

PCD – Plasmid Computer Decoder

Brenda Abrunhosa, João Henrique, Matheus Cunha, Sarah Santos Silva.

Ilum, Escola de Ciência, Centro Nacional de Pesquisa em Energia e Materiais (CNPem), Brasil

RESUMO

Neste trabalho, desenvolvemos um código capaz de receber uma sequência de DNA plasmidial e retornar para o usuário informações relevantes laboratorialmente para o uso desse plasmídeo, como o número de pares de base, a posição de sítios de clivagem de enzimas de restrição, genes de resistência, temperatura de melting etc. Foi possível criar uma interface gráfica que recebesse os dados em .txt e .FASTA.

Palavras-chave: código; plasmídeo; interface.

1 INTRODUÇÃO

Em práticas laboratoriais de biologia molecular, é comum o uso de vetores plasmidiais de clonagem para a replicação e seleção de fragmentos de DNA de interesse dentro de células bacterianas hospedeiras. Cada plasmídeo utilizado como vetor possui uma origem de replicação (ori), sítios de restrição em regiões de clonagem e genes de seleção, como genes de resistência a antibióticos.

Um sítio de restrição é uma sequência específica de nucleotídeos reconhecida por uma enzima de restrição, que cliva o DNA exatamente nesse local. Alguns desses cortes geram extremidades compatíveis que permitem a remoção de um fragmento do plasmídeo e a subsequente inserção de um fragmento de DNA de interesse.

Com esse método, é possível fazer com que a célula hospedeira replique o DNA inserido, produzindo múltiplas cópias do gene de interesse (como, por exemplo, o gene da insulina), por meio da clonagem do plasmídeo recombinante.

Para garantir que apenas as bactérias que receberam o plasmídeo recombinante sejam selecionadas, utiliza-se um gene de resistência a antibióticos como marcador de seleção. Por exemplo, um gene de resistência à ampicilina confere à bactéria a capacidade de crescer em meio contendo esse antibiótico. Assim, apenas as bactérias que incorporaram o plasmídeo sobreviverão, facilitando a identificação e isolamento das colônias transformadas.

Dessa maneira, é evidente a importância do conhecimento dos sítios de restrição e dos genes de resistência que são ofertados pelo plasmídeo para que o biólogo molecular faça a melhor escolha no mercado.

2 MATERIAIS E MÉTODOS

Com o objetivo de entregar um programa que fosse capaz de receber a sequência genética de um plasmídeo e retornar para o usuário uma interface que retornasse para ele as informações de interesse como enzima de restrição, gene de resistência, temperatura de Melting, tradução e gráficos com a quantidade de bases e de pares de base. Desenvolvemos o projeto seguindo uma série de etapas descritas a seguir.

2.1 Interface gráfica

A interface gráfica do programa foi construída utilizando a biblioteca Tkinter, que faz parte da biblioteca padrão do Python como base. A interface foi projetada para ser simples e intuitiva, permitindo ao usuário inserir dados e executar as funções desejadas utilizando botões, sendo esses elementos distribuídos em uma janela principal com um layout interativo.

2.2 Validação de sequência

Como não se sabe a formatação da sequência inserida pelo usuário, foi utilizado o método `.replace()` para remover símbolos, espaços e quebras de linha inesperadas, e padronizou-se a sequência com `.upper()`. Ainda, caso a sequência inserida apresente uma letra não corresponde às bases nem do RNA nem do DNA ou tanto a base exclusiva do RNA quanto a exclusiva do DNA, a sequência será considerada inválida. Caso as bases do RNA sejam reconhecidas, `.replace()` será novamente utilizado para substituir as bases U por T, permitindo a sequência seja tratada como uma de DNA.

2.3 Contagem de bases

Para realizar a contagem de bases foram criadas algumas funções que pudessem ser usadas posteriormente para os cálculos da Temperatura de Melting de gráficos para o usuário. Esse código foi construído contando principalmente com o método `.count()`, para fazer dos pares de base at e cg foi interessante o cálculo de porcentagem usando a função `len()`.

2.4 Tradução e identificação de ORFs

Para a tradução de alguma eventual proteína situada após o promotor, utilizamos uma sequência RBS (sítio de ligação do ribossomo). Nesse ponto, se houver um códon iniciador (ATG), o código captará a sequência de bases em trincas e as armazenará. Após isso, essas trincas serão comparadas com os valores do dicionário que corresponde ao código genético, e a função retornará à sequência de aminoácidos correspondentes.

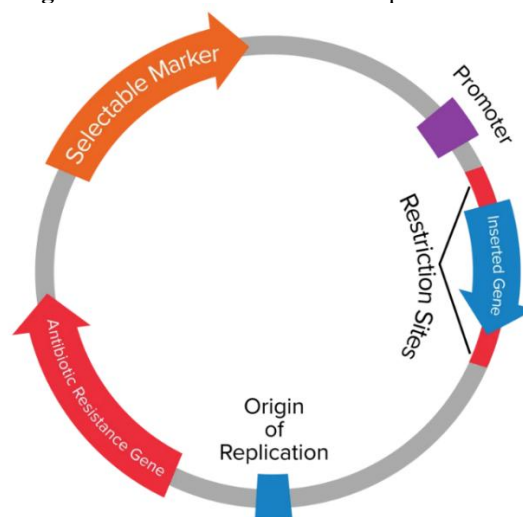
2.5 Enzima de restrição

Algumas sequências do DNA plasmidial são reconhecidas pelas já mencionadas enzimas de restrição. Armazenamos as enzimas de restrição encontradas em plasmídeos em um dicionário contendo o nome da enzima e a sequência correspondente. Com base nisso, a função `enzimas_de_restricao()` procura na sequência inserida as bases de cada enzima. Caso haja correspondência, o método `.count()` é então utilizado para encontrar a frequência da enzima. Esses dados são armazenados em um arquivo `.txt` que abre ao selecionar o botão dessa função.

2.6 Gene de resistência

Os genes de resistência são de interesse por ajudarem na identificação dos antibióticos que seriam úteis, por isso o código foi construído para identificar se tem o gene de resistência na parte do código genético adicionado, retornando à posição de início do gene, seu tamanho e o antibiótico para ser usado nesse caso. O método `.find()` foi crucial para encontrar o gene de resistência dentro do gene recebido, a identificação foi feita a partir da biblioteca construída pelo grupo. Na figura 1 é possível observar como o gene de resistência se encontra no plasmídeo, sendo a seta vermelha.

Figura 1 – Gene de resistência do plasmídeo



Fonte: Monroe, 2020

2.7 Temperatura de Melting

A temperatura de Melting é o ponto de temperatura em que 50% das moléculas de fita dupla de DNA são desnaturadas (rompem as ligações de hidrogênio e tornam-se fitas simples). Isso é importante para fazer PCR com o plasmídeo, por exemplo. Como essa temperatura está ligada à energia de ligação das moléculas, ela é diretamente proporcional ao número de ligações

de hidrogênio que há na dupla-fita. Uma vez que ligações GC possuem 3 ligações de hidrogênio e as ligações AT, duas, um plasmídeo que possui mais pares de base C e G terá uma maior temperatura de melting.

2.8 Gráficos de apoio

Os gráficos são uma forma mais visual de retornar ao usuário as contagens de base e a porcentagem de pares de base, a biblioteca escolhida para realização foi a Matplotlib e os dados pegos das funções de contagem de bases.

2.9 Output do programa

Os resultados obtidos com o programa após a execução de suas funções são disponibilizados ao usuário em dois formatos principais. O primeiro consiste em arquivos de texto, salvos na pasta “arquivos de saída”, facilitando o acesso dos dados que estão disponibilizados de maneira prática. O segundo formato são os gráficos produzidos com a biblioteca Matplotlib, esses gráficos são exibidos na interface e podem ser salvos manualmente pelo usuário, caso ele deseje armazená-los para uso posterior.

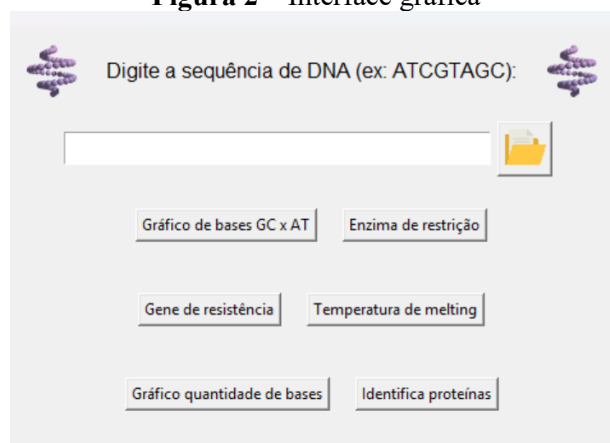
3 RESULTADOS E DISCUSSÃO

Abaixo estão descritos os resultados de cada etapa do código que permitiu a conclusão do trabalho.

3.1 Interface gráfica

Abaixo na figura 2 pode-se observar a interfase gráfica feita conforme planejado na metodologia.

Figura 2 – Interface gráfica



Fonte: Acervo pessoal, 2025.

3.2 Validação de sequência

Caso uma sequência com letras não correspondentes às bases esperadas seja inserida, nesse caso a letra b (figura 3):

Figura 3 – Sequência inválida

Digite a sequência de DNA (ex: ATCGTAGC):

agctagctbacgt

Gráfico de bases GC x AT Enzima de restrição

Gene de resistência Temperatura de melting

Gráfico quantidade de bases Identifica proteínas

Sequência inválida.

Fonte: Acervo pessoal, 2025.

Caso uma sequência com símbolos variados além de letras seja inserida (figuras 4 e 5):

Figura 4 – Símbolos estranhos 1

Digite a sequência de DNA (ex: ATCGTAGC):

a T ; g c

Gráfico de bases GC x AT Enzima de restrição

Gene de resistência Temperatura de melting

Gráfico quantidade de bases Identifica proteínas

Sequência de DNA armazenada:
ATGC

Fonte: Acervo pessoal, 2025.

Figura 5 – Símbolos estranhos 2

Digite a sequência de DNA (ex: ATCGTAGC):

a:.....ccccg,...

Gráfico de bases GC x AT Enzima de restrição

Gene de resistência Temperatura de melting

Gráfico quantidade de bases Identifica proteínas

Sequência de DNA armazenada:
ACCCCG

Fonte: Acervo pessoal, 2025.

Caso uma sequência de RNA seja inserida (figura 6):

Figura 6 – sequência de RNA

Digite a sequência de DNA (ex: ATCGTAGC):

acgu

Gráfico de bases GC x AT Enzima de restrição

Gene de resistência Temperatura de melting

Gráfico quantidade de bases Identifica proteínas

Sequência de RNA identificada e armazenada como sequência de DNA:
ACGT

Fonte: Acervo pessoal, 2025.

Percebe-se que as sequências foram devidamente interpretadas e, se validas, tratadas e armazenadas.

3.3 Tradução e identificação de ORFs

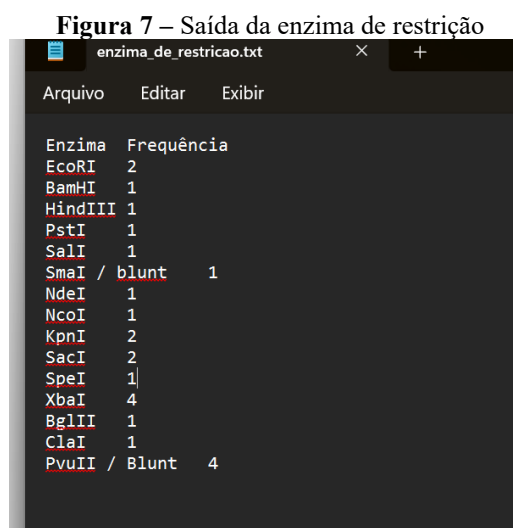
O retorno da tradução é na forma de texto, como no exemplo:

```
{331: ['Metionina', 'Glicina', 'Leucina', 'Leucina', 'Cisteína', 'Ácido aspártico']}
```

Esse é exatamente o retorno esperado, pois mostra a posição (início da sequência) junto com a sequência com o nome dos aminoácidos traduzidos.

3.4 Enzima de restrição

O retorno da enzima de restrição é no formato de arquivo de texto, como pode ser visto na figura 7:



Fonte: Acervo pessoal, 2025.

Esse retorno é o desejado para essa função, visto que as enzimas de restrição são dispostas em um formato de colunas junto com a sua frequência.

3.5 Gene de resistência

O retorno do gene de resistência é na forma de texto, como no exemplo:

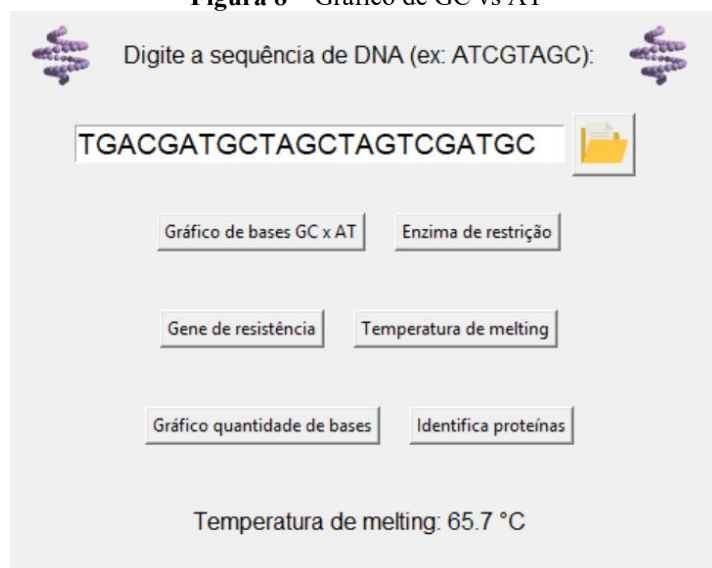
```
[{'gene': 'bla', 'posição': 975, 'tamanho': 424, 'antibiótico': 'Ampicilina'}]
```


Esse é exatamente o retorno desejado, com o nome do gene, a posição dele na sequência (seu início), seu tamanho e há também o antibiótico ao qual ele é resistente.

3.6 Temperatura de Melting

A temperatura de melting é retornada para o usuário na interface com uma casa decimal de precisão (figura 8).

Figura 8 – Gráfico de GC vs AT

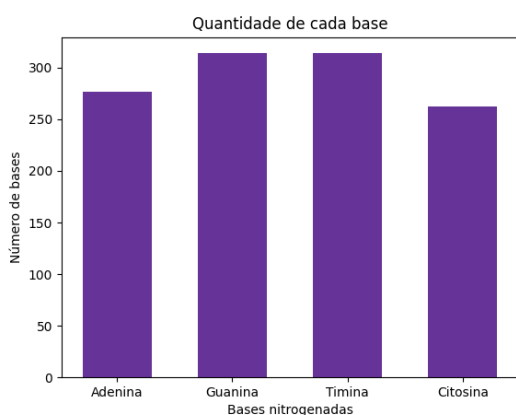


Fonte: Acervo pessoal, 2025.

3.7 Gráficos de apoio

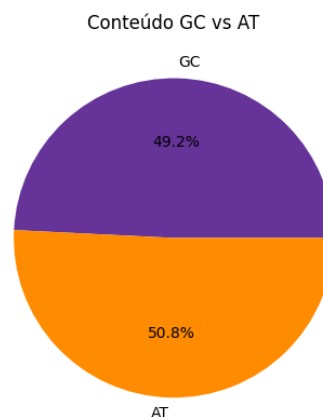
Os gráficos foram construídos conforme planejado na metodologia e apresentaram a seguinte aparência para o usuário conforme as figuras 9 e 10.

Figura 9 – Gráfico de quantidade de cada base



Fonte: Acervo pessoal, 2025.

Figura 10 – Gráfico de GC vs AT






Fonte: Acervo pessoal, 2025.

Usamos argumentos para personalizar nossos gráficos com as cores da Ilum, o que foi interessante para nosso aprendizado de tratamento de dados. Sua funcionalidade também é uma indicação de que as funções de contagem estão funcionando.

3.8 Output do programa

Como mencionado anteriormente, o programa possui dois tipos de saída: arquivos de texto, que são automaticamente salvos na pasta “arquivos de saída” (figura 11), e resultados visuais, como a temperatura de Melting ou gráficos gerados com o Matplotlib. Esses resultados também podem ser visualizados diretamente na interface, permitindo que o usuário salve os gráficos manualmente, caso deseje.

Figura 11 – Arquivos de texto

Nome	Status	Data de modificação	Tipo	Tamanho
 aminoacidos	✓	24/06/2025 17:30	Documento de Texto	1 KB
 enzima_de_restricao	✓	24/06/2025 18:58	Documento de Texto	1 KB
 gene_de_resistencia	✓	24/06/2025 17:29	Documento de Texto	1 KB

Fonte: Acervo pessoal, 2025.

4 CONCLUSÕES

O projeto foi uma experiencia interessante, pois permitiu os integrantes do grupo aproveitarem de seus conhecimentos adquiridos durante toda disciplina de ciência de dados. Unido a isso estudamos e nos preparamos para áreas da biologia que iremos explorar no curso posteriormente, provando a interdisciplinaridade da atividade. Portanto, ficamos felizes em observar que todos os objetivos que estabelecemos foram alcançados com a entrega de um Github completo e informativo, e um programa que recebe uma sequência de plasmídeo e retorna enzima de restrição, gene de resistência, temperatura de Melting, tradução e gráficos para o usuário.

REFERÊNCIAS

ALBERTS, Bruce; *et al.* **Biologia Molecular da Célula**. 6. ed. Porto Alegre: ArtMed, 2017. p. 467-469.

ALBERTS, Bruce; *et al.* **Biologia Molecular da Célula**. 6. ed. Porto Alegre: ArtMed, 2017. p. 483-484.

Monroe, Margo R. “**Plasmids 101: What Is a Plasmid?**” Addgene.org, 2 Apr. 2020, Disponível em: <blog.addgene.org/plasmids-101-what-is-a-plasmid>. Acesso em: 18 jun. 2024

PYTHON SOFTWARE FOUNDATION. **Tkinter** — **Python interface to Tcl/Tk**. Disponível em: <docs.python.org/3/library/tkinter>. Acesso em: 24 jun. 2025.

GRIFFITHS, Mark. **TkDocs**. Disponível em: <tkdocs.com>. Acesso em: 24 jun. 2025.