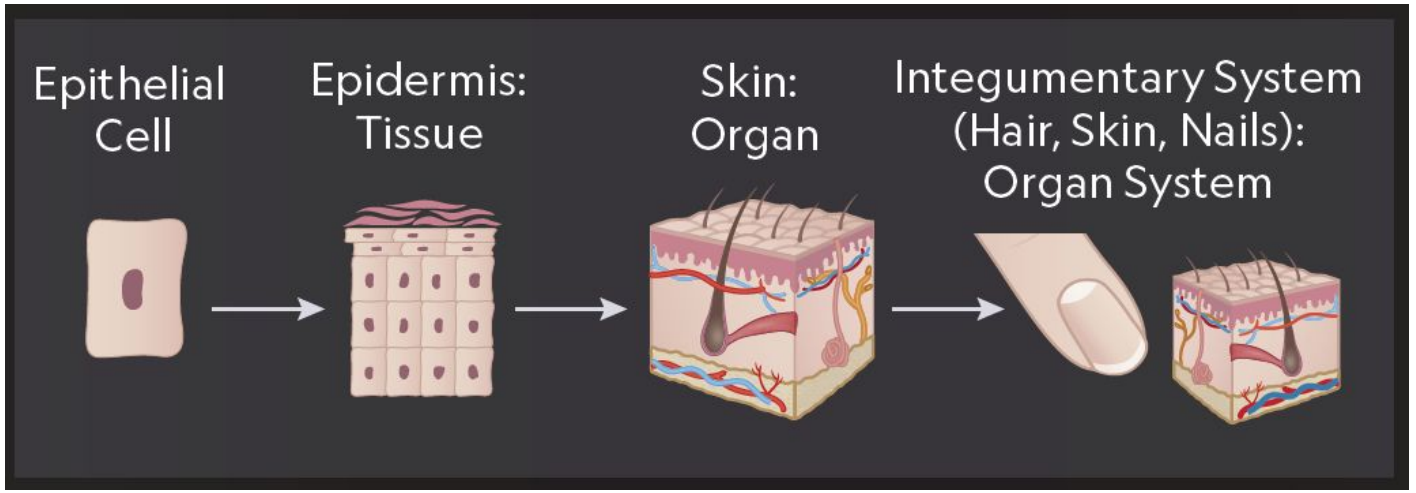
Cornberg, Markus, and Michael P. Manns. 2018. “No Cure for Hepatitis B and D without Targeting Integrated Viral DNA?” Nature Reviews Gastroenterology & Hepatology 15 (4): 195–96. https://doi.org/10.1038/nrgastro.2017.185.

tttcacagctttccaacaagccctacaagatcccagagtcaggggcatatatcttcctgctggtggctccagttcaggaa
cactcaaccctgttccaaatattgcctctcacatctcgtcaatctcctcgaggactggggaccctgcgttgaacatggag
aacatcacatcaggattcctaggacccctgctcgtgttacaggcggggtttttcttgttgacaagaatcctcacaataccc
gcagagtctagactcgtggtggacttctctcaattttctagggttccacccgtgtgtcgtggccaaaattcgcagtccc
caacctccaatcactcaccaacctcctgtcctccaatttgtcctggttatcgcttgatgtgtctgcggcgttttatcata
ttcctcttcatcctgctgctatgcctcatcttcttgttggttcttctggattaccaaggtatgttgcccgtttgtcctct
aattccaggatccacaacaaccagtacgggaccctgcaaaacctgcacgactcctgctcaaggcaactctatgtttccct
cctgttgctgtacaaaacctacggatggaaattgcacctgtattcccatcccatcatcttgggctttcgtaaaatacctа
tgggagtgggcctcagtccgtttctcttggctcagtttactagtgccatttgttcagtggttcgtagggctttcccccac
tgtttggctttcagttatatggatgatgtggtattgggggccaaatctgtacaacatcttgagtccctttataccgctgt
taccaattttcttttgtctttgggtatacatttaaaccctaacaaaacaaagagatggggttattccctaaacttcatgg
gatacgtaattggaagttggggtacattgccacaggatcatattgtacaaaaaatcaaacactgttttaggaaacttcct
gtcaatcgacctattgattggaaagtatgtcaaagaattgtgggtcttttgggctttgccgctccatttacacaatgtgg
ttaccctgccttaatgcctttgtatgcatgtatacaagcgaaacaggcttttactttctcgccaacttacaaggcctttc
taagtaaacagtatatgaacctttaccccgttgcccggcaacggcctggtctgtgccaagtgtttgctgacgcaacccccc
actggttggggcttggctatcggccatcagcgcatgcgtggaacctttgtggctcctctgccgatccatactgcggaact
cctagctgcttgttttgctcgcagccggtctggagcgaaactcattgggactgataattctgtcgtcctttctcggaaat
atacatcatttccatggctgctaggttgtgctgccaactggattcttcgcggaacgtcctttgtttacgtcccgtcggcg
ctgaatcccgcggacgacccctcccggggccgcttgggactctatcgtcccctctccgtctgccgtaccgtccgaccac
ggggcgcacctctctttacgcggtctccccgtctgtgccttctcatctgccggtccgtgtgcacttcgcttcacctctgc
acgttgcatggagaccaccgtgaacgcccatcagagcctgcccaaggtcttacataagaggactcttggactcccagcaa
tgtcaacgaccgaccttgaggcctacttcaaagactgtgtgtttaaagactgggaggagttgggggaggagattaggtta
atgatctttgtattaggaggctgtaggcataaattggtctgcgcaccatcatcatgcaacttttcacctctgcctaatc
atctcttgtacatgtcccactattcaagcctccaagctgtgccttggctggctttggggcatggacattgacccttataa
agaatttggagctagtgtggagttactctcgtttttgccttctgacttctttccttcagtccgggatctacttgatacag
cctcagctctgtatcgggaggccttagagtctccggagcattgctcacctcaccatacagcactcaggcaagccattctc
tgctgggtggaattaacgactctagctacctgggtgggtaataatttggaagatcatgcatccagggacctagtagtcaa
ttatgtaaatgataatatgggactaaagctcagacaactattgtggtttcatatttcttgccttacttttggaaaacaaa
ctgtccttgagtatttggtctccttcggagtgtggattcgcactcctccagcctatcgaccaccaaatgcccctatctta
tcaacacttccggaaactactgttgttagacgaagagaccgaggcgggtcccctagaagaagaactccctcgcctcgcag
acgaagatctcaatcgccgcgtcgcagaagatctcaatctcgggaatctcaatgttagtattccttggactcataaggtg
ggaaattttgttgggctttattcctctactgtccctatctttaatcctgaatggcaaactccttcctttcctaagatcca
tttacatgaggacattattaataggtgtcagcaatttgtaggccctctcactgtaaatgaaaaaagaagattgaaattaa
ttatgcctgctagattttatcctaaccgcactaaatatttgcctctagacaaagggattaaaccttattattctgatcaa
gtagttaatcattacttccagacccgacattatttacatactctttggaaggctgggattctatataagagggaaactac
acgtagcgcctcattttgcgggtcaccatattcttgggaacaagagctacatcatgggaggttggttaccaaaacctcgc
aaaggcatggggacgaatctttctgttcccaaccctctgggattctttcccgatcatcagttggaccctgcattcggagc
caactcaaacaatccagactgggacttcaaccccatcaaggaccgctggccacaagccaaccaggtaggagtgggagcgt
tcggcccagggttcactcccccacacggaggtgttttggggtggaaccctcaggctcagggcatattgactacagtgcca
gcagttcctcctcctgcctccaccaatcggcagtcagggaggcagcctactcccatctctccacctctaagagacagtca
tcctcaggccgtgcagtggaa

Today's goals

- The central dogma of molecular biology
- Applications of biological sequence analysis in drug discovery
  - Deciphering encoding of biological information
  - Comparing between genes and between species
  - Developing new drugs
- Mathematical concepts: Edit distance, Dynamic Programming

The human biological system is hierarchical



Cells: basic building blocks, variable morphologies and functions

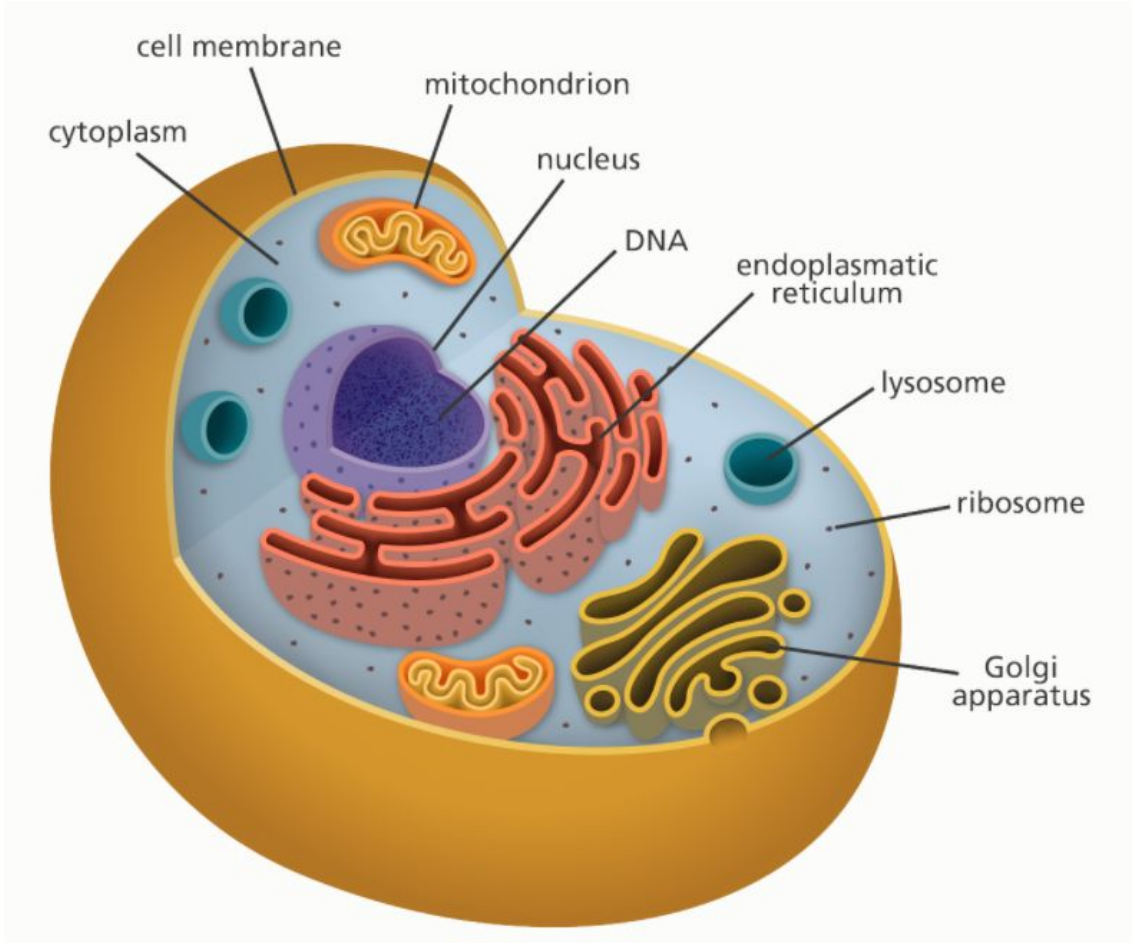
Tissues: groups of specialized cells that communicate and collaborate

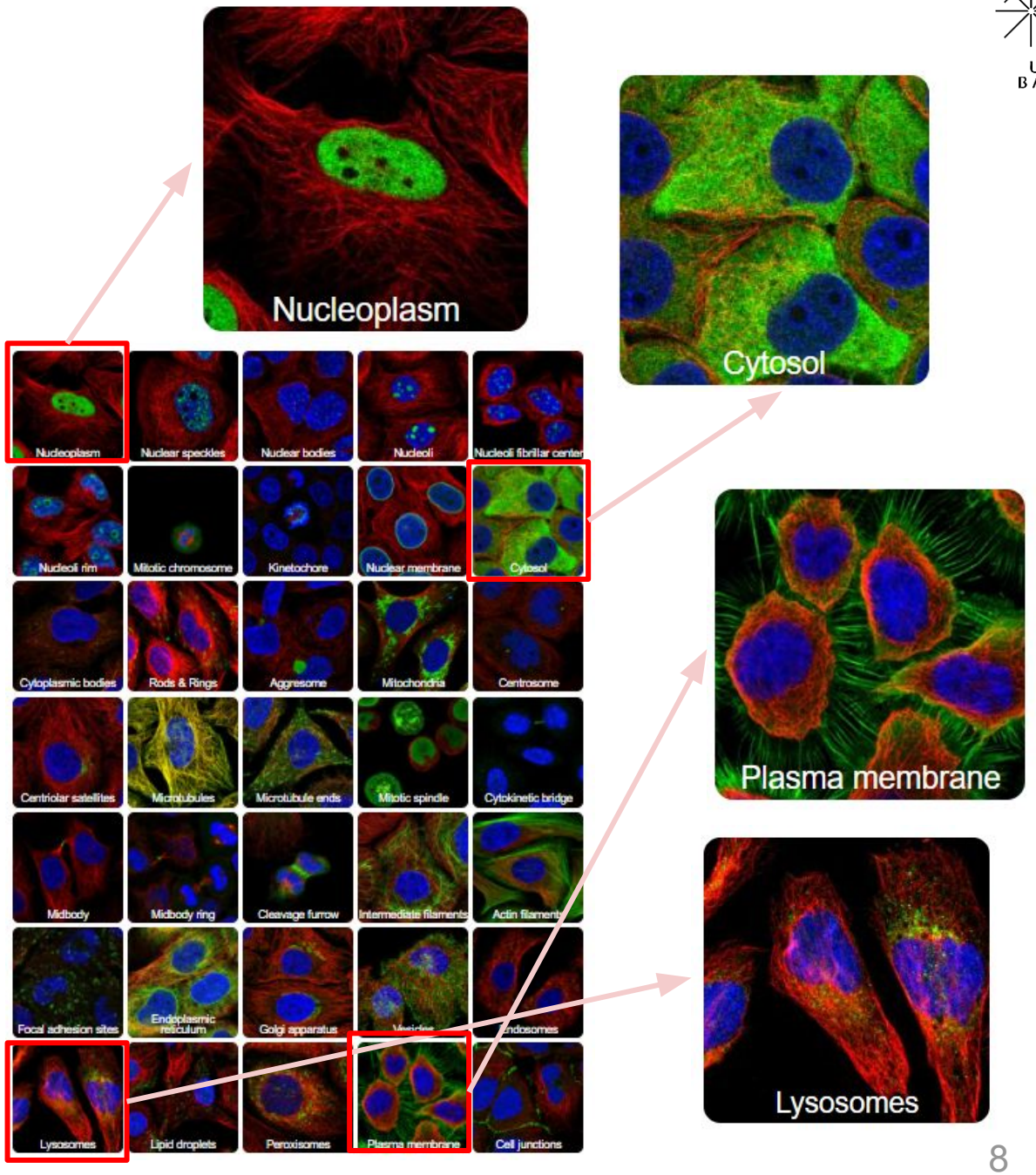
Organ: group of tissues to perform specific functions

Organ systems: group of organs and tissues

The human cell as a material entity

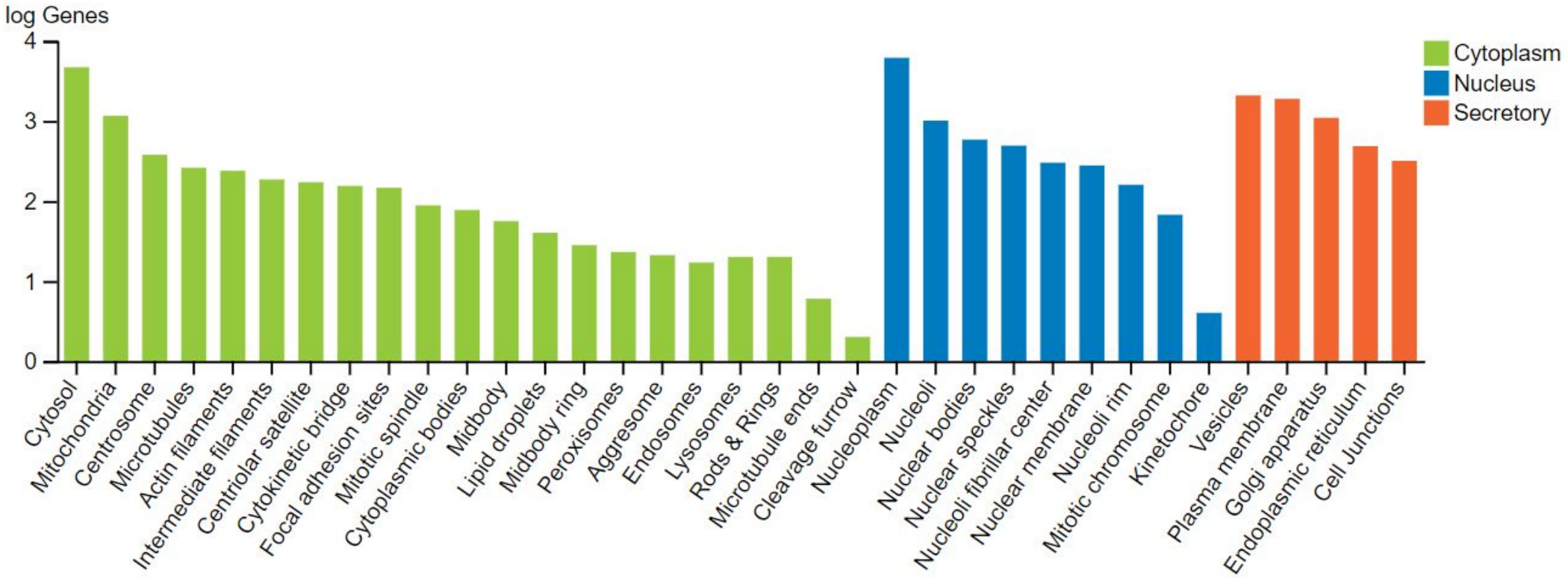
The Physics





Left: Illustration showing the structures of an animal cell. Image credit: Genome Research Limited. Right: [Figure from The Human Protein Atlas]

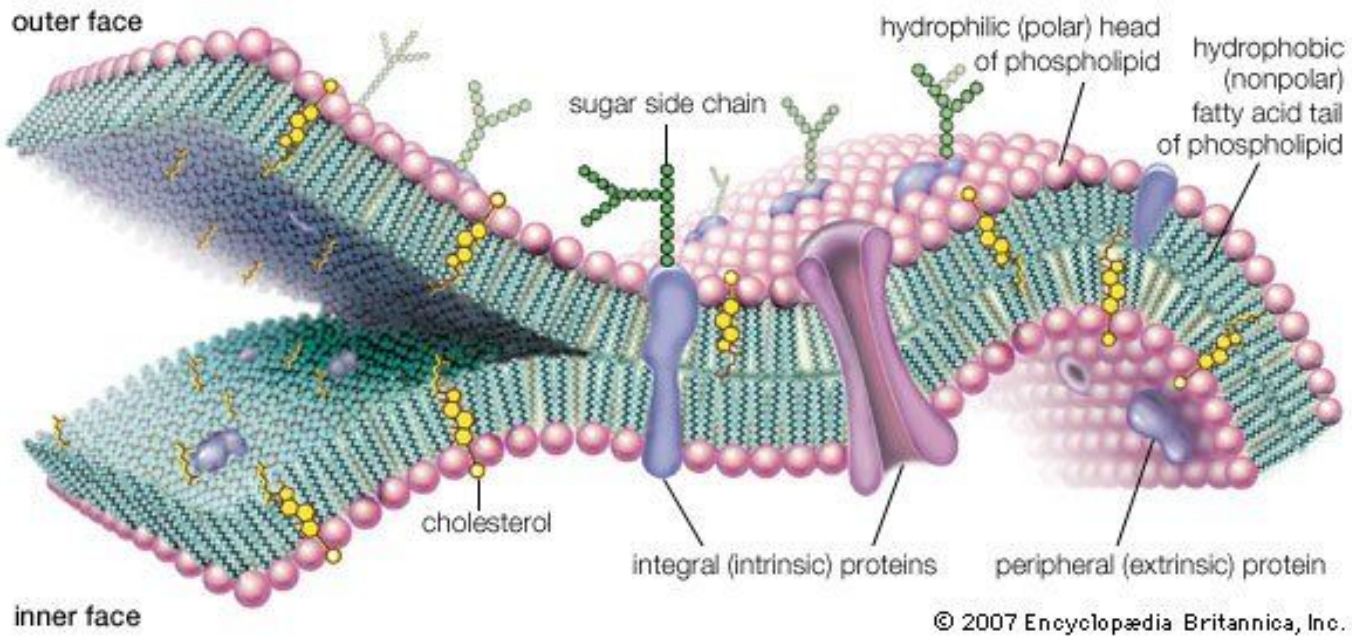
Protein distribution in the human cell

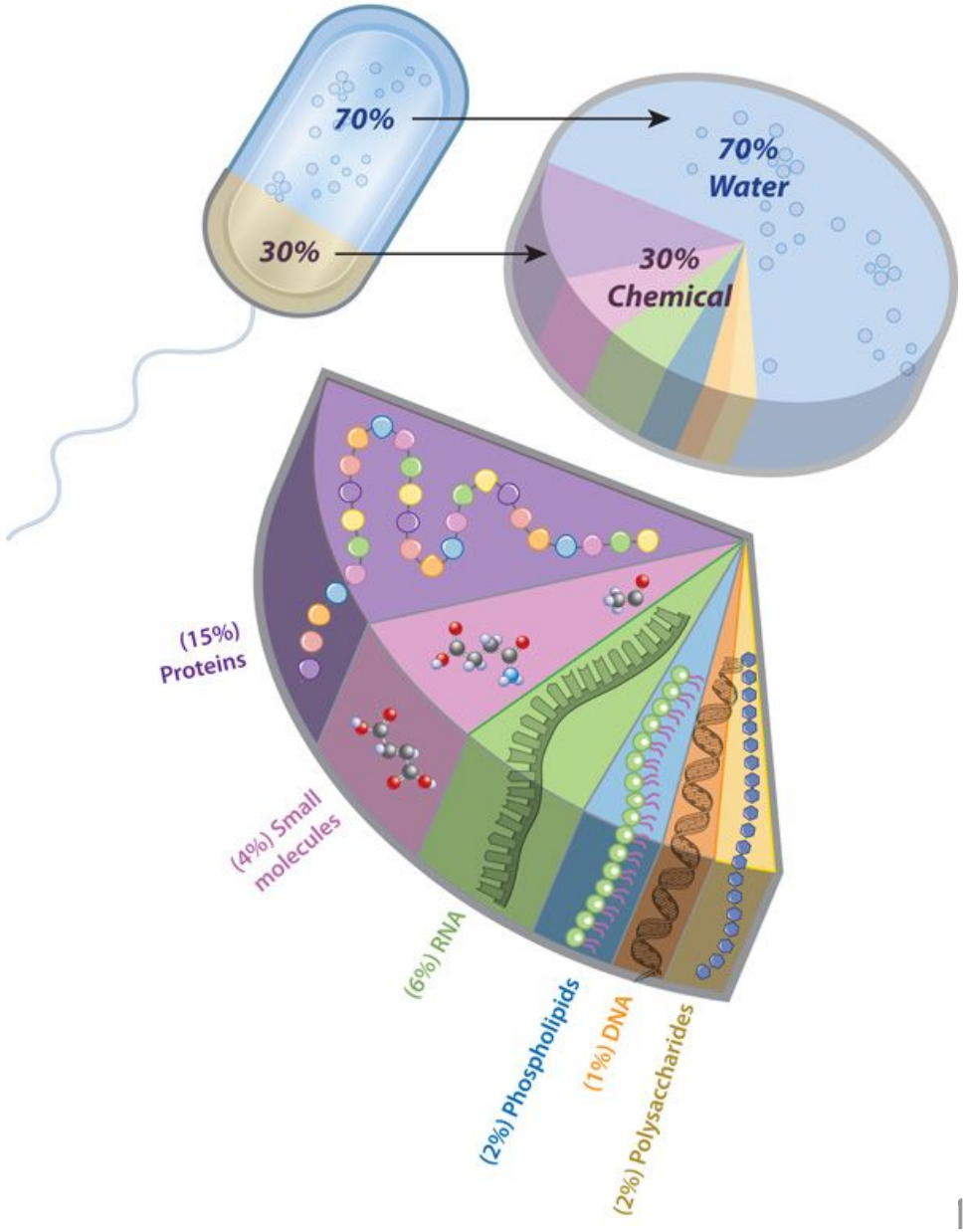


Source: The Human Protein Atlas. Among N=12813 cells, 55% (n=7106) of the proteins were detected in more than one location (*multilocalizing* proteins), and 25% (n=3141) displayed single-cell variation in expression level or spatial distribution.

The human cell as a material entity

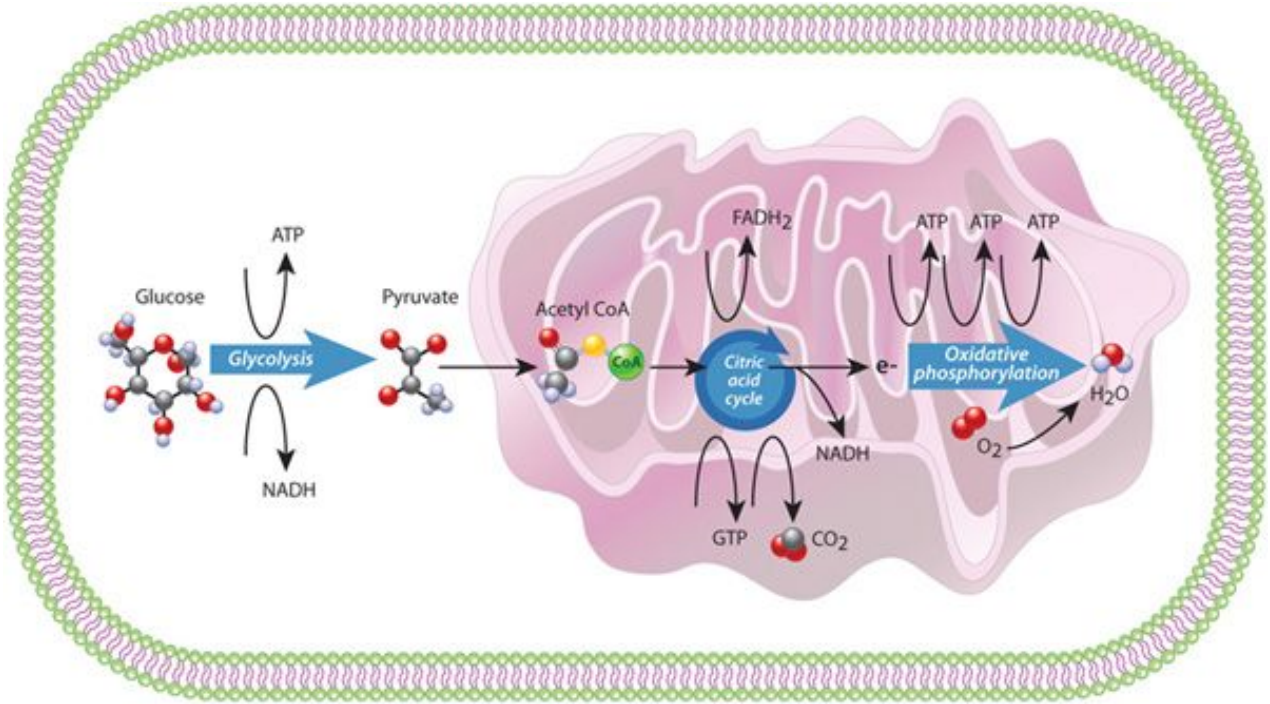
The Chemistry





Left: Cell membrane, copyright of Encylopedia Britannica, Inc.
Right: Chemical composition of a human cell, by Scitable Nature Education.

The human cell as an energy producer and consumer



Energy metabolism: glycolysis takes place in the cytoplasm. Within the mitochondrion, the citric acid cycle occurs in the mitochondrial matrix, and oxidative metabolism occurs at the internal folded mitochondrial membranes (cristae). Source: Nature Education.

tissue	protein synthesis	Na^+/K^+ ATPase	Ca^{+2} ATPase	other
liver	20%	5-10%	5%	gluconeogenesis (15-40%), substrate recycling (20%), proton leak (20%), urea synthesis (12%)
kidney	6%	40-70%	-	gluconeogenesis (5%)
heart	3%	1-5%	15-30%	actinomyosin ATPase (40-50%), proton leak (15% max)
brain	5%	50-60%	significant	a single cortical action potential was estimated to require 10^8-10^9 ATP, BNID 111183)
skeletal muscle	17%	5-10%	5%	proton leak (50%), nonmitochondrial (14%)

Distribution of major oxygen-consuming processes to total oxygen consumption rate of rate tissues in standard state, from Cell Biology By The Numbers. The total energy production rate is about 100W (or ~1W/kg) at rest.

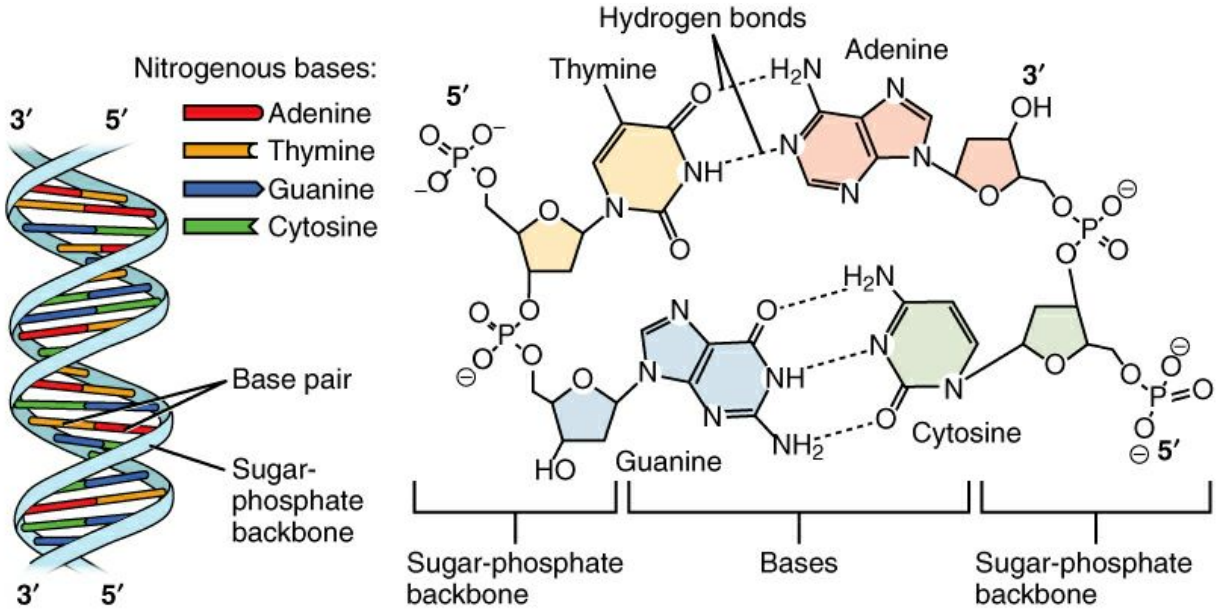
The central dogma of molecular biology: the human cell as an information vehicle

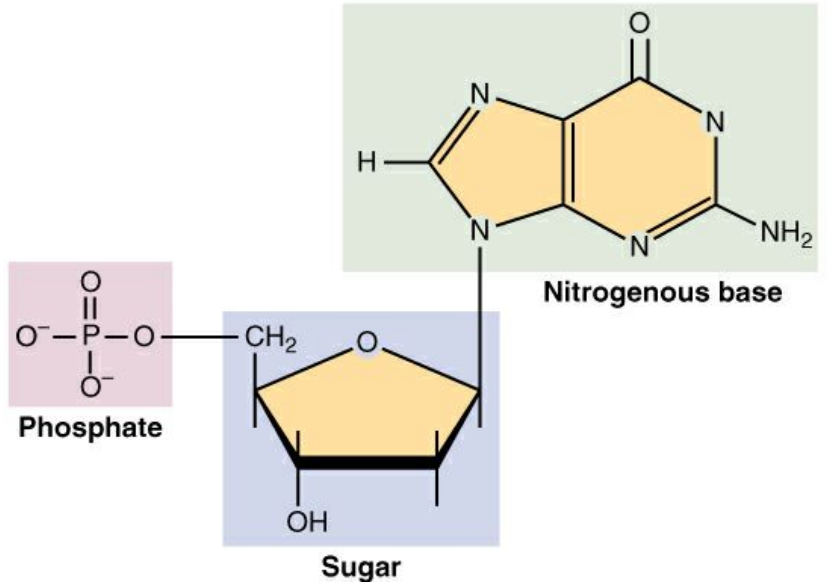
The Central Dogma can be represented by a graph of chemical information vehicles (nodes) and biological information flows (edges)

DNA: sequences and structure

Photo 51, X-ray diffraction image of DNA

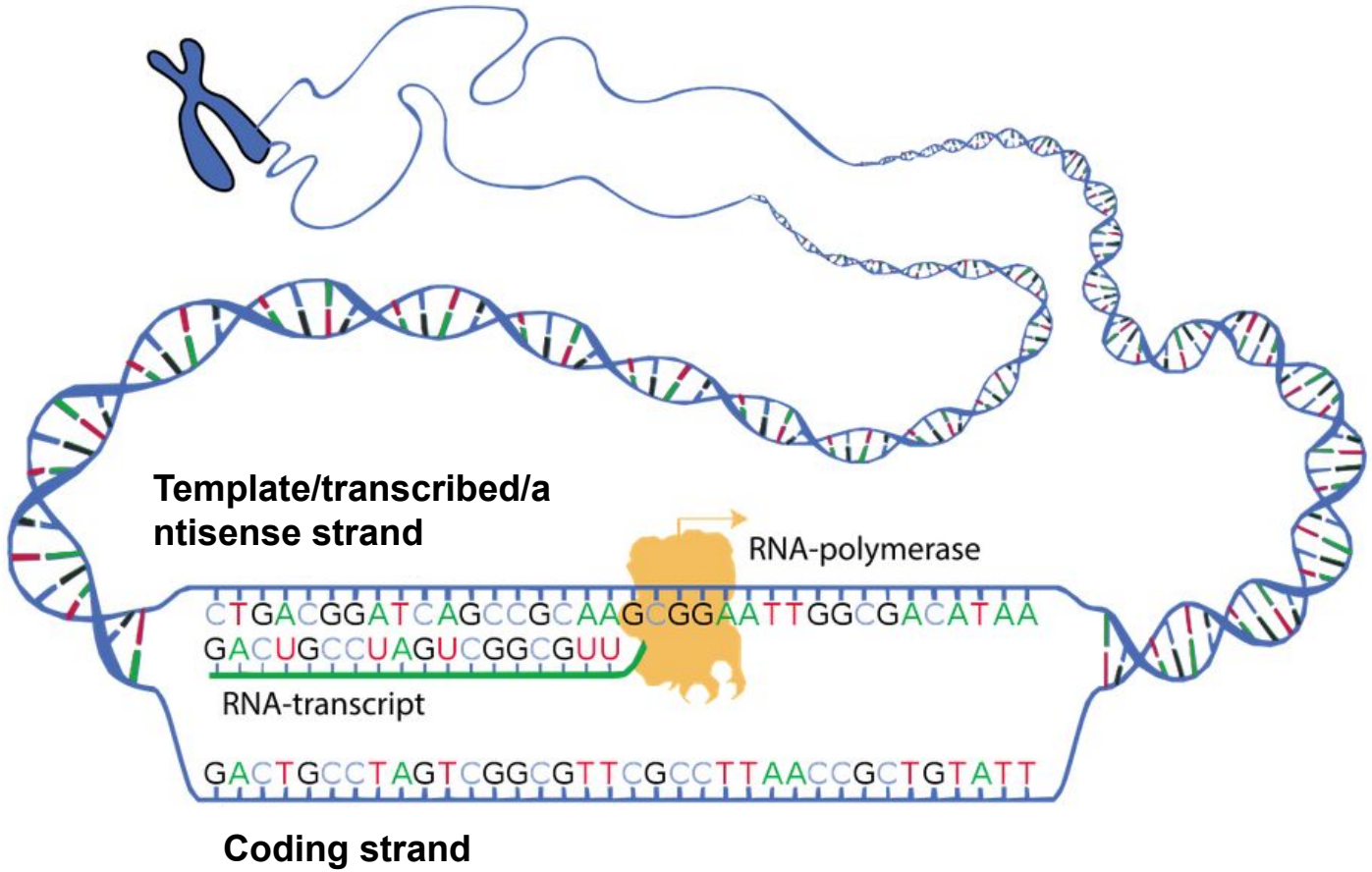
Franklin R, Gosling RG (1953) "Molecular Configuration in Sodium Thymonucleate". ***Nature*** 171**: 740–741.**

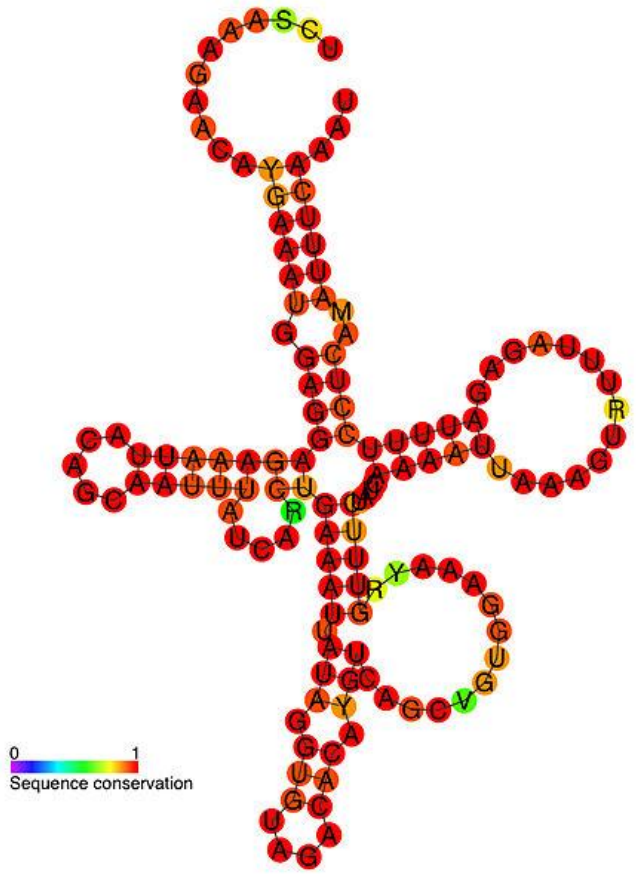




From the textbook OpenStax Anatomy and Physiology, discovered through Wikimedia, reused under the CC license.

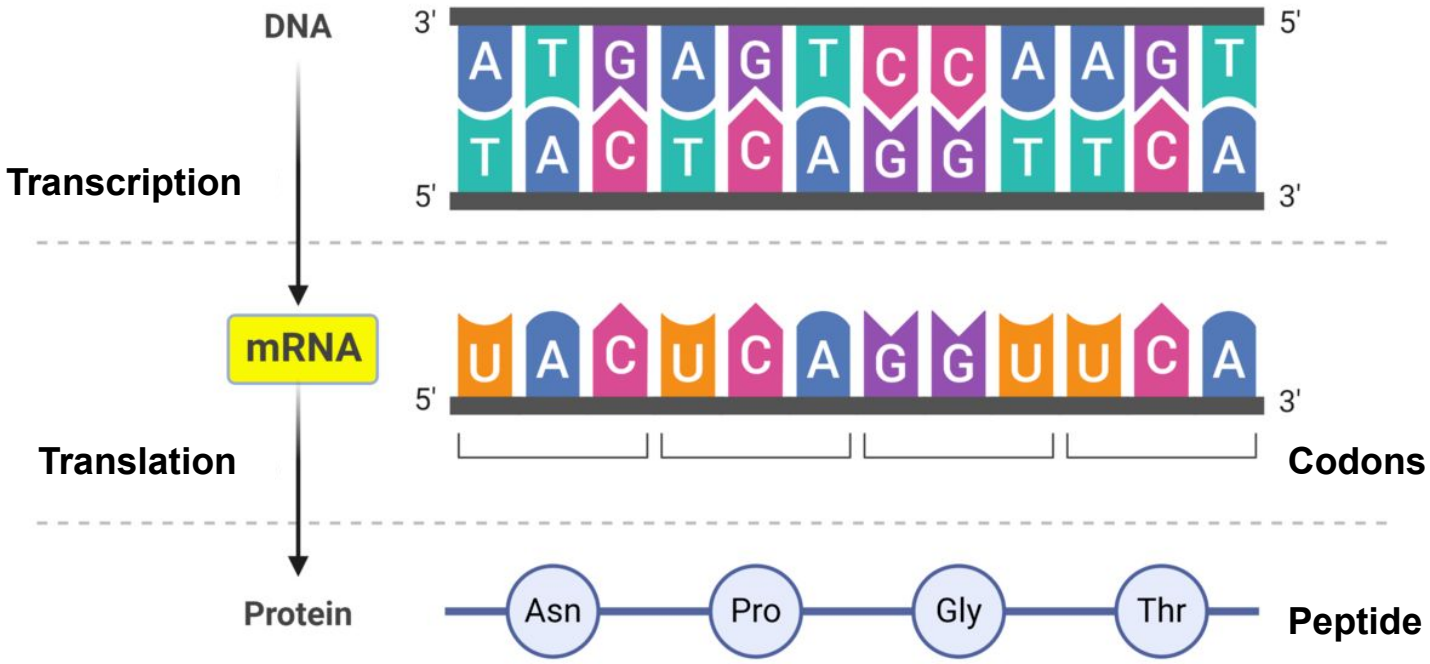
RNA: transcription and the secondary structure



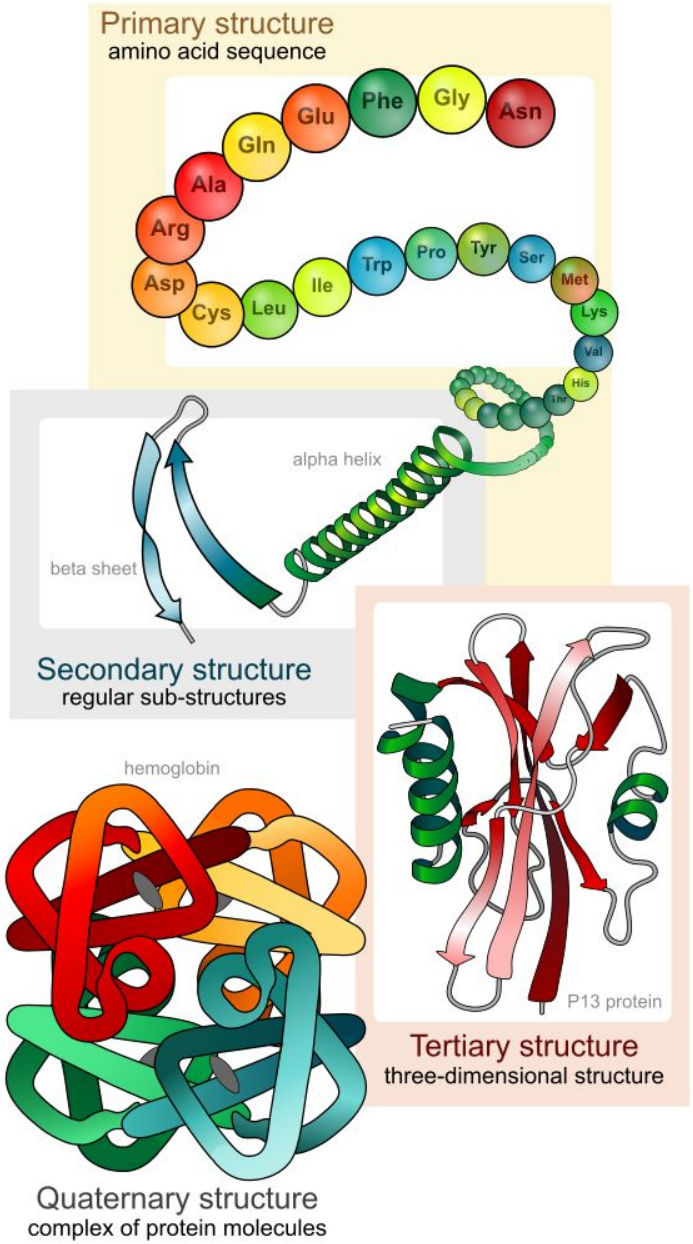


Downloaded and adapted from https://commons.wikimedia.org/wiki/File:DNA_transcriptie.svg and https://en.m.wikipedia.org/wiki/File%3AHAR1F_RF00635_rna_secondary_structure.jpg. Original work by wikipedia user: OrgreBot and user:Ppgardne. Used under CC-SA 3.0 license.

Translation of RNA into proteins

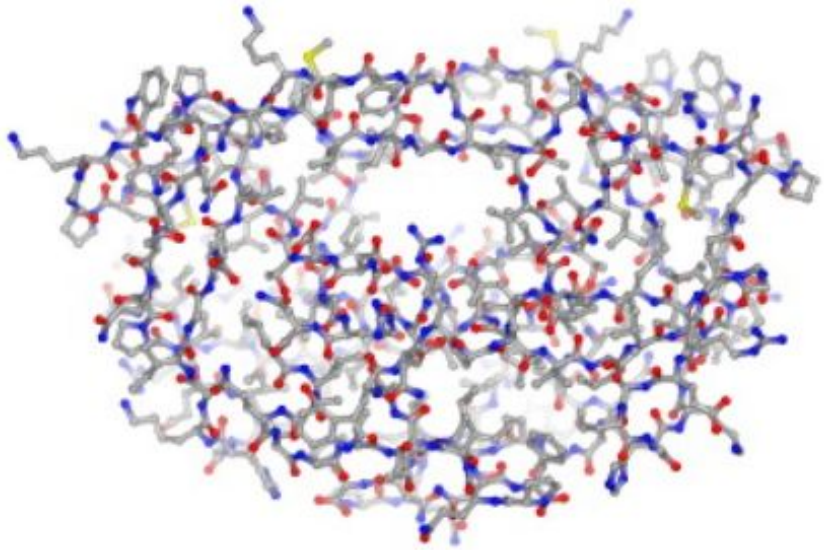


By Lene Martinsen/BioRender. License: CC BY NC SA 3.0

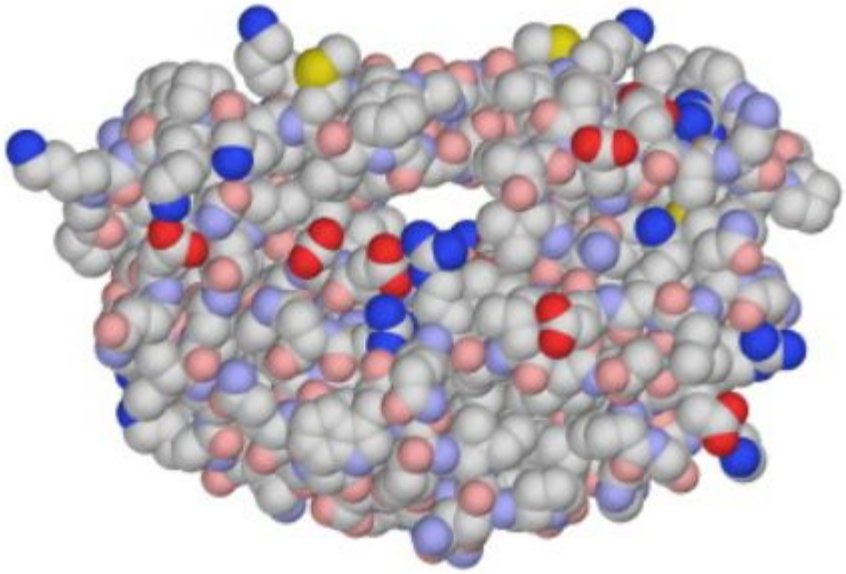


https://de.wikipedia.org/wiki/Datei:Main_protein_structure_levels_en.svg

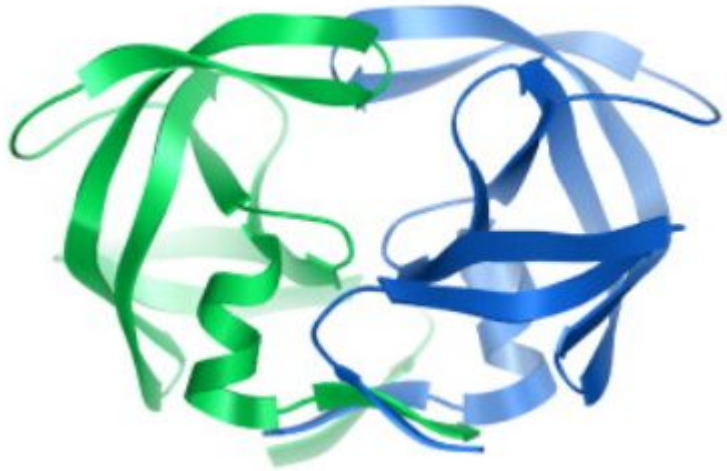
Three basic visual metaphors to display proteins



The bond diagram (balls-and-sticks)



The space-filling diagram



The ribbon diagram

Color codes: **Charged Nitrogen**, **Charged Oxygen**, **Uncharged Nitrogen**, **Unchard Oxygen**, **Carbon** (gray/white), **Sulfur**

Goodsell, David S. “Visual Methods from Atoms to Cells.” Structure 13, no. 3 (March 1, 2005): 347–54. https://doi.org/10.1016/j.str.2005.01.012.

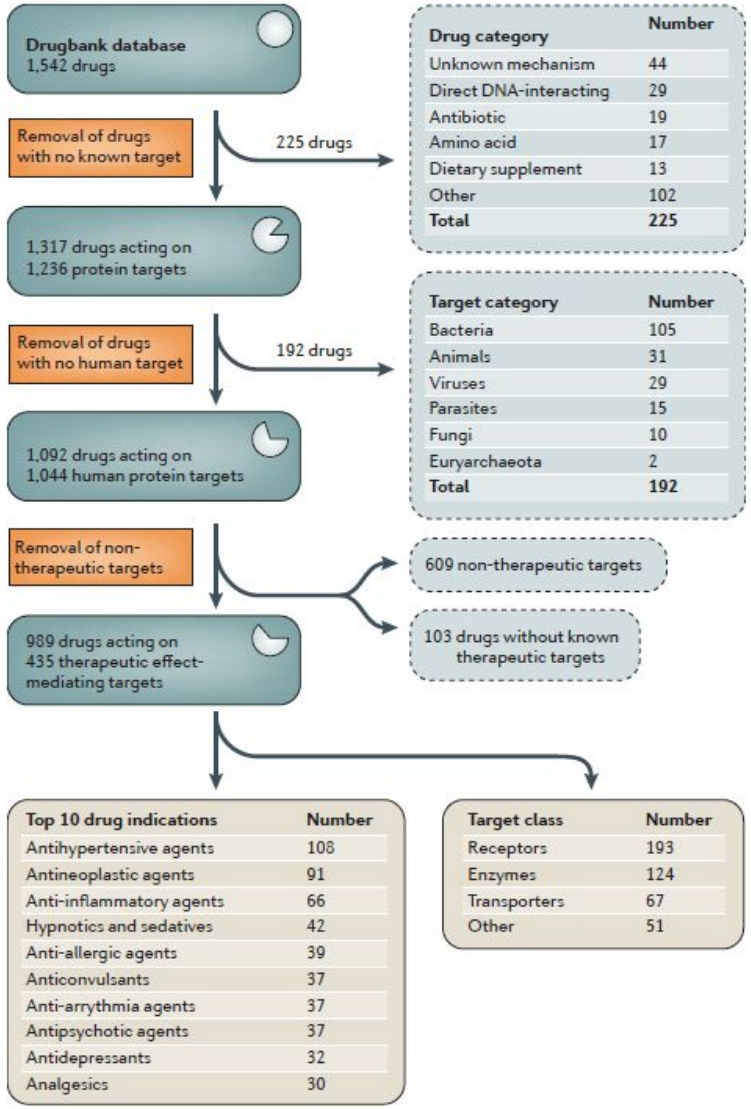
Drugs work by targeting nodes or edges of the central dogma

Target	Example drugs or therapeutic candidates
Protein	• Most small-molecules, for instance GPCR modulators, kinase inhibitors, ion channel inhibitors • Most large-molecules (antibodies)
Translation	• Antimicrobial protein synthesis inhibitors • mTOR-pathway modulating drugs such as rapamycin
RNA	• Anti-sense oligonucleotides (ASO), for instance siRNA (silencing RNA) or locked nucleotide acids (LNA)
Transcription	• Antimicrobials such as actinomycin D and α-Amanitin • Evrysdi (Risdiplam, SMN2 splicing modulator)
Reverse transcription	• Reverse transcriptase inhibitors such as AZT (Zidovudine)
DNA	• Genome-editing therapies such as chimeric activated receptors in T-cells (CAR-T) or CRISPR-CAS9
DNA replication	• Topoisomerase inhibitors such quinolones • Chemotherapies

Break-out

1. Think of **three** drugs that you have used and/or used.
2. Look for their chemical structures, and identify whether they belong to small molecule, antibodies, oligonucleotides, or others.
3. Try to look after their pharmacological targets, and tell which part of the central dogma do they target?
4. (Optional) Try to learn their mechanism-of-action or mode-of-action (MoA), *i.e.* how do they work, and their indications, *i.e.* the diseases they try to cure.

Most drugs so far target proteins



Drug category	Number
Unknown mechanism	44
Direct DNA-interacting	29
Antibiotic	19
Amino acid	17
Dietary supplement	13
Other	102
Total	**225**

Target category	Number
Bacteria	105
Animals	31
Viruses	29
Parasites	15
Fungi	10
Euryarchaeota	2
Total	**192**

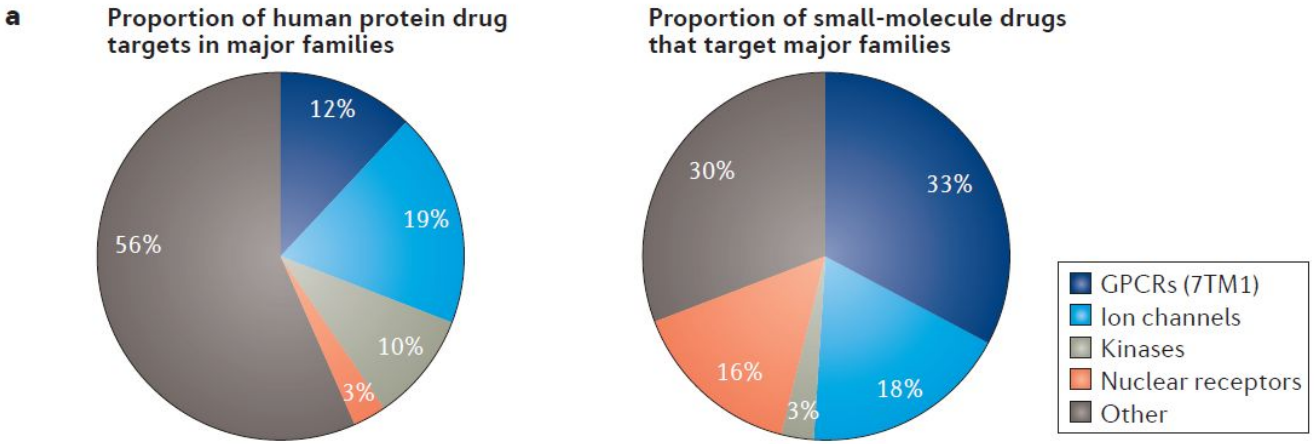
Top 10 drug indications	Number
Antihypertensive agents	108
Antineoplastic agents	91
Anti-inflammatory agents	66
Hypnotics and sedatives	42
Anti-allergic agents	39
Anticonvulsants	37
Anti-arrythmia agents	37
Antipsychotic agents	37
Antidepressants	32
Analgesics	30

Target class	Number
Receptors	193
Enzymes	124
Transporters	67
Other	51

Table 1 | **Molecular targets of FDA-approved drugs**

Drug target class	Targets			Drugs		
	Total targets	Small-molecule drug targets	Biologic drug targets	Total drugs	Small molecules	Biologics
Human protein	667	549	146	1,194	999	195
Pathogen protein	189	184	7	220	215	5
Other human biomolecules	28	9	22	98	63	35
Other pathogen biomolecules	9	7	4	79	71	8

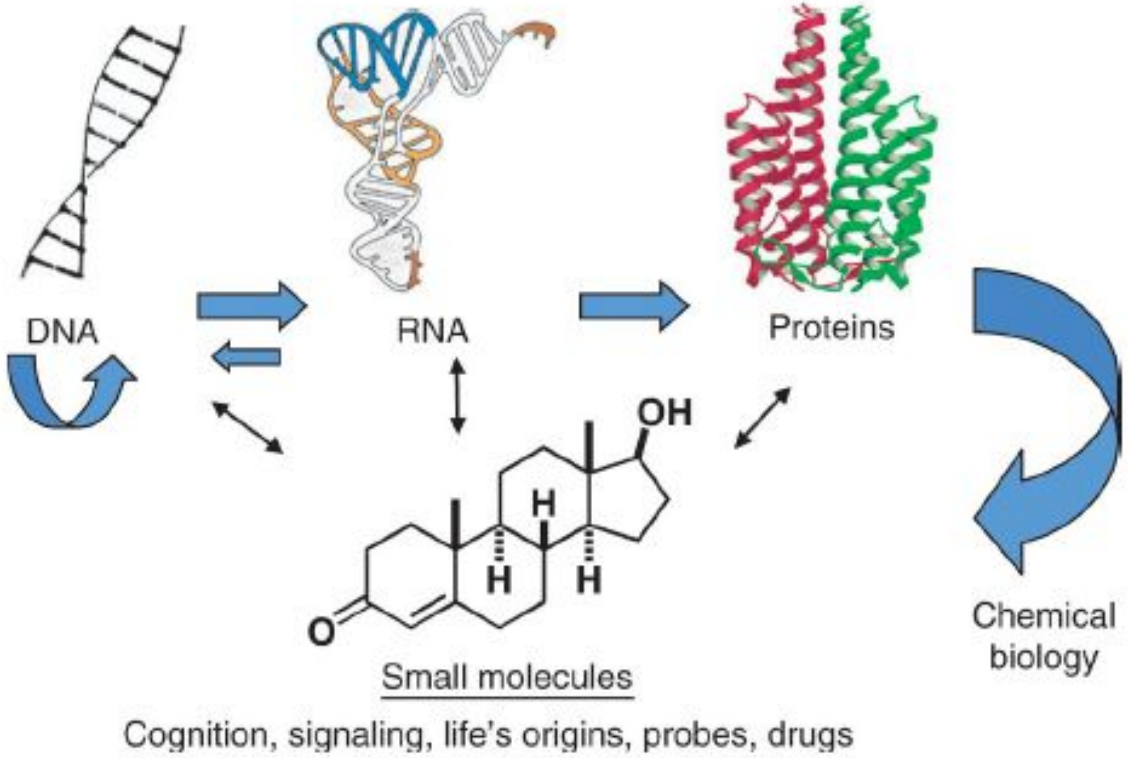
The list also includes antimalarial drugs approved elsewhere in the world.



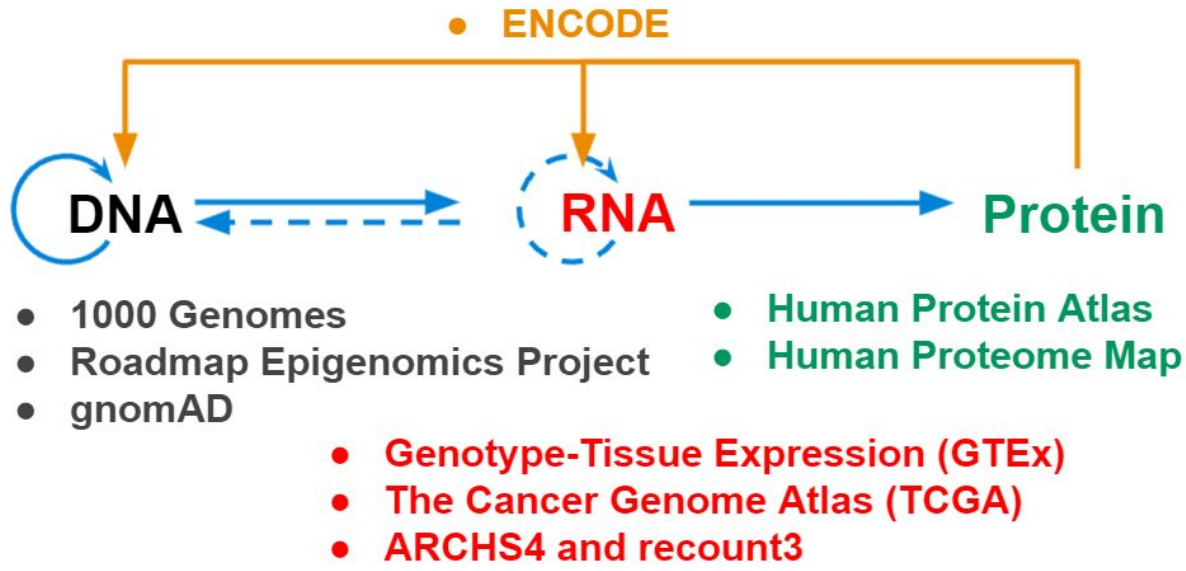
Left: Rask-Andersen, Mathias, Markus Sällman Almén, and Helgi B. Schiöth. 2011. “Trends in the Exploitation of Novel Drug Targets.” Nature Reviews Drug Discovery 10 (8): 579–90. https://doi.org/10.1038/nrd3478.

Right: Santos, Rita, Oleg Ursu, Anna Gaulton, A. Patrícia Bento, Ramesh S. Donadi, Cristian G. Bologa, Anneli Karlsson, et al. 2017. “A Comprehensive Map of Molecular Drug Targets.” *Nature Reviews Drug Discovery* 16 (1): 19–34. https://doi.org/10.1038/nrd.2016.230.

Extending the Central Dogma with small molecules, feedback regulations, and Big Data



Schreiber, Stuart L. “Small Molecules: The Missing Link in the Central Dogma.” Nature Chemical Biology 1, no. 2 (July 2005): 64–66. https://doi.org/10.1038/nchembio0705-64.



Many references. Two for ENCODE are selected here: Moore, Jill E., Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, et al. “Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes.” Nature 583, no. 7818 (July 2020): 699–710. https://doi.org/10.1038/s41586-020-2493-4; Van Nostrand, Eric L., Peter Freese, Gabriel A. Pratt, Xiaofeng Wang, Xintao Wei, Rui Xiao, Steven M. Blue, et al. “A Large-Scale Binding and Functional Map of Human RNA-Binding Proteins.” Nature 583, no. 7818 (July 2020): 711–19. https://doi.org/10.1038/s41586-020-2077-3.

Questions about Bollag *et al.*, Nature 2010

1. What is the **indication** of *PLX4032*?
2. What is the **gene target** of *PLX4032*?
3. The malignancy depends on which biological **pathway***?*
4. What is the **Mechanism of Action** of *PLX4032?*
5. What went wrong in the first **Phase I clinical trial**? And how was it solved?
6. What was the **dosing regimen** in the final Phase I clinical trial, and what is the **response rate**?

Questions for further thinking

- In the video that you watched offline, Susan Desmond-Hellmann summarizes great drug development in four key concepts: (1) Having a deep understanding of the basic science and the characteristics of the drug. (2) Target the right patients. (3) Set a high bar in the clinic. (4) Work effectively with key regulatory decision makers. What parts of this abstract reflect these points?
- Susan Desmond-Hellmann emphasized the importance of collaboration. Is that true when you consider this abstract?
- How do you like the abstract? Anything that you can learn from it about writing?

Offline activities

- Fill the anonymous survey #2: link
- Read the paper Bollag *et al.*, 2010, and answer questions here in Offline Activities.
- Do the exercise for the Levenshtein distance in the Handout.
- Optional: use either Python, R, or any C-family or Lisp-family languages to
  - (basic) Implement a procedure to calculate the Levenshtein distance
  - (advanced) Implement a program or website to display the Dynamic Programming procedure to calculate the Levenshtein distance

Questions for Bollag et al., 2010

1. We learned that many drugs target one of the four protein types: GPCRs, ion channels, kinases, and nuclear receptors. Which type does the target of PLX4032 belong to?
2. How was the efficacy of PLX4032 tested?
3. Why was PLX4032 chosen for further development, but not PLX4720?
4. How was the exposure of PLX4032 in the blood quantified? Which mathematical operation was used?
5. How was the final dosing regimen (960-mg BID) determined?
6. How did patients with the V600K mutation in BRAF respond?
7. What measures were taken to demonstrate the effect of BRAF inhibition in patient biopsies?
8. What side effects of PLX4032 were reported?
9. What measures were taken against side effects and safety concerns of PLX4032?
10. Where do you think mathematics and informatics is used in the discovery and development of PLX4032?

A single-amino-acid difference in BRAF gene may mean longer survival of melanoma patients given the correct treatment

McArthur, Grant A., Paul B. Chapman, Caroline Robert, James Larkin, John B. Haanen, Reinhard Dummer, Antoni Ribas, *et al.*

Safety and Efficacy of Vemurafenib in BRAFV600E and BRAFV600K Mutation-Positive Melanoma (BRIM-3): Extended Follow-up of a Phase 3, Randomised, Open-Label Study

The Lancet Oncology 15, Nr. 3 (1. März 2014): 323–32. https://doi.org/10.1016/S1470-2045(14)70012-9.

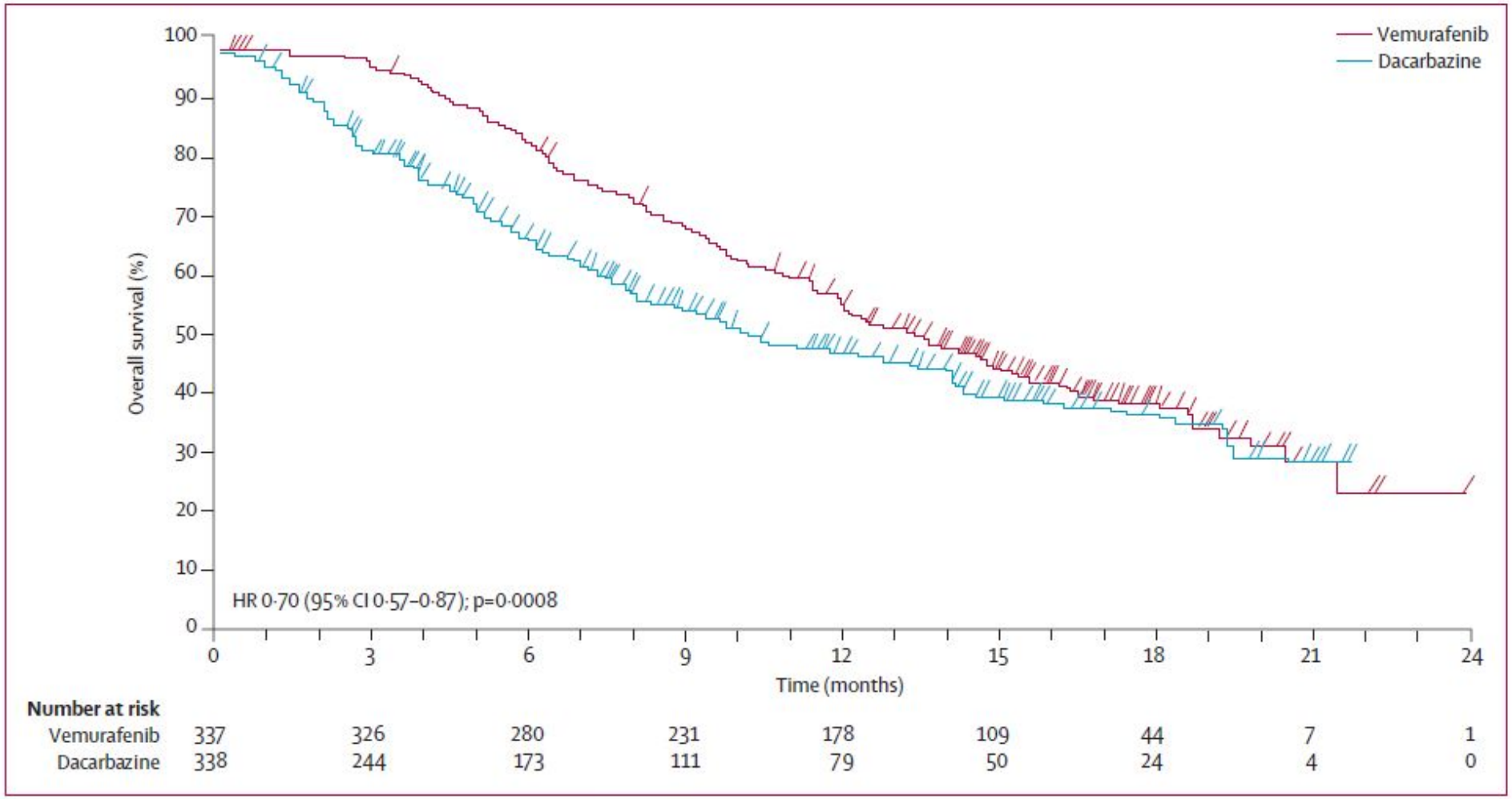
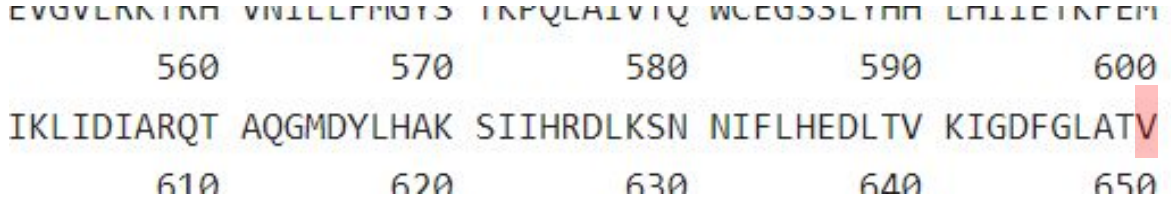


Figure 2: Overall survival (randomised population; censored at crossover) for patients randomly assigned to vemurafenib or to dacarbazine (cutoff Feb 1, 2012)

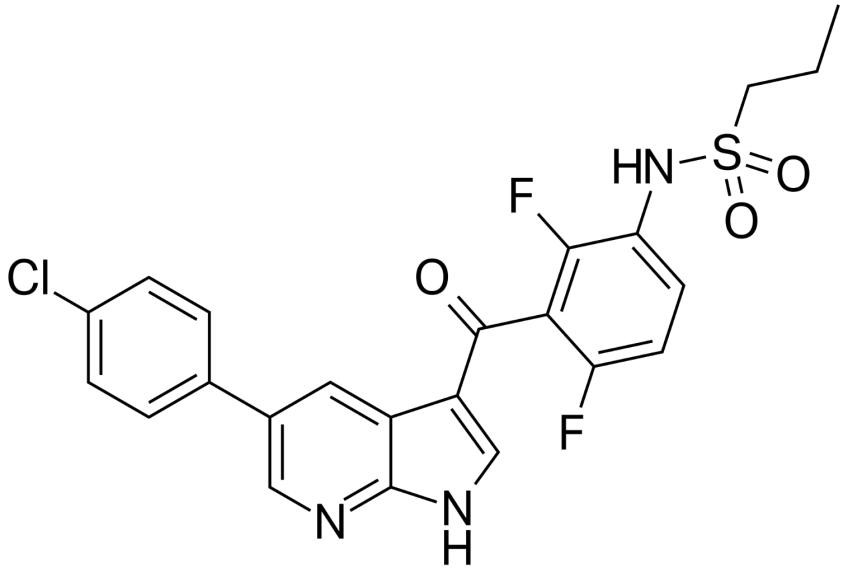
Vemurafenib (Zelboraf, PLX4032)
V600E mutated BRAF inhibition

- V600E: Valine (V) on the amino-acid position 600 is substituted by glutamic acid (E).



Fragment of BRAF protein. Source: UniProtKB, P15056 (BRAF_HUMAN)

- View the 3D structure of the molecule at PDB ligand database
- View the X-ray structure of BRAF in complex with PLX4032 on PDB: accession number 3OG7.
- Find more information about the discovery and clinical efficacy of vemurafenib in the handout.



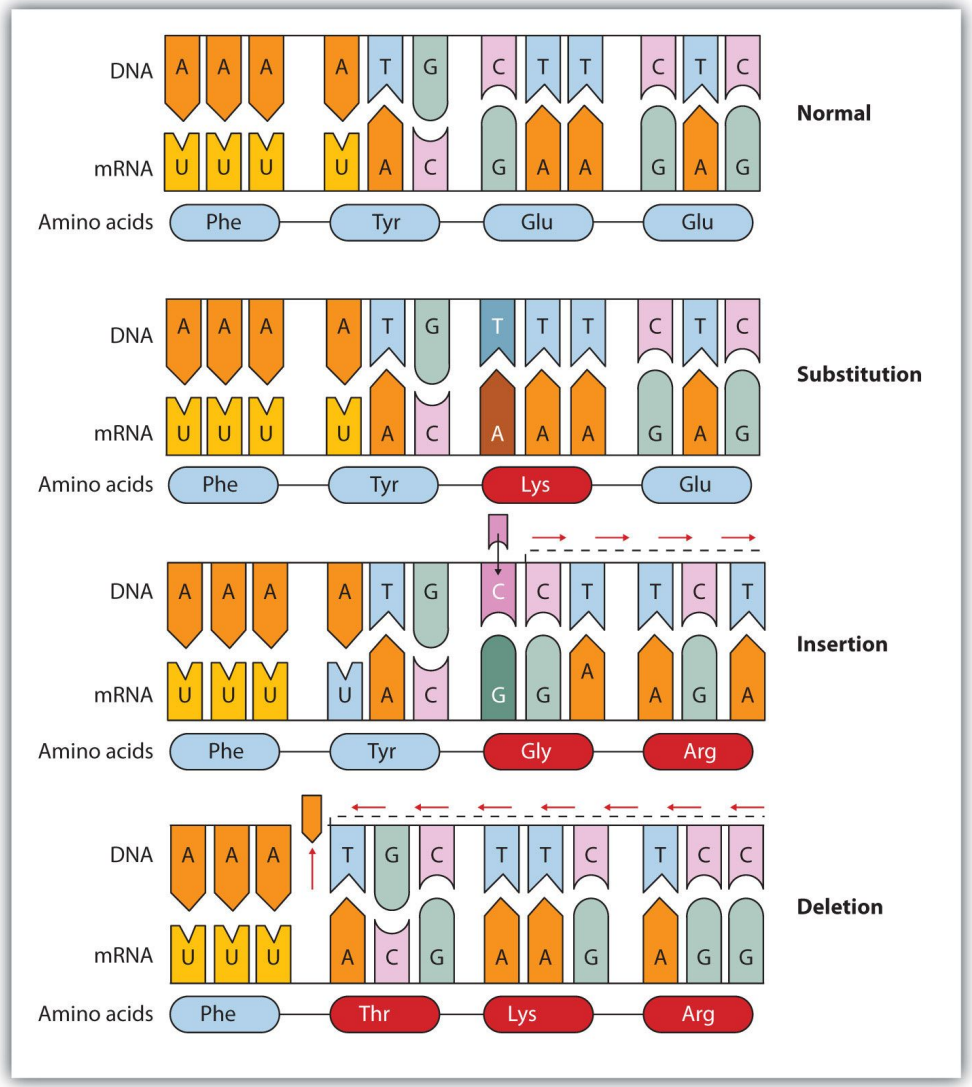
Source: https://commons.wikimedia.org/wiki/File:Vemurafenib_structure.svg

Edit distance: a deterministic view of distance between two sequences

	Insertion	Deletion	Substitution	Transposition	Note
The Levenshtein distance	Allowed	Allowed	Allowed	Not allowed	
The longest common subsequence (LCS) distance	Allowed	Allowed	Not allowed	Not allowed	
The Hamming distance	Not allowed	Not allowed	Allowed	Not allowed	
The Damerau-Levenshtein distance	Allowed	Allowed	Allowed	Allowed (adjacent characters)	Not a distance metric, because triangle inequality is not satisfied
The Jaro-Winkler distance	Not allowed	Not allowed	Not allowed	Allowed	Not a distance metric

Discussion: which distance is mostly used for biological sequence analysis? Why?

Chemistry and biology of point mutation



Disease	Responsible Protein or Enzyme
alkaptonuria	homogentisic acid oxidase
galactosemia	galactose 1-phosphate uridyl transferase, galactokinase, or UDP galactose epimerase
Gaucher disease	glucocerebrosidase
gout and Lesch-Nyhan syndrome	hypoxanthine-guanine phosphoribosyl transferase
hemophilia	antihemophilic factor (factor VIII) or Christmas factor (factor IX)
homocystinuria	cystathionine synthetase
maple syrup urine disease	branched chain α-keto acid dehydrogenase complex
McArdle syndrome	muscle phosphorylase
Niemann-Pick disease	sphingomyelinase
phenylketonuria (PKU)	phenylalanine hydroxylase
sickle cell anemia	hemoglobin
Tay-Sachs disease	hexosaminidase A
tyrosinemia	fumarylacetoacetate hydrolase or tyrosine aminotransferase
von Gierke disease	glucose 6-phosphatase
Wilson disease	Wilson disease protein

Three types of point mutations in DNA, The Basics of General, Organic, and Biological Chemistry

With *Levenshtein distance* we can compare any two pieces of DNA

Levenshtein distance: The minimum number of operations required to transform string a to string b with following operations:

- Insertion, *e.g.* **bat → bait**
- Deletion, *e.g.* **boat → bot**
- Substitution, *e.g.* **pig → big**

The Levenshtein distance between two strings a, b of length $|a|$ and $|b|$ respectively is given by $\mathrm{lev}_{a,b}(|a|, |b|)$ where

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

where $1_{(a_i \neq b_j)}$ is the indicator function equal to 0 when $a_i = b_j$ and equal to 1 otherwise, and $\mathrm{lev}_{a,b}(i,j)$ is the distance between the first i characters of a and the first j characters of b.

Calculate the Levenshtein distance with dynamic programming

- What is the Levenshtein distance between `ATGC` and `AGC`?

		A	T	G	C
A					
G					
C					

		A	T	G	C
A					
G					
C					

- Solution: 1

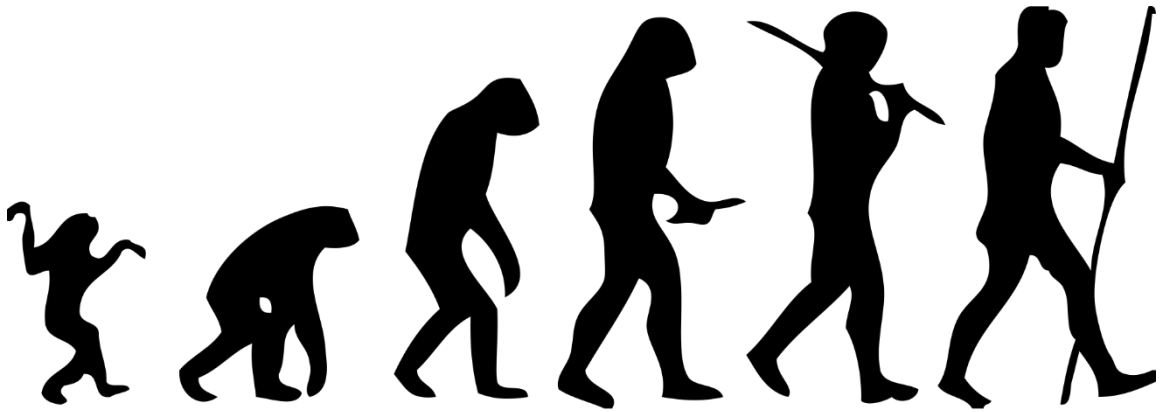
```
ATGC
A-GC
```

Calculate the Levenshtein distance with dynamic programming

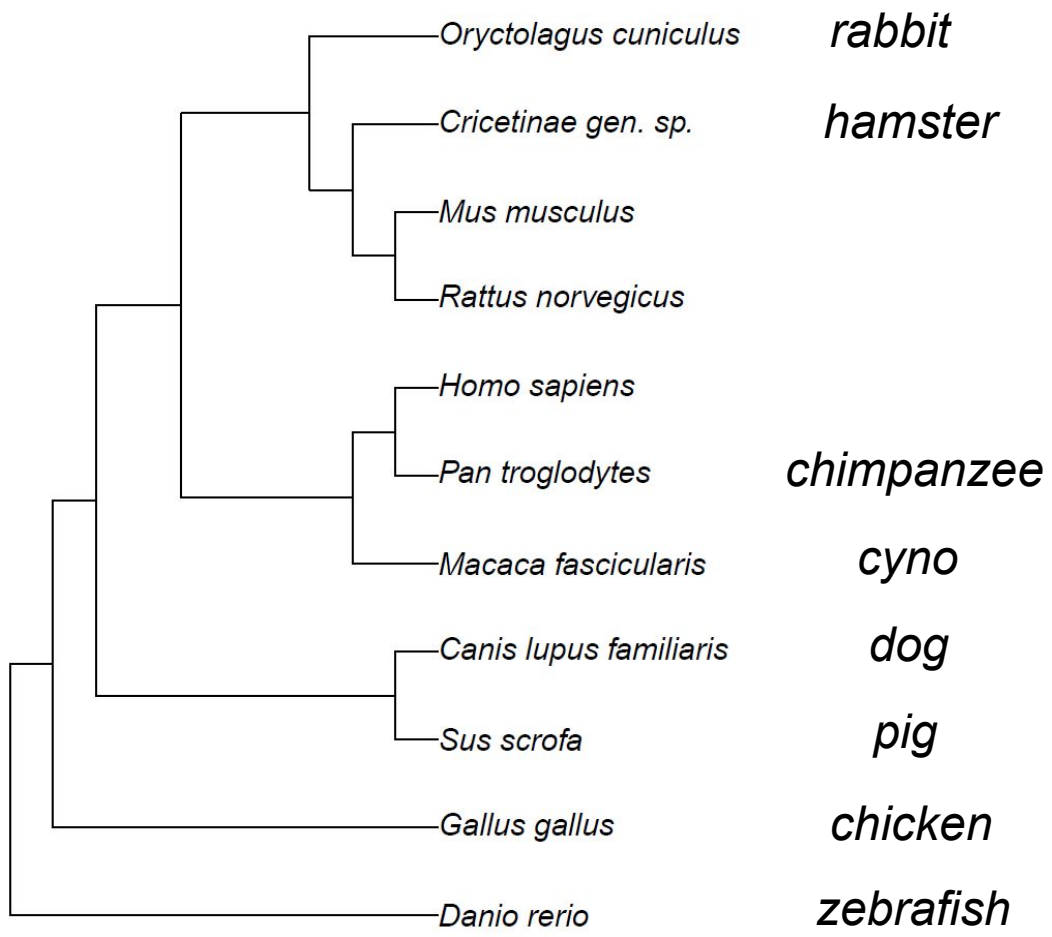
- What is the Levenshtein distance between `ACTGCTT` and `ACATT`?
- Beyond bioinformatics, the Levenshtein distance is often used in computational linguistics and natural language processing. For instance, check out How to Write a Spelling Corrector by Peter Norvig.

		A	C	T	G	C	T	T
A								
C								
A								
T								
T								

Evolution: what is wrong with this figure?



Phylogeny of commonly used species for animal studies



Tree structure retrieved from https://itol.embl.de/ (iTOL, Interactive Tree of Life), visualized with the *FigTree* software developed by Andrew Rambaut

Richard Bell on the origin of the name *Dynamic Programming*

I spent the Fall quarter (of 1950) at RAND. My first task was to find a name for multistage decision processes. An interesting question is, Where did the name, dynamic programming, come from?

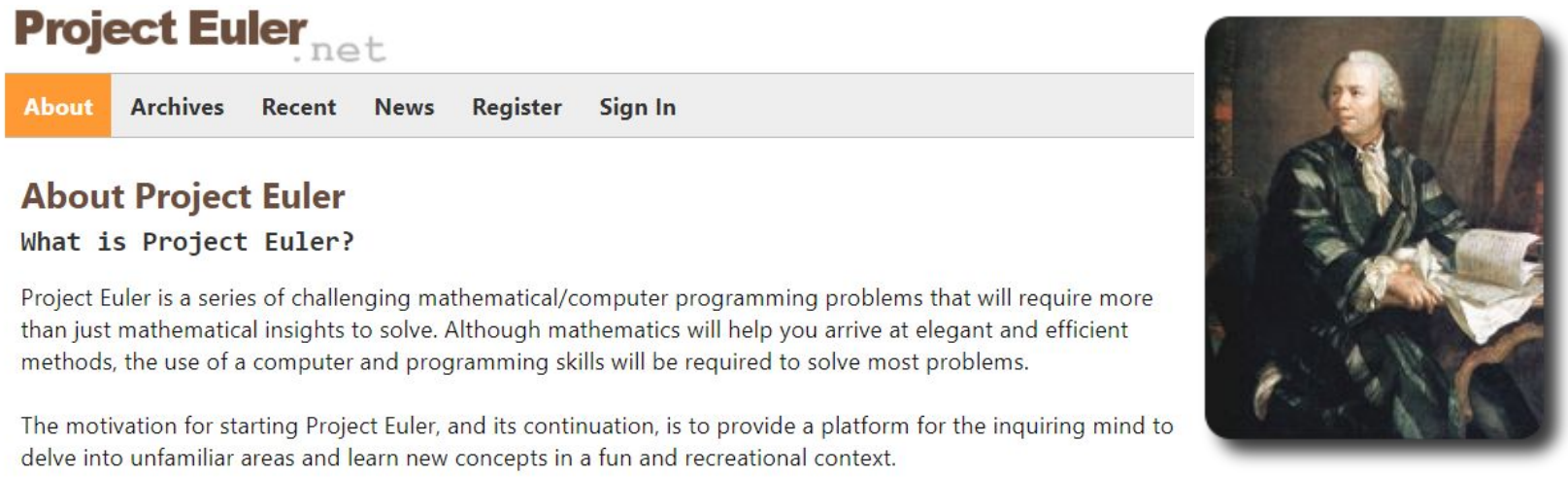
The 1950s were not good years for mathematical research. We had a very interesting gentleman in Washington named Wilson. He was Secretary of Defense, and he actually had a pathological fear and hatred of the word, research. I'm not using the term lightly; I'm using it precisely. His face would suffuse, he would turn red, and he would get violent if people used the term, research, in his presence. You can imagine how he felt, then, about the term, mathematical. The RAND Corporation was employed by the Air Force, and the Air Force had Wilson as its boss, essentially. Hence, I felt I had to do something to shield Wilson and the Air Force from the fact that I was really doing mathematics inside the RAND Corporation. What title, what name, could I choose? In the first place I was interested in planning, in decision making, in thinking. But planning, is not a good word for various reasons. I decided therefore to use the word, "programming" I wanted to get across the idea that this was dynamic, this was multistage, this was time-varying I thought, lets kill two birds with one stone. Lets take a word that has an absolutely precise meaning, namely dynamic, in the classical physical sense. It also has a very interesting property as an adjective, and that is its impossible to use the word, dynamic, in a pejorative sense. Try thinking of some combination that will possibly give it a pejorative meaning. Its impossible. Thus, I thought dynamic programming was a good name. It was something not even a Congressman could object to. So I used it as an umbrella for my activities.

Dreyfus, Stuart. "Richard Bellman on the Birth of Dynamic Programming." Operations Research 50, no. 1 (February 2002): 48–51. https://doi.org/10.1287/opre.50.1.48.17791.

Software tools

- **General biological sequence analysis**
  - EMBOSS software suite: http://emboss.sourceforge.net/, also available online at European Bioinformatics Institute (EBI): https://www.ebi.ac.uk/services
  - BLAST (=Basic Local Alignment Search Tool) can be run at many places, for instances from EBI and National Center for Biotechnology Information (NCBI): https://blast.ncbi.nlm.nih.gov/Blast.cgi
  - Programming access, for instance the Biopython project: https://biopython.org
- **RNA biology**
  - ViennaRNA package (https://www.tbi.univie.ac.at/RNA/)
  - RNA processing tools available at U Bielefeld, for instance RNAhybrid, which finds minimum free energy hybridization using dynamic programming (https://bibiserv.cebitec.uni-bielefeld.de/rnahybrid)
- **Profile Hidden Markov Models (HMMs)**
  - The HMMER package: http://hmmer.org/

The Euler Project



https://projecteuler.net/

- **Learning by problem-solving**
- **Free**
- **Math + CS**

Problem 1: Multiples of 3 and 5

If we list all the natural numbers below 10 that are multiples of 3 or 5, we get 3, 5, 6 and 9. The sum of these multiples is 23.

Find the sum of all the multiples of 3 or 5 below 1000.

Rosalind: a great scientist, and a platform for learning bioinformatics and programming through problem solving

ROSALIND http://rosalind.info/problems/locations/

Rosalind Elsie Franklin

1920-1958

A Rapid Introduction to Molecular Biology click to expand

Problem

A **string** is simply an ordered collection of symbols selected from some **alphabet** and formed into a word; the **length** of a string is the number of symbols that it contains.

An example of a length 21 **DNA string** (whose alphabet contains the symbols 'A', 'C', 'G', and 'T') is "ATGCTTCAGAAAGGTCTTACG."

Given: A DNA string s of length at most 1000 nt.

Return: Four integers (separated by spaces) counting the respective number of times that the symbols 'A', 'C', 'G', and 'T' occur in s.

Sample Dataset

```
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC
```

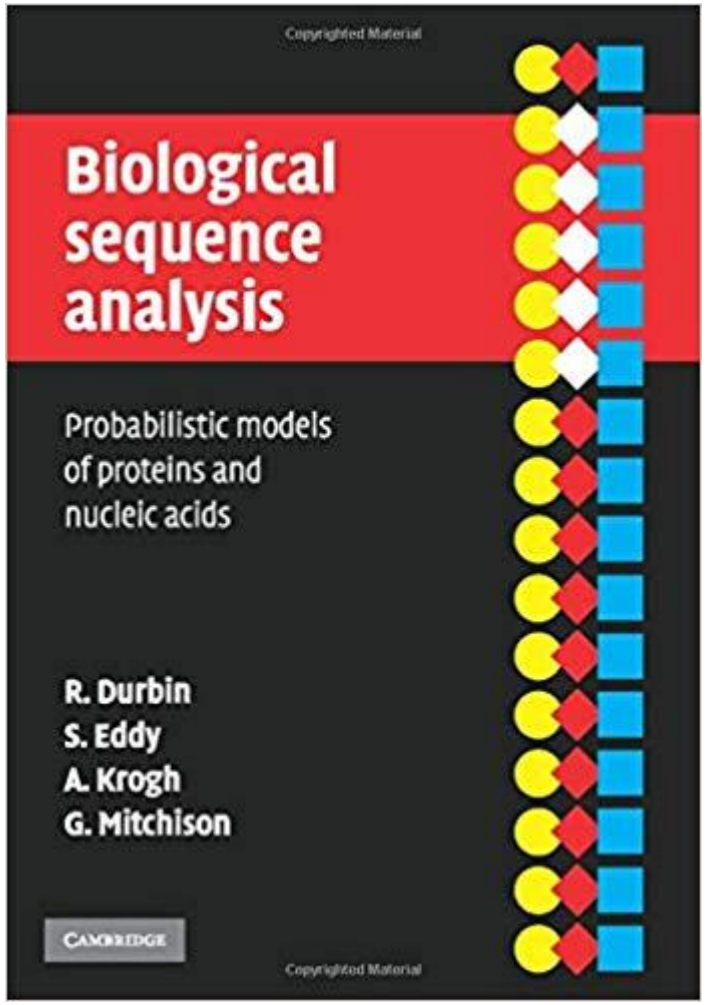
Sample Output

```
20 12 17 21
```

Please login to solve this problem.

Further resources

***Biological Sequence Analysis* by Durbin, Eddy, Krogh, and Mitchison**



Teaching RNA algorithms by the Backofen Lab at U Freiburg, with source codes available on GitHub.

The website hosts among others an interactive tool to visualize how dynamic programming (DP) helps to predict RNA secondary structure.

For a gentle introduction, see also *How Do RNA Folding Algorithms Work?* by Eddy, Sean R, *Nature Biotechnology* 22, Nr. 11 (November 2004): 1457–58. https://doi.org/10.1038/nbt1104-1457.

An Introduction to Applied Bioinformatics by Greg Caporaso (NAU)

The tutorial is written in Python using Jupyter. It introduces concepts in (a) pairwise sequence alignment, (b) sequence homology searching, (c) generalized dynamic programming for multiple sequence alignment, (d) phylogenetic reconstruction, (e) sequence mapping and clustering, as well as (f) machine learning in bioinformatics. Applications and exercises are also available.

Interaction of drug and target: an example with HIV-1 Protease Inhibitor

Protein atoms: ball and stick, in blue and green

The small-molecule drug: ball and stick with traditional atomic coloration

Water: small red spheres