# Diamond code

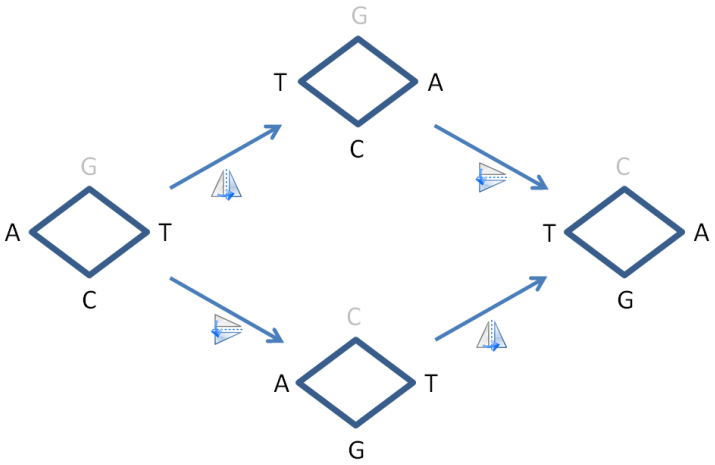
The Russian astrophysicist George Gamow is seen as the father of the Big Bang theory. A title he owes to his prediction of the cosmic microwave background radiation. Gamow was a creative thinker who felt quite at home taking a sidestep into another discipline. His contribution to cracking the genetic code is seen as "perhaps the last example of amateurism in scientific work on grand scale". After all, it was Gamow's idea that the base sequence in DNA might be the code for protein synthesis.

In the spring of 1953, Francis Crick and James Watson decoded the structure of deoxyribonucleic acid (DNA). They discovered that DNA — the molecular basis of heredity — consists of two intertwined strands that run in opposite directions. Each strand is a long molecule comprising a string of sugars, phosphate groups, and one of the following four bases: adenine (A), thymine (T), cytosine (C) or guanine (G). The burning question was soon raised: "How is the information in DNA converted into the production of amino acids, the building blocks of proteins?".

The first step in finding a solution came from an unexpected quarter. After reading the work of Watson and Crick, George Gamow wrote them a letter in the summer of 1953. He suggested that the base sequence in DNA might be the code for protein synthesis. As a physicist, Gamow's idea took the world of biology by storm. He had changed what had, until then, been seen as a chemical problem into purely a question of information storage and transfer. The underlying chemistry was of secondary importance.

Gamow had reduced the problem to the question: "How can a language of four letters provide a code for 20 amino acids?". It soon became clear that the four different bases had to be grouped in threes — in this context these triplets are often called *codons* — to make a unique code for each of the 20 amino acids possible. Groups of two only allow for 16 ($4 \times 4$) possibilities, while triplets provide 64 ($4\times 4 \times 4$) possibilities, which is more than enough.

Gamow himself made the first proposal, what is known as the **diamond code**. He thought that the protein synthesis occurred directly between the two strands of DNA. The four bases form a space in which an amino acid fits perfectly. Which acid that is, depends on the bases of the four corner points, hence the name *diamond*. The bases at the left and right corners of the diamond lie on the same strand, separated by a single base. This base and its complement on the opposite strand constitute the top and bottom corners of the diamond (A is complementary to T and C is complementary to G). In essence, Gamow's code was a three-letter code, as the top and bottom corners were complementary, so that only one of the two actually contained information.




The canonical representation of the codon ACT is determined by rotating the codon on the horizontal and/or vertical axis of the diamond representation of the codon. This results in three alternative codons: TCA (rotation on the vertical axis), AGT (rotation on the horizontal axis) and TGA (rotation on both the horizontal and vertical axis). The alphabetically first ranked of these four variants is called the canonical representation of the codon. In this case, the canonical representation is the codon ACT itself.

Gamow's diamond was also an overlapping code. Each base was part of three sequential codons. For example, the base sequence ATCGAT consisted of the four codons ATC, TCG, CGA and GAT. Gamow came up with an original solution for the 64 possible codons for only 20 amino acids. He suggested that the diamonds could, as it were, be rotated on both axes without that having any significance. If the ACT codon were rotated on the vertical axis, it would become TCA. Rotating it on the horizontal axis would replace the middle base with its complement, making it AGT. If all these symmetries are fully worked out, you end up with 20 unique combinations. The exact number Gamow was looking for.

## Assignment

In this exercise we will represent both DNA and protein sequences as strings that only contain uppercase letters. DNA sequences are limited to the letters A, C, G and T, that represent the possible nucleotides. A codon is a DNA sequence that has three letters. Protein sequences may contain each letter of the alphabet (in practice only 20 letters are used), which now represent the possible amino acids. Your task is to convert DNA sequences into their corresponding protein sequence according to the principles of Gamow's diamond code. Follow these steps to accomplish this task:

- Write a function canonical that returns the canonical representation of the codon that is passed as an argument to the function. The canonical representation of a given codon is determined by rotating it on the horizontal and/or vertical axis of the diamond representation of the codon. The canonical representation is the alphabetically first ranked of the (up to) four codons that result from these rotations.
- Use the function canonical to write a function codon2aa that takes a codon as its argument. The function must return a single letter that represents the amino acid corresponding to the given codon. The letter should be determined in the following way:
    1. Determine the canonical representation $b_1b_2b_3$ of the given codon.
    2. Compute \[ p = (w_1 + 4w_2 + 16w_3)\!\!\!\!\mod 25 \] In this, $w_i$ corresponds to the value of the nucleotide $b_i$ ($1 \leq i \leq 3$), with nucleotide G having value 0, T

having value 1, C having value 2 and A having value 3.
3. The value $p$ gives the position in the alphabet of the letter of the amino acid that corresponds to the given codon. Positions in the alphabet are indexed from zero, so that A is at position 0, B at position 1, C at position 2, …

- Use the function codon2aa to write a function dna2protein. This function should be passed a DNA sequence that has at least three nucleic acids. The function must return the corresponding protein sequence according to Gamow's diamond code. We remind you once more that this is an overlapping code.

## Example

```
>>> canonical('ACT')
'ACT'
>>> canonical('CGC')
'CCC'
>>> canonical('GTC')
'CAG'
>>> len(set([canonical(a + b + c) for a in 'ACGT' for b in 'ACGT' for c in 'ACGT']))
20

>>> codon2aa('ACT')
'C'
>>> codon2aa('CGC')
'R'
>>> codon2aa('GTC')
'O'
>>> len(set([codon2aa(a + b + c) for a in 'ACGT' for b in 'ACGT' for c in 'ACGT']))
20

>>> dna2protein('ATCGAT')
'WYSD'
>>> dna2protein('CCCTCCATCTAGTGCGTGTTCTGTCCGAAGGTATGTCATATCAC')
'RBVBSFWAWDCMBIBMADFAOAOBKSPPLYPEPAOCFNEWCV'
>>> dna2protein('ATTTAACGAATCTACCCGGAGTGGCAACTCAGGAGGACTCTTG')
'GEGGWLSPGWAWFSRKKLMCMYKLWWCVCOLLMLLOCVAFD'
```

## Epilogue

By the time of his trip to biology Gamow had already turned fifty and had a long academic career behind him. He was most famous for his work on quantum mechanics and nuclear physics. Gamow came to some remarkable predictions for his time, simply by applying accepted laws of nature to unusual situations. As such, he predicted in 1948 that there should be a measurable amount of cosmic background radiation if the universe had a hot and compact beginning. Almost twenty years later, the existence of cosmic background radiation was indeed confirmed experimentally.

After he had proposed the diamond code, however, Gamow soon realised that this code was not the correct solution. This was just as well, since it was very sensitive to mutations. With an overlapping code, mutation of one base can impact three successive amino acids. In the meantime, others had become convinced that protein synthesis did not directly occur in DNA, but that ribonucleic acid (RNA) acted as an intermediary. RNA is very similar to DNA, but consists of a single strand of sugars, phosphates, and bases. It also contains the base uracil (U) instead of thymine.

Although his diamond code proved incorrect, Gamow was not ready to throw in the towel. He had formed an informal group of scientists who were more or less involved in addressing the code problem. His *RNA Tie Club* had 20 regular members, one for each amino acid, and four honorary members, one for each base. Gamow himself was alanine (ALA), Watson was proline (PRO), and Crick tyrosine (TYR). The other members were mainly biologists, like Max Delbrück (tryptophan) and Erwing Chargaff (lysine), but Gamow did not repudiate his own background, enlisting a number of leading physicists, including Edward Teller (leucine) and Richard Feynman (glycine). Each member received a specially designed tie bearing a double helix and a tiepin with the acronym of their own personal amino acid. The RNA Tie Club's official notepaper carried the motto "Do or die, or don't try".

After the diamond code, Gamow came up with two alternative codes, one of which he devised together with Feynman. Even Teller, a nuclear physicist *pur sang*, took the time to propose an interesting scheme, in which each amino acid was encoded by two bases and the preceding amino acid. In 1957, Syndey Brenner (valine) abruptly put a stop to all overlapping codes, when they proved incompatible with his analysis of the sequence of amino acids in a number of proteins.

That same year, Crick launched an ingenious non-overlapping code. He claimed that there was only one way in which the base sequence could be read. Imagine that the base sequence AGACGAUUA coded for AGA, CGA and UUA. According to Crick, the triplets of the other two overlapping codes were "nonsense codons", with no significance at all. In this case, therefore, GAC and GAU on the one hand, and ACG and AUU on the other hand, would be nonsense codons. Crick's code was incorrect, but was called "the most elegant biological theory ever to be proposed and proved wrong".

With hindsight, the RNA Tie Club had been too focused on finding a neat explanation of why there are 64 codes for only 20 amino acids. They were brought down to earth in 1961 when Marshall Nirenberg and Heinrich Matthaei — neither members of the *RNA Tie Club* — announced that they were able to produce proteins with artificial RNA. The first RNA they tested was poly-U, a sequence or uracil bases. They discovered that UUU coded for the amino acid phenylalanine. Four years later, the whole coding problem was solved. Compared to the solutions proposed earlier, nature's solution seemed like a rather messy workaround. Some amino acids have only one codon, while others have four, and some even six. Although the real solution was less refined mathematically than his own idea, Gamow admitted that it had one great advantage: it was true.

## References


- **Sanger F, Tuppy H (1951)**. The amino acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochemical Journal* **49**, 463-481.
- **Watson JD, Crick FHC (1953)**. A structure of deoxyribose nucleic acid. *Nature* **171**, 737-738.
- **Gamow G (1954)**. Possible relation between deoxyribonucleic acid and protein structures. *Nature* **173**, 318.
- **Brenner S (1957)**. On the impossibility of all overlapping triplet codes in information transfer from nucleic acid to proteins. *Proceedings of the National Academy of Sciences of the USA* **43**, 687-694.

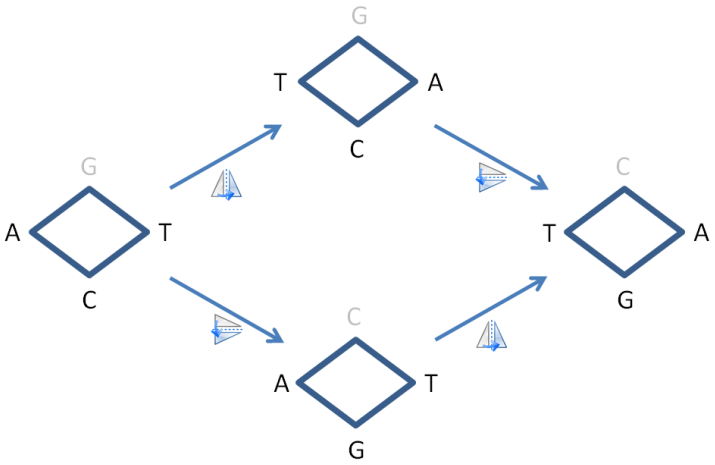
- **Crick FHC, Griffith JS, Orgel LE (1957)**. Codes without commans. *Proceedings of the National Academy of Sciences of the USA* **43**, 416-421.
- **Marshall NW, Matthaei J (1961)**. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences of the USA* **47**, 1588-1602.
- **Hayes B (1998)**. The invention of the genetic code. *American Scientist* **86**, 814.
- **Patel A (2001)**. Why genetic information processing could have a quantum basis. *Journal of Biosciences* **26(2)**, 145-151.
- **Sarabhai A (2003)**. After DNA at the MRC. *Journal of Biosciences* **28(6)**, 665-669.
- **Freeland SJ, Hurst LD (2004)**. Evolution encoded. *Scientific American* **290(4)**, 84-91.


De Russische astrofysicus George Gamow geldt als de vader van de oerknal. Die titel dankt hij aan zijn voorspelling van de kosmische achtergrondstraling. Gamow was een creatieve denker die niet vies was van een uitstapje naar een ander vakgebied. Zijn bijdrage aan de oplossing van de genetische code is wel eens benoemd als "het laatste voorbeeld van amateurisme in groots wetenschappelijk werk". Het was namelijk Gamows idee dat de basenvolgorde in DNA de code is achter de eiwitsynthese.

In het voorjaar van 1953 ontrafelden James Watson en Francis Crick de structuur van DNA. Zij ontdekten dat DNA — de moleculaire basis van erfelijkheid — bestaat uit twee in elkaar gedraaide ketens die in tegenovergestelde richting lopen. Iedere keten is een lange molecule bestaande uit een aaneenschakeling van eenheden die gevormd worden door een suiker, een fosfaatgroep en één van de volgende vier basen: adenine (A), thymine (T) , cytosine (C) of guanine (G). Maar de volgende prangende vraag diende zich al snel aan: "hoe wordt de informatie in DNA vertaald in de aanmaak van eiwitten?".

Nadat de astrofysicus George Gamow het werk van Watson en Crick had gelezen, schrijft hij het tweetal in de zomer van 1953 een brief. Hij oppert het idee dat de basenvolgorde in DNA de code vormt voor de eiwitsynthese. Gamows idee slaat in de wereld van de biologen in als een bom. Wat tot dan toe door iedereen als een chemisch probleem werd benaderd, verandert Gamow in een informatietheoretisch vraagstuk. De onderliggende chemie is daarbij niet belangrijk. Gamow had het probleem gereduceerd tot de volgende vraag: "hoe kan een taal van vier letters coderen voor twintig aminozuren?". Men bedacht al snel dat de vier verschillende basen in drietallen gegroepeerd moesten zijn — in deze context worden dergelijke tripletten *codons* genoemd — om een unieke codering voor elk van de twintig aminozuren mogelijk te maken. Tweetallen geven immers maar 16 ($4 \times 4$) mogelijkheden, en codons geven 64 ($4\times 4 \times 4$) mogelijkheden.

Gamow kwam zelf met het eerste voorstel, de zogenaamde **diamantcode**. In zijn gedachtengang vond de eiwitsynthese plaats direct tussen de twee strengen van DNA. Vier basen vormen een ruimte waar volgens Gamow precies één aminozuur in zou passen. Het type aminozuur was dan afhankelijk van de basen op de vier hoekpunten, vandaar de naam 'diamant'. De basen die het linker- en rechterhoekpunt van de diamant vormen liggen op dezelfde streng, gescheiden door één andere base. Die laatste vormt samen met haar complement op de tegenoverliggende streng het onderste en bovenste hoekpunt (A is complementair met T, en C is complementair met G) . In essentie was Gamows code een drielettercode omdat het onderste en bovenste hoekpunt complementair zijn, zodat van dat tweetal slechts één base werkelijk informatie draagt.




De canonische voorstelling van een gegeven codon ACT wordt bepaald door het codon te spiegelen over de horizontale en/of verticale as van de diamantvoorstelling van het codon. Dit levert drie alternatieve codons op: TCA (spiegeling over verticale as), AGT (spiegeling over horizontale as) en TGA (spiegeling over horizontale en verticale as). De alfabetisch eerst gerangschikte van deze vier varianten wordt de canonische voorstelling genoemd. In dit geval is dit dus het codon ACT zelf.

Gamows diamantcode was ook een overlappende code. Iedere base was onderdeel van drie opeenvolgende codons. De basenvolgorde ATCGAT bestond bijvoorbeeld uit de vier codons ATC, TCG, CGA en GAT. Voor het probleem van de 64 mogelijke codons voor slechts 20 aminozuren bedacht Gamow een originele oplossing. Hij stelde dat de diamanten als het ware gedraaid konden worden over beide assen, zonder dat de betekenis zou veranderen. Het codon ACT wordt bij spiegeling over de verticale as omgezet in TCA. Bij spiegeling over de horizontale as verandert de middelste base in haar complement, en ontstaat dus AGT. Indien al deze symmetriën worden uitgewerkt, kom je uit op twintig mogelijkheden. Precies wat Gamow zocht.

## Opgave

Zowel DNA- als eiwitsequenties worden in deze opgave voorgesteld als strings die enkel bestaan uit hoofdletters. Bij DNA beperkt de reeks van letters zich tot A, C, G en T, die in dit geval nucleotiden voorstellen. Een codon is dan een DNA-sequentie die bestaat uit drie letters. Eiwitsequenties kunnen alle mogelijke hoofdletters bevatten, die in dit geval aminozuren voorstellen. Je opdracht bestaat erin DNA-sequenties om te zetten naar eiwitsequenties volgens het principe van Gamows diamantcode. Hiervoor ga je als volgt te werk:

- Schrijf een functie canonisch die de canonische voorstelling teruggeeft van het codon dat als argument aan de functie wordt doorgegeven. De canonische voorstelling van een gegeven codon wordt bepaald door het codon te spiegelen over de horizontale en/of verticale as van de diamantvoorstelling van het codon. De canonische voorstelling is de alfabetisch eerst gerangschikte van de (maximaal) vier codons die hieruit resulteren.
- Gebruik de functie canonisch om een functie codon2aa te schrijven waaraan een codon als argument moet doorgegeven worden. De functie moet één enkele letter teruggeven die het aminozuur voorstelt dat correspondeert met het gegeven codon. Deze letter moet op de volgende manier bepaald worden:
    1. Bepaal de canonische voorstelling $b_1b_2b_3$ van het gegeven codon.
    2. Bereken \[ p = (w_1 + 4w_2 + 16w_3)\!\!\!\!\mod 25 \] Hierbij stelt $w_i$ de waarde voor van de nucleotide $b_i$ ($1 \leq i \leq 3)$, waarbij de nucleotide G de waarde 0

heeft, T de waarde 1, C de waarde 2 en A de waarde 3.
3. De waarde $p$ geeft de positie in het alfabet aan van de letter die het gezochte aminozuur voorstelt. De posities in het alfabet worden hierbij genummerd vanaf nul, dus staat A op positie 0, B op positie 1, C op positie 2, …

- Gebruik de functie codon2aa om een functie dna2eiwit te schrijven. Aan deze functie moet een DNA-sequentie doorgegeven worden die uit minstens drie nucleotiden bestaat. De functie moet de corresponderende eiwitsequentie teruggeven die men bekomt door Gamows diamantcode toe te passen. We herinneren je er nogmaals aan dat dit een overlappende code is.

## Voorbeeld

```
>>> canonisch('ACT')
'ACT'
>>> canonisch('CGC')
'CCC'
>>> canonisch('GTC')
'CAG'

>>> codon2aa('ACT')
'C'
>>> codon2aa('CGC')
'R'
>>> codon2aa('GTC')
'O'

>>> dna2eiwit('ATCGAT')
'WYSD'
>>> dna2eiwit('CCCTCCATCTAGTGCGTGTTCTGTCCGAAGGTATGTCATATCAC')
'RBVBSFWAWDCMBIBMADFAOAOBKSPPLYPEPAOCFNEWCV'
>>> dna2eiwit('ATTTAACGAATCTACCCGGAGTGGCAACTCAGGAGGACTCTTG')
'GEGGWLSPGWAWFSRKKLMCMYKLWWCVCOLLMLLOCVAFD'
```

## Epiloog

Ten tijde van zijn uitstapje naar de biologie was Gamow vijftig jaar en had hij al een hele wetenschappelijke carrière achter de rug. Hij was vooral beroemd vanwege zijn werk over kwantummechanica en nucleire fysica. Gamow kwam voor die tijd tot opmerkelijke voorspellingen, simpelweg door het toepassen van geaccepteerde natuurwetten op ongebruikelijke situaties. Zo voorspelde hij in 1948 dat er een meetbare hoeveelheid kosmische achtergrondstraling aanwezig moet zijn indien het universum een heet en compact begin heeft gekend. Bijna twintig jaar later werd het bestaan van de kosmische achtergrondstraling inderdaad met metingen bevestigd.

Al vrij snel realiseerde Gamow zich dat zijn diamantcode niet de juiste oplossing was. En gelukkig maar voor ons, want een groot nadeel van zijn idee was de grote gevoeligheid voor mutaties. Bij een overlappende code zal één mutatie van een base namelijk doorwerken in drie opeenvolgende aminozuren. Maar Gamow liet zich niet snel uit het veld slaan. Inmiddels had hij een informeel forum gevormd van wetenschappers die zich in meer of mindere mate bezig hielden met het codeerprobleem, de *RNA Tie Club*. De club bestond uit twintig gewone leden, één voor elk aminozuur, en vier ereleden, één voor elke base. Gamow zelf was alanine (ALA), Watson was proline (PRO) en Crick tyrosine (TYR). Ieder lid kreeg een speciaal ontworpen

stropdas met een afbeelding van een dubbele helix, en een bijhorende dasspeld met de afkorting van zijn eigen persoonlijk aminozuur.

Na het struikelen van de diamantcode passeerden diverse coderingen de revue, de één nog mooier dan de ander. Ondertussen was men ervan overtuigd geraakt dat de eiwitsynthese niet direct aan het DNA plaatsvond, maar dat RNA een intermediair was tussen DNA en eiwitsynthese. Gamow kwam zelf nog met een driehoekscode op de proppen, terwijl Crick het idee lanceerde van een zogenaamde kommavrije code, een niet-overlappende drielettercode. Achteraf gezien heeft men zich te veel vastgebeten in het zoeken naar een fraaie oplossing die impliciet zou verklaren waarom er 64 codes zijn voor maar twintig aminozuren.

De grote ontnuchtering kwam in 1961 toen Marshall Nirenberg en Heinrich Matthaei — beiden geen lid van de *RNA Tie Club* — aankondigden dat ze in staat waren om eiwit te produceren met kunstmatig geproduceerd RNA. Het eerste RNA dat ze testten was poly-U, een aaneenschakeling van uracilbasen (in RNA neemt uracil de plaats in van thymine in DNA). Daaruit bleek dat UUU codeerde voor het aminozuur fenylalanine (PHE). Vier jaar later was het hele codeerprobleem opgelost. In vergelijking met de eerdere oplossingen was de echte oplossing maar een slordig geheel. Sommige aminozuren hebben één codon, andere aminozuren vier en sommige zelfs zes. Hoewel de werkelijke oplossing wiskundig minder geraffineerd is dan zijn eigen idee, had het — aldus Gamow — wel het grote voordeel de waarheid te zijn.

## Bronnen


- **Sanger F, Tuppy H (1951)**. The amino acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochemical Journal* **49**, 463-481.
- **Watson JD, Crick FHC (1953)**. A structure of deoxyribose nucleic acid. *Nature* **171**, 737-738.
- **Gamow G (1954)**. Possible relation between deoxyribonucleic acid and protein structures. *Nature* **173**, 318.
- **Brenner S (1957)**. On the impossibility of all overlapping triplet codes in information transfer from nucleic acid to proteins. *Proceedings of the National Academy of Sciences of the USA* **43**, 687-694.
- **Crick FHC, Griffith JS, Orgel LE (1957)**. Codes without commans. *Proceedings of the National Academy of Sciences of the USA* **43**, 416-421.
- **Marshall NW, Matthaei J (1961)**. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences of the USA* **47**, 1588-1602.
- **Hayes B (1998)**. The invention of the genetic code. *American Scientist* **86**, 814.
- **Patel A (2001)**. Why genetic information processing could have a quantum basis. *Journal of Biosciences* **26(2)**, 145-151.
- **Sarabhai A (2003)**. After DNA at the MRC. *Journal of Biosciences* **28(6)**, 665-669.
- **Freeland SJ, Hurst LD (2004)**. Evolution encoded. *Scientific American* **290(4)**, 84-91.