

Computação na genética

Do  ao nucleotídeo 

Luiz Pedro Petroski



Luiz Pedro Petroski

Formação

- Doutorando em Ciências da Saúde - PUC PR
- Engenharia de Computação - UEPG
- Mestrado em Computação Aplicada - UEPG

Profissional

- Desenvolvedor
- Analista de Sistemas
- Professor



Quem é você?

- **Curso?**
- **Ano/periodo?**
- **Qual/quais linguagens sabe?**



Image by upklyak on Freepik

Vamos abstrair

Níveis de abstração

- Do físico (BIT) ao conceitual
(Esquema)

Analogias

- Árvore
- Fila/Pilha
- Processo
- Memória
- Banco de dados
- Rede Neural

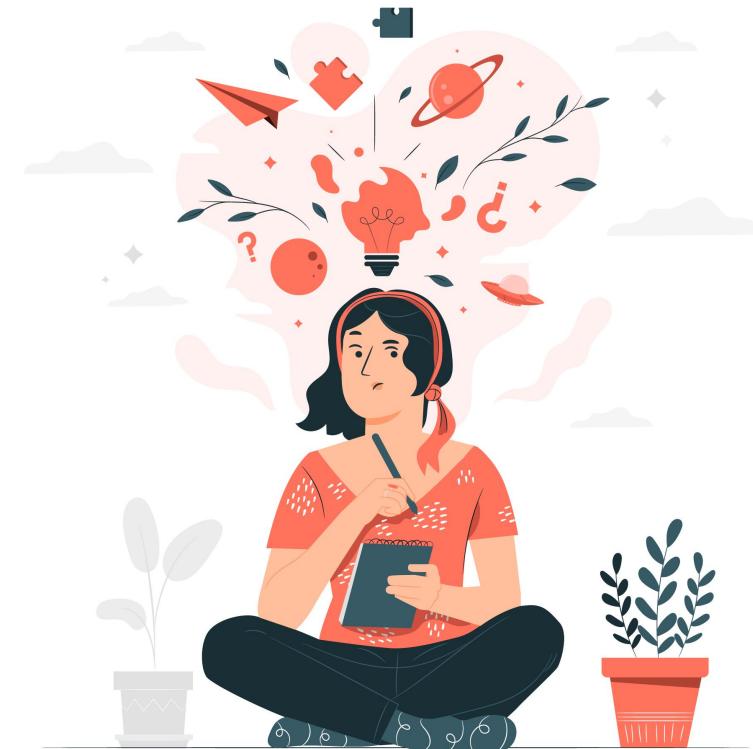


Image by storyset on Freepik

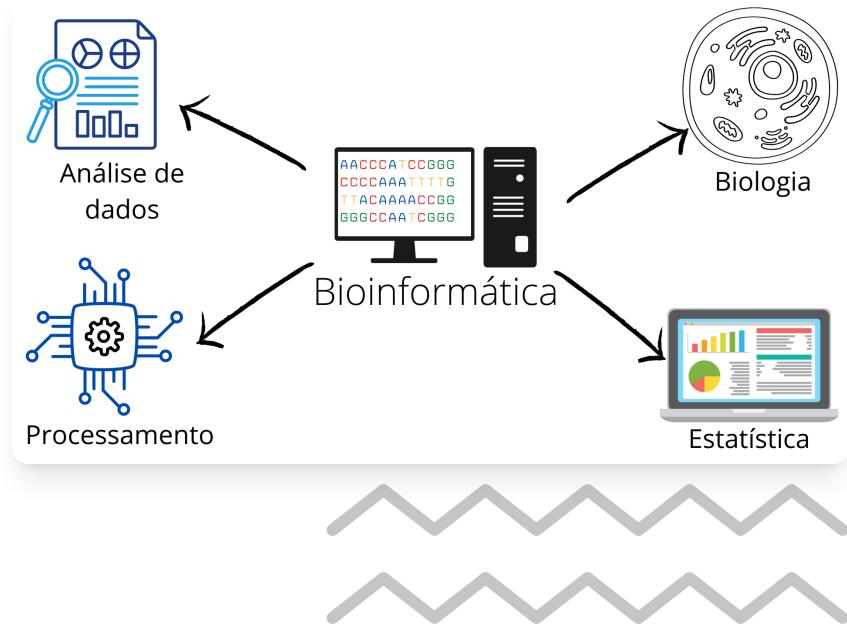
O que é a bio+informática

Aplicação de estatística e Ferramentas de computação em dados biológicos

Sistemas de informação

Dados das Multi-ômicas

- Genoma
- Transcriptoma
- Proteoma
- Metaboloma



O que é a informática?

- Qual é a menor unidade da computação
- bit, 0/1, ligado/desligado, sim/não, true/false
- Com 8 bits temos?
- Um Byte(B) $2^8=256$ combinações
- Instruções
- Opcode: add, sub, mul, inc, and, nor, ld1, st1
- Linguagens de alto nível
- C, Java, Python, PHP, TypeScript, R, C#
- Compiladores, transpiladores, interpretadores
- Modelos lógicos e conceituais
- Diagramas (Casos de uso (UML), histórias de usuário, diagramas de bloco)
- Programa completo, Sistema de Informação

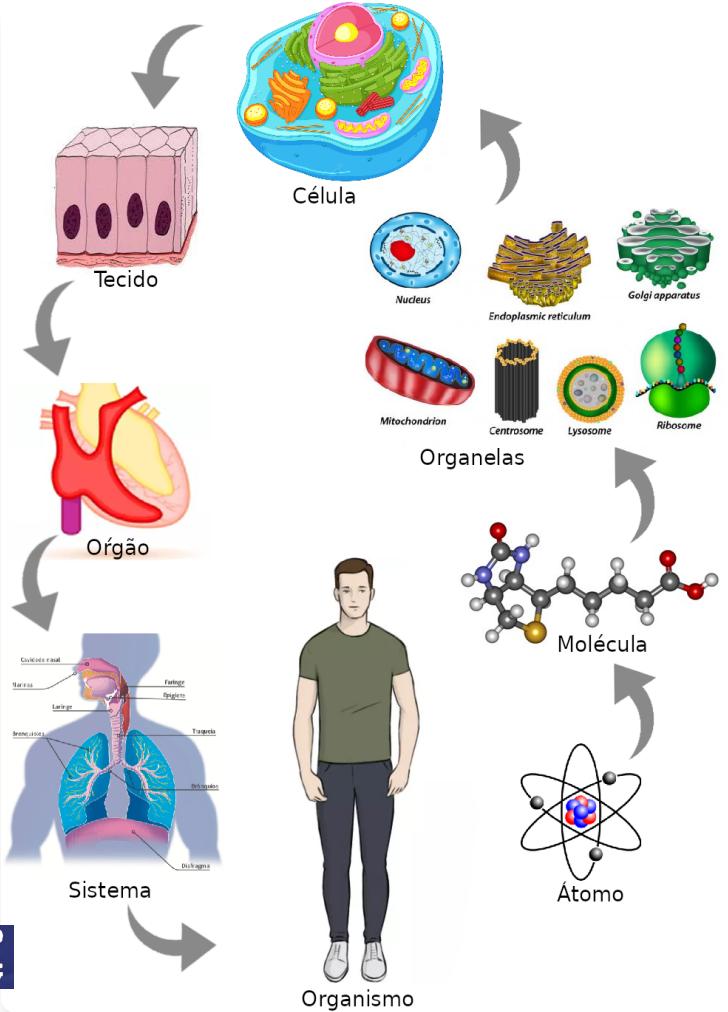


E a bio?

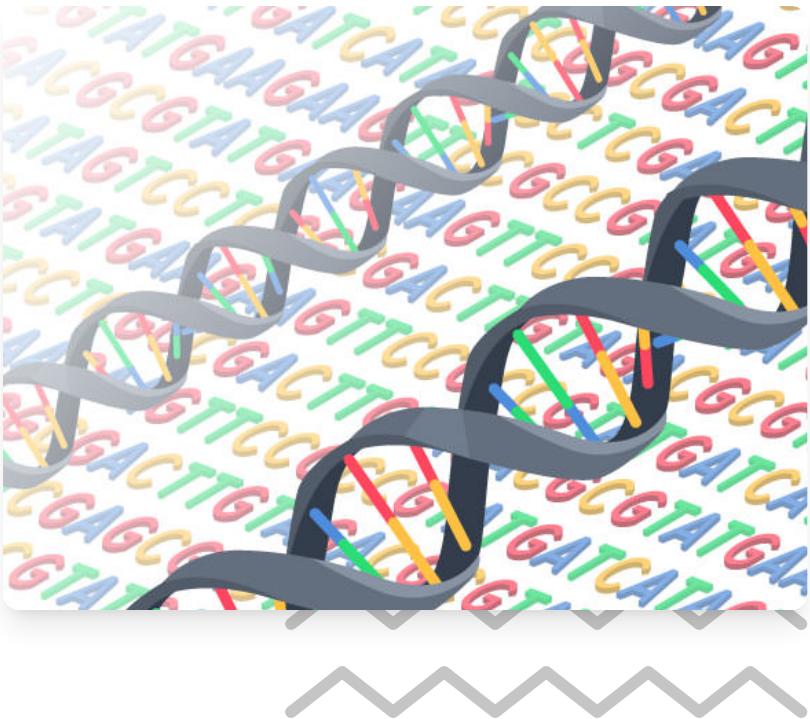
Biologia: Ciência que estuda a vida (organismos vivos)!

Exemplo: você!

Do que você é formado?



Conseguimos fazer alguma relação com a computação?

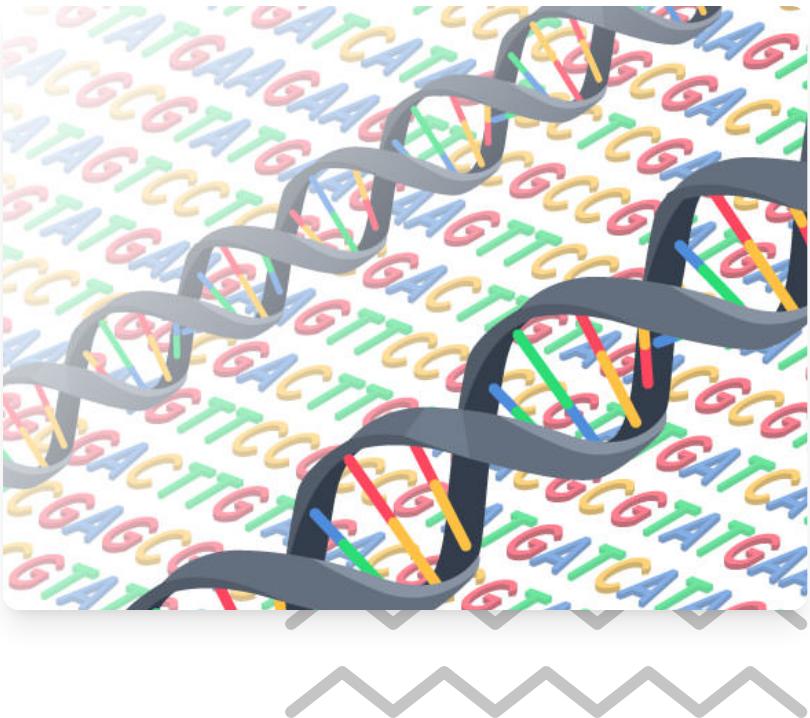


**Seria possível fazer
um DEBUG de um
ser vivo?**

Calma, calma!!

**Quais são os dados que
conseguimos coletar de um
ser vivo?**

- DNA
- RNA
- Proteína



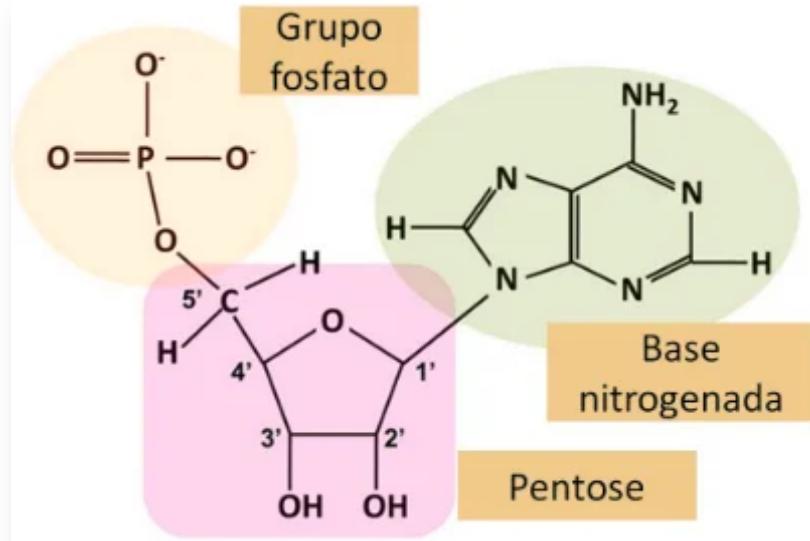
**Seria possível fazer
um DEBUG de um
ser vivo?**

Calma, calma!!

**Quais são os dados que
conseguimos coletar de um
ser vivo?**

- DNA
- RNA
- Proteína

Menor unidade de dados dos seres vivos



O nucleotídeo

Base nitrogenada: Purinas -> adenina (A) e guanina (G) e Pirimidinas -> citosina (C), uracila (U) e timina (T).

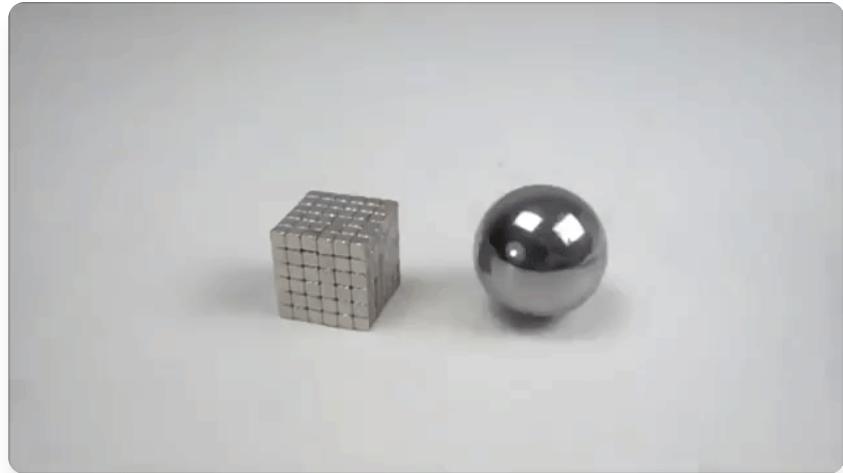
Grupo fosfato (HPO₄): Derivado do ácido fosfórico.

Pentose: Um açúcar de 5 carbonos. No DNA temos a

Estrutura

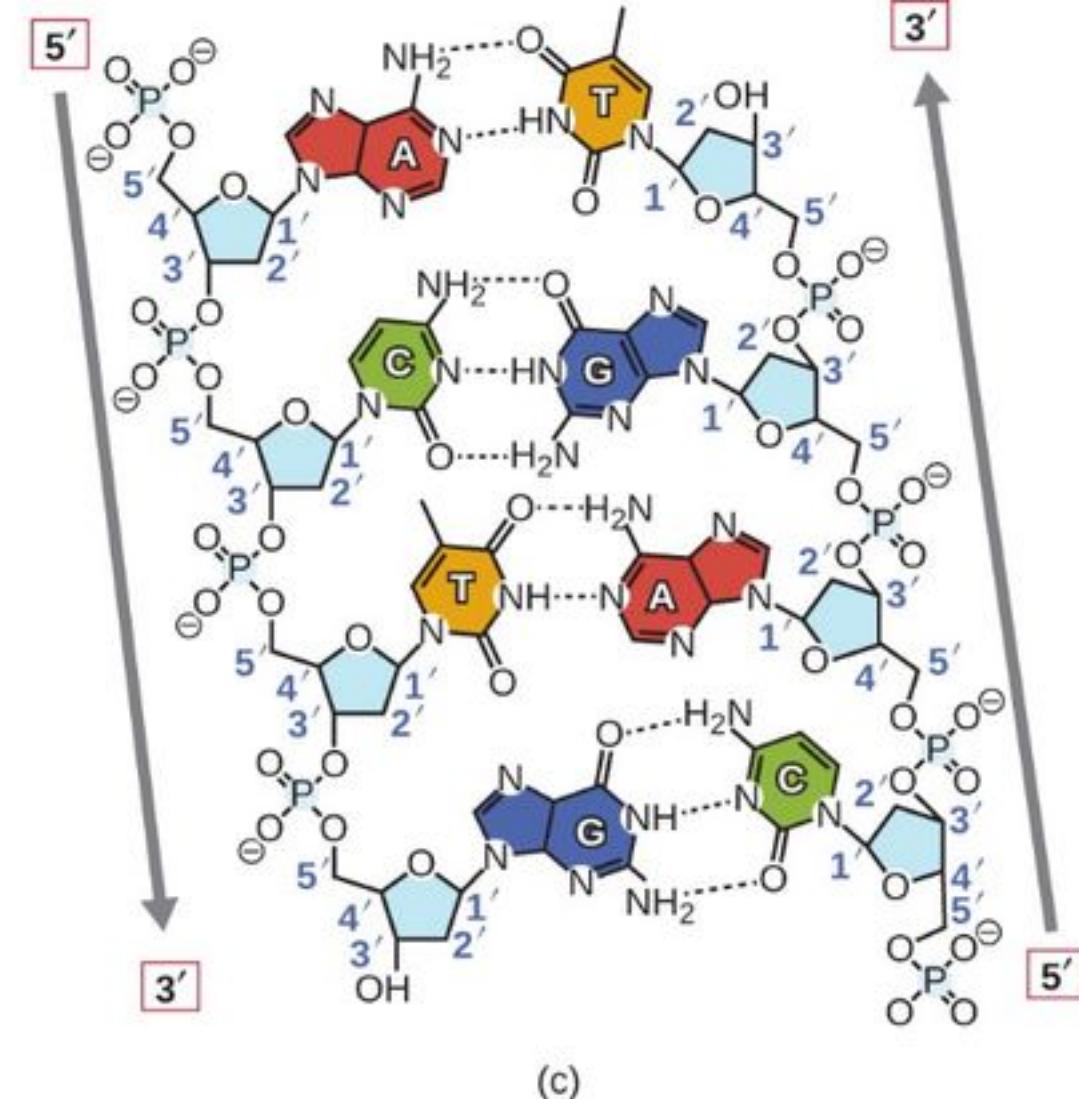
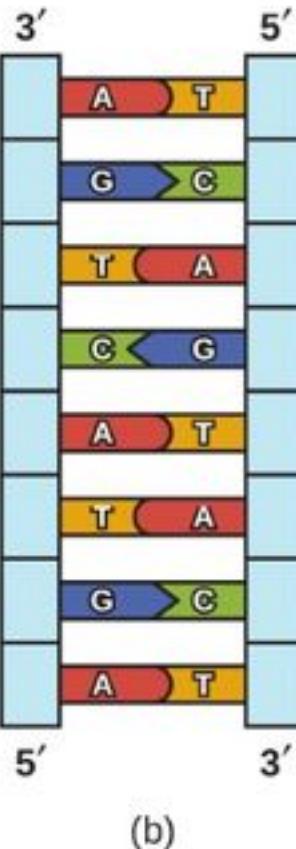
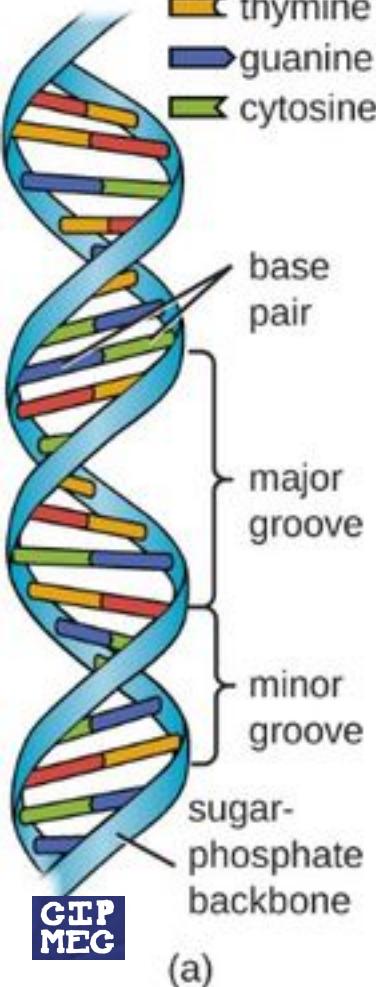
O DNA pode formar uma dupla hélice, constituída por duas fileiras de polinucleotídeos, unidas entre si por ligações de hidrogênio.

A ligação de hidrogênio promove a forma e a estabilidade do DNA para proteger o código genético, mas também prevê a fácil ruptura das ligações (“Unzip”) através da ação de enzimas para a replicação do DNA.



nitrogenous bases:

- adenine
- thymine
- guanine
- cytosine



Pareamento das bases

Cada tipo de base nitrogenada pode interagir com uma outra base complementar, formando ligações de hidrogênio.

- Guanina pareia com Citosina
- Adenina pareia com Timina

Processo de replicação semiconservativo

Os nucleotídeos são o alfabeto do código da vida

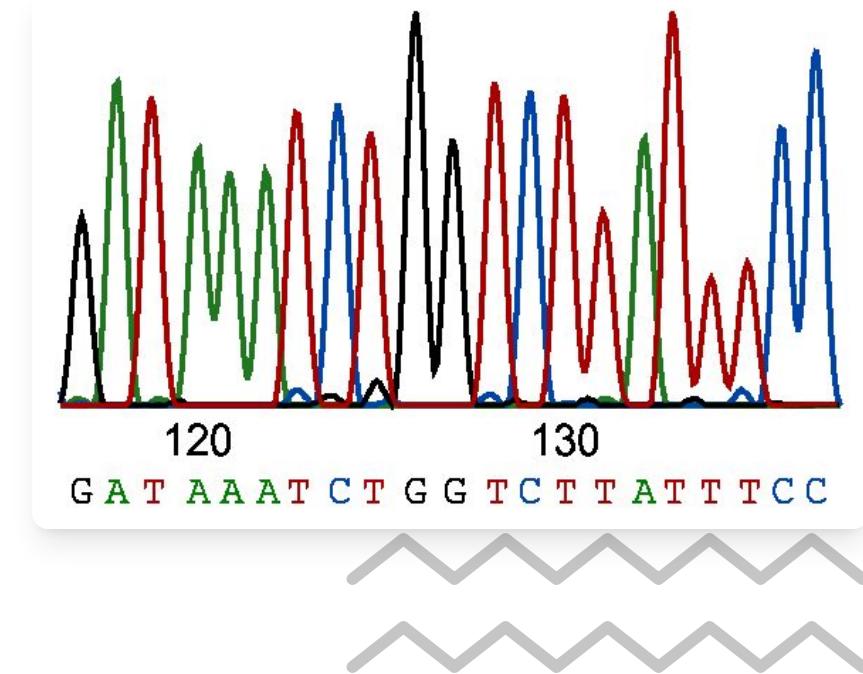
Como funciona o código no corpo humano?

Como o código é compilado?

Podemos copiar esse código para o computador?

Podemos editar esse código?

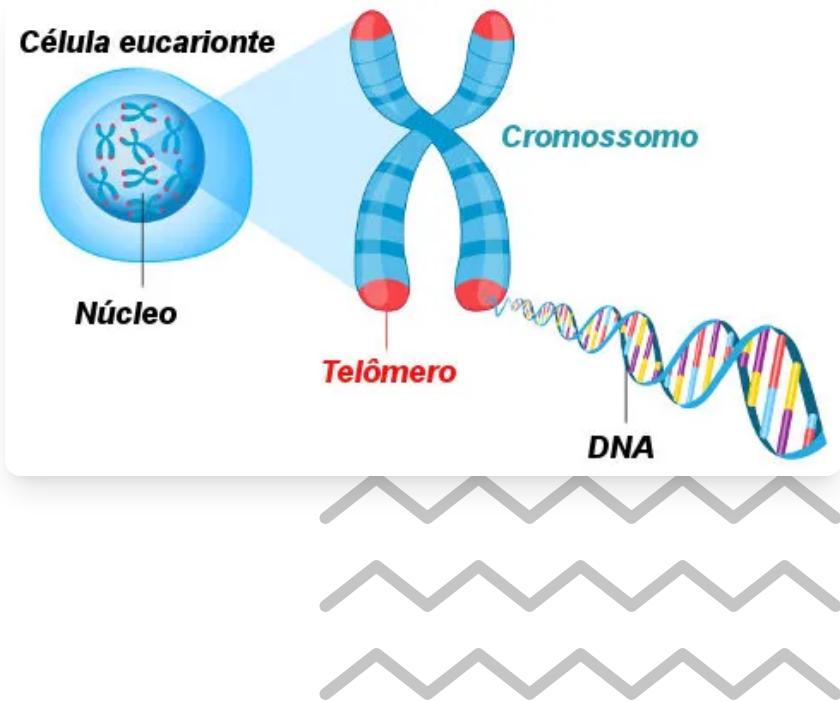
Podemos simular/emular um organismo?



Como funciona o código no corpo humano?

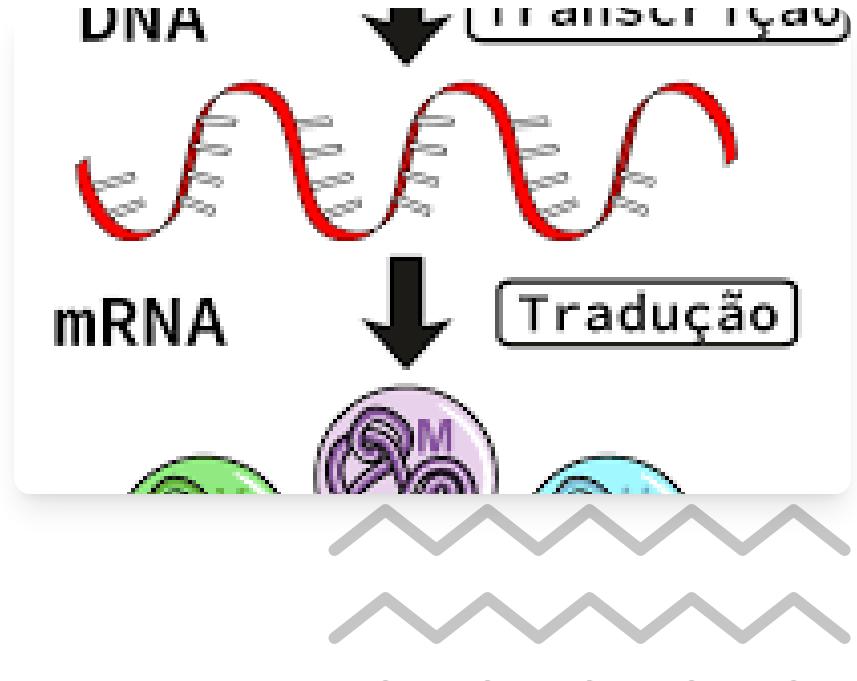
O DNA fica empacotado no núcleo de cada célula

- Segmento responsável pelas características herdadas: gene
- Forma os cromossomos (Cromatina)
- Cada célula tem uma cópia
- É como um repositório de código descentralizado



O DNA é compilado?

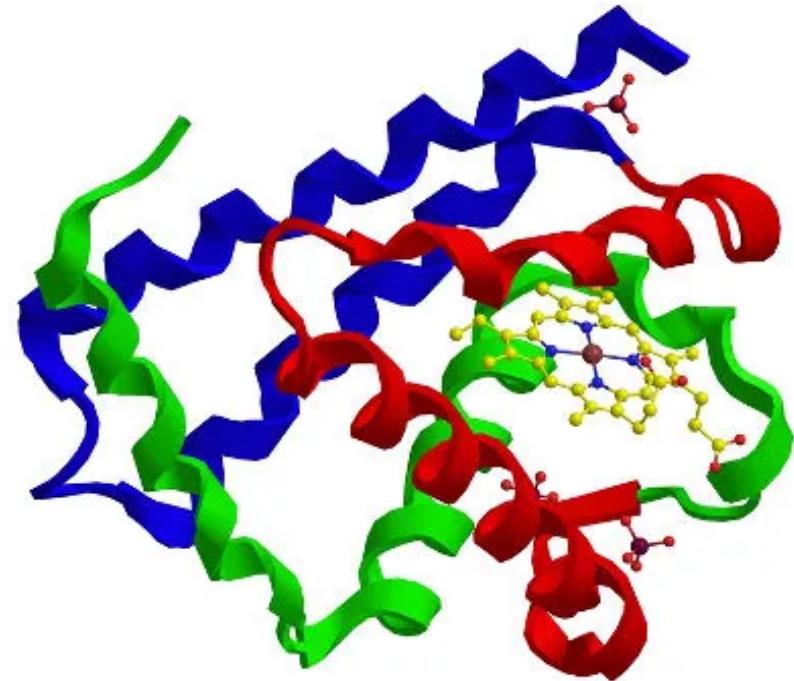
- Sim!! (analogia)
 - Esse processo é chamado de transcrição e tradução
 - O processo de transcrição pode ser entendido como uma espécie de compilação do código-fonte (DNA) em código objeto (RNA).
 - A tradução pode ser entendido como a transformação do código



A proteína é o executável de um sistema biológico

A maioria das funções biológicas é executada por proteínas

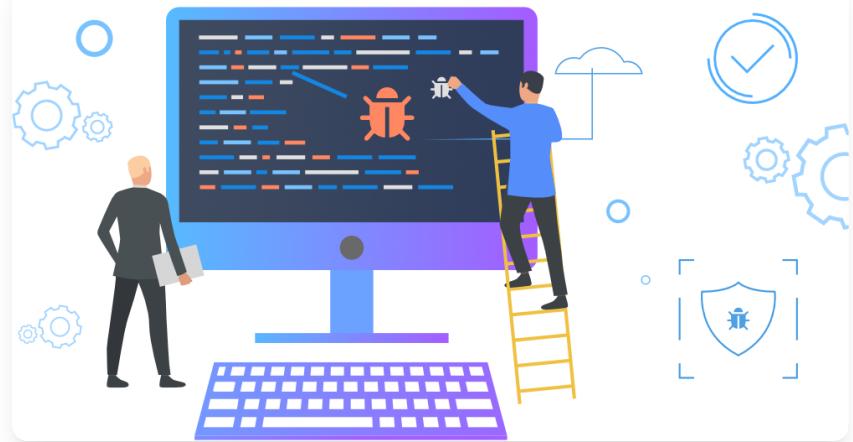
Caso ocorra algum problema com a proteína (ou com o processo de formação da proteína) pode causar problemas no organismo (doenças)



Conseguimos copiar esse código?

Assim poderíamos analisar e procurar onde estão ocorrendo os erros nos executáveis (proteínas)

- Mais uma vez sim!!
 - Podemos fazer o sequenciamento do DNA (genoma)
 - Também podemos fazer o sequenciamento do RNA (transcriptoma)



Qual a diferença entre DNA e RNA?

- O DNA é estático!!
 - O mesmo em todas as células
 - Não sofre alteração
 - Nem todo o código é executado em todos os lugares
- O RNA é dinâmico!!
 - É variável por cada tecido
 - Regulado por outros genes e fatores externos (ex medicamentos)
 - Mostra qual é o código está sendo executado naquele momento e lugar (monitor de processos)

É possível editar o DNA?

- Sim!!
 - CRISPR/Cas9
 - Promissor para diversas doenças como HIV e cancer
 - Mas é necessário um alvo bem definido e preciso

Então precisamos conhecer a função de cada gene. Qual a dificuldade?

- Alguma vez você mudou uma variável e nada mais funcionou?



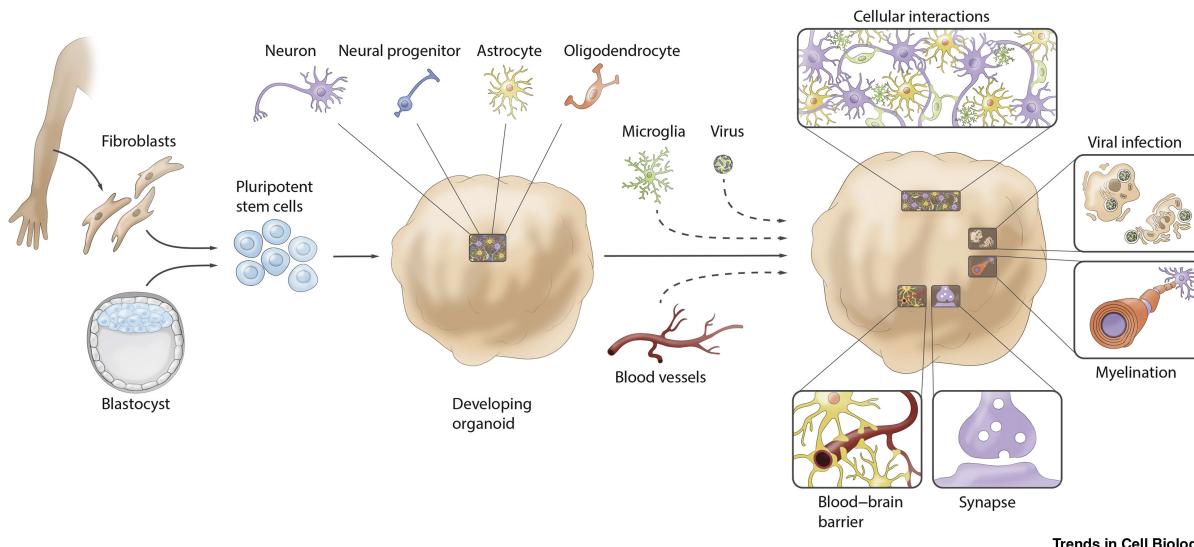
Qual o papel de cada gene?

- Temos 2 cópias da maioria dos genes (pai e mãe)
- Os genes são ativados ou não de acordo com a necessidade (célula especializada)
 - Isso pode ser visto pelo RNA diferente em cada tecido
 - É chamado de regulação da expressão gênica
- Por isso pode ocorrer de possuir um gene relacionado a alguma doença, mas nunca manifestar
- Ou por uma regulação inadequada (super expressão ou inibição) apresentar uma condição sindrômica

Tipos de amostras

- Retirada de tecido/biopsia
- Modelos animais
- Amostras post mortem
- Modelos in vitro

Modelo de organóide



Sequenciamento RNA-seq

- Coletar uma amostra
 - Ler as bases A, T, C, G
 - O primeiro método foi criado por Frederick Sanger (década de 70)
 - Hoje é utilizado as técnicas de NGS (Sequenciadores de Nova Geração)
- Fazer uma análise da qualidade das leituras
 - Arquivos chamados .fastq
 - O resultado do sequenciamento são de 20 a 50 milhões de "reads" com 50 a 150 bp cada
 - Cada bp tem um score de qualidade de leitura do equipamento

FASTA format

Identifier | @HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
Sequence | TTAATTGGTAAATAATCTCCTAATAGCTTAGATNTTACCTNNNNNNNNNTAGTTCTTGAGA
+ sign & identifier | +HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
Quality scores | efcfffffffceffffcfffffff`feed] `]_Ba_ ^ __[YBBBBBBBBBBRTT\\]] [] dddd`

Base T
phred Quality] = 29

Processamento RNA-seq

- Alinhamento
 - Alinhar os reads contra um genoma de referência
 - Há diversas bases de genomas de referência
 - Por exemplo o genoma humano mais utilizado é o GRC38 e o T2T
 - O resultado é o arquivo SAM (Sequence Alignment/Map)

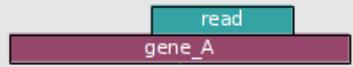
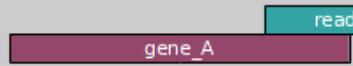
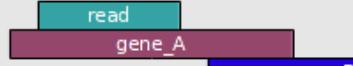
SAM format

QHD VN:1.5 SO:coordinate							Header section
@SQ SN:ref LN:45							Alignment section
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *							
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *							
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;							
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *							
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;							
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1							
							Optional fields in the format of TAG:TYPE:VALUE
							QUAL: read quality; * meaning such information is not available
							SEQ: read sequence
							TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.
							PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.
							RNAME: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.
							CIGAR: summary of alignment, e.g. insertion, deletion
							MAPQ: mapping quality
							POS: 1-based position
							RNAME: reference sequence name, e.g. chromosome/transcript id
							FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.
							QNAME: query template name, aka. read ID

Alinhamento das reads



Quantificação da expressão de cada gene

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A

Quantificação da expressão de cada gene

Quantificação de expressão gênica

The diagram illustrates the relationship between three tables: Genes, ID Amostras, and SRA Design. A bracket labeled 'Genes' groups the first two tables. Another bracket labeled 'ID Amostras' groups the middle and rightmost tables. An arrow labeled 'Condição' points from the ID Amostras table to the SRA Design table.

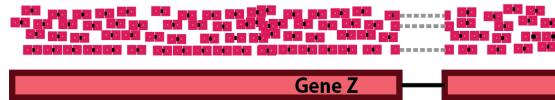
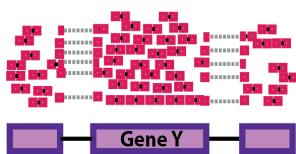
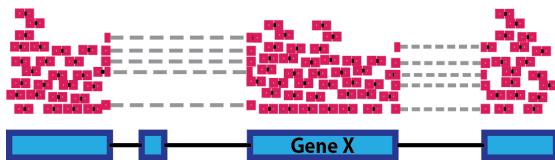
Genes		ID Amostras			Condição	
		ensgene	SRR11684499	SRR11684500		
		<chr>	<dbl>	<dbl>	<dbl>	SRA Design
	ENSG00000000003		1264	1435	1036	SRR11684510 NE_FXS
	ENSG00000000005		21	2	2	SRR11684509 NE_FXS
	ENSG00000000419		415	543	396	SRR11684508 NE_FXS
	ENSG00000000457		231	268	214	SRR11684507 NE_FXS
	ENSG00000000460		110	164	112	SRR11684506 NE_FXS
	ENSG00000000938		0	0	0	SRR11684505 NE_FXS
⋮	⋮	⋮	⋮	⋮	⋮	SRR11684504 NE_CTR
⋮	⋮	⋮	⋮	⋮	⋮	SRR11684503 NE_CTR
⋮	⋮	⋮	⋮	⋮	⋮	SRR11684502 NE_CTR

Problemas em comparar quantificação entre amostras

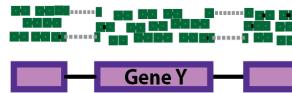
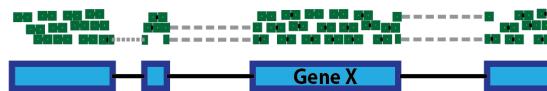
1. Diferença no tamanho da biblioteca de sequenciamento (número de reads)
2. Diferença na composição da biblioteca (comparação entre diferentes tecidos ou mesmo tecido com fator de transcrição com knock out)

Normalização

Sample A Reads



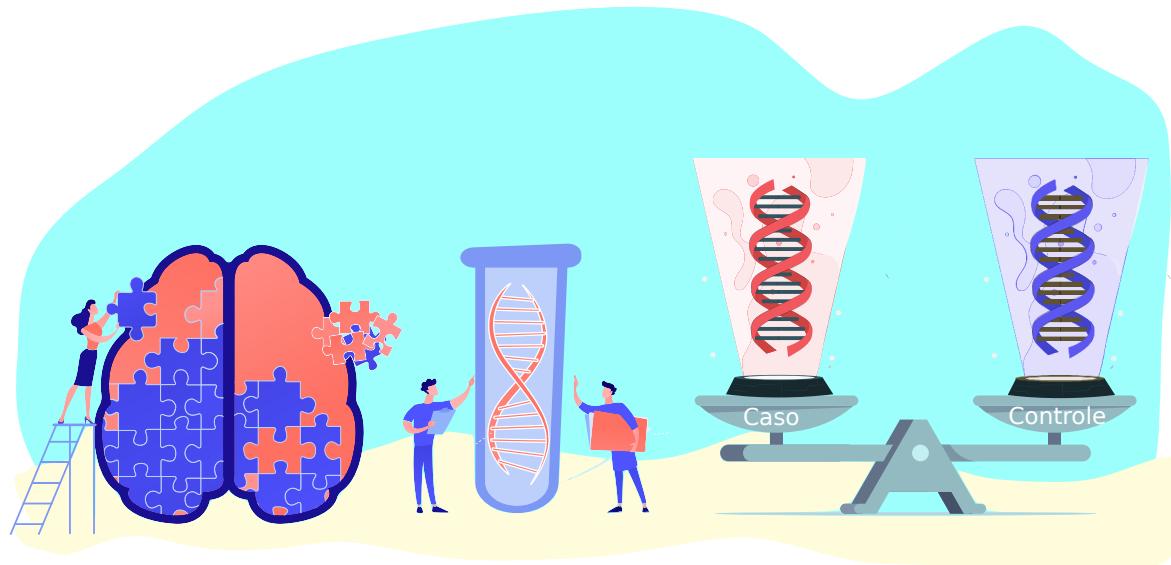
Sample B Reads



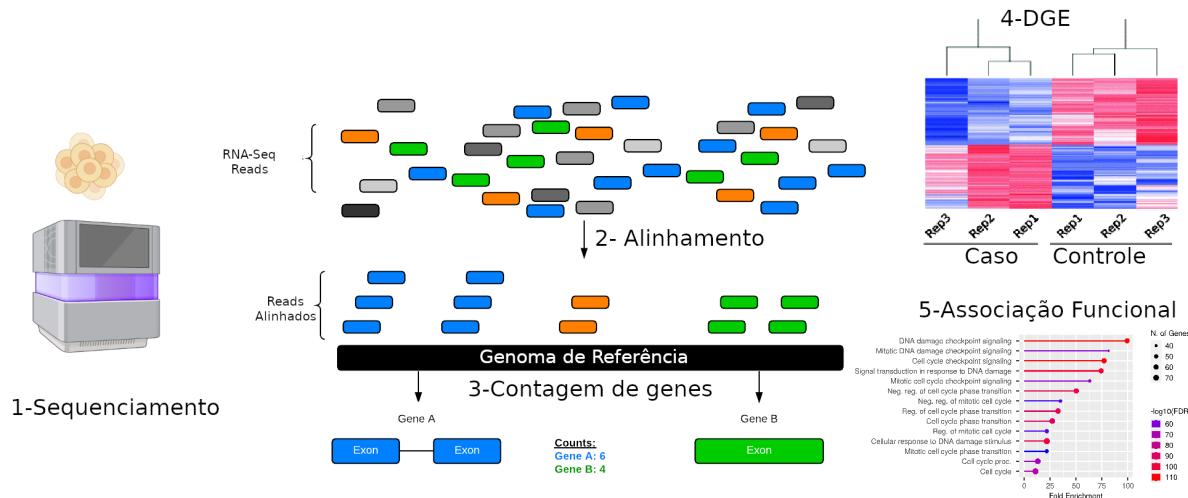
Normalização

- TPM (transcripts per kilobase million): Contagem pelo tamanho do transcrito (kb) por milhão de reads mapeados
- RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped): Similar ao TPM
- DESeq2's median of ratios: usa uma média ponderada das razões de expressão logarítmica entre as amostras
- Z-score: Normalização do valor de TPM

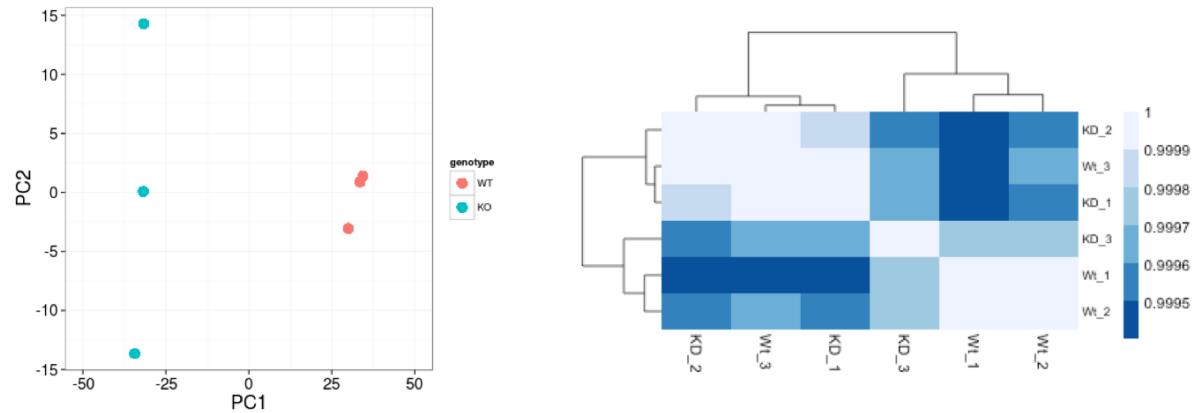
Estudos Caso/Controle



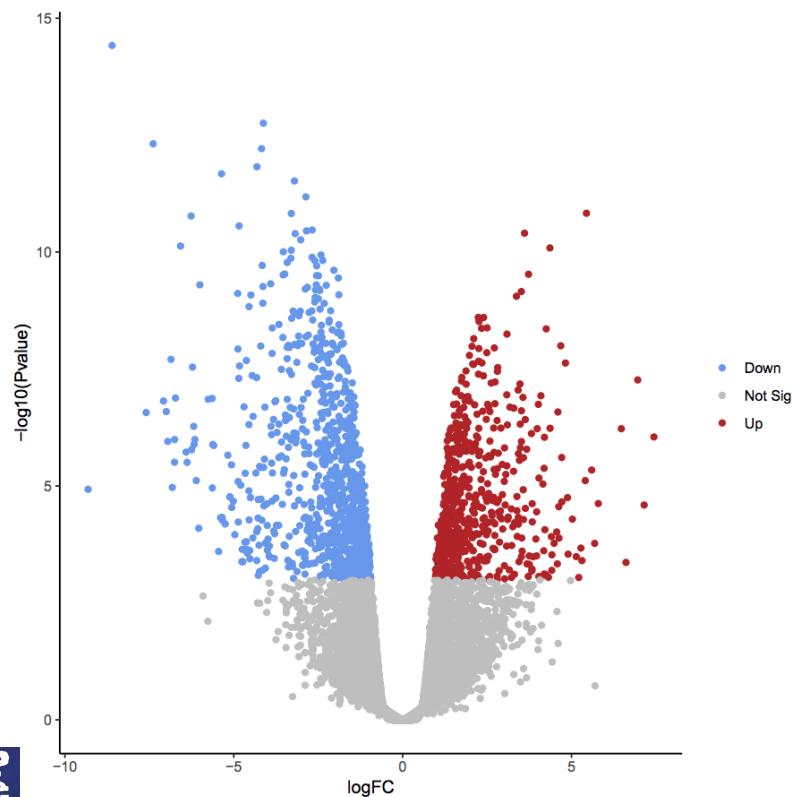
Estudos Caso/Controle



Análise de expressão diferencial de genes (DGE)



Análise de expressão diferencial de genes (DGE)





Uso de Machine Learning em Sequenciamento



Correção de erro de sequenciamento

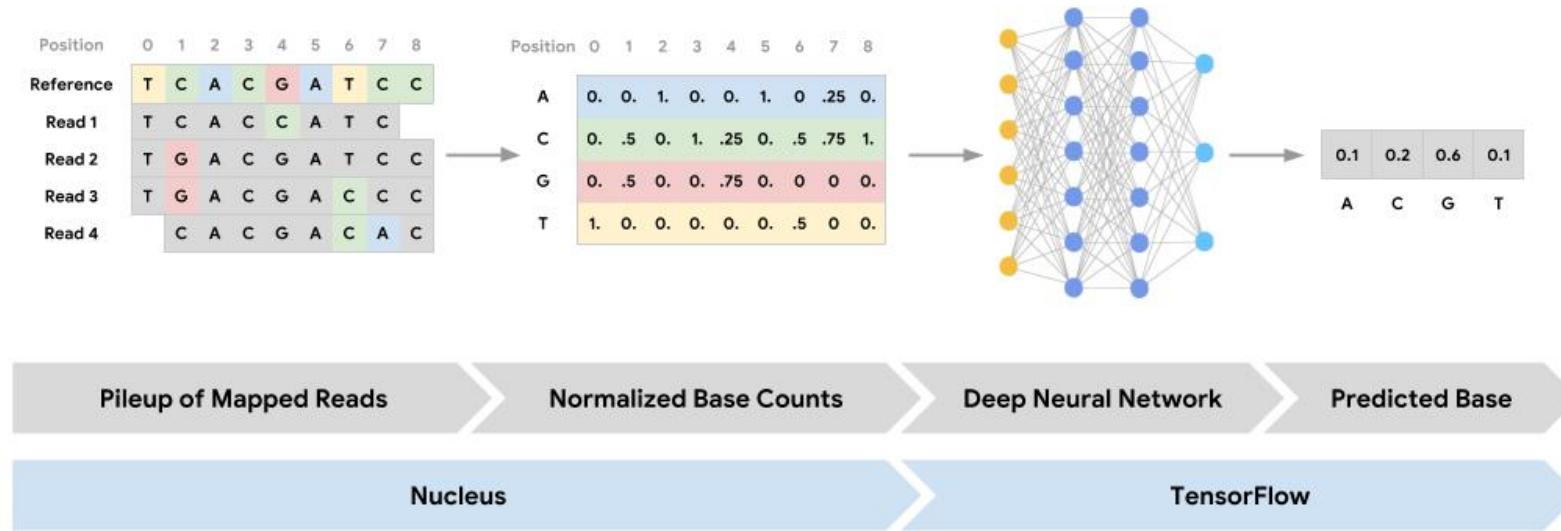
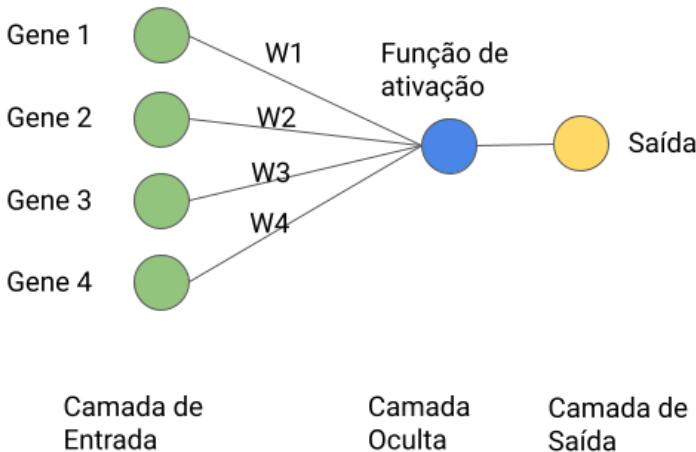


Figure 1: We formulate consensus-based DNA sequencing error correction as a multiclass classification problem. Using Nucleus, we construct a matrix of normalized base counts in a genomic range. TensorFlow allows us to train a neural network that can predict the correct base at the middle position of the window.

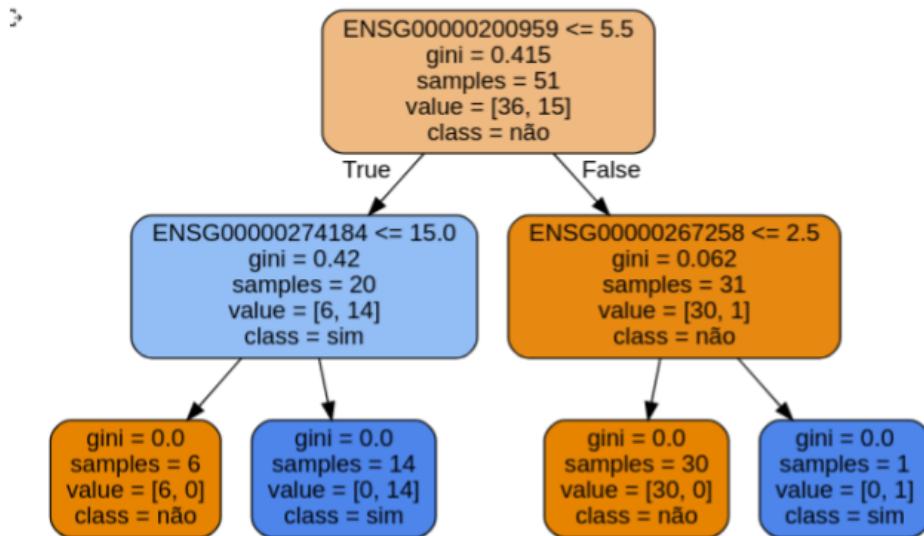
<https://google.github.io/deepvariant/posts/2019-01-31-using-nucleus-and-tensorflow-for-dna-sequencing-error-correction/>

Predição de condição por expressão gênica

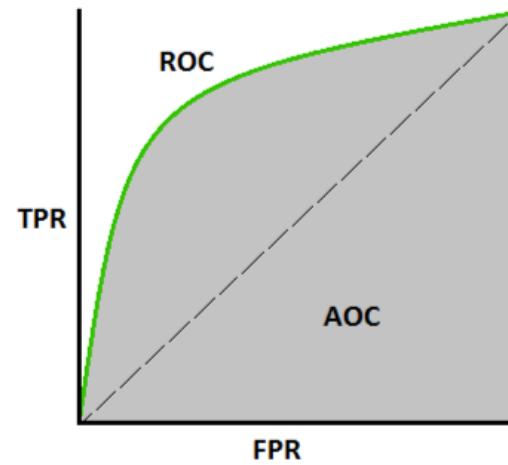
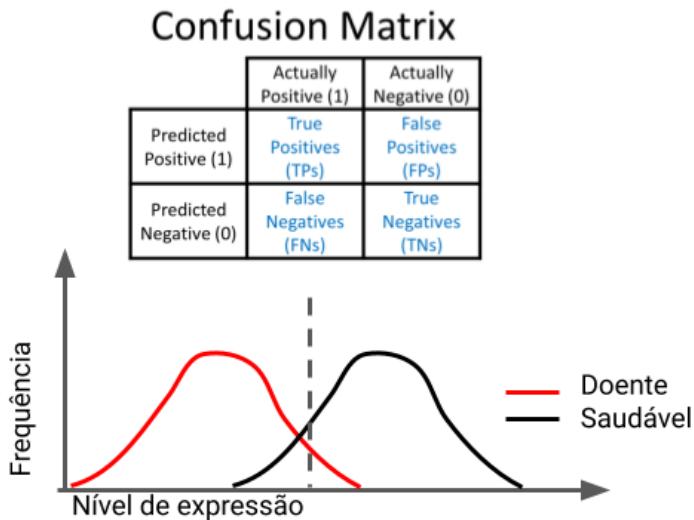
Gene 1	Gene 2	Gene 3	Gene 4	Condição
10	1	5	7	CTL
8	2	6	6	CTL
1	9	5	3	AF
0	10	4	2	AF
...



Predição de condição por expressão gênica



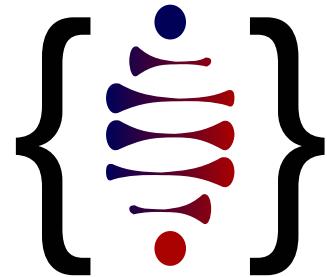
Performance de predição



Perspectivas

- Determinar relações complexas de regulação dos genes
- Determinar genes alvos para diagnóstico e tratamento
- Validar com experimentos PCR (RNA) e Western Blotting (Proteína)
- Experimentos com knockout de genes
- Modelos experimental com organóides

Perguntas?



petroskilp@gmail.com