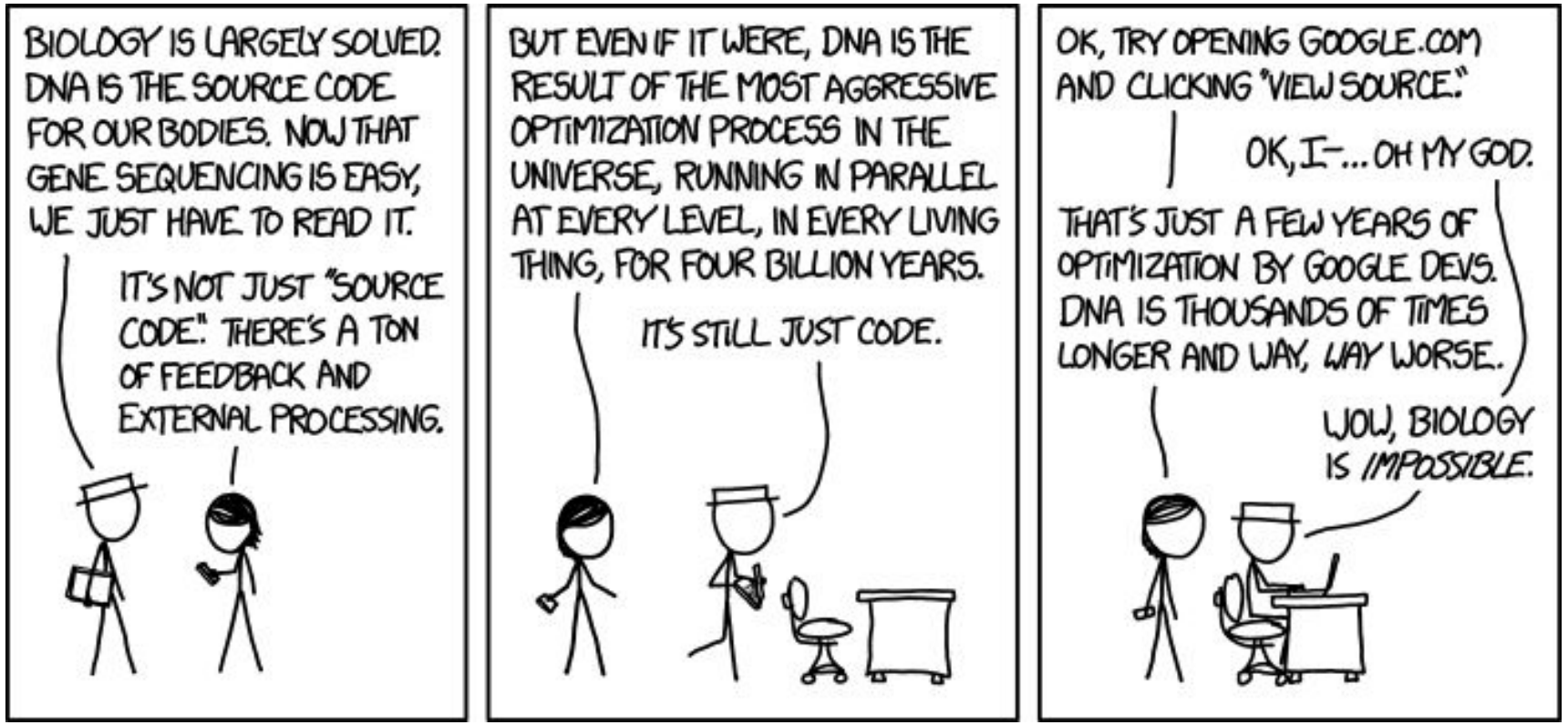
Follow-up of survey and offline-activities of Lecture 1

AMIDD Lecture 2: Biological Sequence Analysis



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

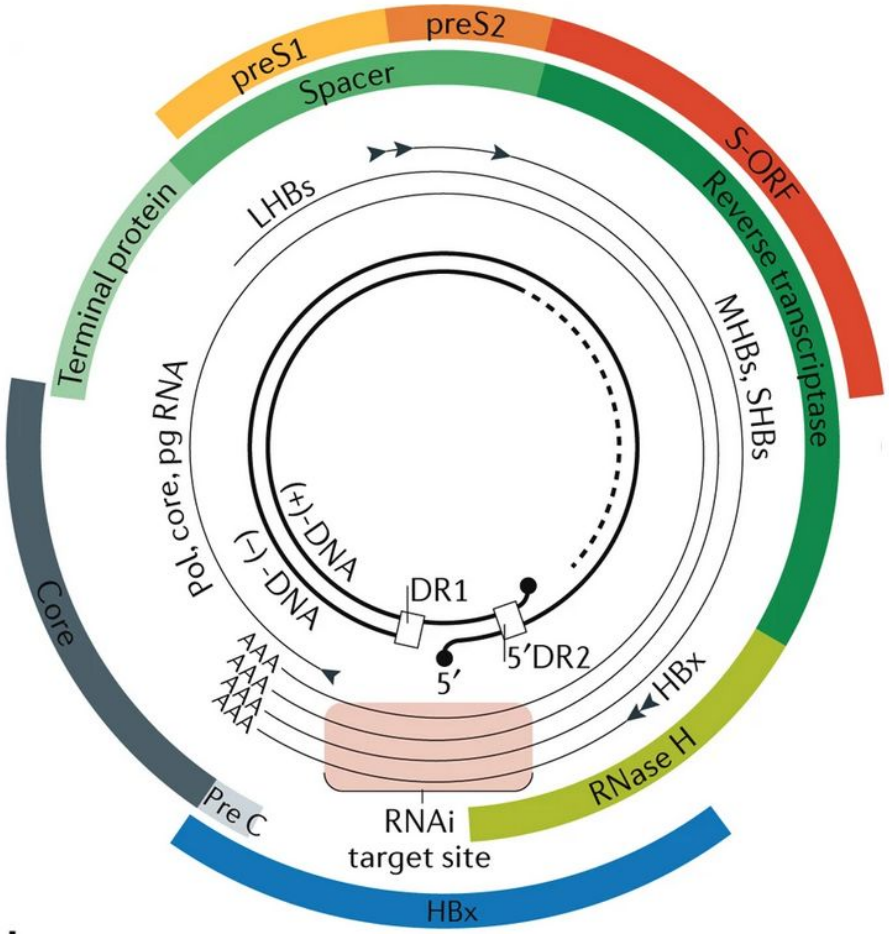
Part of the source code of *Google.com*

As of 24.09.2020


```
try{
_.Yq=function(a){_.z(this,a,0,-1,null,null)};_.v(_.Yq,_.y);_.Yq.prototype.bb=function(){retu
_.cr=function(a,b,c,d,e){_.K.call(this);this.B=b;this.W=d;this.F=e;this.M=!1;this.A={};this.
"";this.j=_.xc(_.E(a,17,1),1);a=0;for(b=c[a];a<c.length;a++,b=c[a])this.A[b]=!0,this.o[b]=!0
var dr=function(a,b,c,d){var e=_.Td("SCRIPT");e.async=!0;e.type="text/javascript";e.charset=
e.onload=function(){k()};e.onreadystatechange=function(){e.onreadystatechange=null;l(e)};e.c
var fr=function(a,b){var c=_.Td("LINK");c.setAttribute("rel","stylesheet");c.setAttribute("t
_.cr.prototype.D=function(a,b){if(!this.M)if(void 0!=b)window.setTimeout((0,_.r)(this.D,this

}catch(e){_._DumpException(e)}
try{
var gr=function(a){_.z(this,a,0,-1,null,null)},nr,pr;_.v(gr,_.y);
var hr=[1,2,3,4,5,6,9,10,11,13,14,28,29,30,34,35,37,38,39,40,42,43,48,49,50,51,52,53,62,500]
(this.data.ved=f.ved,delete f.ved);a=[];for(var g in f)0!=a.length&&a.push(","),a.push(ir(g)
mr.prototype.log=function(a,b){try{if(this.j||(kr(a)?this.o:this.D)){var c=new lr(this.A,thi

}catch(e){_._DumpException(e)}
try{
var qr=function(){_.fm.A(_.Bc)},rr=function(a,b){var c=_.bm();c=_.Kq(c,qr);a.addEventListene
_.h=wr.prototype;_.h.gf=function(a){a&&xr(this)&&a!=xr(this)&&xr(this).Ve(!1);this.A=a};_.h.
_.h.Jk=function(a,b){this.C[a]=b};_.h.Uh=function(a){return!this.C[a.hc()]};_.h.Aj=function(
var yr=function(a){_.K.call(this);this.D=a;this.A=this.j=null;this.F=0;this.C={};this.o=!1;a
var zr=function(a,b,c){if(!a.o)if(c instanceof Array){c=_.ka(c);for(var d=c.next();!d.done;d
yr.prototype.B=function(a,b){if(this.o)return null;if(b instanceof Array){var c=null;b=_.ka(
yr.prototype.G=function(a,b){this.j=b;this.A=a;b.preventDefault?b.preventDefault():b.returnV
Ar.prototype.init=function(a,b,c){window.gapi={};var d=window.___jsl={};d.h=_.J(_.B(a,1));nu
(function(){var a;window.gbar&&window.gbar._LDD?a=window.gbar._LDD:a=[];var b=_.Rd();ur(wind
_.u("gbar.ldb",_.r(_.fm.A,_.fm,_.Bc));
_.u("gbar.mls",function(){});
_.Hc("eq",new yr(_.bm()));
_.Hc("gs",(new Ar).init(_.vd(),_.G(_.L(),_.Zq,5)||new _.Zq,_.G(_.L(),_.Yq,6)||new _.Yq));
(function(){for(var a=function(e){return function(){var f={n:e};_.or().log(44,f)}},b=0;b<_.X
var Br=function(a){_.gm(function(){var b=document.querySelector("."+a);b&&(b=b.querySelector
var Cr=document.querySelector(".gb_C"),Dr=/(\s+|^)gb_ig(\s+|$)/;Cr&&!Dr.test(Cr.className)&&
var Er=new wr(_.bm());_.Hc("dd",Er);_.u("gbar.close",(0,_.r)(Er.Cf,Er));_.u("gbar.cls",(0,_.
Br("gb_Sa");
_.gm(function(){var a=document.querySelector(".gb_uc");a&&zr(_.Qd("eq"),a,"click")});
_.u("gbar.qfgw",(0,_.r)(document.getElementById,document,"gbqfqw"));_.u("gbar.qfgq",(0,_.r)(

}catch(e){_._DumpException(e)}
})(this.gbar_);
// Google Inc.
</script><div class="gb_Fa"><div class="gb_5a gb_F gb_1 gb_7a" aria-label="Kontoinformatione
var a=m;window.W_jd=window.W_jd||{};for(var b=0;b<a.length;b+=2)window.W_jd[a[b]]=JSON.parse
var k=this||self,l=function(){},m=function(a){var b=typeof a;return"object"==b&&null!=a||"fu
K}I=!J}var ia=I,ja=function(){if(!k.addEventListener||!Object.defineProperty)return!1;var a=
e?b=a.fromElement:"mouseout"==e&&(b=a.toElement);this.relatedTarget=b;c?(this.clientX=void 0
a.metaKey;this.pointerId=a.pointerId||0;this.pointerType="string"===typeof a.pointerType?a.p
b.concat(),a=0;a<b.length;a++){var f=b[a];f&&f.capture==e&&!f.h&&(f=ya(f,c),d=d&&!1!==f)}ret
c.length-1;0<=d;d--){b.a=c[d];var f=za(c[d],a,!0,b);e=e&&f}for(d=0;d<c.length;d++)b.a=c[d],f
(function(){var c=google.time();if(google.timers&&google.timers.load.t){for(var a=document.g
```

~190k characters (excluding spaces)

Hepatitis B virus genome: 3.2kb, infecting ~290 million people



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
gcagagtctagactcgtggtggacttctctcaatttttctaggggttccacccgtgtgtcgtggccaaaattcgcagtccc
caacctccaatcactcaccaacctcctgtcctccaatttgtcctggttatcgcttgatgtgtctgcggcgttttatcata
ttcctcttcatcctgctgctatgcctcatcttcttgttggttcttctggattaccaaggtatgttgcccgtttgtcctct
aattccaggatccacaacaaccagtacgggaccctgcaaaacctgcacgactcctgctcaaggcaactctatgtttccct
cctgttgctgtacaaaacctacggatggaaattgcacctgtattcccatcccatcatcttgggctttcgtaaaatacccta
tgggagtgggcctcagtccgtttctcttggctcagtttactagtgccatttgttcagtggttcgtagggctttcccccac
tgtttggctttcagttatatggatgatgtggtattgggggccaaatctgtacaacatcttgagtccctttataccgctgt
taccaattttcttttgtctttgggtatacatttaaaccctaacaaaacaaagagatggggttattccctaaacttcatgg
gatacgtaattggaagttggggtacattgccacaggatcatattgtacaaaaaatcaaacactgttttaggaaacttcct
gtcaatcgacctattgattggaaagtatgtcaaagaattgtgggtcttttgggctttgccgctccatttacacaatgtgg
ttaccctgccttaatgcctttgtatgcatgtatacaagcgaaacaggcttttactttctcgccaacttacaaggcctttc
taagtaaacagtatatgaacctttaccccgttgcccggcaacggcctggtctgtgccaagtgtttgctgacgcaaccccc
actggttggggcttggctatcggccatcagcgcatgcgtggaacctttgtggctcctctgccgatccatactgcggaact
cctagctgcttgttttgctcgcagccggtctggagcgaaactcattgggactgataattctgtcgtcctttctcggaaat
atacatcatttccatggctgctaggttgtgctgccaactggattcttcgcggaacgtcctttgtttacgtcccgtcggcg
ctgaatcccgcggacgacccctcccggggccgcttgggactctatcgtccccttctccgtctgccgtaccgtccgaccac
ggggcgcacctctctttacgcggtctccccgtctgtgccttctcatctgccggtccgtgtgcacttcgcttcacctctgc
acgttgcatggagaccaccgtgaacgcccatcagagcctgcccaaggtcttacataagaggactcttggactcccagcaa
tgtcaacgaccgaccttgaggcctacttcaaagactgtgtgtttaaagactgggaggagttgggggaggagattaggtta
atgatctttgtattaggaggctgtaggcataaattggtctgcgcaccatcatcatgcaacttttcacctctgcctaatc
atctcttgtacatgtcccactattcaagcctccaagctgtgccttggctggctttggggcatggacattgacccttataa
agaatttggagctagtgtggagttactctcgttttgccttctgacttctttccttcagtccgggatctacttgatacag
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 and Dynamic Programming

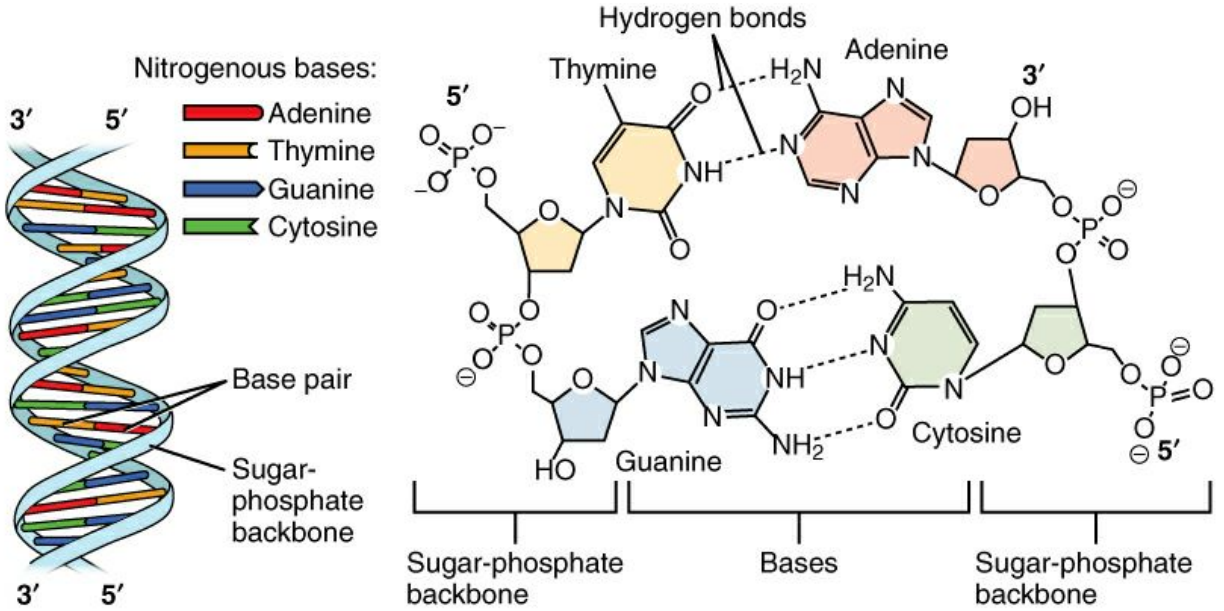
The central dogma of molecular biology

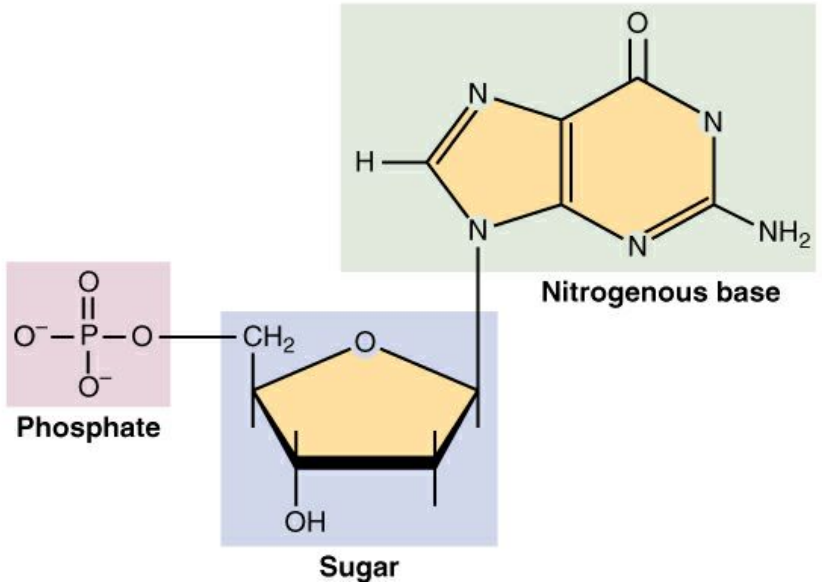
The Central Dogma can be represented by a graph of chemical information vehicles (nodes) and biological information flows (edges)

DNA

Photo 51, X-ray diffraction image of DNA

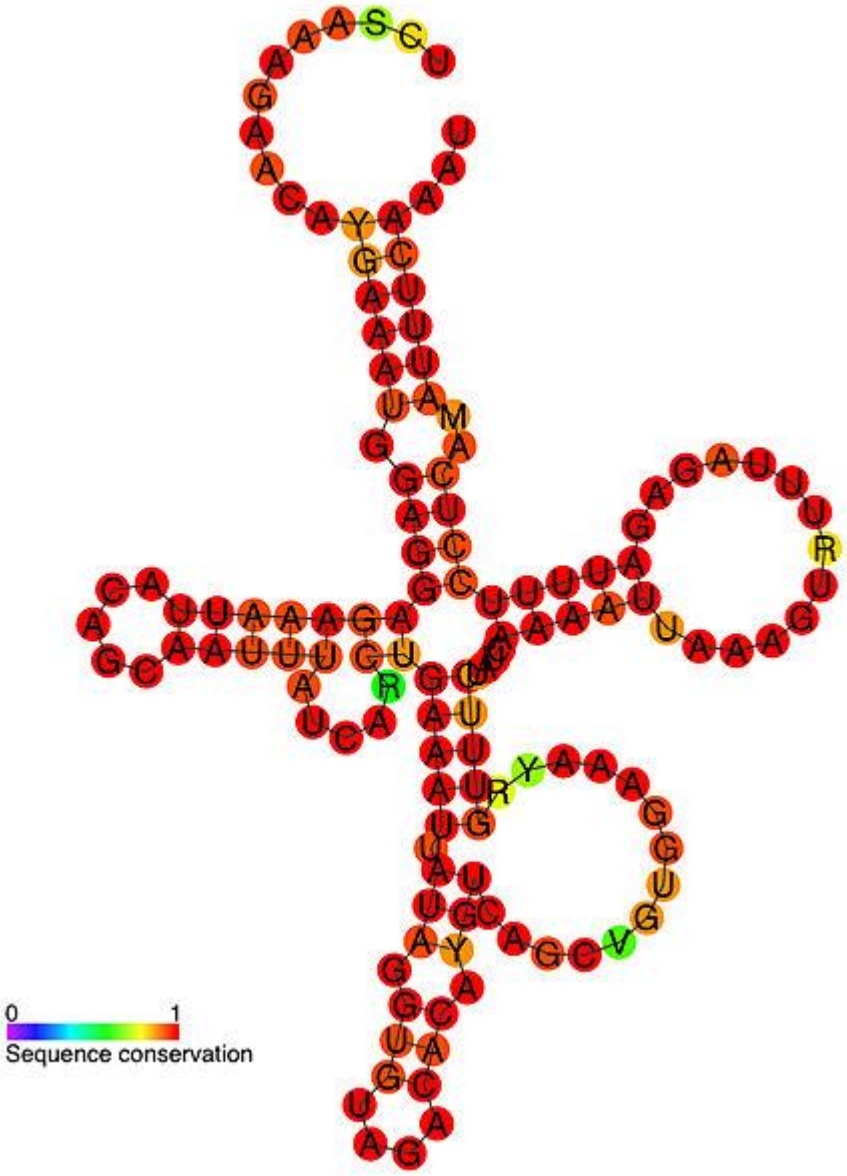
Franklin R, Gosling RG (1953) "Molecular Configuration in Sodium Thymonucleate". ***Nature*** 171**: 740–741.**





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

RNA structure

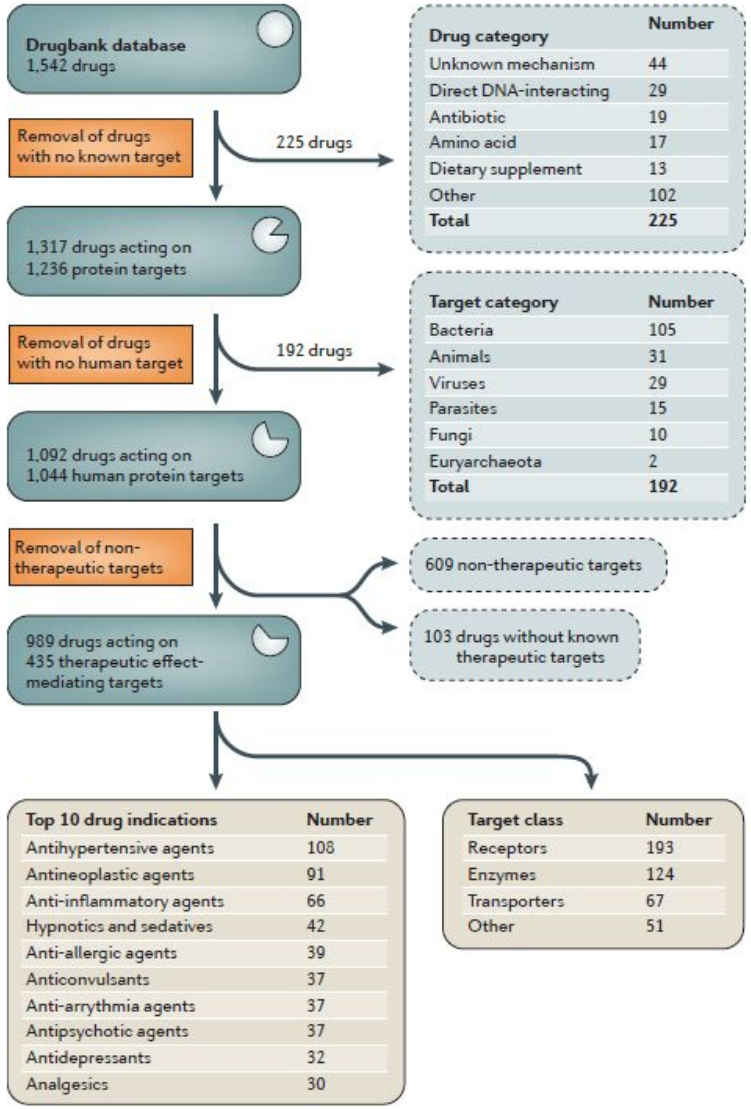


Downloaded from https://en.m.wikipedia.org/wiki/File%3AHAR1F_RF00635_rna_secondary_structure.jpg. Original work by wikipedia user:Ppgardne. Used under CC-SA 3.0 license.

Drugs work by targeting nodes or edges of the central dogma

Target	Example drugs or therapeutic candidates
Protein	• Most small-molecules, for instance GPCR agonists or antagonists, 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

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**

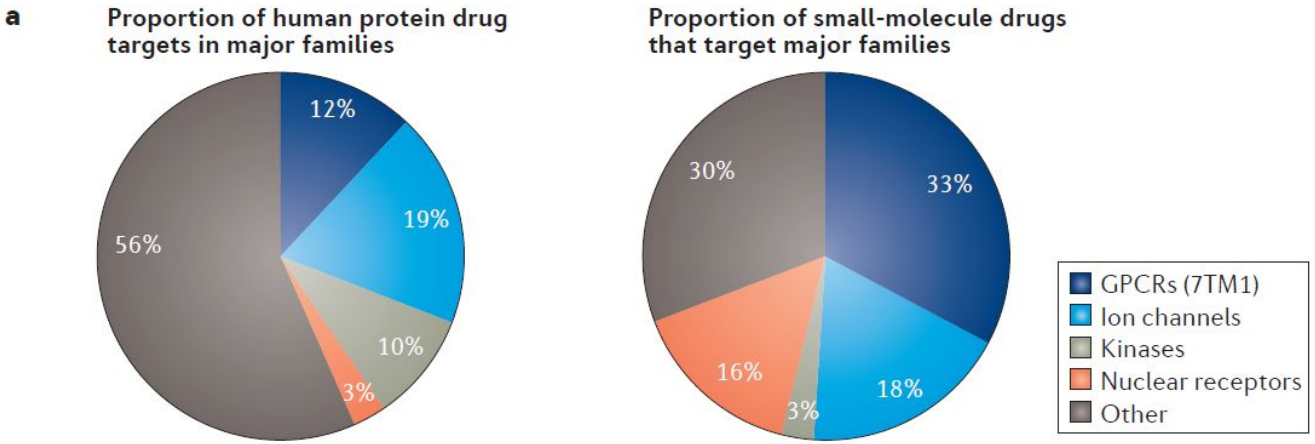
Top 10 drug indications	Number
Antihypertensive agents	108
Antineoplastic agents	91
Anti-inflammatory agents	66
Hypnotics and sedatives	42
Anti-allergic agents	39
Anticonvulsants	37
Anti-arrythmia agents	37
Antipsychotic agents	37
Antidepressants	32
Analgesics	30

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.

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?

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.

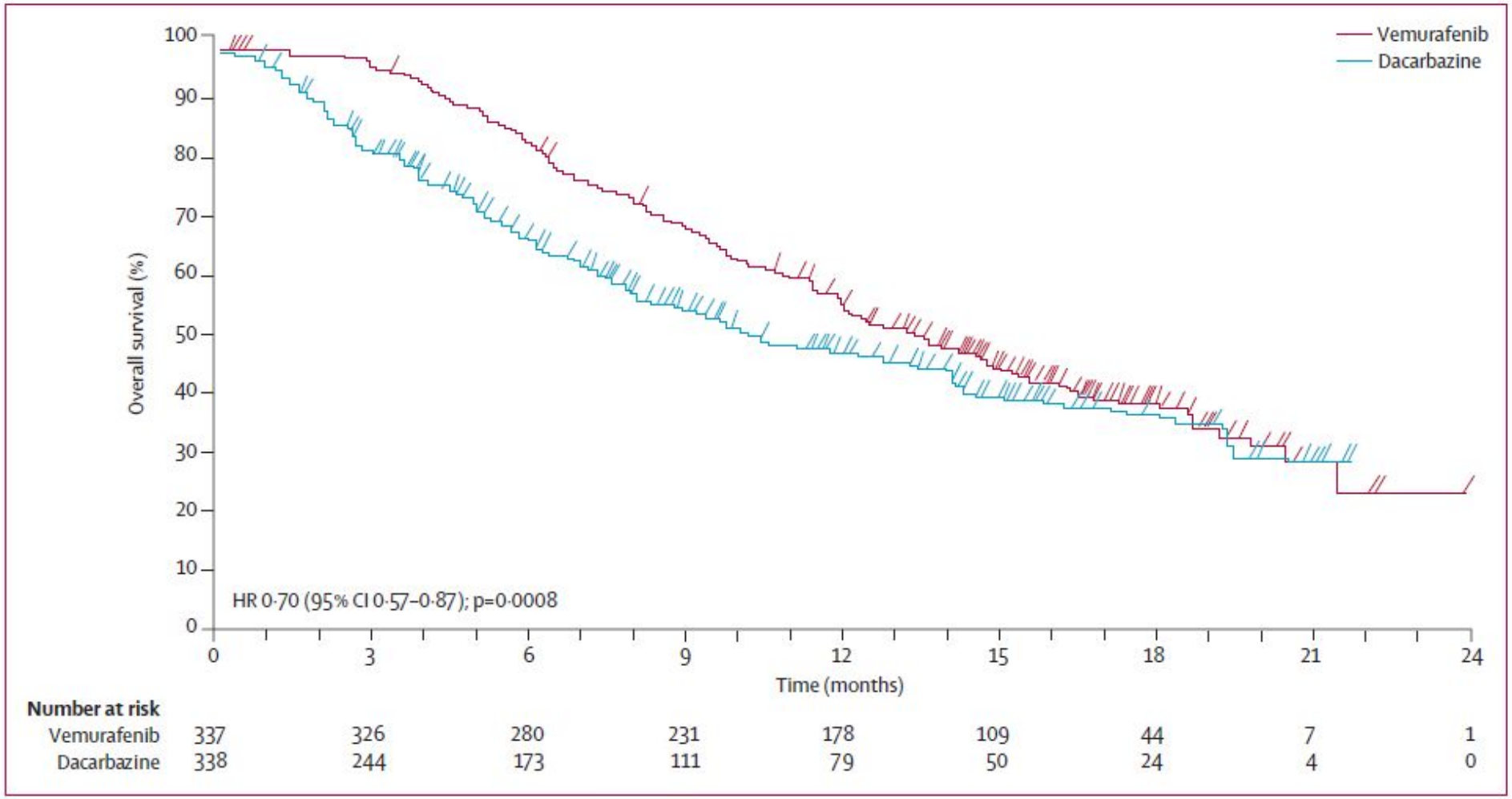
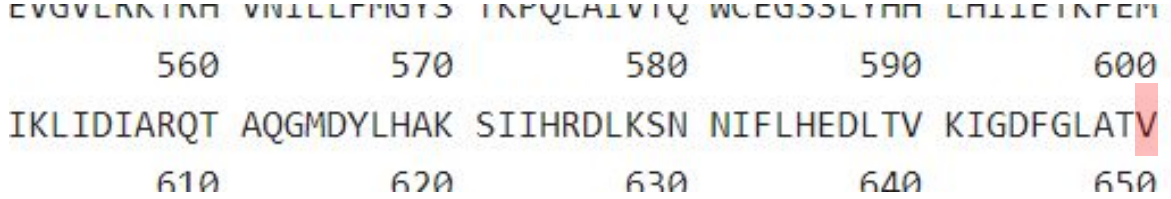


Figure 2: Overall survival (randomised population; censored at crossover) for patients randomly assigned to vemurafenib or to dacarbazine (cutoff Feb 1, 2012)

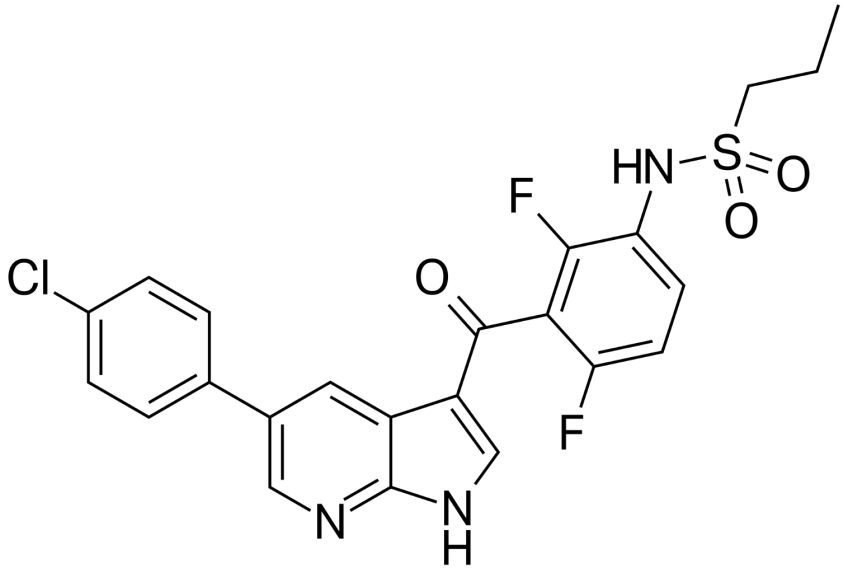
Vemurafenib (Zelboraf, PLX4032)
V600E mutated BRAF inhibition

- 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.



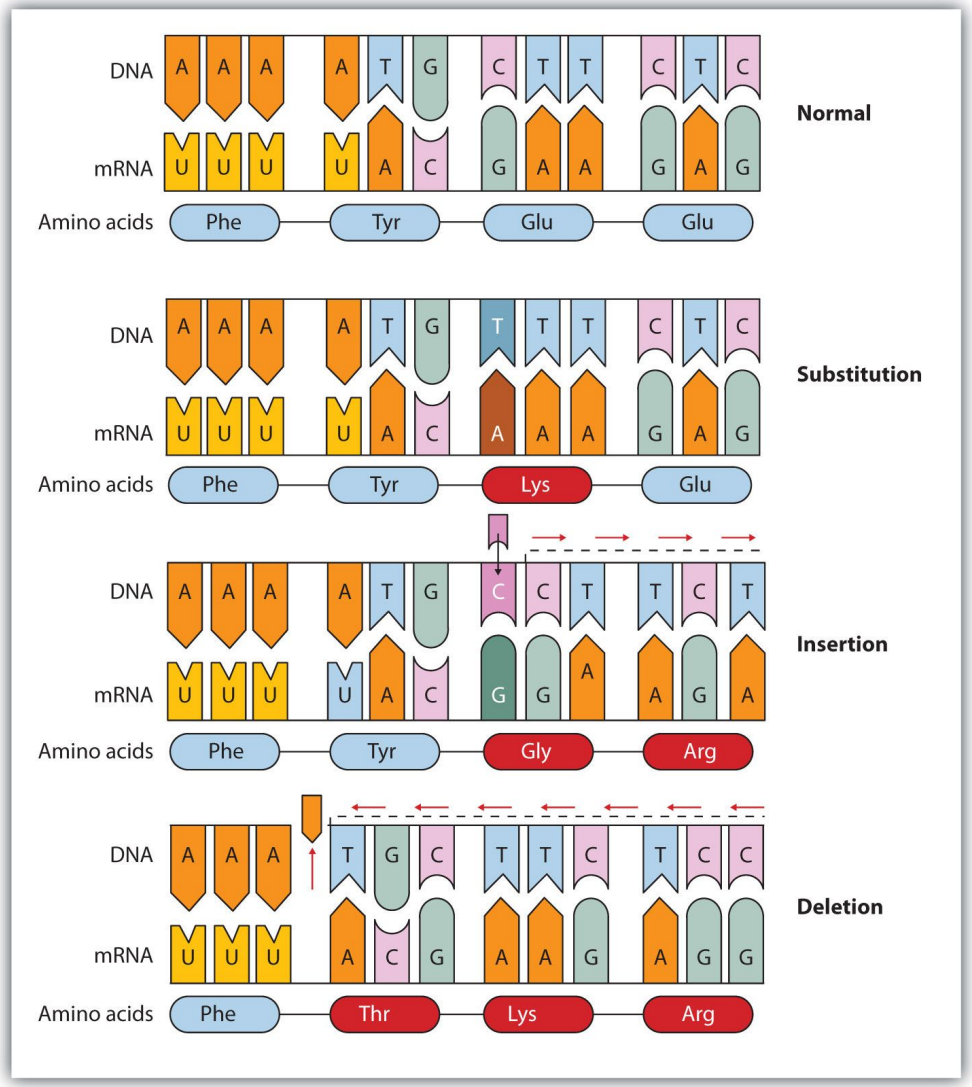
Source: https://commons.wikimedia.org/wiki/File:Vemurafenib_structure.svg

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

The Levenshtein distance

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

- Insertion, for instance **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

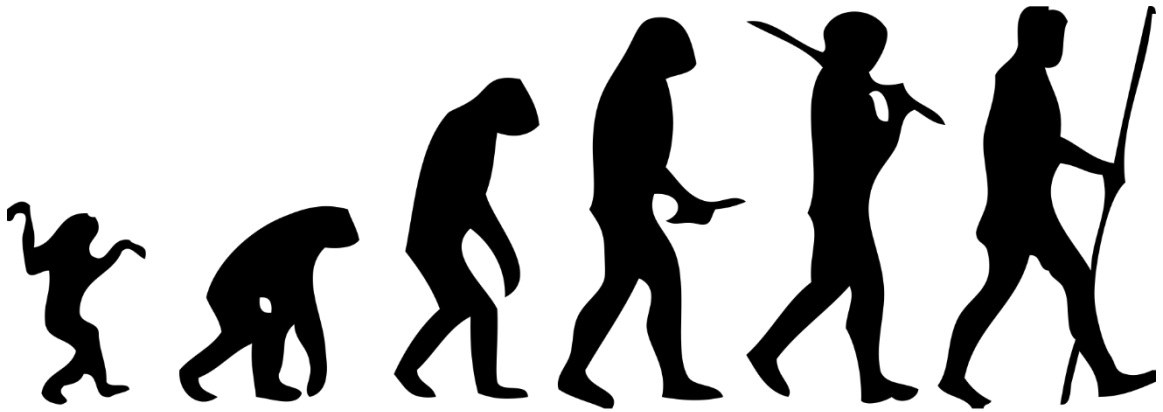
```
ATGC
A-GC
```

Calculate the Levenshtein distance with dynamic programming

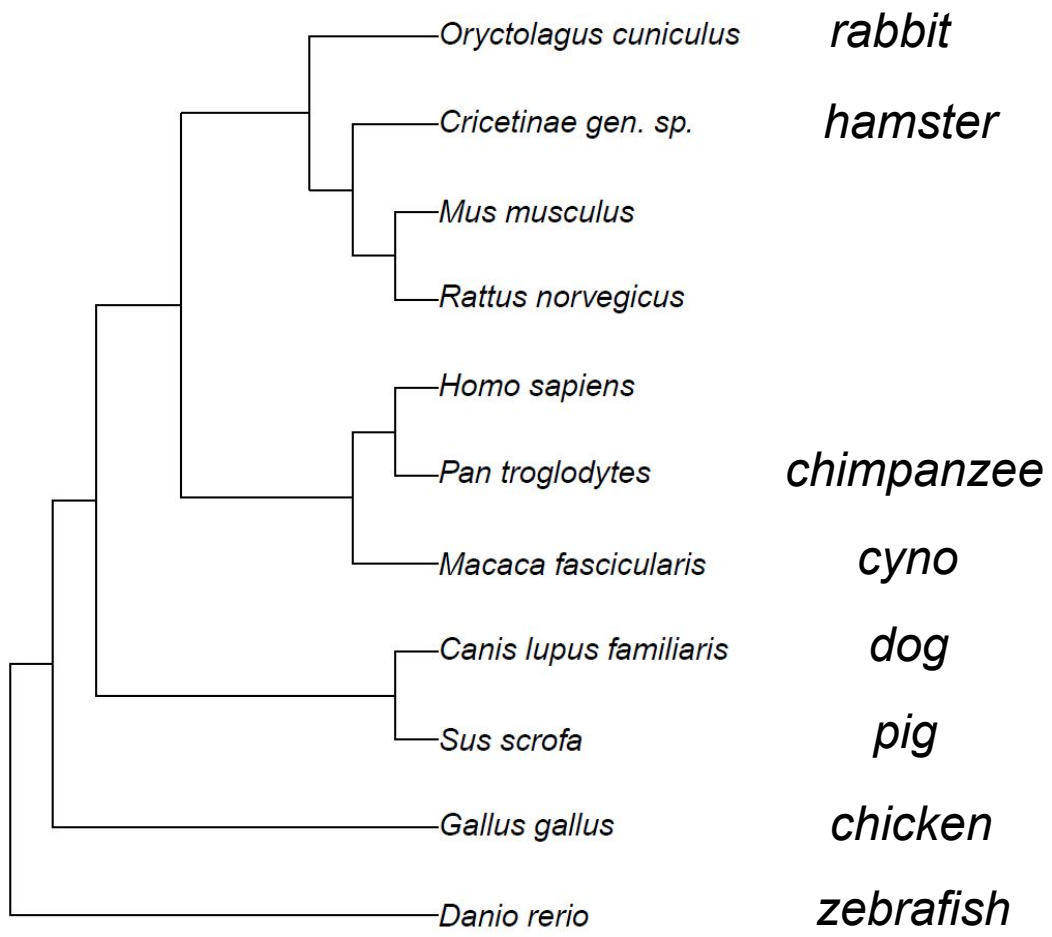
- What is the Levenshtein distance between `ACTGCTT` and `ACATT`?
- Beyond bioinformatics, the Levenshtein distance is often used in computational linguistics and natural language processing. For instance, check out How to Write a Spelling Corrector by Peter Norvig.

		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

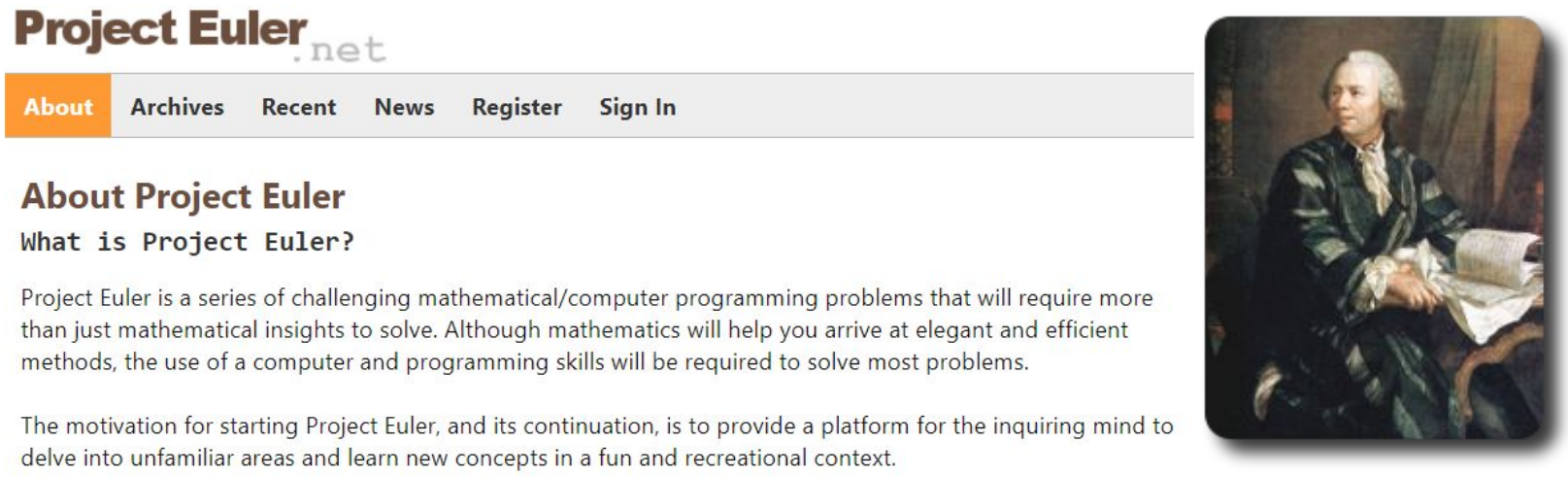


Tree structure retrieved from https://itol.embl.de/ (iTOL, Interactive Tree of Life), visualized with the *FigTree* software developed by Andrew Rambaut

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 Elsie Franklin

1920-1958

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

→

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
```

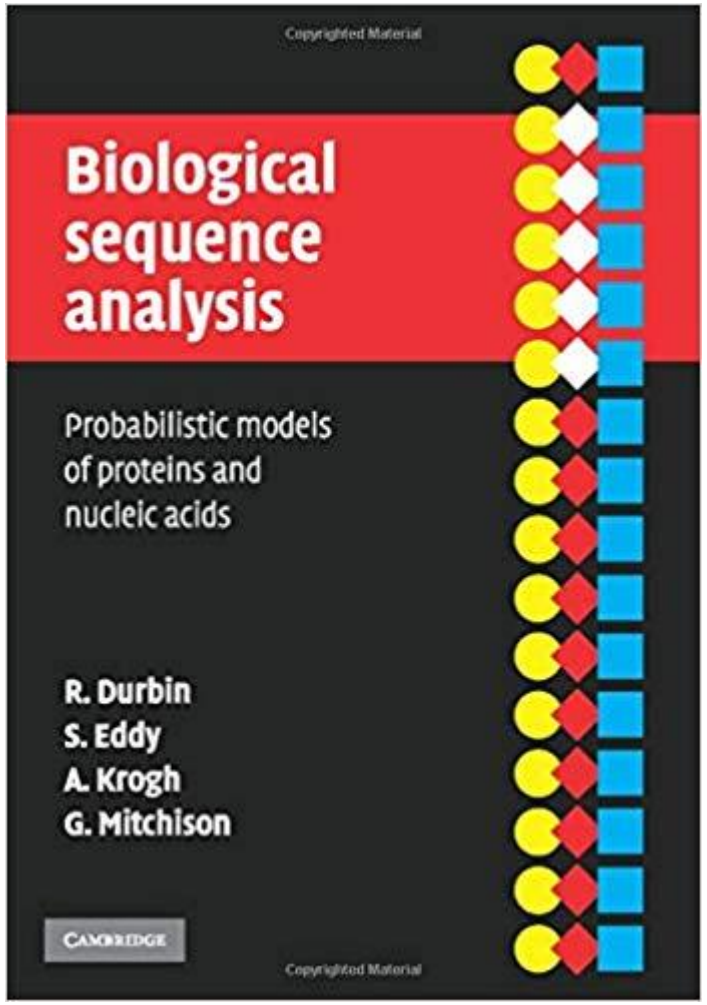
Sample Output

```
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.

Summary and Q&A