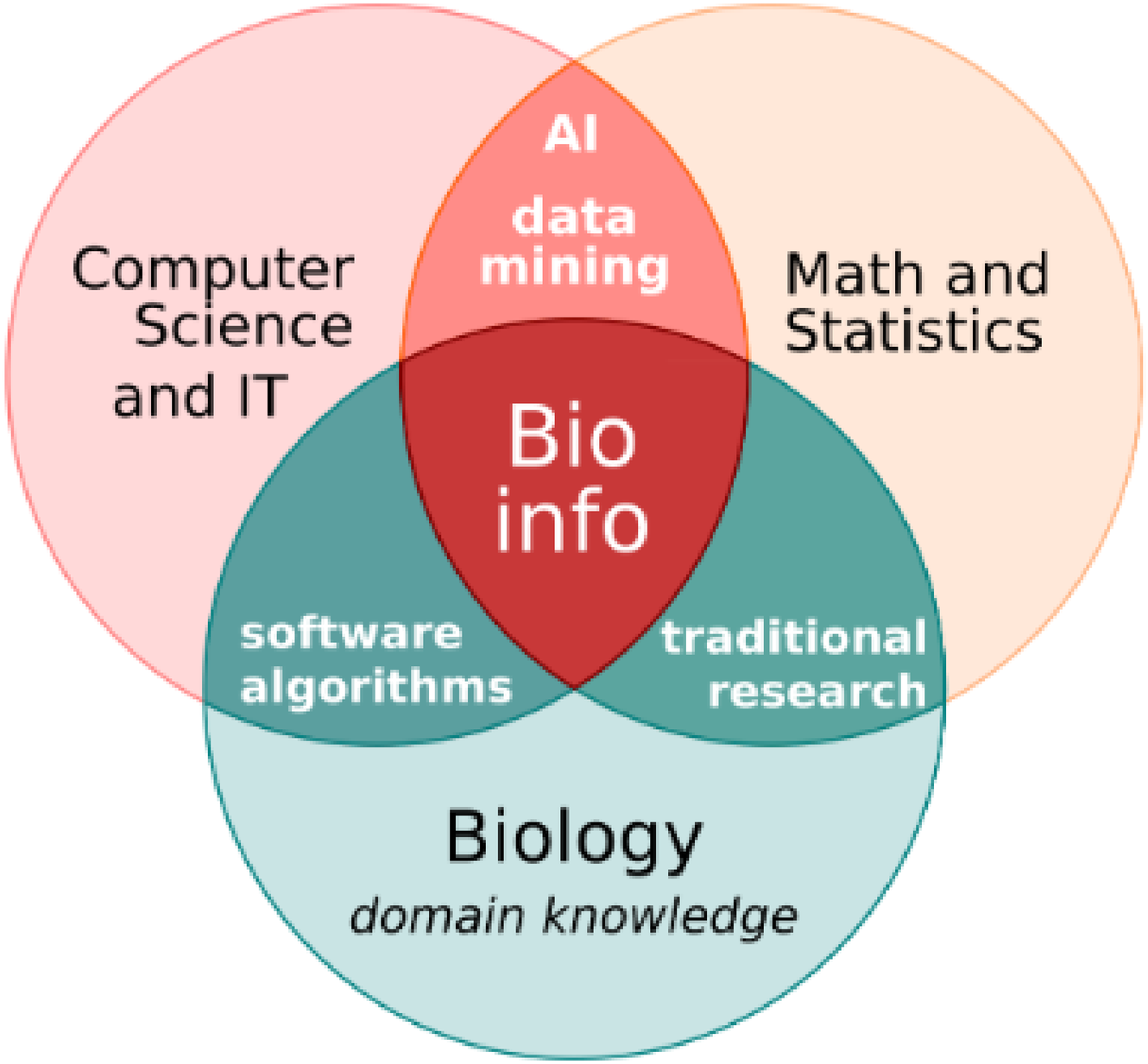
Abordagens didáticas no ensino de bioinformática

Alison Henrique Ferreira Julio

30.01.2025

edital 054 / vaga RP-001 - Biologia celular e do desenvolvimento: Bioinformática

O que é bioinformatica? Uma área multidisciplinar entre:

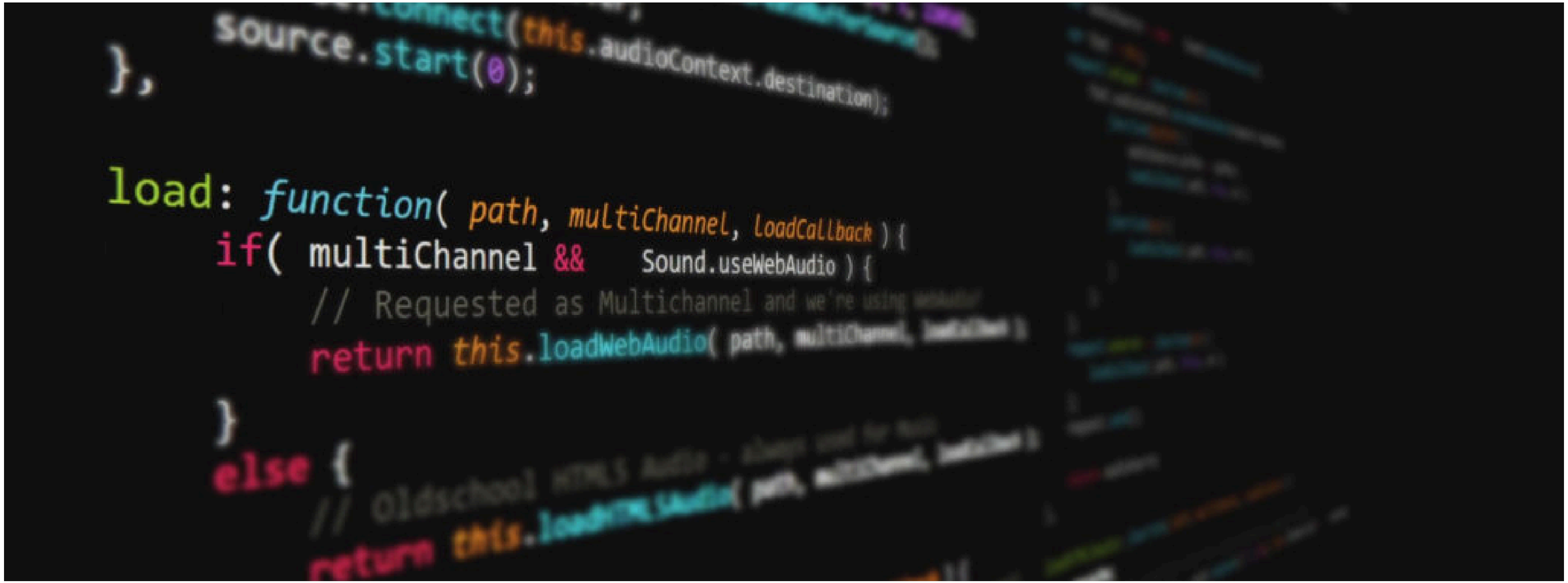


Com a matemática e a estatística no seu núcleo, a bioinformática aplica estes campos para tornar os diversos e complexos dados das ciências da vida mais compreensíveis e úteis, para descobrir novos conhecimentos biológicos e para fornecer novas perspectivas para discernir princípios unificadores.

UA LITTLE ROCK
Donaghey College of Science, Technology, Engineering, and Mathematics

O que é bioinformatica?

Ciência da computação e da informação – A bioinformática depende fortemente de estratégias para adquirir, armazenar, organizar, arquivar, analisar e visualizar dados.

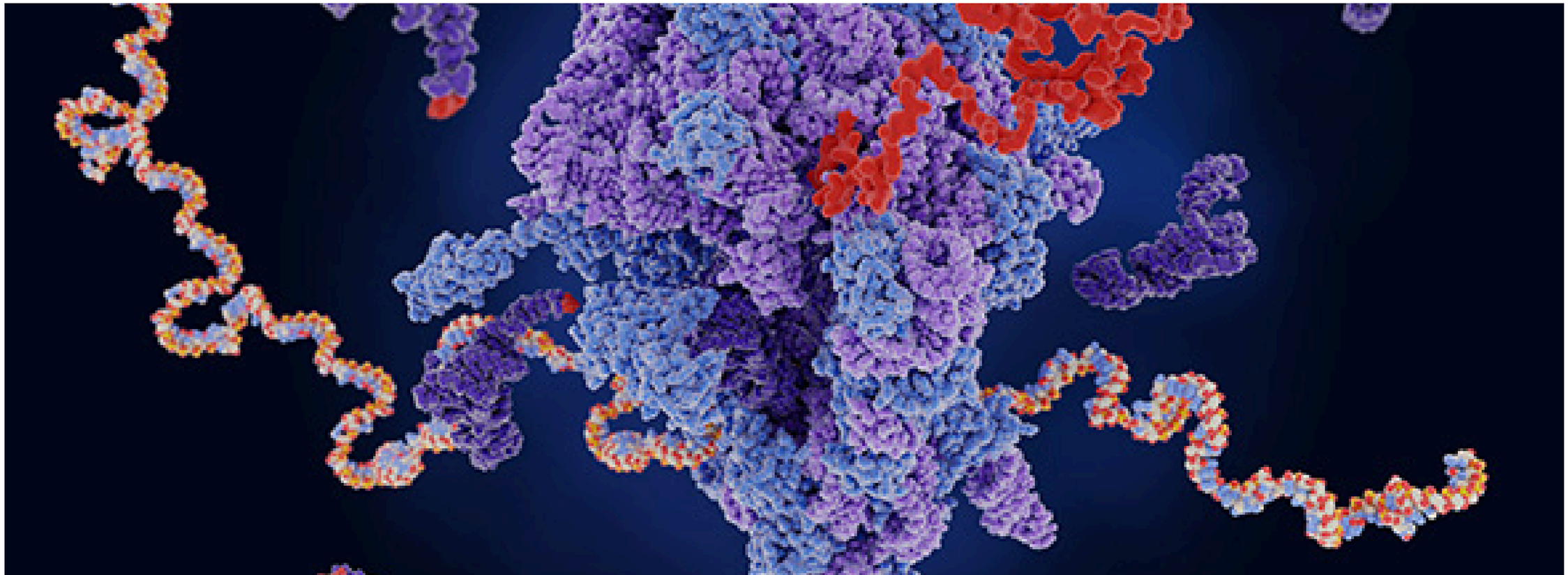


O que é bioinformatica?

Ciências da vida, saúde e médicas – A bioinformática apoia a informática médica; mapeamento genético em pedigrees e estudos populacionais; funcional, estrutural e farmacogenômica; proteômica e dezenas de outras “-ômicas” em evolução.

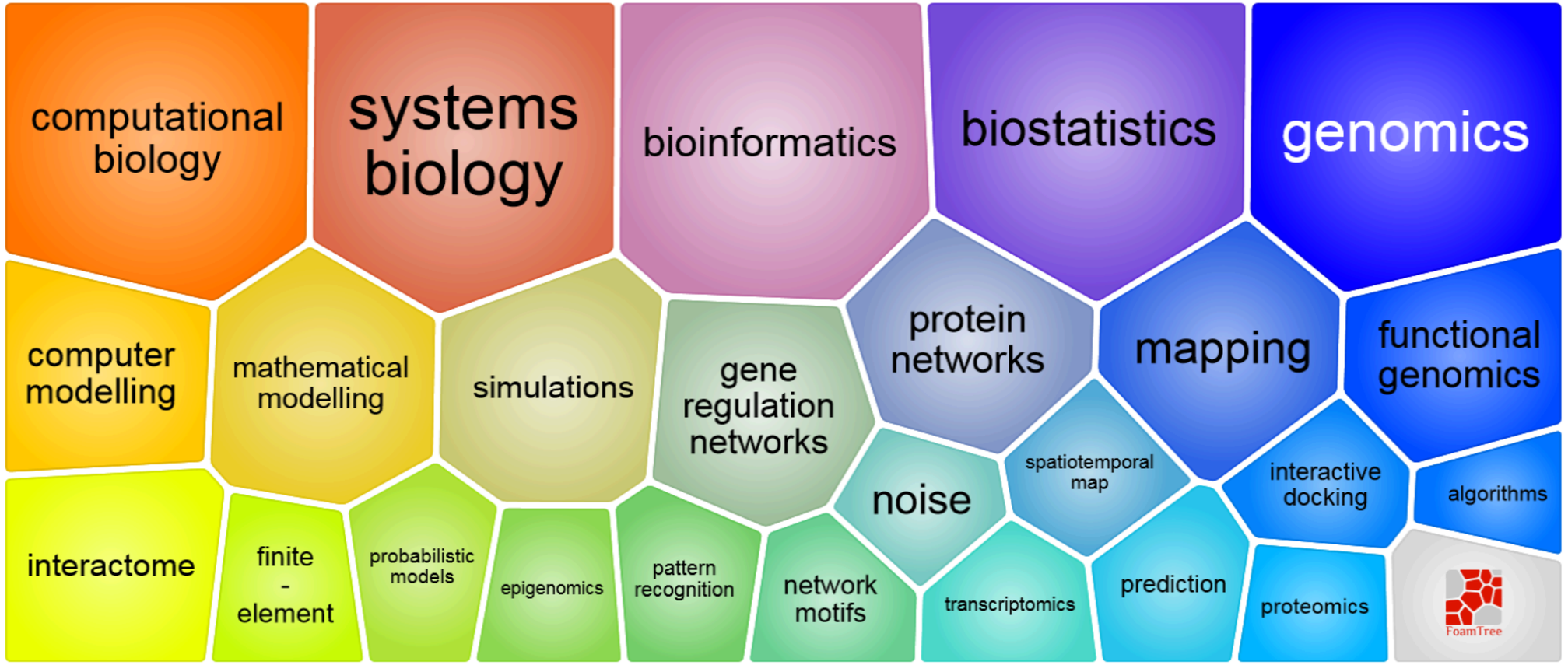
O que é bioinformatica?

Ciências básicas – A bioinformática depende de uma base sólida de química, bioquímica, biofísica, biologia, genética e biologia molecular que permite a interpretação de dados biológicos em um contexto significativo.



O que é bioinformatica?

Biologia computacional - A bioinformática abrange o desenvolvimento e aplicação de métodos analíticos e teóricos de dados, modelagem matemática e técnicas de simulação computacional para o estudo de sistemas biológicos, comportamentais e sociais.



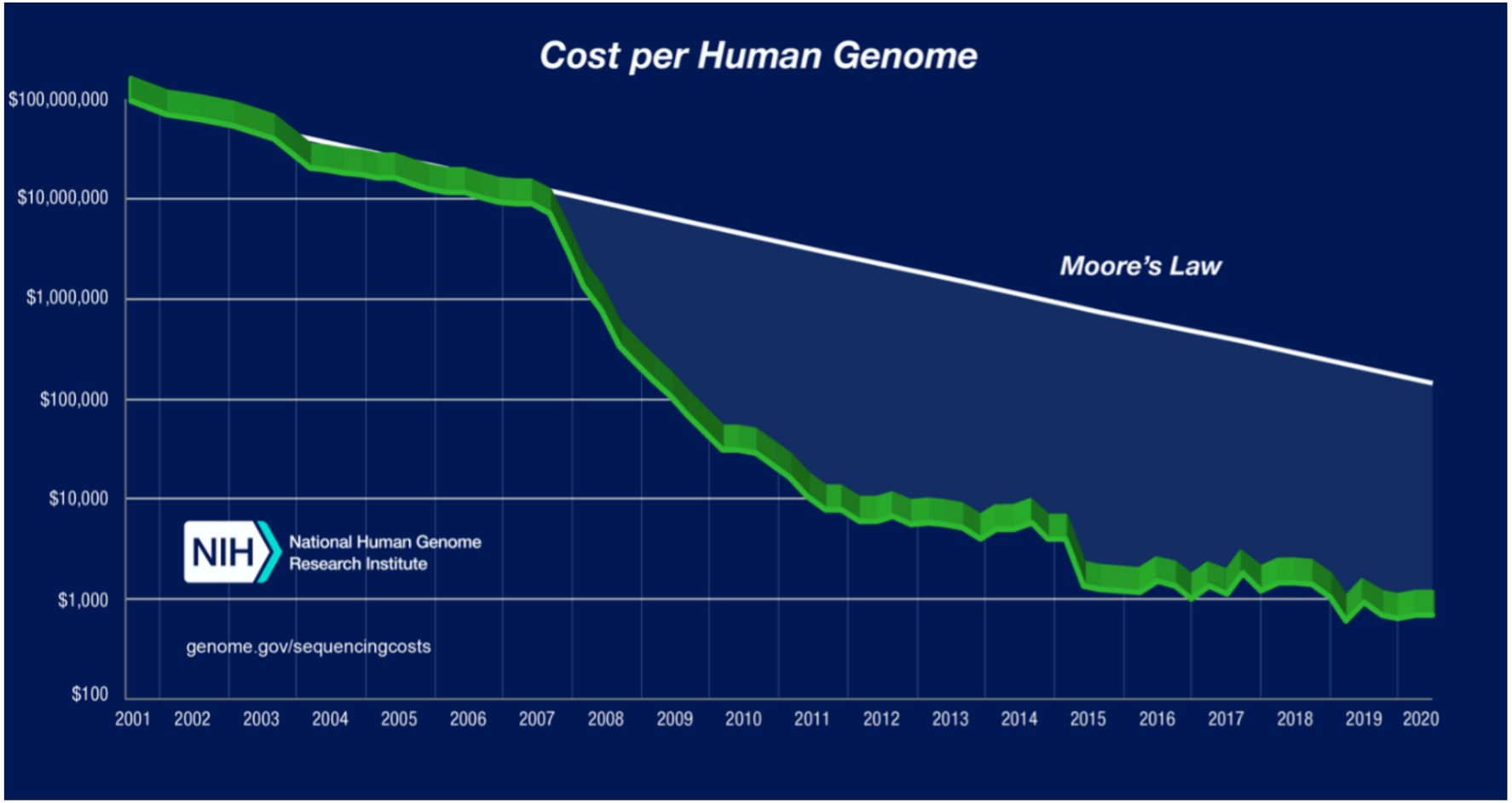
Histórico

A primeira definição do termo bioinformática foi cunhada por *Paulien Hogeweg* e *Ben Hesper* em 1970, para se referir ao estudo de processos de informação em sistemas biológicos. Esta definição colocou a bioinformática como um campo paralelo à bioquímica.

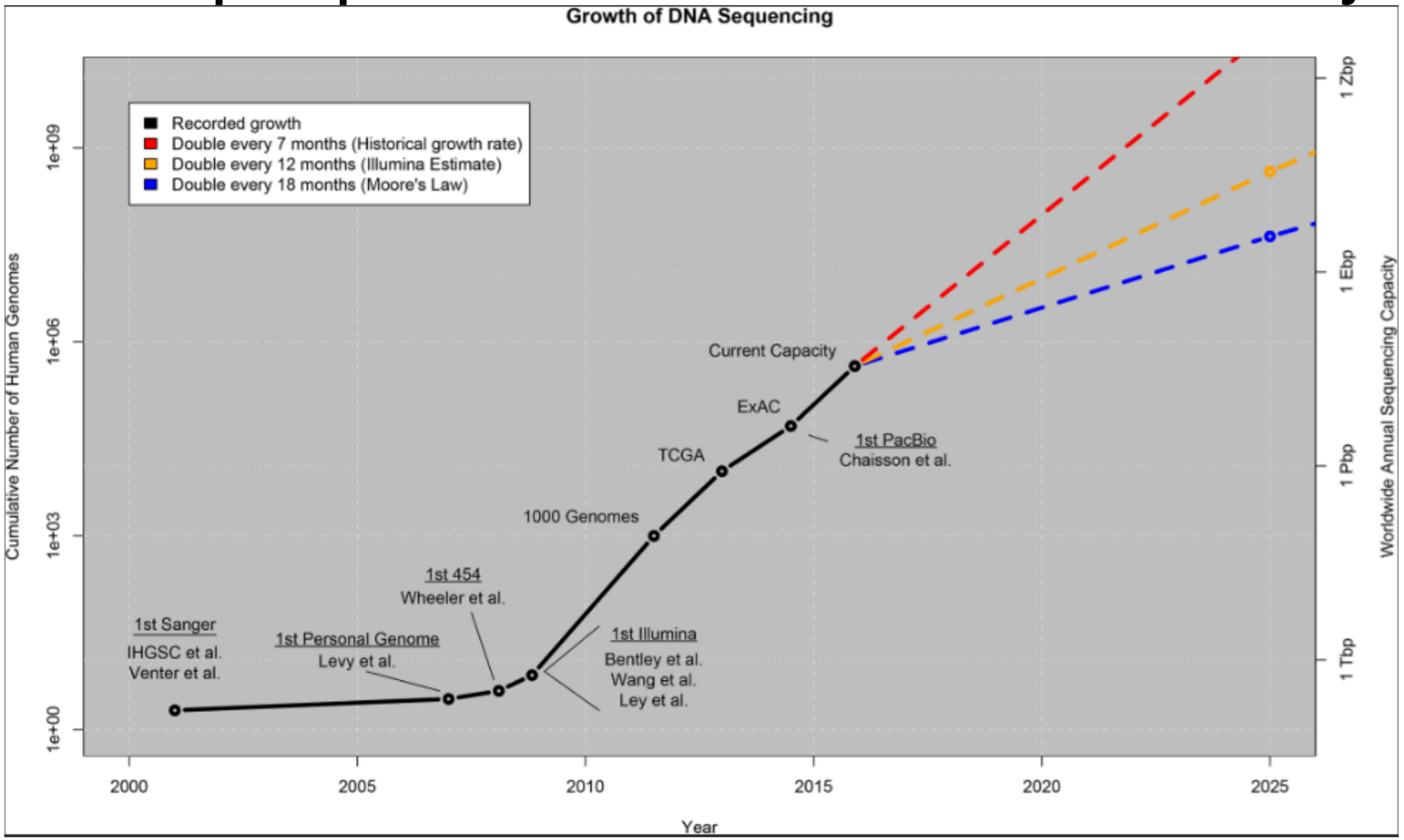
Paulien Hogeweg

Ben Hesper

A bioinformática traz uma perspectiva multidisciplinar para muitos dos problemas críticos que a profissão das ciências da saúde enfrenta hoje

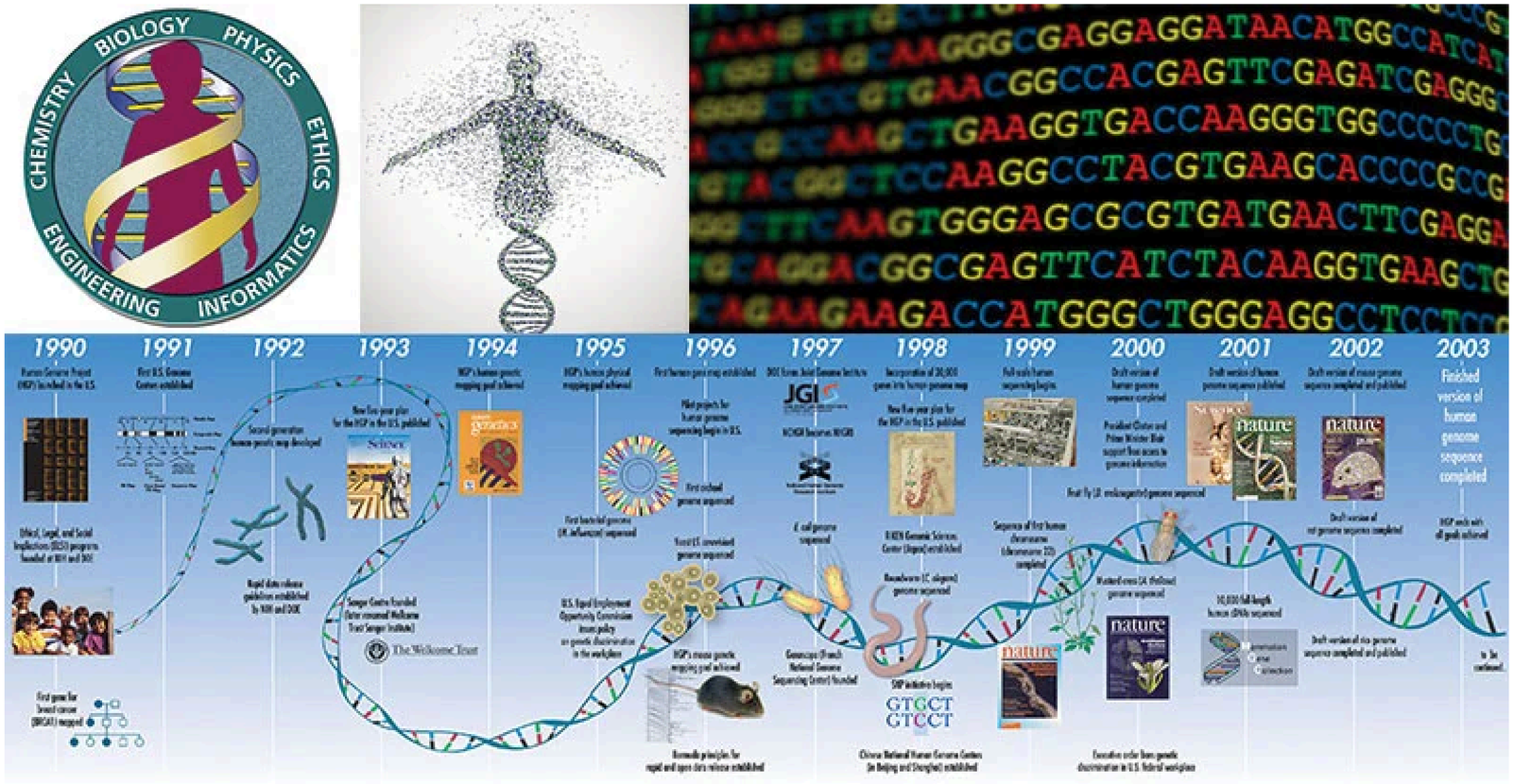


A bioinformática traz uma perspectiva multidisciplinar para muitos dos problemas críticos que a profissão das ciências da saúde enfrenta hoje



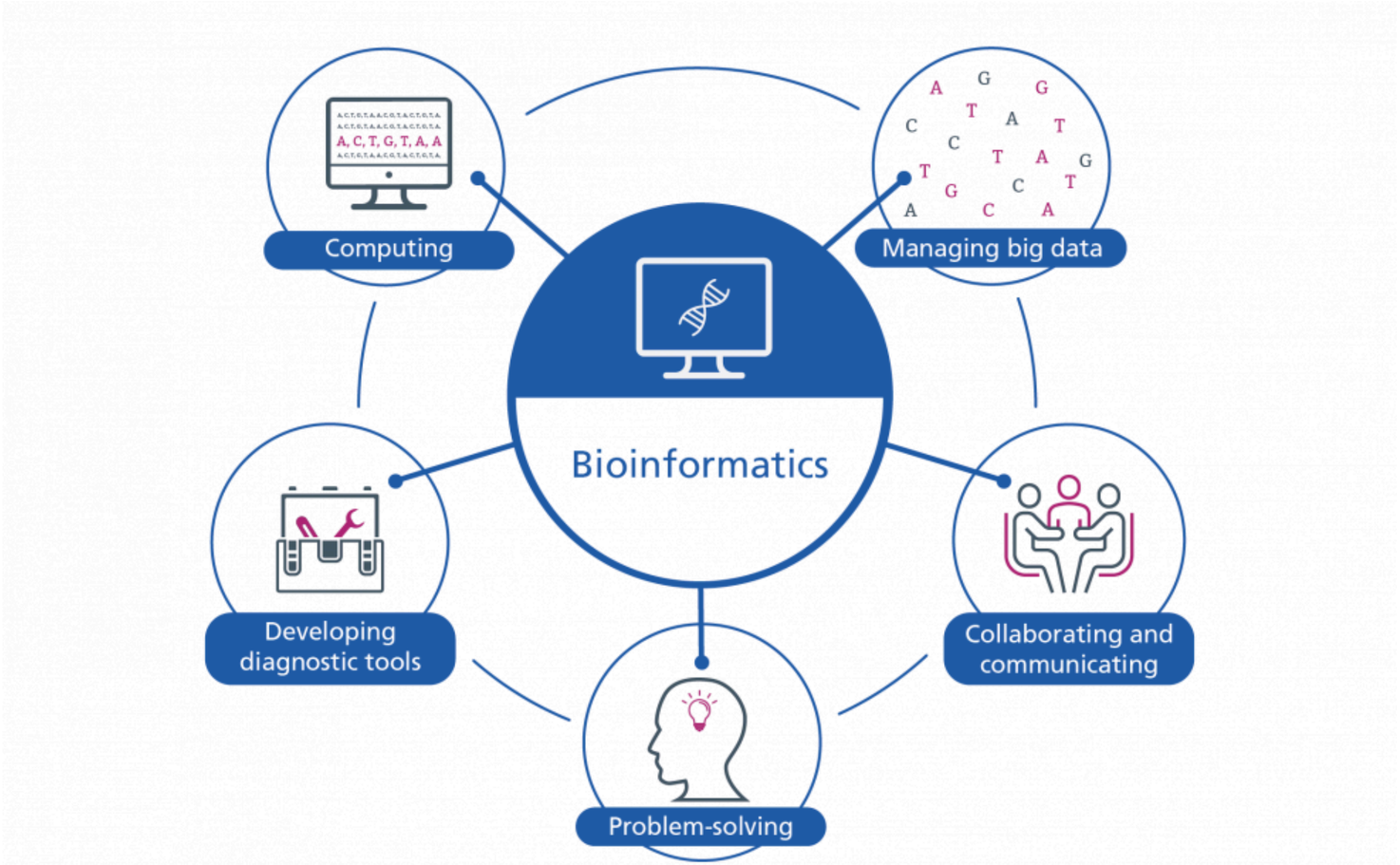
A bioinformática e a biologia computacional envolveram a análise de dados biológicos, particularmente DNA, RNA e sequências de proteínas.

O campo da bioinformática experimentou um crescimento a partir do **Projeto Genoma Humano** e pelos rápidos avanços na tecnologia de sequenciamento de DNA

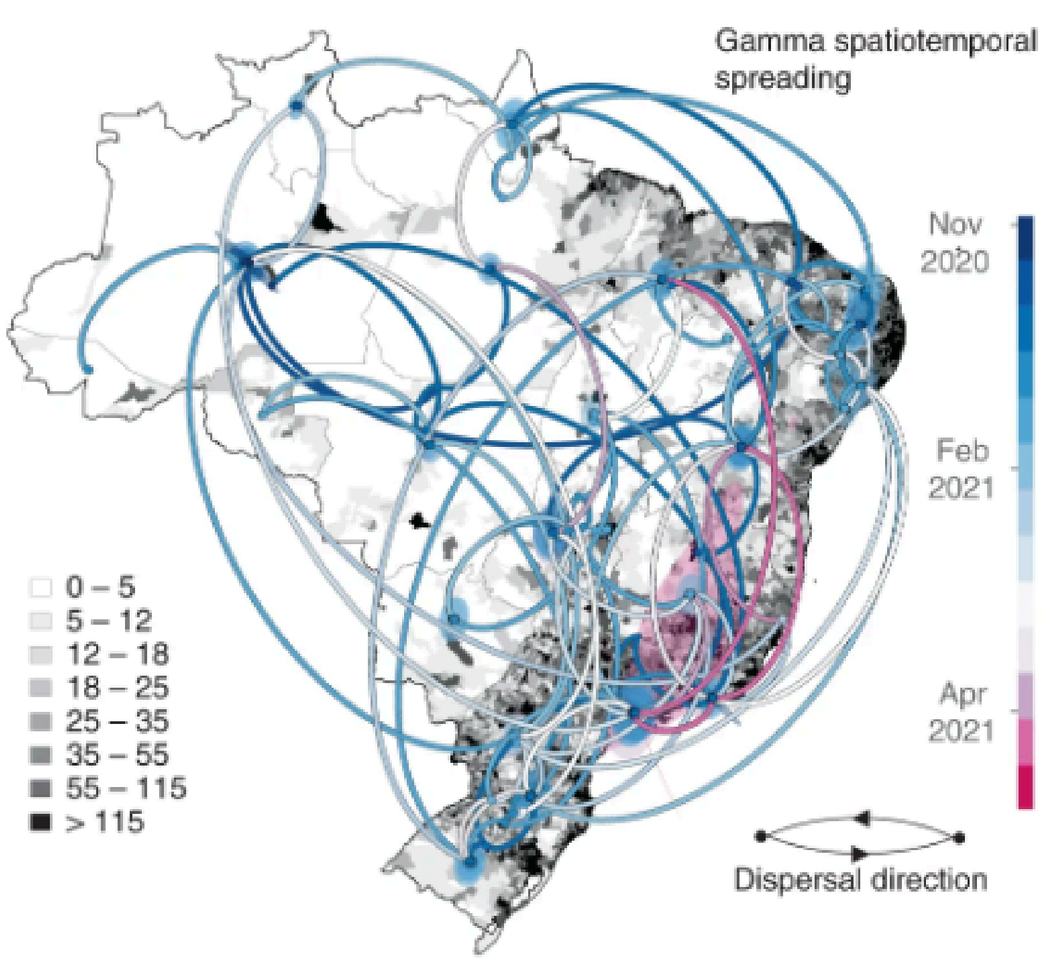


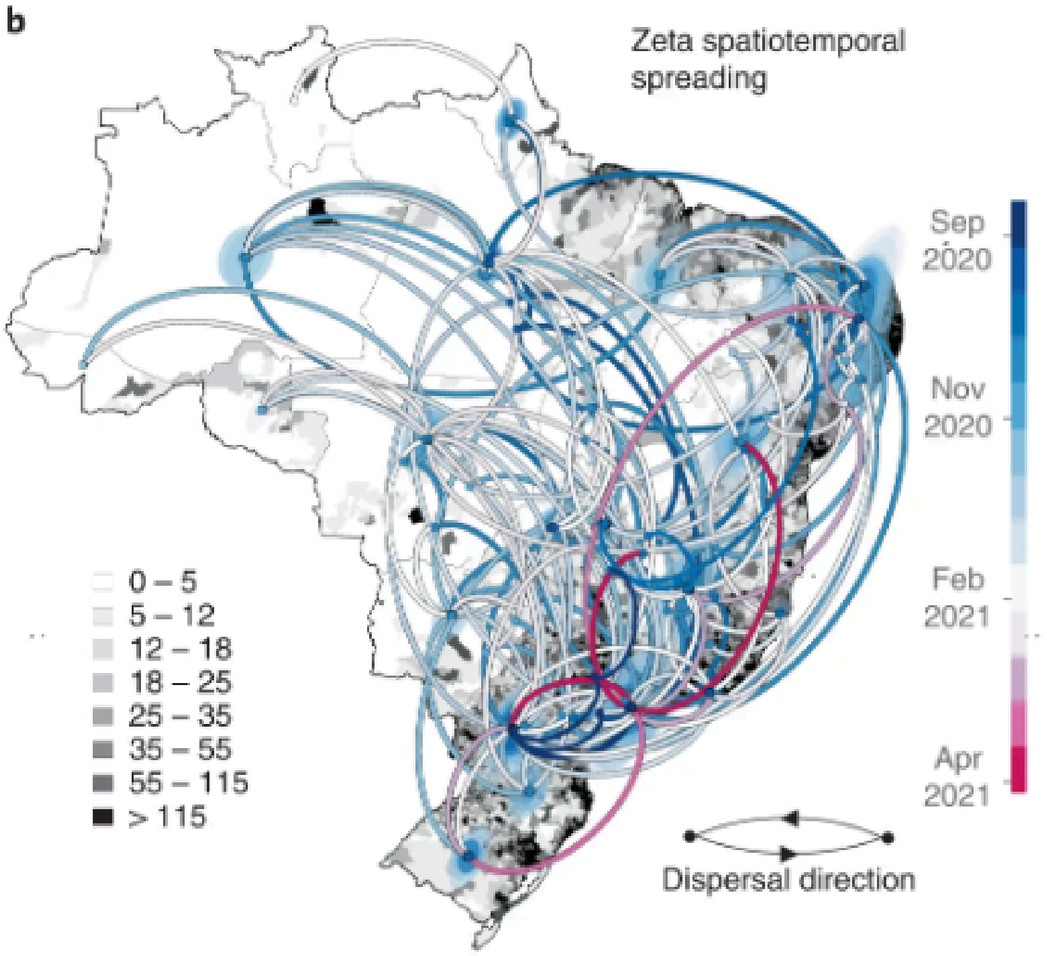
Bioinformática usa computadores para **organizar, coletar, analisar e compreender** dados biológicos

A partir disso, une computadores e pesquisadores em ciências da vida, e tem um enorme potencial em múltiplas áreas, já que a maioria dos programas e dados está disponível gratuitamente.



Impacto da bioinformática nos tempos atuais - Pandemia de COVID-19





Giovaneti et al.Nat. Microbio

Desafíos no Ensino de Bioinformática

Grande Interdisciplinaridade: Unindo biologia, ciência da computação e estatística

Acompanhando os avanços rápidos: atualizando os currículos regularmente

Origens diversificadas dos alunos: abordando diversos níveis de especialização

Infraestrutura computacional: fundamental para o aprendizado prático

Métodos de ensino tradicionais

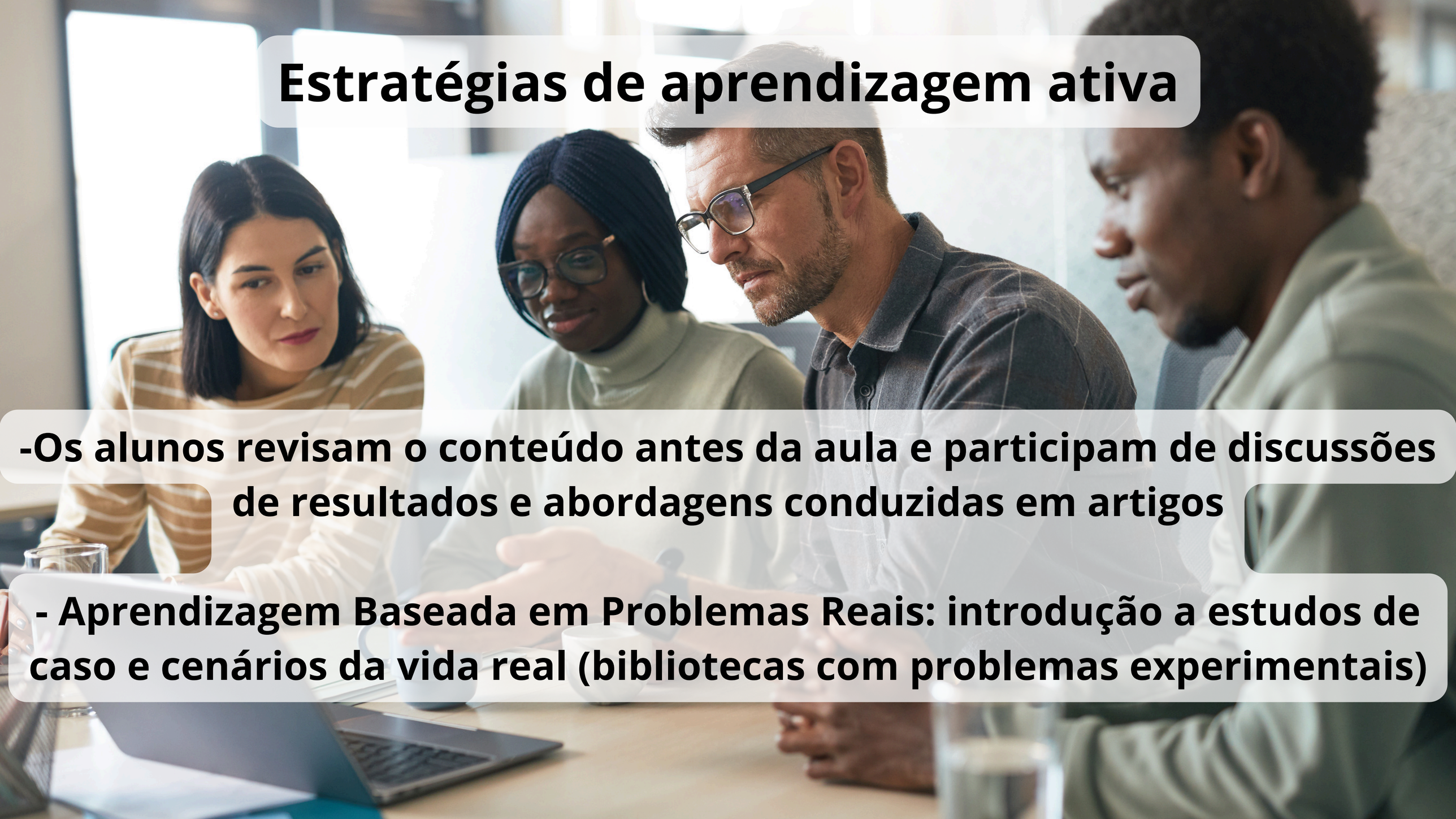
Aulas expositivas: Conceitos fundamentais explicados através de apresentações

Livros didáticos e materiais de leitura: fontes de conhecimento fundamental

Exames e Avaliações: Avaliando a compreensão teórica

Livros-texto em português

Estratégias de aprendizagem ativa

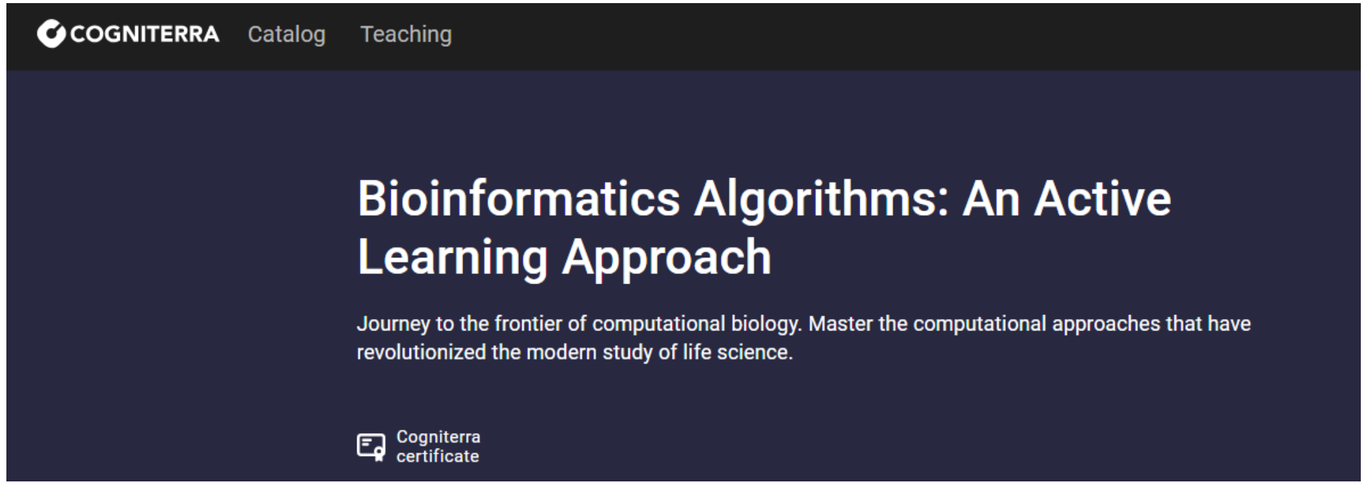


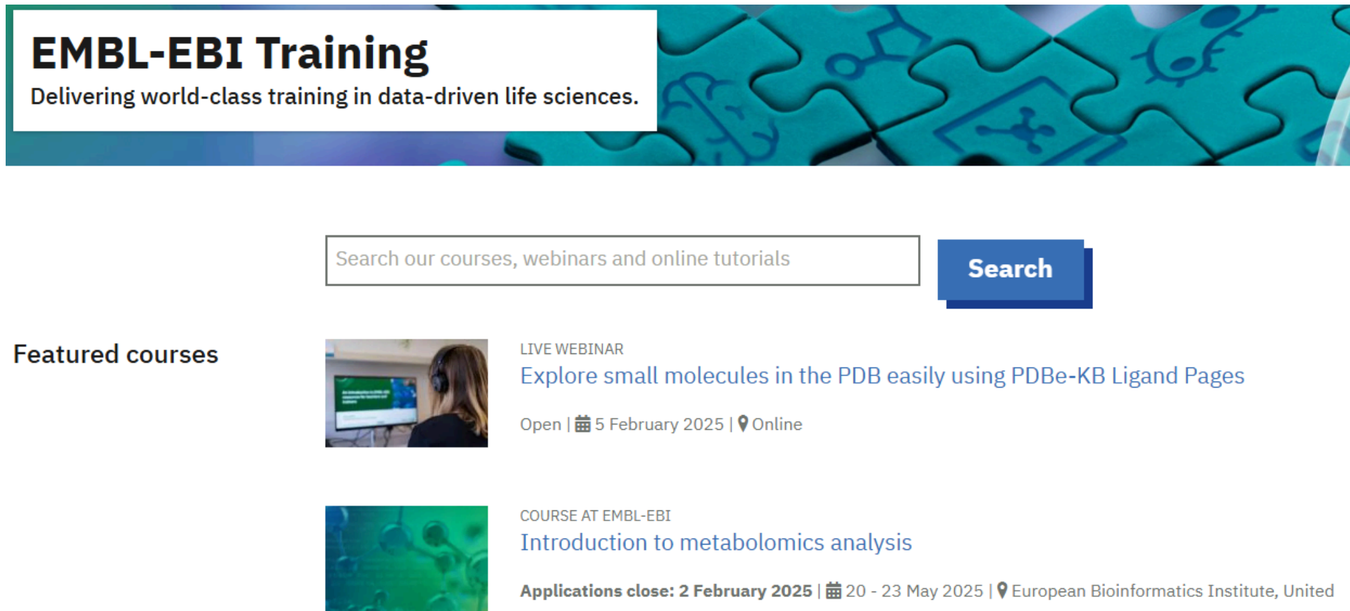
-Os alunos revisam o conteúdo antes da aula e participam de discussões de resultados e abordagens conduzidas em artigos

- Aprendizagem Baseada em Problemas Reais: introdução a estudos de caso e cenários da vida real (bibliotecas com problemas experimentais)

Aprendizagem on-line e combinada

Cursos on-line: plataformas como COGNITERRA, EMBL-EBI,





Conceitos fundamentais para graduação e pós-graduação

Identificação do nível de conhecimento dos estudantes

Abordagens computacionais e práticas de programação:

Ensino de linguagens de programação (Python, R)

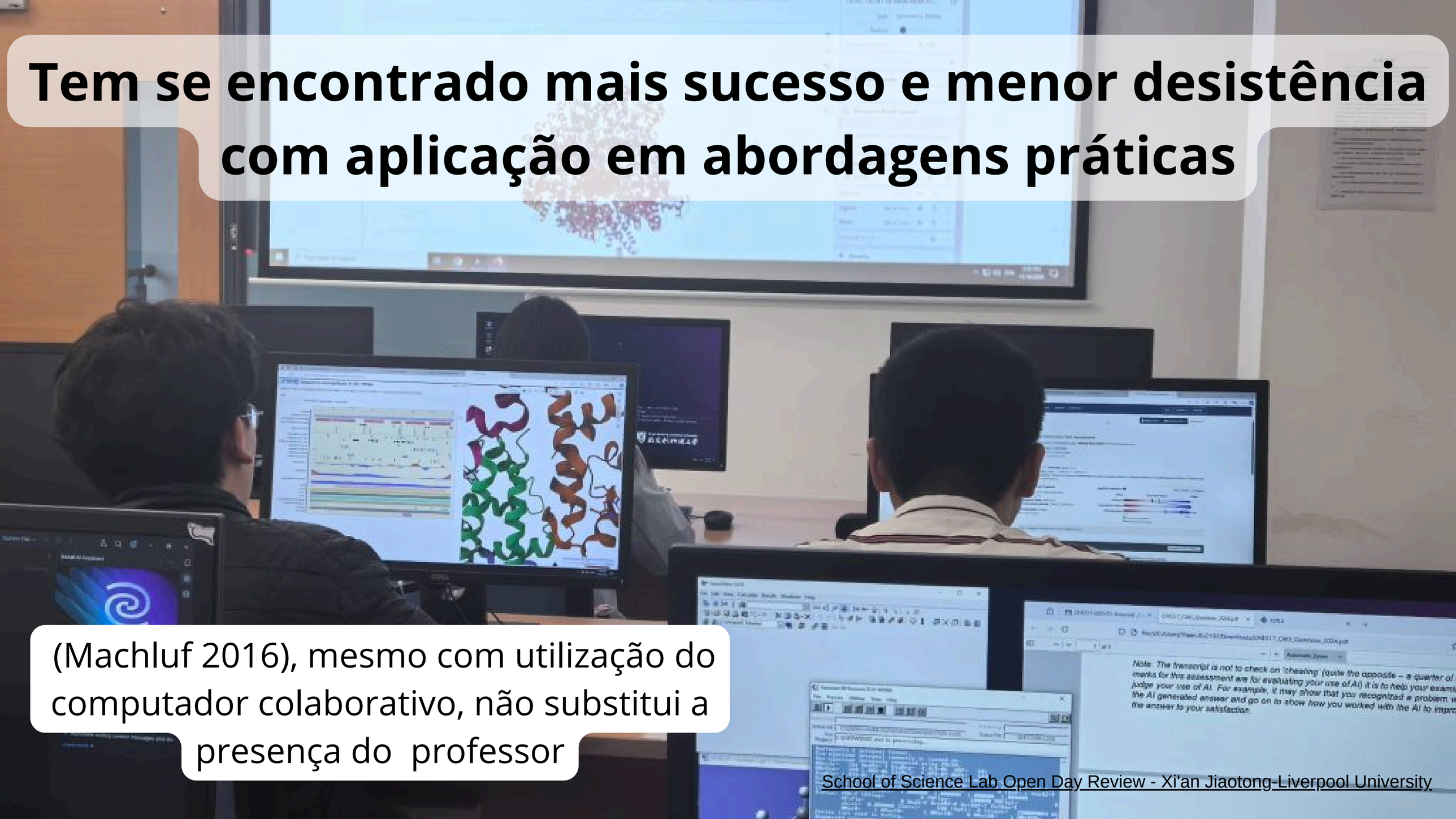
Exploração de banco de dados:

NCBI, Ensembl, UniProt para dados biológicos

Desenvolvimento de Algoritmos:

Projetando soluções computacionais para problemas biológicos

Tem se encontrado mais sucesso e menor desistência com aplicação em abordagens práticas



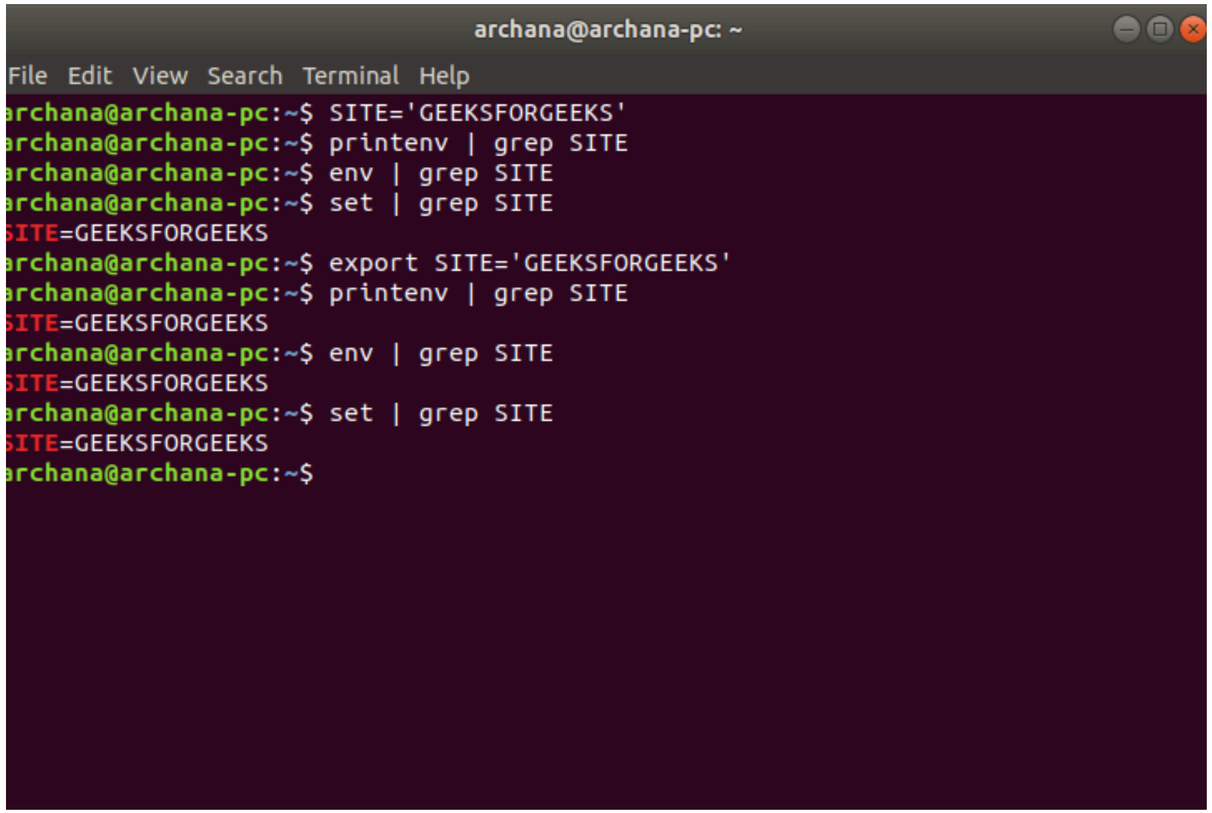
(Machluf 2016), mesmo com utilização do computador colaborativo, não substitui a presença do professor

School of Science Lab Open Day Review - Xi'an Jiaotong-Liverpool University

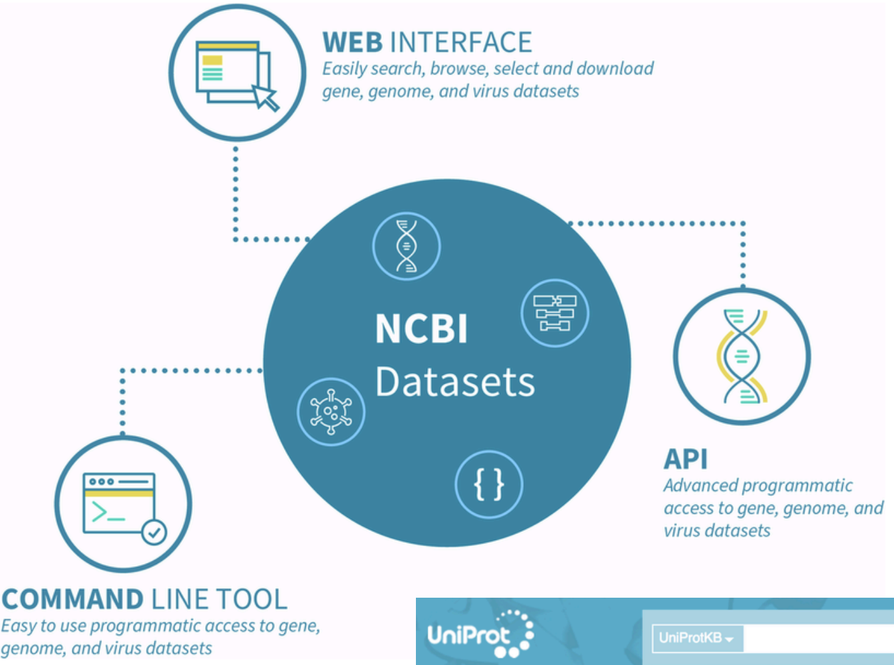
Abordagens computacionais e programação prática: Ensino de linguagens de programação (Python, R, scripting Bash)

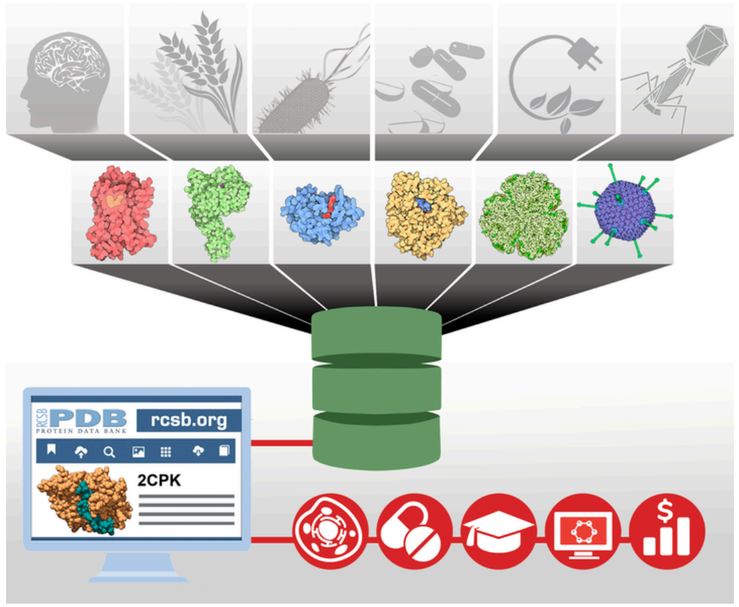
Step-by-step

Ambiente Linux com Bash



Exploração de banco de dados: NCBI, Ensembl, UniProt, PDB (Protein data base), AFDB, StringDB





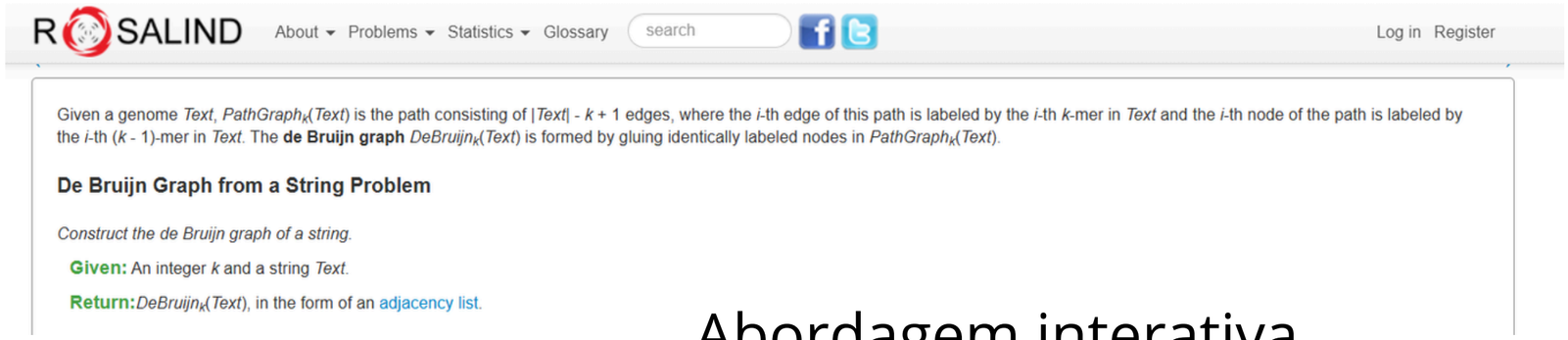
Exposição dos alunos a desenvolvimento de Algoritmos: Projetando soluções computacionais para problemas biológicos

Conceitos fundamentais

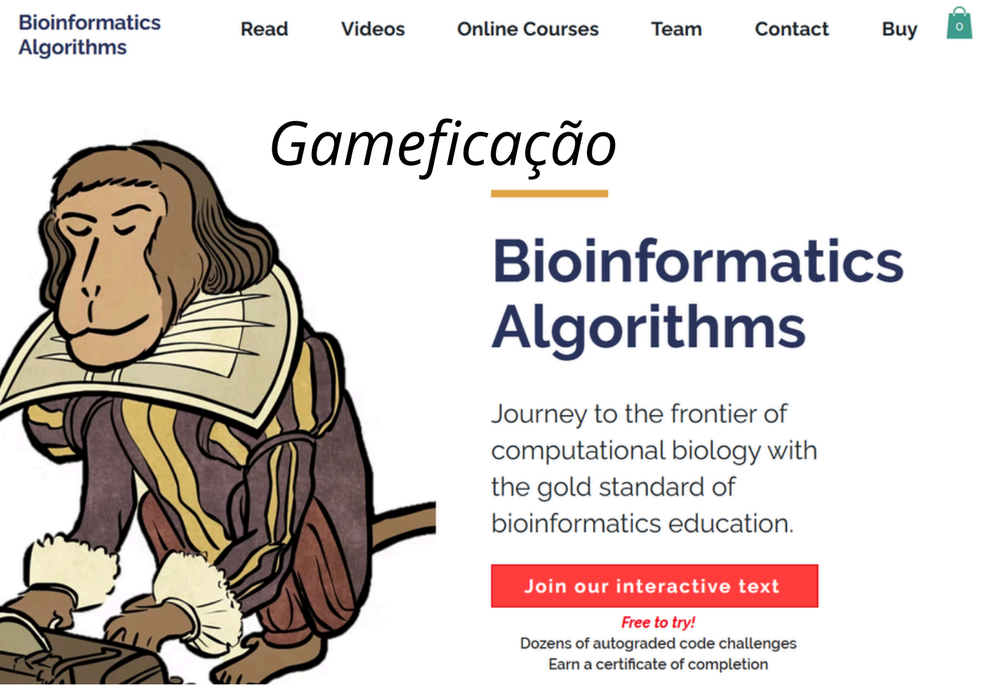
Needleman-Wunsch

match = 1 mismatch = -1 gap = -1

		G	C	A	T	G	C	G
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

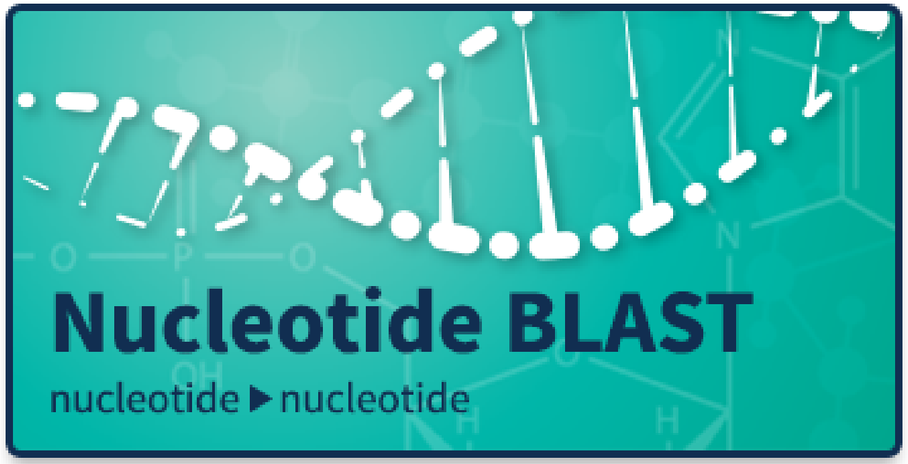


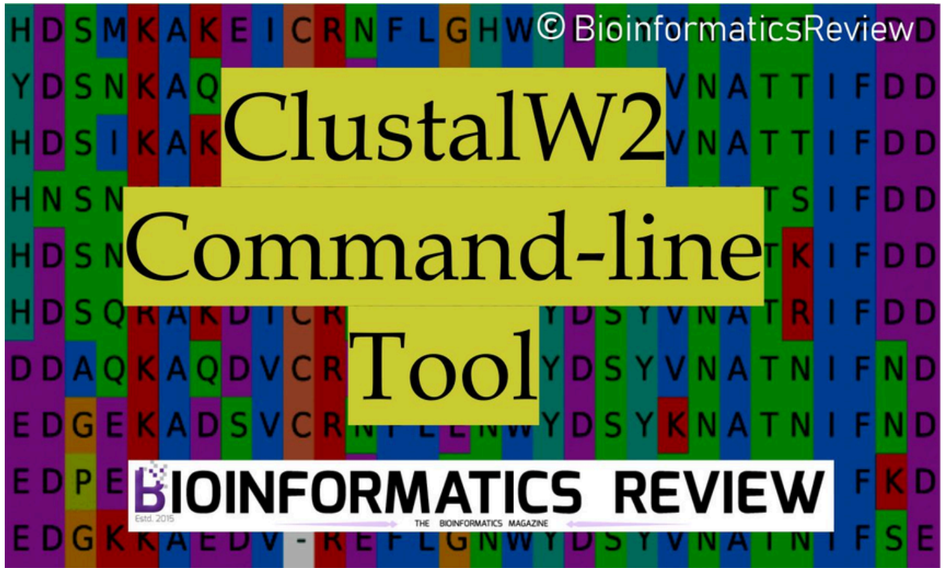
Abordagem interativa

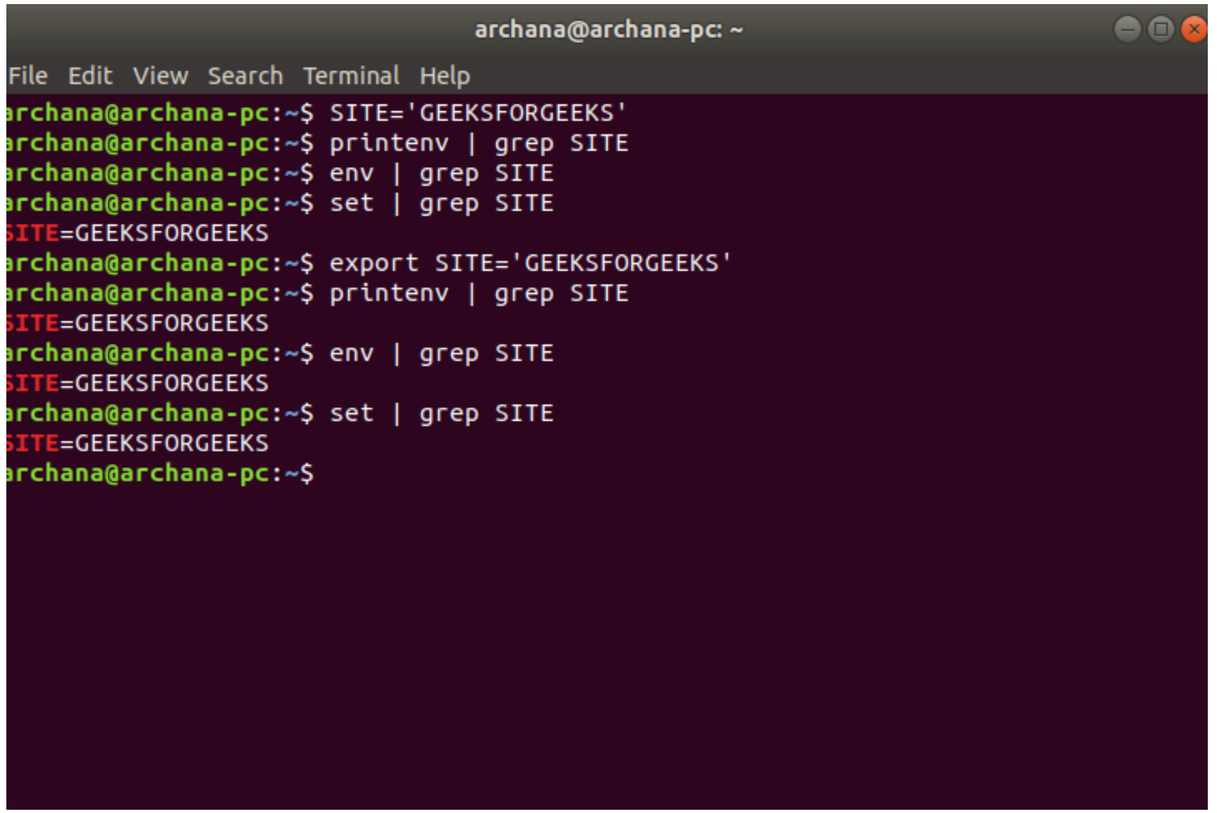


Software e Ferramentas na Educação em Bioinformática

BLAST & Clustal Omega: ferramentas de alinhamento de sequência

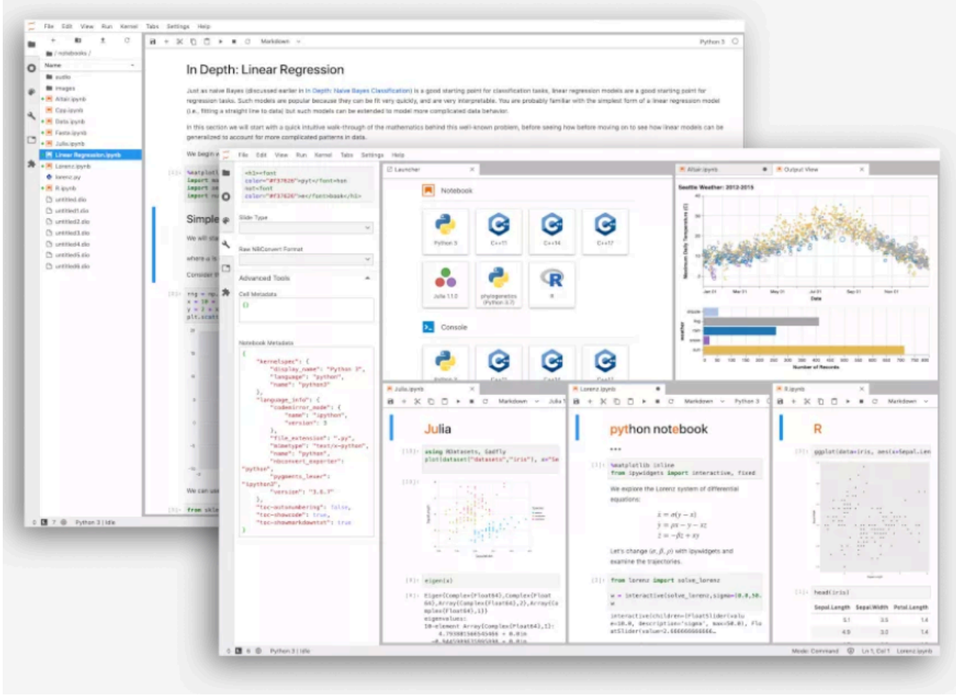






Bioconductor & Galaxy: plataformas de análise de dados

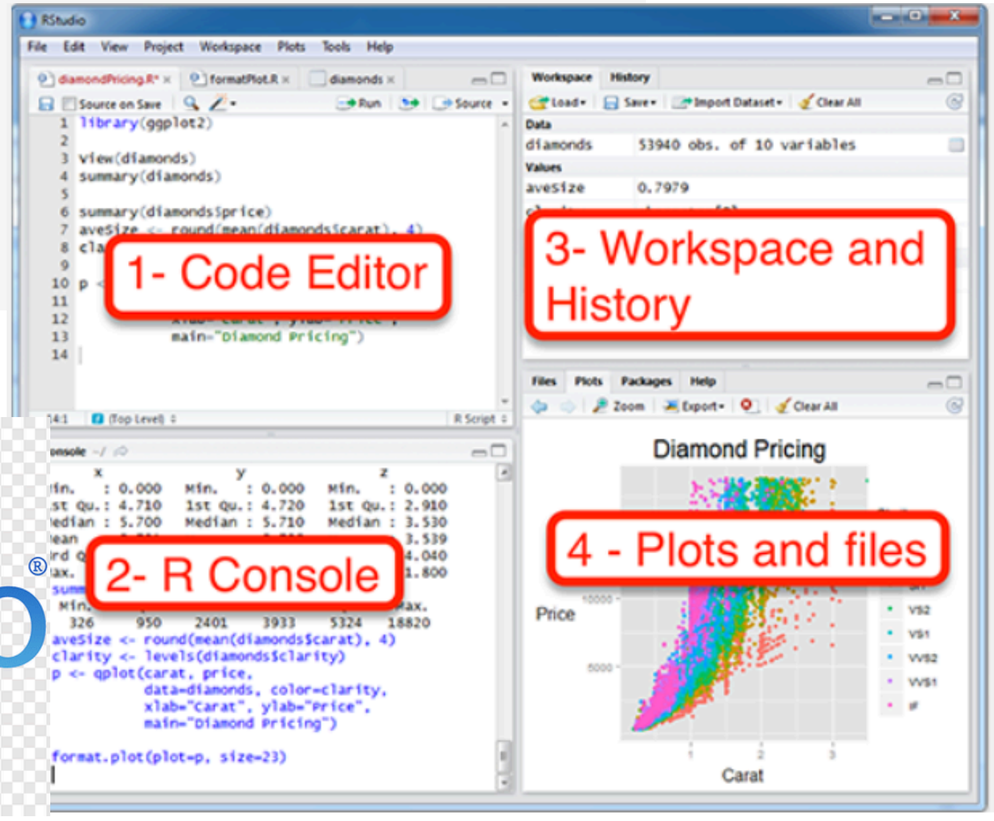
Jupyter Notebooks e RStudio: ambientes de codificação interativos



JupyterLab: A Next-Generation Notebook Interface

JupyterLab is the latest web-based interactive development environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning. A modular design invites extensions to expand and enrich functionality.

Try it in your browser | Install JupyterLab



Aprendizagem Baseada em Projetos em grupo ou individuais

Projetos em grupo: solução colaborativa de problemas

Conjuntos de dados do mundo real: análise de dados genômicos, proteômicos, transcriptômicos e estruturais

Primeiras abordagens práticas de bioinformática na graduação

Wood and Gebhardt 2013
Dados reais e treinamento prático

Início simples e básico

Bioinformatics teaching resources

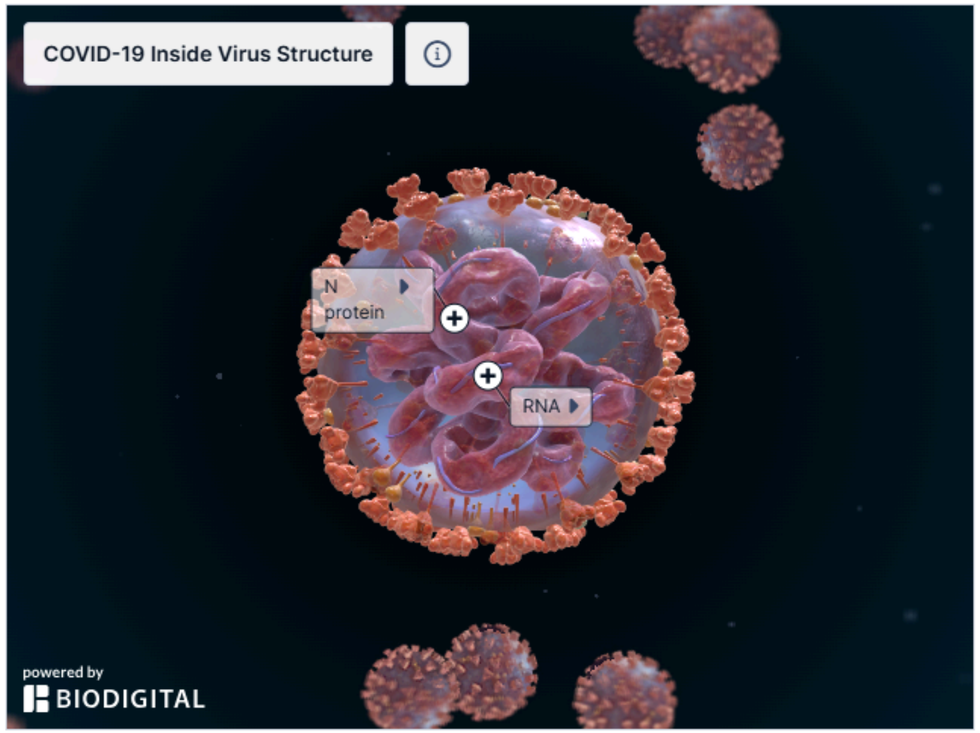
Worksheets and lesson ideas to challenge students aged 11 to 16 to use bioinformatics in the classroom (GCSE and Key Stage 3)

Bioinformatics is the process of using computers to organise, collect, analyse and understand biological data. The subject unites computers and biologists, and has huge potential in the science classroom as the majority of programs are freely available.

- BLAST to search genomes and proteomes for genes and proteins
- NCBI to find sequences for proteins and genes
- PDB to visualise 3D protein structures

Bioinformatics and the coronavirus

Download exploring the genome of SARS-CoV-2 using bioinformatics. This session introduces students (aged 15-18) to a number of tools used in bioinformatics to explore the genome of SARS-CoV-2 (coronavirus). There are a number of questions that students answer as they move through the session and there is an assignment at the end for them to complete. Answers are provided to questions in the notes section. The following link allows students to explore interactive models of SARS-CoV-2.



Science as a community

This presentation takes you through what is currently known about the genome of the coronavirus isolated from China in 2019.

We have access to this data thanks to the work of a community of scientists who are committed to making their data publicly available.

UA LITTLE ROCK
Donaghey College of Science, Technology, Engineering, and Mathematics

Before you begin…

Today you will look at data represented in something called FASTA format.

This is a text based format for representing DNA, RNA or protein sequences.

Read more about this here.

Do you understand FASTA format?

CAA59436.1 band 7 integral membrane
MAEKRHTRDSEAQRLPDSFKDSPSKGLGPCGWI

Is this a protein or DNA sequence?

What does M represent?

What is missing from this FASTA format?

How coronaviruses get inside cells and replicate

1. The S protein of the coronavirus binds to surface proteins on the outside of the host cell
2. The virus genome now gains access to inside the cell
3. When inside, the single stranded RNA genome attaches to host ribosomes and gets them making virus proteins
4. These proteins then make more copies of the virus genome
5. After assembly, virus particles are transported to the cell membrane and are released going on to infect other cells

Q. What does the 'S' stand for in S protein and where is it located?

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4369385

The genome of SARS-CoV-2

- The 30 kilobase genome of the Wuhan seafood market strain is publicly available
- You are now going to explore its genome today
- Go to the NCBI GenBank page to look at this sequence and answer the questions from the next slide

Questions about the genome

- How long is the viral genome?
- What organism was this virus isolated from?
- Find the *S gene* that codes for the S protein. What are the nucleotide positions that this gene is located in?
- For comparison, how long is a) the human genome and b) the *E.coli* genome?

How similar is this virus to other viruses?

To find out you will search the GenBank database using a program called BLAST. It's a bit like Google but for biological sequences.

Step 1. Copy the complete genome sequence of the Wuhan isolate in FASTA format

Step 2. Paste this into the BLAST query sequence box

Step 3. Press BLAST and wait for results to show

Reviewing your results

- Look at the results – or matches. These list other sequences in the database that are similar to the sequence you searched with. Take some time to explore this page.
- Which other countries have sequenced SARS-CoV-2 genomes that match the Wuhan isolate?
- Find a match that is less than 81 % identical. Which host did this virus come from?

Reflections

How has your understanding of SARS-CoV-2 changed as a result of this session?

How has your understanding of bioinformatics changed as a result of this session?

Computational thinking - (Goodman and Dekhtyak 2014)

Assignment

Using the tools and information you have learnt about in this session:

1. Explore the genome of human immunodeficiency virus (GenBank page)

2. Compare and contrast SARS-CoV-2 with Human immunodeficiency virus (GenBank page)

Bioinformática estrutural - início simples e básico

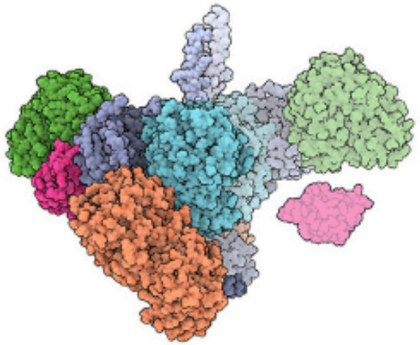
RCSB PDB-101 *Molecular explorations through biology and medicine*

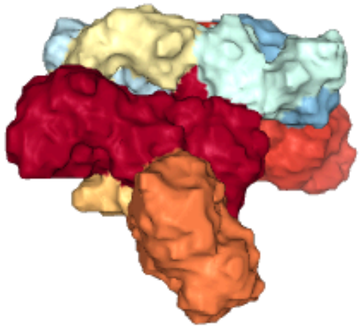
Search Molecule of the M

Training and outreach portal of RCSB PDB PROTEIN DATA BANK

Molecule of the Month

January 2025





Assembly Line Polyketide Synthases

Large multienzyme complexes that synthesize diverse small molecules in a stepwise manner.

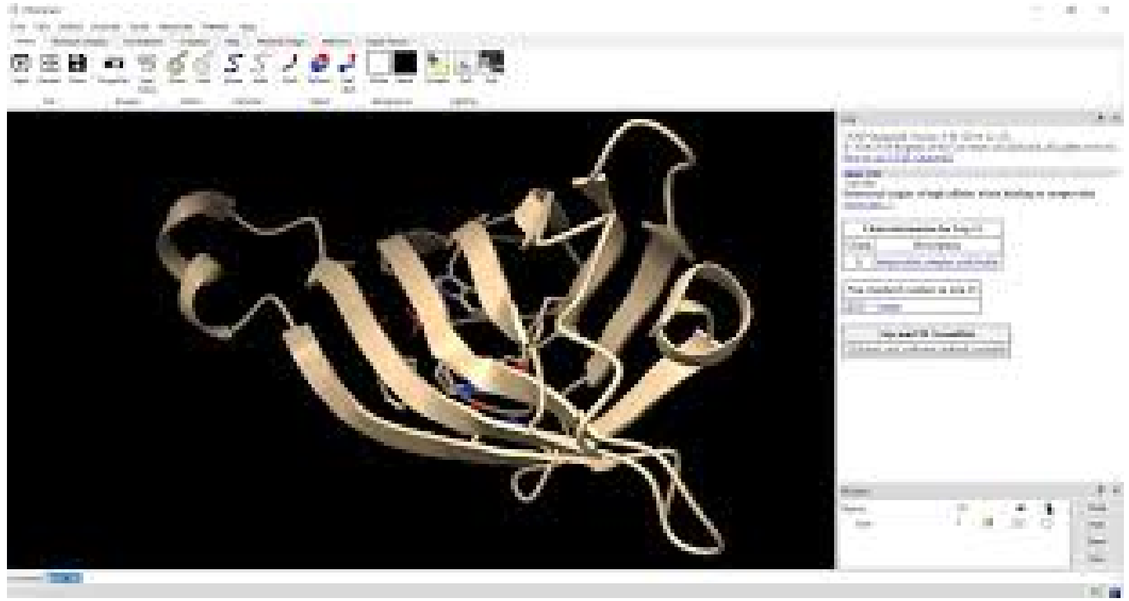
More

3D View: 7S6C

Style	Color	Spin
○ Cartoon	○ Rainbow	● On
○ Spheres	● Chain	○ Off
● Surface	○ Structure	

All articles: By Date | By Category | By Title

Vincular atividades a currículos científicos pré-existentes



Abordagens práticas de bioinformática na pós-graduação baseadas em projetos práticos

Aprofundando

Projetos em grupo: solução colaborativa de problemas

Conjuntos de dados do mundo real: análise de dados genômicos, proteômicos, transcriptômicos e estruturais

Requer pré-requesitos

1 Como julgar se os experimentos conduzidos na bancada foram feitos adequadamente

2 Design experimental em bioinformatica aplicada a omicas

Ao usar métodos de alto rendimento, como RNAseq pode traçar o perfil - na resolução de um nucleotídeo únicos (isoformas) - da abundância de dezenas de milhares de transcritos distintos codificados em genomas de eucariotos

Adaptações: Por razões de poder computacional disponível, pode-se selecionar cromossomos especifícos por exemplo

Akalin 2020_Computational Genomics with R

```r
library(pheatmap)
# get the expression data for the gene set of interest
M <- normalizedCounts[rownames(normalizedCounts) %in% geneSet1, ]
# log transform the counts for visualization scaling by row helps visualizing
# relative change of expression of a gene in multiple conditions
pheatmap(log2(M+1),
         annotation_col = colData,
         show_rownames = TRUE,
         fontsize_row = 8,
         scale = 'row',
         cutree_cols = 2,
         cutree_rows = 2)
```

Qual a relevancia biológica?
Criticidade sobre os resultados

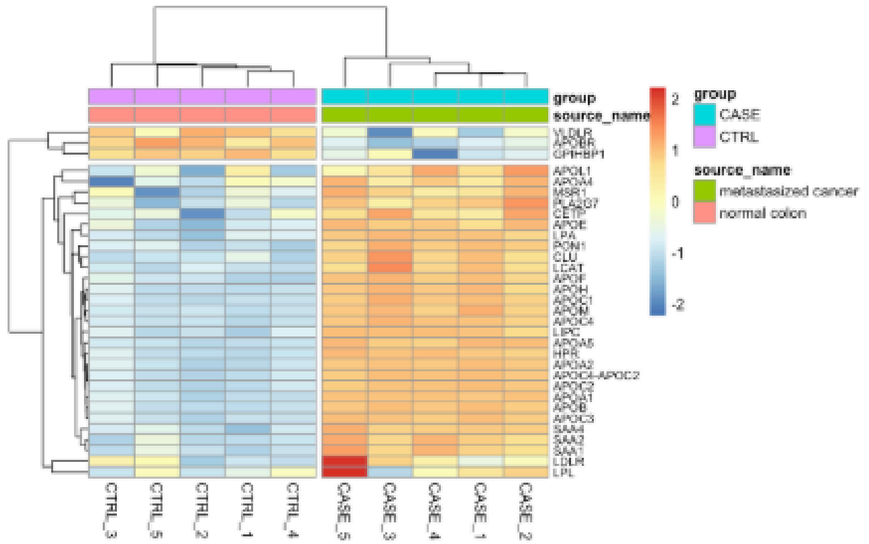


FIGURE 8.11: Heatmap of expression value from the genes with the top GO term.

```r
library(stats)
library(ggplot2)
#transpose the matrix
M <- t(tpm[selectedGenes,])
# transform the counts to log2 scale
M <- log2(M + 1)
#compute PCA
pcaResults <- prcomp(M)

#plot PCA results making use of ggplot2's autoplot function
#ggfortify is needed to let ggplot2 know about PCA data structure.
autoplot(pcaResults, data = colData, colour = 'group')
```

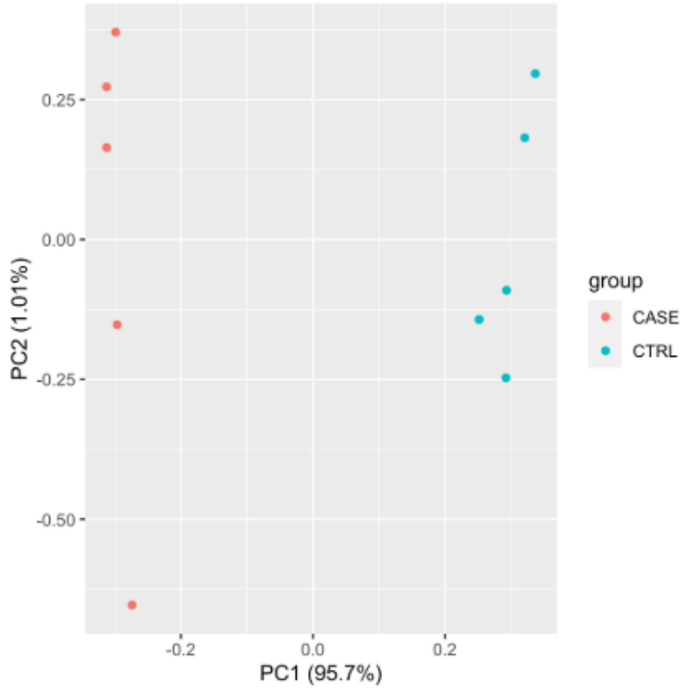


FIGURE 8.3: PCA plot of samples using TPM counts.

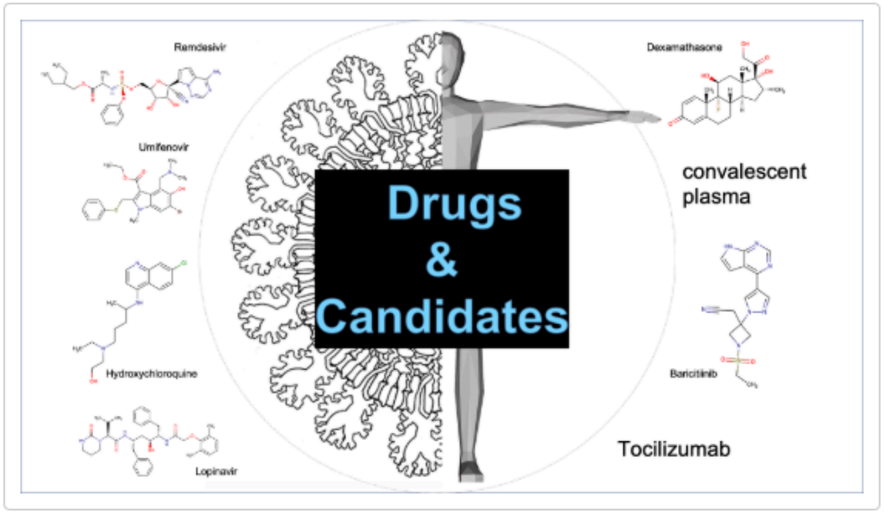
Bioinformática estrutural aprofundando nas práticas

PDB-101 | Molecule of the Month | Browse | Learn | Train | Teach | Global Health | SciArt | Events

COVID-19 in Molecular Detail

- Getting Started: Hand Washing
- SARS-CoV-2 Life Cycle
- The Main Protease Enzyme
- Evolution of SARS-CoV-2
- SARS-CoV-2 Genome and its Expression
- Infection: The Spike story
- The Disease: COVID-19

Treatment: Drugs



Blocking Replication with Remdesivir

Learning Objective: To explore the Structure of RNA dependent RNA Polymerase (RdRP) bound to Remdesivir using RCSB Mol* to understand how it can block replication.

Exploration

A. Explore the structure and action of Remdesivir

In order to understand how Remdesivir works it may be helpful to first learn about its chemical properties. This part of the exercise will introduce you to a resource called DrugBank - a free online data resource with information about drugs, their targets, and many other details.

- Go to the DrugBank home page (https://www.drugbank.ca/) and type the name "Remdesivir" in the top search bar.

ZINC | Substances | Catalogs | Tranches | Biological | More

ZINC20

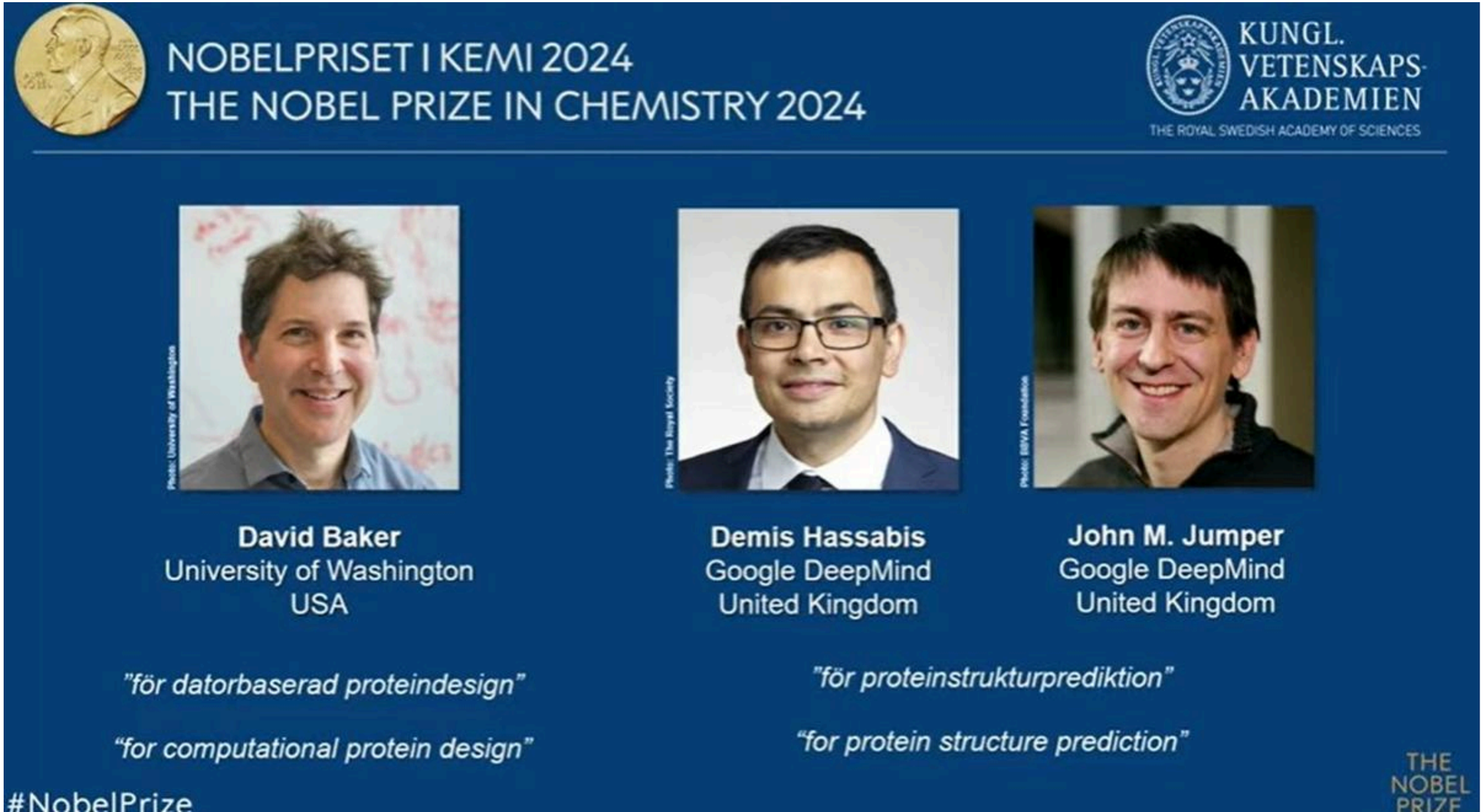
Welcome to ZINC, a free database of commercially-available compounds for virtual screening. ZINC contains over 230 million purchasable compounds in ready-to-dock, 3D formats. ZINC also contains over 750 million purchasable compounds you can search for analogs in under a minute.

B2. Do you recognize the Remdesivir in this structure? Is it part of the polymer or listed as a small molecule?

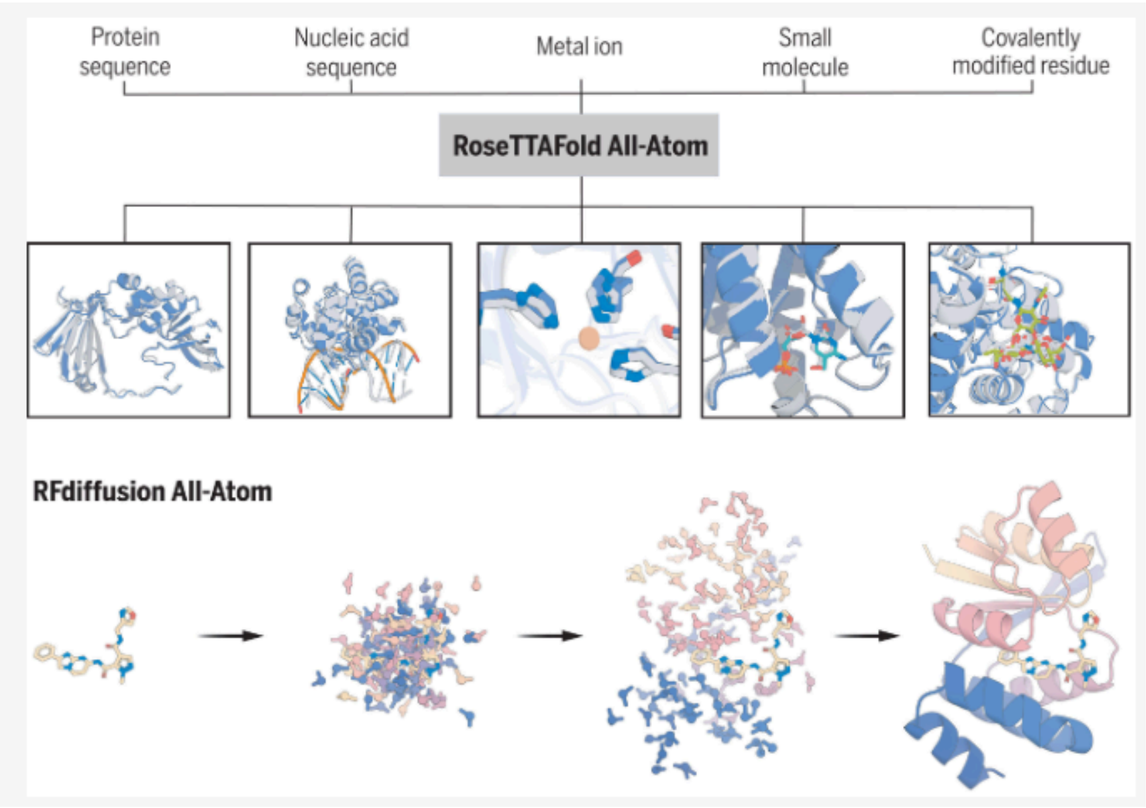
Hint: The picture that you saved in section A is that of the Prodrug. This molecule is processed and only the nucleotide analog remains in this structure.

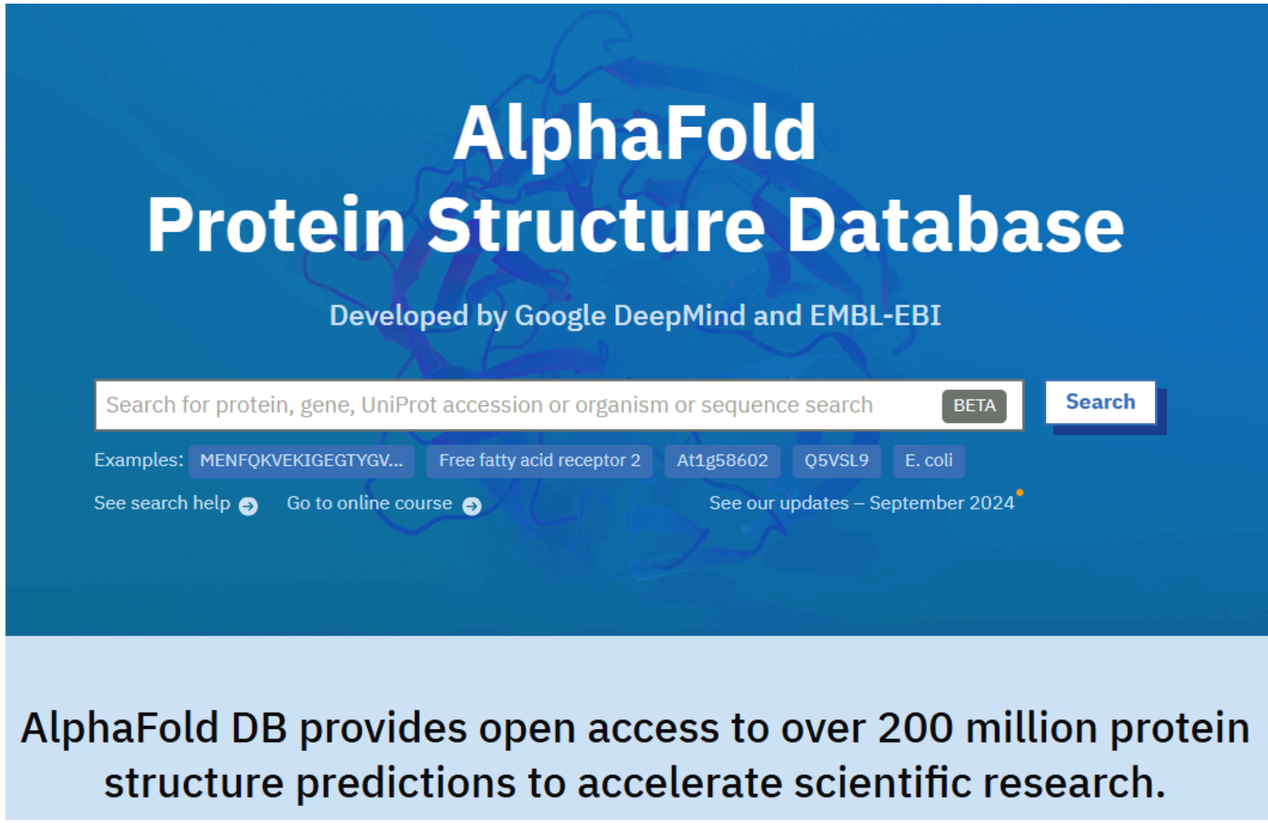
Direções Futuras na pesquisa e educação de Bioinformática

Inteligência Artificial e Integração de Aprendizado de Máquina



Direções Futuras na Educação em Bioinformática
Inteligência Artificial e Integração de Aprendizado de Máquina





AlphaFold DB provides open access to over 200 million protein structure predictions to accelerate scientific research.

São predições, as quais precisam ser rigorosamente avaliadas

Article | Open access | Published: 30 November 2023

AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination

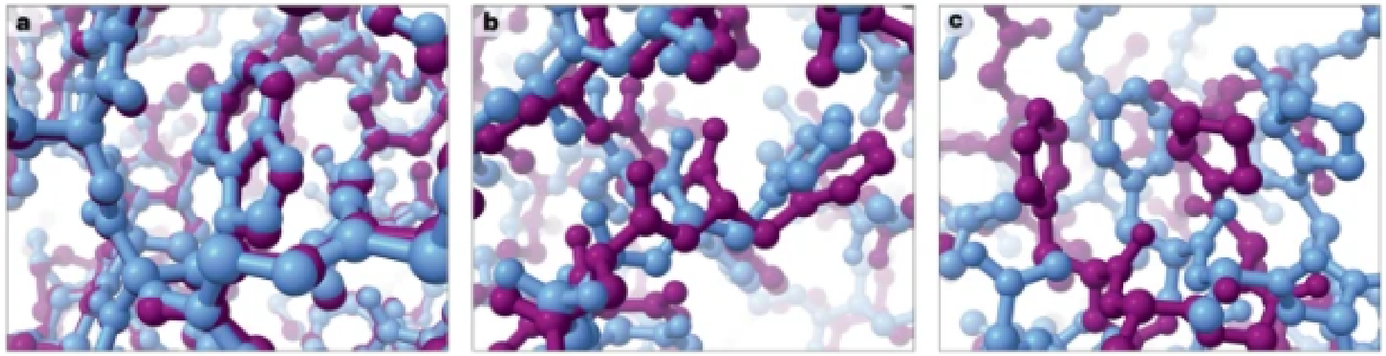
Thomas C. Terwilliger, Dorothee Liebschner, Tristan I. Croll, Christopher J. Williams, Airlie J. McCoy, Billy K. Poon, Pavel V. Afonine, Robert D. Oeffner, Jane S. Richardson, Randy J. Read & Paul D. Adams

Nature Methods **21**, 110–116 (2024) | Cite this article

60k Accesses | **81** Citations | **189** Altmetric | Metrics

A Bioinformática na era das AIs

Fig. 1: Comparison between X-ray crystallography and Alphafold.



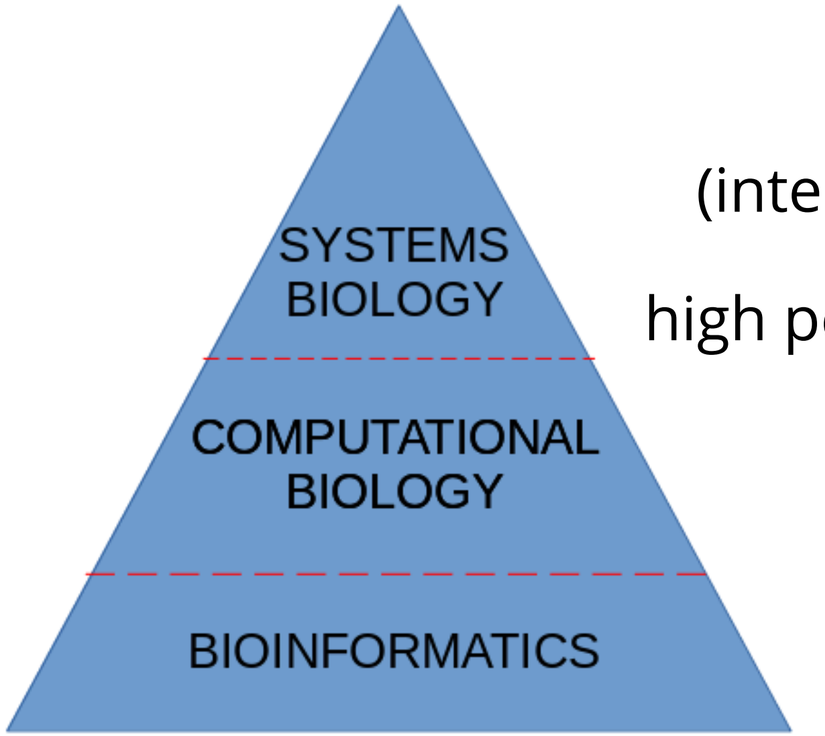
Three examples of the comparison between PDB structures determined by X-ray crystallography (blue) and AlphaFold predicted models of the same protein (purple) showing a region with excellent local correlation in 7WAA (**a**), a region with incorrect prediction in 7SFL (**b**) and a region with displacement and distortion in 7NAZ (**c**). Data from ref. [7].

Conclusão

Importância da inovação contínua no ensino de bioinformática

Próximo passo

Redes e biologia de sistemas



Realidade
virtual
(interativo e colaborativo)

high performance computing
(HPC)

Parke 2013

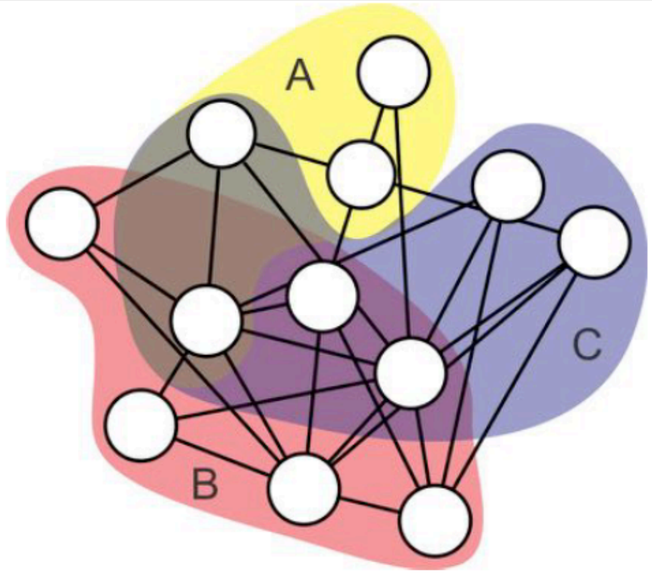


Figura 9-6: Representação de um hipergrafo. As regiões destacadas em várias cores caracterizam as diferentes propriedades ou atividades bioquímicas representadas na rede. Assim, cada cor estaria representando diferentes vias metabólicas (A, B e C). Os nós da rede indicam componentes presentes em cada uma das vias metabólicas e/ou participando de vias distintas nas regiões intersectadas.

Verli et al 2014 Bioinformática