**Instituto de Ciências Biomédicas / Universidade de São Paulo**

**Elisabete José Vicente**

pasted-image.tiff

**Introdução à Bioinformática - do DNA à proteína (Roteiro de aulas práticas)**

pasted-image.tiff

**Fig 1**: **Dogma central da Biologia:** O fluxo de informação genética do [DNA](https://www.ncbi.nlm.nih.gov/books/n/mboc4/A4754/def-item/A5084/) para o [RNA](https://www.ncbi.nlm.nih.gov/books/n/mboc4/A4754/def-item/A5756/) (**transcrição**) e do RNA para a [protein](https://www.ncbi.nlm.nih.gov/books/n/mboc4/A4754/def-item/A5688/)**a** (**tradução**) ocorre em todas as células vivas. [Copyright](https://www.ncbi.nlm.nih.gov/books/about/copyright/) © 2002, Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter; Copyright © 1983, 1989, 1994, Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, and James D. Watson

(Fonte: <https://www.ncbi.nlm.nih.gov/books/NBK21050/figure/A974/> )

**Reutilização deste material:** Salvo indicação em contrário, os conteúdos podem ser reutilizados não comercialmente sem o pedido de permissão, lembrando sempre de fazer a devida citação.

2019.05.17

**CAPÍTULO 1: INTRODUÇÃO**

**1. Regras gramaticais para representação de uma sequência de DNA**

**A seguir você aprenderá como escrever sequências de DNA e se familiarizará com as convenções usadas para sua representação.**

O **DNA** (ácido desoxirribonucleico) e o **RNA** (ácido ribonucleico) são os dois tipos de moléculas conhecidas por registrar e transmitir informações genéticas hereditárias de uma geração para a seguinte (Fig 1).

O uso de **RNA** para esse propósito é limitado a uma família de vírus chamada **RNA vírus** - **o restante dos organismos biológicos conhecidos usa o DNA para armazenar e transmitir informações genéticas**.

Com poucas exceções (que incluem gametas, certas células imunes e células tumorais), o conteúdo de DNA de todas as células de um organismo é idêntico.

A informação codificada no **DNA** forma o **mapa genético** para a **produção de proteínas** e outros componentes importantes da maquinaria celular. Estruturalmente, forma uma grande seqüência de nucleotídeos ligados covalentemente uns aos outros, formando uma espinha dorsal de anéis de desoxirribose (açúcar) e grupos fosfato aos quais as bases estão ligadas, como mostrado no diagrama abaixo. Existem três bases que são encontradas no DNA e no RNA: **adenina**, **citosina** e **guanina**. A **timina** é encontrada exclusivamente no **DNA** e é substituída pelo **uracila** exclusivamente no **RNA.**

|  |  |
| --- | --- |
| pasted-image.tiff | **Figura 2:** O **DNA** e seus blocos construtores: O [DNA](https://www.ncbi.nlm.nih.gov/books/n/mboc4/A4754/def-item/A5084/) é feito de 4 tipos de nucleotídios (adenina (**A**), citosina (**C)**, guanina (**G)** e timina (**T**) ) que são ligados covalentemente em uma cadeia polinucleotídica (“DNA strand”) com um esqueleto de açúcar ([sugar](https://www.ncbi.nlm.nih.gov/books/n/mboc4/A4754/def-item/A5842/)-phosphate backbone) no qual as bases (A, C, [G](https://www.ncbi.nlm.nih.gov/books/n/mboc4/A4754/def-item/A5198/), and T) estão ligadas. A molécula de DNA é composta por duas fitas de DNA mantidas juntas por ligações do tipo pontes de hidrogênio entre as bases pareadas. As setas nas extremidades das fitas de DNA indicam as polaridades de cada uma das duas fitas, que correm [antiparallel](https://www.ncbi.nlm.nih.gov/books/n/mboc4/A4754/def-item/A4834/)as uma a outra.    [Copyright](https://www.ncbi.nlm.nih.gov/books/about/copyright/) © 2002, Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter; Copyright © 1983, 1989, 1994, Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, and James D. Watson. (Fonte:<https://www.ncbi.nlm.nih.gov/books/NBK21050/figure/A974/> |

O DNA genômico (com exceção de alguns vírus) é de cadeia dupla. Isto significa que para cada cadeia de DNA com uma determinada sequência, há uma cadeia oposta que contém a **sequência complementar** onde **A complementa T** e **G complementa C**. Essa complementaridade é chamada de “**regra de pareamento de Watson-Crick**” em honra dos dois pesquisadores. que publicaram o primeiro artigo descrevendo o pareamento de bases encontrado na estrutura do DNA (<https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397>).

Ao representar os dois filamentos da molécula de DNA, precisamos escrever os dois lados, seguindo a regra de Watson-Crick de **A = T** e **G = C**:

**5’**-ATGCGATCGGACAGTCGAGTCCAGTAGACGATC-**3’** forward strand

**3’**-TACGCTAGCCTGTCAGCTCAGGTCATCTGCTAG-**5’** reverse strand

Neste exemplo, observe que, além da seqüência de nucleotídeos, encontramos os **números 5’ e 3’** (chamamos então de **cinco primos** e **três primos**) no início e no final de cada um dos filamentos de DNA. Observe também que a **segunda seqüência mostrada é a fita reversa** (“reverse strand”) que é **complementar** à fita **5’—> 3’** (“forward strand”).

A fita de DNA em sentido direto (“**forward strand**”) é escrita na orientação 5’-3’, e a fita em sentido reverso (“**reverse strand**”), ou fita complementar, é escrita na orientação 3'-5'. Isso representa a maneira como essas duas moléculas seriam vistas em um fragmento de DNA. Esses números não são arbitrários, eles representam os carbonos na estrutura química do anel de açúcar dos nucleotídeos que estão envolvidos nas reações químicas que formam o esqueleto do DNA fosfato-açúcar. *In vivo*, a síntese de DNA ocorre pela adição de nucleotídeos livres através de sua extremidade 5‘ ao extremo 3' de uma molécula de DNA nascente. Isso produz direcionalidade (enzimas só podem sintetizar DNA na direção 5‘ para 3’). Assim, os pesquisadores adotaram esta convenção para a escrita de seqüências de ácidos nucléicos. Isto significa que, salvo indicação em contrário, todas as sequências de ácidos nucléicos são escritas na direção de 5‘ para 3'.

Apesar de ser uma dupla hélice de sequências de DNA complementares, o DNA é quase sempre representado como uma única sequência.

5’-ATGCGATCGGACAGTCGAGTCCAGTAGACGATC-3’

A anotação 5 '-> 3' é uma convenção que todos os pesquisadores em genética seguem, portanto, não é mais preciso especificá-la. Assim, a sequencia de DNA acima pode ser simplesmente escrita como:

ATGCGATCGGACAGTCGAGTCCAGTAGACGATC

**……………………………………………………………………………………………………**

Vamos agora testar e fixar o que aprendemos!

**Questão 1:**

Representação da fita reversa do DNA (“reverse strand”).

Com base no que aprendemos até agora sobre convenções, qual é a maneira correta de representar a cadeia reversa da sequência: ATGCATGC ?

Lembre-se que as seqüências de nucleotídeos são sempre escritas na orientação 5 '-> 3’.

Para chegar à resposta, você pode seguir estas etapas:

1. Anote a sequência em um pedaço de papel/ou na tela,

2. Escreva as bases complementares (um T para cada A, um G para cada C, e vice-versa) sob cada letra,

3. Pegue a segunda linha e inverta a ordem.

**Resposta:** A representação correta da cadeia acima é: 5’-GCATGCAT-3’

**Questão 2:**

Imagine que você tenha que escrever, em uma linha, uma seqüência de DNA. Qual das afirmações seguintes é a correta?

1. Eu escrevo a sequência de DNA de 5 '-> 3' se for a “forward strand”

2. Eu escrevo a sequência de DNA de 3 '-> 5' se for a “reverse strand”

3. Eu escrevo ambas as sequencias dos filamentos de DNA, a “forward strand” como 5 '-> 3' e a “reverse strand” diretamente abaixo como 3 '-> 5'

**Resposta:** Você deve sempre escrever a sequência de DNA de 5 '-> 3' se for a “forward strand” ou a se for a “reverse strand” . Assim, a melhor resposta é a 1 acima.

**Questão 3:**

Qual é a seqüência reversa complementar dessas 4 letras da cadeia de DNA? **ACTG**

1. GTCA

2. TGAC

3. CAGT

**Resposta:** A sequencia reversa é: CAGT, a 3 acima.



**…………………………………………………………………………………………………………..**

**2. Do DNA a proteína: Como uma sequência de proteína é gerada a partir de uma sequência de DNA.**

Nos sistemas biológicos, o processo de **transcrição** “transcreve”, ou seja copia, o DNA em RNA. É essa molécula de RNA que servirá de modelo para a produção de proteínas. O processo de produzir proteínas a partir de um modelo de RNA é chamado de "**tradução**" e é realizado pelo ribossomo. A molécula de RNA que leva a mensagem do DNA para o ribossomo é chamada de RNA mensageiro ou **mRNA**.

Os blocos de construção de proteínas são aminoácidos. Semelhante ao DNA, existe uma convenção que determina como a cadeia de aminoácidos nas proteínas é representada. As sequências proteicas são representadas **a partir do seu terminal amino** ou **N** para o **terminal carboxi** ou **C**. Esta é a direção na qual eles são lidos a partir do mRNA e sintetizados pelo ribossomo. Assim, N-terminais de aminoácidos livres são quimicamente ligados ao C-terminal da proteína nascente (**N-terminal** **—> C-terminal**) . Isso corresponde à direção em que eles aparecem em seu projeto de DNA ( **5’—> 3’**).

Cada aminoácido é codificado por um grupo de três nucleotídeos no mRNA. Cada palavra de três letras é chamada de **códon** porque corresponde ao código (“encodes for”) de um aminoácido. Esse código, ou correspondência entre códons e aminoácidos, é conhecido como **código genético**.

Um aminoácido pode ser codificado por mais de um códon e, por isto, **o código genético é degenerado**. **Tabelas de códons** podem ser usadas para decifrar o código; e, essas tabelas podem representar os códons de DNA ou RNA, com a única diferença sendo que na tabela de códon de RNA “T” é substituído por “U”.

Espécies diferentes podem ter códigos genéticos diferentes, mas todas seguem a mesma regra: cada códon corresponde sempre ao mesmo aminoácido.

Embora essa tabela descreva os códons de DNA, lembre-se de que **o DNA é transcrito em mRNA**, e que este **é traduzido em aminoácidos** que formam proteínas. **A previsão de uma sequência de aminoácidos** baseada na sua sequência nucleotídica é conhecida como **tradução conceitual**. Uma tradução conceitual é uma previsão da sequência de aminoácidos baseada na sequência de nucleotídeos e no código genético conhecido.

Para exemplificar, abaixo está representada uma pequena sequência de DNA e a sequência de aminoácidos codificada (os números dos códons são apenas uma referência relativa):

Codon number 1 2 3 4 5 6 7 8 9 10 11

Nucleotide sequence **ATG** CGA TCG GAC AGT CGA GTC CAG TAG ACG ATC

Amino-acid sequence **M** R S D S R V Q - T I

Observe que: o 1o códon (**ATG**) codifica **metionina** (**M**); o 9º códon (**TAG**) codifica um **sinal STOP;** e queo **3o** e o **5o** códons são diferentes, mas ambos codificam para **serina** (**S**).

**Quadros de Leitura:**

Como mencionado anteriormente, o **código genético** é lido em **códons** de palavras de **três letras**. Portanto, significa que para uma sequência de DNA de orientação conhecida, **existem três possíveis traduções conceituais**: a **primeira** começando na primeira base, a **segunda** começando na segunda base e finalmente a **terceira** começando na terceira base. Estes são referidos como **três “quadros de leitura”:**

ATGCGATCGGACAGTCGAGTCCAGTAGACGATC nucleotide sequence

M R S D S R V Q - T I 1st reading frame

C D R T V E S S R R 2nd reading frame

A I G Q S S P V D D 3rd reading frame

No exemplo, o quadro de primeira leitura começa com uma **metionina** (**M**) codificada pelo códon **ATG**, mas se considerarmos o segundo quadro de leitura e, portanto, começarmos a “ler” o código a partir da segunda base da seqüência de nucleotídeos, o primeiro aminoácido a ser lido seria (**C**) codificado pelo códon **TGC**.

Além disso, se não soubéssemos a orientação da sequência nucleotídica, a tradução conceitual poderia ser lida tanto na fita de DNA 5 '-> 3' quanto na reversa (3 '-> 5'), **dando um adicional de três possibilidades para se ler o código**.

Uma ferramenta útil para se fazer a **previsão da tradução conceitual** de uma seqüência de nucleotídeos é a “**ferramenta de tradução do ExPASy**” ([“ExPASy translate tool”](https://web.expasy.org/translate/) ). Este servidor fornece uma maneira rápida e fácil de encontrar a seqüência de aminoácidos correspondente à uma seqüência de nucleotídeos em todos os seis possíveis quadros de leitura.

**…………………………………………………………………………………………………………..**

Vamos agora testar e fixar o que aprendemos!

****

**Questão 1:**

Por que não experimentar e verificar se as três seqüências de aminoácidos oferecidas como primeiro, segundo e terceiro quadros de leitura na figura acima estão corretas?

**Questão 2:**

Você consegue descobrir quais são as sequências de aminoácidos que podem ser codificadas a cadeia reversa?

**…………………………………………………………………………………………………………..**

**O Código Genético** (“coding sequence” – CDS )

Um **códon** é uma série de três nucleotídeos (um triplete) que codifica um resíduo de aminoácido específico em uma cadeia polipeptídica ou a terminação da tradução (códon de terminação).

**Você sabia que o código genético tem significados diferentes?**

Sim. A mensagem contida no DNA (códons) é traduzida para aminoácidos distintos em diferentes replicons. Vamos ver como é isto.

Inicialmente:

Os **códigos genéticos** são tradicionalmente representados em uma **Tabela de códons** **de DNA** que serão **transcritos** em **mRNA**, sempre seguindo o **sentido 5'-3'.**

No site do National Center for Biotechnological Information – U. S. National Library of Medicine ([**https://www.ncbi.nlm.nih.gov/**](https://www.ncbi.nlm.nih.gov/)**)** são apresentadas **várias tabelas de códigos genéticos** usados por diferentes células e organelas **([Genetic Codes – NCBI](file:///Users/bete/Documents/Bioinformatica /Genetic Codes – NCBI, https:/www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi ). )**[, https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi ).](file:///Users/bete/Documents/Bioinformatica /Genetic Codes – NCBI, https:/www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi ). )

**O código genético padrão é apresentado no Quadro abaixo:**

**TTT F Phe TCT S Ser TAT Y Tyr TGT C Cys**

**TTC F Phe TCC S Ser TAC Y Tyr TGC C Cys**

**TTA L Leu TCA S Ser TAA \* Ter TGA \* Ter**

**TTG L Leu i TCG S Ser TAG \* Ter TGG W Trp**

**CTT L Leu CCT P Pro CAT H His CGT R Arg**

**CTC L Leu CCC P Pro CAC H His CGC R Arg**

**CTA L Leu CCA P Pro CAA Q Gln CGA R Arg**

**CTG L Leu i CCG P Pro CAG Q Gln CGG R Arg**

**ATT I Ile ACT T Thr AAT N Asn AGT S Ser**

**ATC I Ile ACC T Thr AAC N Asn AGC S Ser**

**ATA I Ile ACA T Thr AAA K Lys AGA R Arg**

**ATG M Met i ACG T Thr AAG K Lys AGG R Arg**

**GTT V Val GCT A Ala GAT D Asp GGT G Gly**

**GTC V Val GCC A Ala GAC D Asp GGC G Gly**

**GTA V Val GCA A Ala GAA E Glu GGA G Gly**

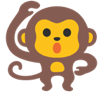
**GTG V Val GCG A Ala GAG E Glu GGG G Gly**

Códon de Iniciação tradicional: **AUG**

Códons de Iniciação Alternativos: O **código padrão** atualmente permite o início em **UUG** e **CUG**, além de **AUG** (veja **i** acima).

Khorana, Holley e Nirenberg receberam o prêmio Nobel 1968 pela elucidação do

Código genético (<http://nobelprize.org/nobel_prizes/medicine/laureates/1968/index.html> )



**IMPORTANTE**

Entre no site acima do NCBI, para:

- **Confirmar** se o gene objeto de seu estudo esta **Tabela de códigos genéticos padrão**;

- VEJA, também, as **demais Tabela de códigos genéticos** seguidas por outros ***replicons***, mesmo que o seu objeto de estudo seja a bactéria *E. coli* (para enriquecer o seu conhecimento).

**Os seguintes códigos genéticos são descritos nos links abaixo (NCBI-NLM)**

* [The Standard Code](https://www.bioinformatics.org/JaMBW/2/3/TranslationTables.html#SG1)
* [The Vertebrate Mitochondrial Code](https://www.bioinformatics.org/JaMBW/2/3/TranslationTables.html#SG2)
* [The Yeast Mitochondrial Code](https://www.bioinformatics.org/JaMBW/2/3/TranslationTables.html#SG3)
* [The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code](https://www.bioinformatics.org/JaMBW/2/3/TranslationTables.html#SG4)
* [The Invertebrate Mitochondrial Code](https://www.bioinformatics.org/JaMBW/2/3/TranslationTables.html#SG5)
* [The Ciliate, Dasycladacean and Hexamita Nuclear Code](https://www.bioinformatics.org/JaMBW/2/3/TranslationTables.html#SG6)
* [The Echinoderm Mitochondrial Code](https://www.bioinformatics.org/JaMBW/2/3/TranslationTables.html#SG9)
* [The Euplotid Nuclear Code](https://www.bioinformatics.org/JaMBW/2/3/TranslationTables.html#SG10)
* [The Bacterial "Code"](https://www.bioinformatics.org/JaMBW/2/3/TranslationTables.html#SG11)
* [The Alternative Yeast Nuclear Code](https://www.bioinformatics.org/JaMBW/2/3/TranslationTables.html#SG12)
* [The Ascidian Mitochondrial Code](https://www.bioinformatics.org/JaMBW/2/3/TranslationTables.html#SG13)
* [The Flatworm Mitochondrial Code](https://www.bioinformatics.org/JaMBW/2/3/TranslationTables.html#SG14)
* [Blepharisma Nuclear Code](https://www.bioinformatics.org/JaMBW/2/3/TranslationTables.html#SG15)

Para exemplificar, o código genético de bactérias é apresentado no Quadro abaixo:

**11. The Bacterial "Code" (transl\_table=11)**

**TTT F Phe TCT S Ser TAT Y Tyr TGT C Cys**

**TTC F Phe TCC S Ser TAC Y Tyr TGC C Cys**

**TTA L Leu TCA S Ser TAA \* Ter TGA \* Ter**

**TTG L Leu i TCG S Ser TAG \* Ter TGG W Trp**

**CTT L Leu CCT P Pro CAT H His CGT R Arg**

**CTC L Leu CCC P Pro CAC H His CGC R Arg**

**CTA L Leu CCA P Pro CAA Q Gln CGA R Arg**

**CTG L Leu i CCG P Pro CAG Q Gln CGG R Arg**

**ATT I Ile i ACT T Thr AAT N Asn AGT S Ser**

**ATC I Ile i ACC T Thr AAC N Asn AGC S Ser**

**ATA I Ile i ACA T Thr AAA K Lys AGA R Arg**

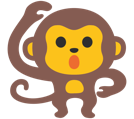
**ATG M Met i ACG T Thr AAG K Lys AGG R Arg**

**GTT V Val GCT A Ala GAT D Asp GGT G Gly**

**GTC V Val GCC A Ala GAC D Asp GGC G Gly**

**GTA V Val GCA A Ala GAA E Glu GGA G Gly**

**GTG V Val i GCG A Ala GAG E Glu GGG G Gly**

Quais são as principais diferenças entre o **código genético padrão** e o

**código genético utilizado por bactérias**? Qual é o significado disto?

**Códons** **preferenciais** (“Codon usage bias”)

Diferentes replicons têm diferenças na frequência de ocorrência de códons sinônimos na codificação do DNA.

There are 64 different codons (61 codons encoding for amino acids plus 3 stop codons) but only 20 different translated amino acids. The overabundance in the number of codons allows many amino acids to be encoded by more than one codon. Because of such redundancy it is said that the [genetic code](https://en.wikipedia.org/wiki/Genetic_code) is degenerate. The genetic codes of different organisms are often biased towards using one of the several codons that encode the same amino acid over the others—that is, a greater frequency of one will be found than expected by chance. How such biases arise is a much debated area of [molecular evolution](https://en.wikipedia.org/wiki/Molecular_evolution). Codon usage tables detailing genomic codon usage bias for most organisms in [GenBank](https://en.wikipedia.org/wiki/GenBank) and [RefSeq](https://en.wikipedia.org/wiki/RefSeq) can be found in the [HIVE-Codon Usage Table database](https://hive.biochemistry.gwu.edu/cuts/about).[[1]](https://en.wikipedia.org/wiki/Codon_usage_bias#cite_note-1)

Existem 64 códons diferentes (61 códons codificam aminoácidos e mais 3 stop códons) mas existem apenas 20 aminoácidos traduzidos diferentes. A superabundância no número de códons permite que muitos aminoácidos sejam codificados por mais de um códon. Por causa de tal **redundância**, diz-se que **o código genético é degenerado**.

Devido à **degeneração do código genético**, a maioria dos aminoácidos pode ser codificada por múltiplos códons sinônimos. Códons sinônimos ocorrem naturalmente com diferentes frequências em diferentes organismos. A escolha dos códons pode afetar a expressão, estrutura e função da proteína. As Tecnologias de DNAs recombinantes geralmente aproveitam a primeira vantagem deste conhecimento implementando uma técnica denominada **otimização de códons**, na qual os códons são substituídos por sinônimos, de modo a aumentar a expressão da proteína recombinante desejada. Esta técnica baseia-se no conhecimento preciso das frequências de uso de códons (Athey et al 2017 – citado abaixo).

Quantificar com precisão o viés de uso de códons nos diferentes organismos é útil não somente para otimização de códons, mas também para **estudos evolutivos** e de **tradução**: **relações filogenéticas de organismos** e **relações de coevolução** patógeno-hospedeiro podem ser exploradas empregando o estudo de semelhanças de uso de códons. Além disso, o uso de códons tem mostrado afetar a estrutura e a função das proteínas através da interferência na cinética de secreção e do dobramento das proteínas excretadas pela célula (“co-translational proteins, translocation of proteins”, [Nyathi et al 2013](https://linkinghub.elsevier.com/retrieve/pii/S0167488913000797) ).

Assim, os **códigos genéticos de diferentes organismos são frequentemente inclinados a usar um dos vários códons** que codificam o **mesmo aminoácido** sobre os outros - isto é, uma maior frequência de um será encontrada do que o esperado por acaso. Como tais vieses surgem é uma área muito debatida da evolução molecular. Tabelas de uso de códons que detalham o viés de uso de códons genômicos para a maioria dos organismos no

[GenBank](https://en.wikipedia.org/wiki/GenBank) e  [RefSeq](https://en.wikipedia.org/wiki/RefSeq)  podem ser encontradas no banco de dados da Tabela [HIVE-Codon Usage Table database](https://hive.biochemistry.gwu.edu/cuts/about) (Athey et al 2017 – veja link abaixo).

Athey, John; Alexaki, Aikaterini; Osipova, Ekaterina; Rostovtsev, Alexandre; Santana-Quintero, Luis V.; Katneni, Upendra; Simonyan, Vahan; Kimchi-Sarfaty, Chava (2017-09-02). ["A new and updated resource for codon usage tables"](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5581930). *BMC Bioinformatics*. **18** (391): 391. [doi](https://en.wikipedia.org/wiki/Digital_object_identifier):[10.1186/s12859-017-1793-7](https://doi.org/10.1186%2Fs12859-017-1793-7). [PMC](https://en.wikipedia.org/wiki/PubMed_Central) [5581930](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5581930). [PMID](https://en.wikipedia.org/wiki/PubMed_Identifier) [28865429](https://www.ncbi.nlm.nih.gov/pubmed/28865429).

**3. Lendo e escrevendo sequencias de DNA e de proteínas**

**Arquivos FASTA -** arquivos em formatos mais básicos para estocar sequências de DNA e de proteína.

Os dados de sequência de DNA são normalmente armazenados em arquivos de texto, às vezes também chamados “**flat files”**. Estes são arquivos podem ser abertos em quase qualquer editor de texto. O tipo mais comum de arquivo é chamado de **arquivo FASTA**, no qual as sequências são armazenadas no **formato** **FASTA**.

O nome **FASTA** deriva de um pacote de software escrito em meados dos anos 80 que pesquisava rapidamente grandes coleções de dados de sequência - o software é chamado de FASTA, mas era também chamado de FAST-N (nucleotídeo) e FAST-P (proteína).

O formato FASTA deve, no mínimo, ter um cabeçalho (sempre precedido por um “>”) na primeira linha do arquivo, e a sequência começando na segunda linha. O cabeçalho inclui algumas informações mínimas sobre a sequência. Por exemplo, o gene ***hpcC*** de ***Escherichia coli****,* com o número de acesso **X81322.1**, pode ser representado da seguinte forma:

>X81322.1 E.coli hpcC gene

GAAGTAGAAGGCGTGGGCCGCCTGGTGAACCGAATTGTTGAGTGAGGAAACAGCGAAATG

AAAAAAGTAAATCATTGGATCAACGGCAAAAATGTTGCAGGTAACGACTACTTCCTGACC

ACCAATCCGGCAACGGGTGAAGTGCTGGCGGATGTGGCCTCTGGCGGTGAAGCGGAGATC

AATCAGGCGGTAGCGACAGCGAAAGAGGCGTTCCCGAAATGGGCCAATCTGCCGATGAAA

GAGCGTGCGCGCCTGATGCGCCGTCTGGGCGATCTGATCGACCAGAACGTGCCAGAGATC

GCCGCGATGGAAACCGCGGACACGGGCCTGCCGATCCATCAGACCAAAAATGTGTTGATC

CCACGCGCTTCTCACAACTTTGAATTTTTCGCGGAAGTCTGCCAGCAGATGAACGGCAAG

ACTTATCCGGTCGACGACAAGATGCTCAACTACACGCTGGTGCAGCCGGTAGGCGTTTGT

GCACTGGTGTCACCGTGGAACGTGCCGTTTATGACCGCCACCTGGAAGGTCGCGCCGTGT

CTGGCGCTGGGCATTACCGCGGTGCTGAAGATGTCCGAACTCTCCCCGCTGACCGCTGAC

CGCCTGGGTGAGCTGGCGCTGGAAGCCGGTATTCCGGCGGGCGTTCTGAACGTGGTACAG

GGCTACGGCGCAACCGCAGGCGATGCGCTGGTCCGTCATCATGACGTGCGTGCCGTGTCG

TTCACCGGCGGTACGGCGACCGGGCGCAATATCATGAAAAACGCCGGGCTGAAAAAATAC

TCCATGGAACTGGGCGGTAAATCGCCGGTGCTGATTTTTGAAGATGCCGATATTGAGCGC

GCGCTGGACGCCGCCCTGTTCACCATCTTCTCGATCAACGGCGAGCGCTGCACCGCCGGT

TCGCGCATCTTTATTCAACAAAGCATCTACCCGGAATTCGTGAAATTTGCCGAACGCGCC

AACCGTGTGCGCGTGGGCGATCCGACCGATCCGAATACCCAGGTTGGGGCGCTTATCAGC

CAGCAACACTGGGAAAAAGTCTCCGGCTATATCCGTCTGGGCATTGAAGAAGGCGCCACC

CTGCTGGCGGGCGGCCCGGATAAACCGTCTGACCTGCCTGCACACCTGAAAGGCGGCAAC

TTCCTGCGCCCAACGGTGCTGGCGGACGTAGATAACCGTATGCGCGTTGCCCAGGAAGAG

ATTTTCGGGCCGGTCGCCTGCCTGCTGCCGTTTAAAGACGAAGCCGAAGCGTTACGCCTG

GCAAACGACGTGGAGTATGGCCTCGCGTCGTACATCTGGACACAGGATGTCAGCAAAGTG

CTGCGTCTGGCGCGCGGCATTGAAGCAGGCATGGTGTTCGTCAACACCCAGTTCGTGCGT

GACCTGCGCCACGCATTTGGCGGCGTAAAACCTCGCACCGGGCGTGAAGGCGGTGGATAC

AGTTCGAAGTGTTCGCGGAAATGAAGAAGAACGTCTGCATTCCATGGCGGACCATCCCA

Você pode acessar a entrada na [base de dados do NCBI usando este link](https://www.ncbi.nlm.nih.gov/nuccore/587109?report=fasta) .

Arquivos contendo sequências **FASTA** são comumente denominados com a extensão **“.fa”** ou **“.fasta”.** Por exemplo, se eu fosse salvar a sequência acima em um arquivo, eu poderia chamá-lo de "**E.coli\_hpcC.fasta"**. Não é obrigatório chamar um arquivo de sequência de DNA **xxx.fasta** ou **xxx.fa** (onde **xxx** representa qualquer combinação de letras ou números usados para nomear um arquivo), e eu poderia chamá-lo **“E.coli\_hpcC.mickeymouse”** se eu quisesse, mas **significaria absolutamente** **nada** para outras pessoas.

O Mundo da bioinformática é cheio de convenções que são regras realmente não escritas. Escolhemos segui-los para facilitar a comunicação e compartilhar dados com outros cientistas.

Ao determinar o tamanho de uma sequência de DNA, falamos em termos de “bases” ou “pares de bases”; a diferença entre eles implica que o último contém ambas as fitas da molécula de DNA. Mas esta nomenclatura não tem consequências em termos de tamanho: uma molécula de 100 bases tem o mesmo comprimento que uma molécula de 100 pares de bases.

É importante notar que as sequências **FASTA** não estão restritas a sequências de DNA, elas também podem ser usadas para representar sequências de proteínas, nas quais cada letra representa um único aminoácido. Abaixo está apresentado um exemplo de um arquivo FASTA de uma sequência de proteína.

No exemplo abaixo, esta apresentada a sequência FASTA de uma desidrogenase de *Escherichia coli*. Esta proteína tem o número de acesso “[CAA57102.1](https://www.google.com/search?client=safari&rls=en&q=CAA57102.1&ie=UTF-8&oe=UTF-8)”, conforme mostrado no cabeçalho da entrada do arquivo FASTA.

>CAA57102.1 5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase [Escherichia coli]

MKKVNHWINGKNVAGNDYFLTTNPATGEVLADVASGGEAEINQAVATAKEAFPKWANLPMKERARLMRRL

GDLIDQNVPEIAAMETADTGLPIHQTKNVLIPRASHNFEFFAEVCQQMNGKTYPVDDKMLNYTLVQPVGV

CALVSPWNVPFMTATWKVAPCLALGITAVLKMSELSPLTADRLGELALEAGIPAGVLNVVQGYGATAGDA

LVRHHDVRAVSFTGGTATGRNIMKNAGLKKYSMELGGKSPVLIFEDADIERALDAALFTIFSINGERCTA

GSRIFIQQSIYPEFVKFAERANRVRVGDPTDPNTQVGALISQQHWEKVSGYIRLGIEEGATLLAGGPDKP

SDLPAHLKGGNFLRPTVLADVDNRMRVAQEEIFGPVACLLPFKDEAEALRLANDVEYGLASYIWTQDVSK

VLRLARGIEAGMVFVNTQFVRDLRHAFGGVKPRTGREGGGYSSKCSRK

Abaixo, segue outro exemplo, copia do arquivo [FASTA de amilase de *Bacillus* sp](https://www.google.com/search?client=safari&rls=en&q=CAA57102.1&ie=UTF-8&oe=UTF-8) .

>AAX85453.1 amylase [Bacillus sp. WPD616]

MRKEAIHHRSTDNFAYAYDSETLHLRLQTKKNDVDHVELLFGDPYEWHDGAWQFQTMPMRKTGSDGLFDY

WLAEVKPPYRRLRYGFVLRAGGEKLVYTEKGFYHEAPSDDTAYYFCFPFLHRVDLFQAPDWVKDTVWYQI

FPERFANGNPAISPKGARPWGSEDPTPTSFFGGDLQGIIDHLDYLADLGITGIYLTPIFRAPSNHKYDTA

DYFEIDPHFGDKETLKTLVKRCHEKGIRVMLDAVFNHCGYEFGPFQDVLKNGAASRYKDWFHIREFPLQT

EPRPNYDTFAFVPQMPKLNTAHPEVKRYLLDVATYWIREFDIDGWRLDVANEIDHQFWREFRQAVKALKP

DVYILGEIWHDAMPWLRGDQFDAVMNYQLADAALRFFAKEDMSASEFADRLMHVLHSYPKQVNEAAFNLL

GSHDTPRLLTVCGGDVRKVKLLFLFQLTFTGSPCIYYGDEIGMTGGNDPECRKCMVWDPEKQNKELYEHV

KQLIALRKQYRALRRGDVAFLAADDEVNHLVYAKTDGNETVMIIINRSNEAAEIPMPIDARGKWLVNLLT

GERFAAEAETLCVSLPPYGFVLYAVESW