

Fluxo de Trabalho para Predição de Estrutura de Proteínas

Pedro Henrique do Nascimento Esteves (orientado)¹, Tácio Vinício Amorim Fernandes (co-orientador)², Gregório Kappaun Rocha (orientador)³

¹Universidade Estadual do Norte Fluminense, Campos, RJ, Brasil.

²Instituto Nacional de Metrologia, Qualidade e Tecnologia – INMETRO, Duque de Caxias RJ, Brasil.

³Instituto Federal Fluminense, Macaé, RJ, Brasil.

RESUMO

Analisar modelos, inferir funções, entender o enovelamento e as características físico-químicas de proteínas e enzimas utilizando programas computacionais é uma realidade cada vez mais comum em estudos que investigam o funcionamento de sistemas biológicos. Técnicas *in silico* para predição de estrutura de proteínas representam uma etapa importante neste processo. A predição de estrutura de proteínas é uma área bastante desafiadora que exige conceitos teóricos interdisciplinares e conhecimento de diferentes ferramentas de edição, comparação, modelagem e análise de sequências. O domínio desta diversidade de ferramentas computacionais não é tarefa trivial. Deste modo, o presente trabalho tem como objetivo desenvolver um fluxo de trabalho para a automatização de etapas-chave necessárias para a construção de um modelo proteico a partir de sua sequência genômica. O desenvolvimento de ferramentas mais amigáveis e de rotinas automatizadas são importantes para otimizar o tempo das análises, principalmente devido ao grande volume de dados depositados em bancos biológicos. Além disso, servem para auxiliar no estudo computacional de dados biológicos por pesquisadores ainda pouco familiarizados com as ferramentas de bioinformática. Os *scripts* que compõem o fluxo desenvolvido permitem ao usuário realizar, de maneira automática, tarefas que vão desde a seleção dos alvos moleculares, à filtragem, extração e separação das sequências, alinhamento *online* de sequências, até a construção de modelos proteicos tridimensionais. O genoma do parasita hemoflagelado *Trypanosoma cruzi*, causador da Doença de Chagas, foi selecionado como alvo para avaliação do método em um estudo de caso. Os conjuntos de instruções desenvolvidos neste trabalho estão disponíveis para o uso por qualquer usuário.

Palavras-chave: fluxo de trabalho, predição de estrutura de proteínas, bioinformática, biopython, biologia computacional.

ABSTRACT

Analyzing models, inferring functions, understanding folding and the physical-chemical characteristics of proteins and enzymes using computer programs is an increasingly common reality in studies that investigate the functioning of biological systems. *In silico* techniques for protein structure prediction represent an important step in this process. Protein structure prediction is a very challenging area that requires interdisciplinary theoretical concepts and knowledge of different tools for editing, comparing, modeling and analyzing sequences. Mastering this diversity of computational tools is not a trivial task. In this way, the present work aims to develop a workflow for the automation of key steps necessary for the construction of a protein model based on its genomic sequence. The development of friendlier tools and automated routines are important to optimize the analysis time, mainly due to the large volume of data deposited in biological banks. In addition, they serve to assist in the computational study of biological data by researchers who are still unfamiliar with bioinformatics tools. The scripts that make up the developed flow allow the user to carry out, automatically, tasks ranging from the selection of molecular targets, to filtering, extracting and separating the sequences, online alignment of sequences, to the construction of three-dimensional protein models. The genome of the hemoflagellate parasite *Trypanosoma cruzi*, which causes Chagas disease, was selected as a target for the evaluation of the method in a case study. The instruction sets developed in this work are available for use by any user.

Keywords: workflow, protein structure prediction, bioinformatics, biopython, computational biology.

INTRODUÇÃO

As proteínas são biomoléculas fundamentais para o perfeito funcionamento de um organismo e apresentam-se, inicialmente, em um formato linear primário representado por um polímero de aminoácidos, e apenas desempenha sua função quando adquire a estrutura terciária (conformação nativa), que é formada pelo arranjo global das estruturas secundárias [1]. Logo, prever o arranjo atômico tridimensional (3D) de uma proteína é uma maneira de se obter informações a respeito de sua respectiva função, permitindo a identificação de regiões de interesse, tais como o sítio ativo de uma enzima. Essas informações são fundamentais para quaisquer pesquisas que envolvam o desenvolvimento de fármacos e de vacinas [2].

Devido à alta complexidade envolvida no processo de enovelamento das proteínas, o problema da predição de estrutura de proteínas (PSP) é frequentemente apontado como o mais desafiador na biologia computacional [3]. A PSP busca prever, a partir da sequência primária de aminoácidos, o arranjo 3D dos átomos da estrutura nativa de uma proteína.

Descobrir a estrutura 3D de uma proteína pode ser feito de forma experimental através de técnicas laboratoriais, tais como a cristalografia e difração de raios-X, a difração de nêutrons, a ressonância magnética nuclear, a criomicroscopia eletrônica [4][5] [6] ou com a ajuda de programas de computadores (abordagem conhecida como *in silico*) [7][8].

Devido às limitações das técnicas experimentais, tais como alto custo, tempo elevado, dificuldade de purificação e cristalização de algumas proteínas, aliadas ao

gigantesco e crescente volume de dados biológicos gerados principalmente pelos projetos genomas, observa-se a necessidade de investimento em técnicas *in silico*, que mostram-se fundamentais para que as informações obtidas por estudos genômicos sejam mais bem compreendidas [9].

Algumas dessas ferramentas para PSP, tais como as metodologias aplicadas nos pacotes do Rosetta [10], do AlphaFold [11] e do I-tasser [12] demonstraram bastante sucesso na construção de modelos proteicos com o uso de técnicas modernas de inteligência artificial e aprendizagem de máquina, conforme observado na 13ª edição do CASP (*Critical Assessment of Techniques for Protein Structure Prediction*) (2018) e na 14ª edição CASP (2020). Este evento bienal e mundial avalia através de testes cegos diversos métodos de modelagem de estrutura de proteínas, buscando identificar avanços e gargalos na área, bem como aprimorar a predição de estruturas, tal como a do SARS-CoV-2 com alto grau de similaridade conformacional em tempo recorde [13] [14] [15].

Grande parte destas técnicas de modelagem de estruturas proteicas se baseia em algoritmos e linguagens de programação amplamente difundidas e adaptados especificamente ao problema, tais como algoritmos evolutivos [16], redes neurais profundas [17], método de Monte Carlo [18], entre outras [19]. Verifica-se, ainda, uma grande oferta de rotinas depositadas em bibliotecas para a área da biologia computacional como o BioPerl, Biojava, BioPython, BioRuby e BioSmalltalk [20]. Tais ferramentas permitem a realização de uma grande variedade de análises, possibilitando a realização de tarefas complexas com a ajuda de um computador, tais como: alinhar sequências; converter arquivos entre diferentes formatos; realizar o mapeamento

de coordenadas genéticas; extrair regiões intergênicas de uma sequência; comparar estruturas de proteínas; identificar mutações; entre outras [21]. Alguns métodos automatizados para análise e modelagem de proteínas também são disponibilizados, tais como o Mholline [22] e o Asaprot [23].

Mesmo com muitas ferramentas disponibilizadas para uso *on-line* e/ou via servidor, conectá-las de forma automatizada pode ser um desafio para muitos usuários e um fator limitante para análises de grande volume de dados.

Deste modo, o presente trabalho busca desenvolver um fluxo de trabalho para a automatização de etapas-chave necessárias para o estudo e predição de modelos proteicos a partir de sequências genômicas depositadas no banco de dados GenBank [24]. Busca-se, ainda, apresentar brevemente a usabilidade de algumas ferramentas de bioinformática importantes e ilustrar o uso das mesmas via comandos básicos do BioPython.

Para avaliar a metodologia, um estudo de caso foi realizado com o genoma do parasita hemoflagelado *Trypanosoma cruzi* (*T. cruzi*).

METODOLOGIA

Neste trabalho, o Python foi utilizado como linguagem de programação, pois permite trabalhar rapidamente e integrar sistemas de forma eficaz, com uma sintaxe simples, que pode ser trabalhada de forma estruturada ou orientada a objetos e com farta documentação disponível. Usou-se o BioPython, biblioteca gratuita e específica para biologia computacional, que possui seu código-fonte disponibilizado e compatível com quase todas as licenças do mundo [25].

As ferramentas foram executadas em ambiente *Microsoft Windows 10*, por ser um

sistema operacional popular e comumente utilizado.

Escolha do alvo para o estudo de caso e avaliação da metodologia

Foi selecionado como alvo para avaliação do método, o parasita hemoflagelado *T. cruzi*, causador da Tripanossomíase americana, doença popularmente conhecida como Doença de Chagas.

A Doença de Chagas é uma doença tropical negligenciada (DTN) e a principal causa de insuficiência cardíaca e morte em países endêmicos da América Latina, segundo a Organização Mundial de Saúde (OMS) [24], motivo pelo qual este alvo foi selecionado. O indivíduo que contrai o *T. cruzi* e desenvolve a Doença de Chagas pode sofrer diversos efeitos patológicos, com destaque para os danos nos sistemas cardíaco, nervoso e digestório (Vieira, 2017) [26]. A transmissão do parasita ocorre por diferentes vias, podendo ser através de insetos vetores da família *Triatominae* (conhecidos popularmente como barbeiros), ou por outros mecanismos de transmissão, tais como a transfusão de sangue de doador portador da doença, a transmissão vertical via placenta (de mãe para filho), a ingestão de alimentos contaminados ou acidentalmente em laboratórios (Moreno, 2017) [27]. Todas as etapas apresentadas no fluxo de trabalho proposto serão exemplificadas com este alvo.

As figuras apresentadas ao longo do texto estão disponibilizadas em melhor resolução ao final do artigo.

Configurando o ambiente de trabalho

Para a execução do fluxo de trabalho será necessário configurar o ambiente no computador do usuário. Todo o fluxo poderá ser consultado a partir do fluxograma apresentado na Figura 1.

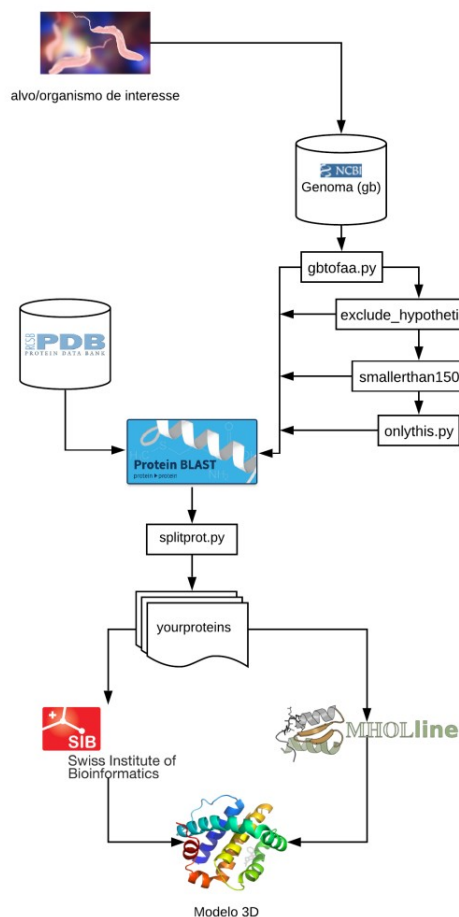


Figura 1| Fluxograma a partir do arquivo *gb* extraído do portal do NCBI.

A preparação do ambiente de trabalho se dá inicialmente pela instalação do *Python* e da biblioteca de ferramentas *Biopython*. Neste estudo, as versões 2.7 e 3.6 do *Python* e 1.78 do *Biopython* foram utilizadas. Os arquivos executáveis necessários à instalação do *Python* poderão ser baixados nos portais *python.org* e do *Biopython*. Basta acessar o diretório de instalação do *Python* pelo terminal do *Windows*, e no diretório *Scripts* executar o comando: *pip install biopython*. Sua instalação é simples, requerendo apenas que os instaladores sejam executados localmente, em suas configurações padrões.

Adicionalmente a este ambiente, pode ser instalada a ferramenta local de alinhamento de sequências BLAST (*Basic Local Alignment*

Search Tool) [28]. O BLAST é uma ferramenta de alinhamento local usada para comparar sequências biológicas. Através do alinhamento é possível encontrar regiões similares entre diferentes sequências, identificar mutações, e até mesmo buscar uma sequência alvo em algum banco de dados. Essa ferramenta, além de diversas funcionalidades, calcula a significância estatística entre os alinhamentos, informação importante para a identificação de estruturas homólogas, que podem ser usadas em métodos de PSP baseados em modelagem comparativa.

Para instalação do BLAST local, basta apenas acessar o portal <https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/> e baixar a versão correspondente ao sistema operacional do seu computador. Neste trabalho foi instalado a versão *ncbi-blast-2.11.0+-win64*. O procedimento é o mesmo das demais ferramentas, bastando apenas baixar o executável e instalar com as configurações padrões.

Também é interessante instalar a ferramenta PyMOL [29]. O PyMOL é um sistema de visualização molecular de código aberto, mantido e distribuído pela Schrödinger. Esta ferramenta permite renderizar e animar as estruturas 3D e visualizá-las a partir dos modelos gerados pela maioria das ferramentas de modelagem disponíveis. Está disponível em pymol.org.

Busca por sequências de interesse e a seleção do arquivo *gb*

O estudo inicia-se com a escolha de um **alvo**, que pode ser um organismo ou uma proteína de interesse.

A seguir, é necessário pesquisar em bancos de dados públicos de genomas, tais como NCBI, EMBL, GISAID as sequências de nucleotídeos

do organismo ou proteína (ou, usar dados laboratoriais próprios).

As informações necessárias podem ser obtidas diretamente do Genbank (Figura 2), que estão organizadas em um arquivo de texto em formato “*gb*”. Este arquivo já encontra-se organizado a partir de um relatório de montagem e anotação do genoma estudado, e armazena, dentre outros dados, a sequência primária de aminoácidos, a identificação e a função da proteína. Esta etapa pode ser executada manualmente e os passos para a seleção do arquivo *gb* serão descritos a seguir.

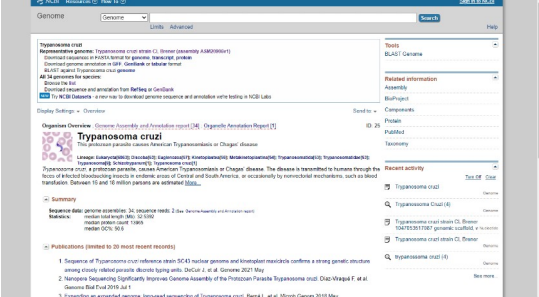


Figura 2 | Portal NCBI. Banco de dados GenBank exibindo informações genômicas sobre o protozoário de interesse de estudo *Trypanosoma cruzi*

Ao pesquisar o genoma do organismo de interesse no GenBank, na opção de “*Genome Assembly and Annotation report*”, é possível listar os projetos submetidos ao portal e os respectivos mapeamentos genômicos e proteômicos da espécie e determinar qual a ser avaliado. Ao clicar no nome do organismo, uma página dará acesso a todas as informações anotadas deste (Figura 3).

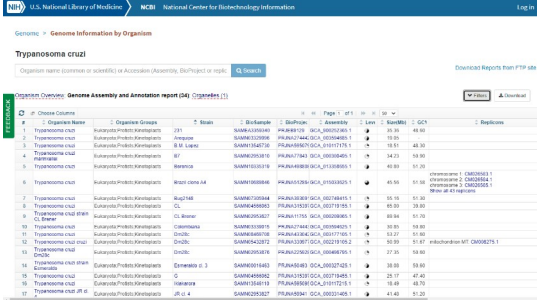


Figura 3 | Portal NCBI Página de exibição de projetos genômicos submetidos por organismo/cepa.

Conforme visto na Figura 3, a página lista os mapeamentos submetidos da espécie, bem como cepa, projetos, tamanho, dentre outras informações disponibilizadas por colunas editáveis. Ao se clicar em *Choose COLUMNS*, demais informações podem ser adicionadas.

Após selecionar o mapeamento desejado, outra página é exibida (Figura 4), contendo uma extensa gama de informações e relatórios e o *link* de acesso ao arquivo *gb* acessado ao clicar em “Go to nucleotide: GenBank” que o redirecionará para a página seguinte (Figura 5).



Figura 4 | Portal NCBI Página de exibição do projeto genômico PRJNA15540 *Trypanosoma cruzi* strain CL Brener 1047053517087 genomic scaffold, whole genome shotgun sequence.

O usuário poderá refinar ainda mais sua busca, filtrando as proteínas pelo seu tamanho. Neste estudo será aplicado um filtro para proteínas com menos de 150 resíduos. Dependendo do estudo, o tamanho das sequências poderá indicar o melhor método a ser usado para a construção dos modelos e, deste modo, interferir na qualidade dos resultados. Para remover sequências com mais de 150 resíduos, basta o usuário executar o *script smallerthan150.py* no mesmo diretório do arquivo *Fasta*, pelo *prompt* de comando, com a seguinte sintaxe:

```
python smallerthan150.py sequencia.fasta  
saida.fasta
```

O conteúdo deste arquivo de saída poderá ser consultado no apêndice.

Apesar deste *script* filtrar apenas proteínas com mais de 150 resíduos, ele pode facilmente ser alterado para qualquer quantidade, bastando apenas editá-lo em qualquer editor de texto, alterando-se o valor contido na linha: *"if((len(line.seq)<150))"*.

Em alguns casos, também poderá ser necessário separar as sequências proteicas por uma determinada classificação. O arquivo *Fasta* pode conter, além dos resíduos, uma breve descrição da proteína, e esta pode ser uma informação determinante, que poderá ser filtrada por este *script*, como por exemplo, filtrar apenas proteínas *heat shock*² e chaperonas². Então, basta o usuário, com o *script onlythis.py* no mesmo diretório do arquivo *Fasta*, executar no *prompt* de comando a seguinte sintaxe:

```
python onlythis.py sequencenohyp.fasta "heat  
shock" >> saida.fasta && python onlythis.py  
sequencenohyp.fasta "chaperone" >>  
saida.fasta
```

O arquivo de saída terá o nome de *saida.fasta*. Adicionalmente, os arquivos de saída mencionados nesta seção poderão ter seu nome alterado na própria linha de comando, de acordo com o critério do usuário. Ao invés de *saida.fasta*, poderá ser, por exemplo, *enoveladoras.fasta* (Apêndice iii).

(iii) Alinhamento local das sequências com o banco de dados de estruturas de proteínas PDB

Para alinhar e comparar uma sequência específica de interesse filtrada na etapa anterior podemos recorrer ao banco de dados de sequências que faz parte do *National Center for Biotechnology Information, U.S. National Library of Medicine - NCBI/NLM* [30] e do banco de dados experimentais *Protein Data Bank – PDB* [31]. Para compararmos as sequências de proteínas que temos como alvo com as depositadas no PDB (que já foram determinadas experimentalmente), o usuário deverá executar o *script blastp.py* com a seguinte sintaxe:

```
blastp.py sequence.fasta
```

Este algoritmo alinhará a sequência escolhida ao banco de dados *online* do PDB. O arquivo de sequência deverá conter apenas uma única sequência - ideal para análise individual ou validação do alinhamento.

Esta etapa pode ser útil para diversos objetivos, como por exemplo, verificar a existência de estruturas tridimensionais já determinadas, evitando assim, a construção de modelos de forma desnecessária (que será descrito em detalhes a seguir); ou ainda, para buscar por estruturas evolutivamente aparentadas que podem servir de arcabouço para métodos de predição *baseados em molde*.

2 proteínas que garantem o correto enovelamento de outras proteínas

(iv) Verificação de modelo preexistente

Antes de iniciar a modelagem, recomenda-se que o usuário verifique se o alvo molecular já não foi determinado experimentalmente. Para tanto, serão usados os dados depositados no banco de dados PDB. O PDB contém uma ampla base de dados de estruturas tridimensionais já determinadas experimentalmente. Estão disponíveis informações sobre a técnica utilizada para determinação, sequência de aminoácidos, forma unidimensional e respectiva sequência primária e secundária, representação gráfica, métrica de qualidade da estrutura, etc.

Neste estudo, como o interesse foi modelar uma estrutura ainda não existente, efetuou-se uma busca comparativa neste portal.

Utilizou-se a comparação da estrutura primária extraída do *Genbank* contra o banco *PDB* a partir do alinhamento local *Blastp* (recurso que permite o alinhamento a partir de sequências de aminoácidos em vez de bases nitrogenadas, tal como em um genoma). Para este fim, é necessário baixar as proteínas do organismo de interesse também no portal do *PDB* e, em seguida, realizar a comparação.

Para baixar as proteínas do organismo basta acessar o portal *rcsb.org* (Figura 7) e digitar na busca o nome do organismo, e filtrar pela cepa, se for o caso. A página exibe 25 modelos por vez. Aqui, o que interessa são as *IDs*.

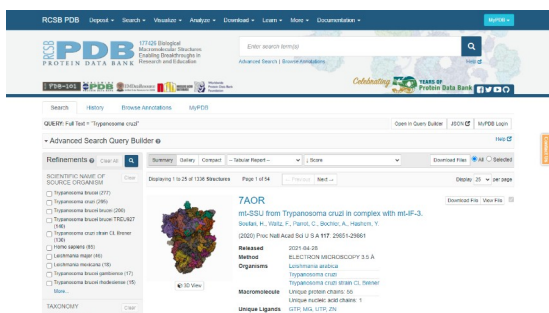


Figura 7 | Portal RCSB PDB. Banco de estruturas moleculares exibindo proteínas modeladas do protozoário de interesse *Trypanosoma cruzi*.

Na guia “*Tabular report*” (Figura 7), selecione “*PDB IDs*”. Serão exibidos apenas os identificadores das proteínas. Esses códigos, compostos por 4 números e letras, serão importantes para se baixar corretamente o arquivo *Fasta* destas proteínas. Basta clicar em “*Download IDs*” e o portal gerará um arquivo txt de nome similar à “*rcsb_pdb_ids_20210330130002*” contendo os *IDs*. Abra o arquivo e copie todo o seu conteúdo.

Agora será necessário, a partir dos *IDs* das proteínas, baixar as sequências correspondentes. Para tanto, deve-se acessar a página do PDB na seção de *downloads* de arquivos *Fasta*, disponível em <https://www.rcsb.org/downloads/fasta> e cole os *IDs*, selecione “*Single FASTA file*” e “*Launch Download*” para gerar o arquivo *Fasta* com as sequências que serão usadas para comparar com as sequências das proteínas do organismo baixado no endereço eletrônico do NCBI.

Uma das técnicas utilizadas neste trabalho para se verificar a existência de modelos preexistentes foi o alinhamento local das sequências utilizando-se o “*protein BLAST*”. Para executar um alinhamento “*protein BLAST*” entre as sequências de resíduos de aminoácidos do organismo e as sequências das proteínas depositadas no PDB, basta executar o comando abaixo no *prompt*, conforme a figura 8.

```
blastp -query C:\Python27\sequencenohyp.fasta -subject C:\Python27\rcsb_pdb_20200811180304.fasta -out PDBxNCBI
```



```

C:\Windows\system32\cmd.exe
Microsoft Windows [versão 10.0.19042.928]
(c) Microsoft Corporation. Todos os direitos reservados.
C:\Users\pedrocd> C:\Python27
C:\Python27>python exclude_hypothetical.py sequence.gb product > sequencenohyp.fasta
C:\Python27>cd C:\NCBI\blast\bin
C:\NCBI\blast\bin>blastp -query C:\Python27\sequencenohyp.fasta -subject C:\Python27\ncsp_pdb_20200811100104.fasta -out PDBxNCBI

```

Figura 8|Alinhamento local. Comandos executados no *prompt* de comando do Windows 10, onde o arquivo *gb*, foi previamente copiado para o diretório local de instalação do Python (padrão é *C:\Python27*) e o seu conteúdo filtrado, retirando-se as proteínas hipotéticas e copiando as sequências da *feature product* para o arquivo *sequencenohyp.fasta*. Em seguida seu conteúdo foi comparado com as proteínas depositadas no *PDB* (também previamente copiados para o diretório local do Python). O arquivo de saída contendo o resultado do alinhamento local, é denominado *PDBxNCBI*.

O objetivo desses *scripts* é identificar as proteínas que já possuem estrutura depositada no *PDB*. Este arquivo de saída apresentará um amplo e valioso relatório, contendo os resíduos das proteínas alinhadas e comparadas entre o arquivo baixado no *NCBI*, o tratado anteriormente (*query*) e os modelos depositados no *NCBI* (*subject*), bem como índices de identidade, similaridade e espaços entre os resíduos alinhados a partir de cálculos estatísticos de significância de correspondências, a fim de inferir relações funcionais e evolutivas e identificar membros de famílias de genes a partir de um *score*. O referido arquivo de saída *PDBxNCBI* pode ser consultado no apêndice.

Alinhamento local vs. *WWWBlast*

Alternativamente ao alinhamento local, o usuário poderá realizar o *Blast* pela internet, utilizando-se de uma biblioteca disponibilizada pelo projeto Biopython e cujo algoritmo utilizado neste trabalho estará disponível no mesmo endereço citado anteriormente. (<http://bityli.com/pspwkf>).

Com o mesmo arquivo de entrada utilizado no *Blast* local, no seu devido formato *Fasta*, execute o script “*splitandblastp.py*” pela sintaxe:

```
python splitandblastp.py sequencenohyp.fasta
> PDBxNCBI
```

Este procedimento é muito demorado, requer conexão com internet estável e pode levar de várias horas a alguns dias.

Apesar de demorado, este procedimento dispensa a instalação prévia do *Blast* e o *download* do arquivo *Fasta* das moléculas depositadas no *PDB*. O arquivo de saída será salvo no mesmo diretório, como no exemplo do *Blast* local.

O alinhamento local pode ser mais rápido, entretanto, dependerá da base de dados a ser utilizada (neste exemplo foram utilizadas como base as estruturas existentes no *PDB* para o organismo utilizado neste trabalho).

Para alinhamento de todas as sequências deste trabalho, foram utilizadas as ferramentas de alinhamento múltiplo citadas nesta seção, porém há outras ferramentas com mesma função, tal como *T-cooffee* [32] ou *Clustal Omega*[33], que possuem diferentes abordagens e parâmetros, a critério do usuário. O *Clustal Omega*, por exemplo, possui recursos como alinhamento local com interface gráfica e o *T-coffee* permite o alinhamento direto em seu portal, disponível em <http://tcoffee.crg.cat/>.

Após o alinhamento, optamos por descartar as sequências com 80% ou mais de identidade (identificado no arquivo de saída *PDBxNCBI* como “*Identities*”). Sugere-se que estas sequências podem ser consideradas proteínas que já possuem modelos determinados. Também optamos por descartar as sequências com identidade abaixo de 30%, uma vez que esse é um limite usado para a detecção de proteínas homólogas em muitos estudos [34]. Entretanto, este raio de corte pode ser definido a critério do usuário. Esta informação fica disponível no arquivo de saída citado, conforme ilustrado na Figura 9.

em 3D e anotações funcionais, dentre elas o *Blast*, já utilizado neste fluxo de trabalho e o *Modeller* – este sendo uma reconhecida ferramenta de modelagem comparativa que produz modelos 3D de estruturas de proteínas [37].

Para submeter a proteína de interesse, basta clicar em “*New Job Submission*” e realizar o *upload* dos arquivos, de acordo com as proteínas que se deseja modelar, geradas pela seção (i) do capítulo de modelagem deste artigo. Será necessário informar um endereço eletrônico para que o portal envie os resultados.

RESULTADOS

Escolhemos a proteína *putative³ heat shock protein DNAJ* como alvo molecular para avaliar este fluxo de trabalho em um estudo de caso.

A escolha para este alvo é em virtude de sua importância para a compreensão e por estarem ligadas a doenças cardíacas. Estas proteínas da família de proteínas de estresse de choque térmico (HSPs), possuem funções de reparo ou degradação de outras proteínas danificadas e atuando como chaperonas moleculares, e estão envolvidas em patologias tais como hipertrofia cardíaca, lesão da parede vascular, cirurgia cardíaca, pré-condicionamento isquêmico, envelhecimento e, possivelmente, mutações em genes que codificam proteínas contráteis e canais iônicos [38].

No relatório de alinhamento, demonstrou-se que suas sequências produziram 21 alinhamentos significativos, com 45% de similaridade com estas proteínas depositadas no PDB, conforme mencionado na seção (iv) - Verificação de modelo preexistente e conforme ilustrado na Figura 9.

3 proteína cuja estrutura foi predita porém não tem evidências experimentais de sua expressão

Foram, então, gerados modelos tanto no *Swiss-Model* (Figura 11), quanto no *MHOLline* (Figura 12) a partir do arquivo de saída *Fasta* da respectiva proteína alvo, conforme mencionado na seção (i) -Separando as proteínas.



Figura 11|Página de resultados do Swiss-Model. Resultados do modelo, estimativas de qualidades local e global, o molde utilizado, o alinhamento, dentre outros dados.



Figura 12|Representação do modelo gerado. Arquivo PDB gerado pelo Modeller a partir do MHOLline e visualizado a partir do aplicativo PyMol.

Como visto, os modelos gerados por ambas as ferramentas tiveram aspectos diferenciados. Apesar de ambas as ferramentas utilizarem técnicas de modelagem a partir de moldes de estruturas homólogas e na modelagem por homologia, a qualidade de um modelo pode ser determinada pela distância evolutiva entre a proteína-alvo e as estruturas de modelo disponíveis, a qualidade destes modelos podem ser avaliados, por exemplo, por uma função de *scoring*, tal como a QMEAN – uma função que avalia padrões estatísticos comparando-se as características geométricas do modelo com estruturas determinadas experimentalmente [39] ou métodos de Estimativa de Precisão de Modelo (EMA), que se utiliza de aprendizagem de máquina intuitiva (*deep learning*) [40]. Alternativamente, o usuário poderá utilizar de ferramentas mais simples e comuns para validação dos modelos, tais como validações baseadas por métodos que analisam a distribuição dos ângulos diedros da cadeia principal em um gráfico conhecido por gráfico de Ramachandran [41] ou sobreposição de modelos para o cálculo de raiz quadrada do desvio médio (RMSD) a partir das diferenças de localização dos átomos Ca e da cadeia principal a fim de medir-se desvios estruturais [42].

Apesar de muitos passos para se chegar ao modelo predito, o tempo dedicado para a execução do mesmo foram de 6 semanas, tendo em vista as diversas plataformas e recursos disponíveis, bem como da farta literatura produzida e disponível. A maior parte deste tempo foi dedicado à leitura da documentação das ferramentas.

CONCLUSÃO

Desenvolver ferramentas que possam atender a uma grande gama de objetivos e a diferentes perfis de usuários é um grande

desafio. Um sistema muito automatizado pode não permitir que se obtenha resultados desejáveis, ou que não atenda a um trabalho específico, ao tempo em que um fluxo de trabalho muito flexível acaba por não prover a automatização necessária para que o trabalho do usuário seja otimizado. Este trabalho procurou abordar, no âmbito da biologia computacional, as ferramentas mais utilizadas na manipulação de sequências biológicas e desenvolver um guia e um fluxo de trabalho que visam facilitar o uso das mesmas em investigações de problemas biológicos. Com isso, pretende-se que usuários com pouca experiência na área possam se familiarizar com estas metodologias, bem como reduzir o tempo gasto na aprendizagem da programação e na condução de suas pesquisas.

A atualização das rotinas apresentadas é necessária, tanto para manter as ferramentas embarcadas atualizadas, como para acompanhar a dinâmica da área, tanto em conhecimento biológico quanto computacional.

Com o fluxo apresentado, o usuário é capaz de realizar de maneira automática, tarefas que vão desde a seleção dos alvos moleculares, a filtragem, extração e separação das sequências, alinhamento online com uma única linha de comando até a construção de modelos proteicos.

Como perspectiva, tem-se o desenvolvimento de um aplicativo modular que reunirá todos esses recursos para processamento local, de forma transparente para o usuário e com a possibilidade de integração com ferramentas existentes e futuras. Esta plataforma pretende contar com interface amigável para que o usuário possa submeter sequências, selecionar os filtros e gerar os resultados de forma ainda mais parcimoniosa, intuitiva e customizada,

sem a necessidade de preparação de ambiente.

REFERÊNCIAS

[1] D. Baker and A. Sali, "Protein structure prediction and structural genomics," *Science*, vol. 294, no. 5540, pp. 93–96, 2001.

[2] Ekins S, Mestres J, Testa B: In silico pharmacology for drug discovery: applications to targets and beyond. *Br J Pharmacol* 2007, 152:21-37.

[3] M. Judy, K. Ravichandran, and K. Murugesan, "A multi-objective evolutionary algorithm for protein structure prediction with immune operators," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 12, no. 4, pp. 407–413, 2009.

[4] AMBROSIO, André Luis Berteli; FRANCHINI, Kleber Gomes. *Cristalografia macromolecular: a biologia sob a ótica dos raios X*. Cienc. Cult., São Paulo, v. 69,n. 3, p. 29-36, 2017.

[5] Franks, N. P., & Lieb, W. R. (1979). The structure of lipid bilayers and the effects of general anaesthetics. *Journal of Molecular Biology*, 133(4), 469–500.

[6] Wang, H. Cryo-electron microscopy for structural biology: current status and future perspectives. *Sci. China Life Sci.* 58, 750–756 (2015)

[7] Brünger, A., & Nilges, M. (1993). Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR spectroscopy. *Quarterly Reviews of Biophysics*, 26(1), 49-125.

[8] Al-Lazikani, B., Jung, J., Xiang, Z., & Honig, B. (2001). Protein structure prediction.

Current Opinion in Chemical Biology, 5(1), 51–56.

[9] Pandey, Gaurav; Kumar, Vipin; Steinbach, Michael. (2006). *Computational Approaches for Protein Function Prediction: A Survey*. Retrieved from the University of Minnesota Digital Conservancy.

[10] Carol A. Rohl, Charlie E.M. Strauss, Kira M.S. Misura, David Baker, "Protein Structure Prediction Using Rosetta", *Methods in Enzymology*, Academic Press, Volume 383, pp. 66-93, 2004.

[11] Mohammed AlQuraishi, AlphaFold at CASP13, *Bioinformatics*, Volume 35, Issue 22, 15, pp. 4862–4865, 2019.

[12] Yang J., and Zhang Y., 2015. Protein structure and function prediction using I-TASSER. *Curr. Protoc. Bioinform.* 52:5.8.1-5.8.15.

[13] Senior, A.W., Evans, R., Jumper, J. et al. Improved protein structure prediction using potentials from deep learning. *Nature* 577, 706–710 (2020).

[14] Kryzhtafovych, Andriy & Schwede, Torsten & Topf, Maya & Fidelis, Krzysztof & Moult, John. (2019). Critical assessment of methods of protein structure prediction (CASP)-Round XIII. *Proteins: Structure, Function, and Bioinformatics*. 87. 10.1002/prot.25823.

[15] Callaway E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature*. 2020.

[16] Inserting Co-evolution Information from Contact Maps into a Multiobjective Genetic Algorithm for Protein Structure Prediction Gregorio K. Rocha, Karina B. dos Santos, Jaqueline S. Angelo, Fabio L. Custodio, Helio J. C. Barbosa, Laurent E. Dardenne, Laboratório Nacional de Computação Científica LNCC/MCTIC, Petrópolis - RJ, Brazil,

Universidade Federal de Juiz de Fora, Juiz de Fora, MG, Brazil

Volume 33, Issue suppl_1, 1 January 2005, pp. D34–D38.

[17] Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., ... Hassabis, D. (2019). Protein structure prediction using multiple deep neural networks in CASP13. *Proteins: Structure, Function, and Bioinformatics*.

[18] Nicholas Metropolis & S. Ulam (1949) The Monte Carlo Method, *Journal of the American Statistical Association*, 44:247, 335-341.

[19] Zhang, Y. (2008). Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, 18(3), 342–348.

[20] Moitra, Dipanjan. (2015). Performance Evaluation of BioPerl, Biojava, BioPython, BioRuby and BioSmalltalk for Executing Bioinformatics Tasks. *INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING*. 3. 157-164.

[21] B. Webb and A. Sali, "Protein structure modeling with modeller," *Protein Structure Prediction*, pp. 1–15, 2014.

[22] Rossi, Artur Duque, et al. "MHOLline 2.0: Workflow for automatic large-scale modeling and analysis of proteins". *Revista Mundi Engenharia, Tecnologia e Gestão* (ISSN: 2525-4782), vol. 5, n 6, agosto de 2020.

[23] DA SILVA, M.L.; BULGARELLI, C.; BISCH, P.M. ASAPROT (Automatic Structural Annotation of Proteins) *Revista da Propriedade Industrial* Seção I, Nº 2392. Instituto Nacional de Propriedade Industrial. Processo BR 51 2016 001404-0, 08 de Novembro de 2016 (A).

[24] Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, David L. Wheeler, GenBank, *Nucleic Acids Research*,

[24] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, Michiel J. L. de Hoon, Biopython: freely available Python tools for computational molecular biology and bioinformatics, *Bioinformatics*, Volume 25, Issue 11, 1 June 2009, pp. 1422–1423.

[25] World Health Organization, organizador. Investing to overcome the global impact of neglected tropical diseases: Third WHO report on neglected tropical diseases. World Health Organization, 2015.

[26] Joseli Lannes-Vieira, Portal da Doença de Chagas, Propostas para explicar a fisiopatogenia da doença de Chagas, Laboratório de Biologia das Interações, Instituto Oswaldo Cruz/Fiocruz. Acesso em 30 de março de 2021, disponível em: <http://chagas.fiocruz.br/patogenia/>

[27] Alejandro M. Hasslocher Moreno, Portal da Doença de Chagas, Mecanismos de transmissão da doença de Chagas, Laboratório de Biologia das Interações, Instituto Oswaldo Cruz/Fiocruz. Acesso em 30 de março de 2021, disponível em: <http://chagas.fiocruz.br/transmissao/>

[28] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids*.

[29] PyMOL, The PyMOL Molecular Graphics System, Version 2.4, Schrödinger, LLC.

[30] Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res*. 2013

- Jan;41(Database issue):D36-42. doi: 10.1093/nar/gks1195. Epub 2012 Nov 27. PMID: 23193287; PMCID: PMC3531190.
- [31] P. W. Rose, A. Prlic, C. Bi, W. F. Bluhm, C. H. Christie, S. Dutta, R. K. Green, D. S. Goodsell, J. D. Westbrook, J. Woo et al., "The rcsb protein data bank: views of structural biology for basic and applied research and education," *Nucleic acids research*, vol. 43, no. D1.
- [32] T-Coffee: A novel method for multiple sequence alignments Notredame, Higgins, Heringa, *JMB*, 302 (205-217).
- [33] Sievers F., Wilm A., Dineen D., Gibson T.J., Karplus K., Li W., Lopez R., McWilliam H., Remmert M., Söding J., Thompson J.D. and Higgins D.G. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.
- [34] Joshi, T., Xu, D. Quantitative assessment of relationship between sequence similarity and function similarity. *BMC Genomics* 8, 222 (2007).
- [35] SWISS-MODEL Workspace/ GMQE Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296-W303 (2018).
- [36] Schwede, T. (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Research*, 31(13), 3381–3385.
- [37] "Structural Modelling and Comparative Analysis of Homologous, Analogous and Specific Proteins from *Trypanosoma cruzi* versus *Homo sapiens*: Putative Drug Targets for Chagas' Disease Treatment". Priscila VSZ Capriles, Ana CR Guimaraes, Thomas D Otto, Antonio B Miranda, Laurent E Dardenne and Wim M Degrave. *BMC Genomics* 2010.
- [38] Benjamin, I. J., & McMillan, D. R. (1998). Stress (Heat Shock) Proteins: Molecular Chaperones in Cardiovascular Biology and Disease. *Circulation Research*, 83(2), 117–132.
- [39] Benkert, P., Biasini, M., Schwede, T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* 27, 343-350 (2011).
- [40] E. C. Lima, F. L. Custódio, G. K. Rocha and L. E. Dardenne, "Estimating Protein Structure Prediction Models Quality Using Convolutional Neural Networks," 2018 International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1-6.
- [41] RAMACHANDRAN GN, RAMAKRISHNAN C, SASISEKHARAN V Stereochemistry of polypeptide chain configurations. *J Mol Biol* 7:95–99, (1963).
- [42] Kufareva, I., & Abagyan, R. (2011). Methods of Protein Structure Comparison. *Homology Modeling*, 231–257.

Apêndice

A seguir, estão ilustrados os arquivos de saída dos *scripts*, bem como o arquivo *gb* extraído do banco de dados *Genbank* disponível no portal do NCBI e o arquivo de saída do alinhamento realizado neste trabalho.

(i) sequence.gb

```
LOCUS       NM_001849447               990720 bp    DNA     linear     CON 28-NOV-2009
DEFINITION   Trypanosoma cruzi strain CL Brener 1047053517087 genomic scaffold,
              whole genome shotgun sequence.
ACCESSION    NM_001849447
VERSION      NM_001849447.1
DBLINK       Project: 15540
BioProject:  PRJNA15540
KEYWORDS     WGS; RefSeq.
SOURCE       Trypanosoma cruzi strain CL Brener
ORGANISM     Trypanosoma cruzi strain CL Brener
Eukaryota; Euzoenozoa; Kinetoplastida; Trypanosomatidae;
Trypanosoma; Schizotrypanum.
REFERENCE    1 (bases 1 to 990720)
AUTHORS      El-Sayed, H.M., Myler, P.J., Bartholomeu, D.C., Nilsson, D.,
              Aggarwal, G., Tran, A.N., Ghedin, E., Worthey, E.A., Delcher, A.L.,
              Blandin, G., Westenberger, S.J., Caler, E., Cerqueira, G.C.,
              Branche, C., Haas, B., Anupama, A., Arner, E., Aslund, L., Attipoe, P.,
              Bontempi, E., Brington, F., Burton, P., Cadag, E., Campbell, D.A.,
              Carrington, M., Crabtree, J., Darban, H., da Silveira, J.F., de
              Jong, P., Edwards, K., Englund, P.T., Fazelina, G., Feldblyum, T.,
              Ferella, M., Frasch, A.C., Gull, K., Horn, D., Hou, L., Huang, Y.,
              Kindlund, E., Klingbeil, M., Kluge, S., Koo, W., Lacerda, D.,
              Levin, M.J., Lorenzi, H., Louie, T., Machado, C.R., McCulloch, R.,
              McKenna, A., Mizuno, Y., Mottram, J.C., Nelson, S., Ochaya, S.,
              Osogawa, R., Pal, G., Parsons, M., Pentony, M., Petersson, D., Pop, M.,
              Ramirez, J.L., Rinta, J., Robertson, L., Salzberg, S.L., Sanchez, D.O.,
              Seyler, A., Sharma, R., Shetty, J., Simpson, A.J., Sisk, E., Tammi, M.T.,
              Taretton, K., Teixeira, S., Van Aken, S., Vogt, C., Ward, P.N.,
              Wickstead, B., Wortman, J., White, O., Fraser, C.M., Stuart, K.D. and
              Anderson, B.
TITLE        The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas
              disease
JOURNAL      Science 309 (5733), 409-415 (2005)
PUBMED       16020725
REFERENCE    2 (bases 1 to 990720)
AUTHORS      El-Sayed, H.M., Myler, P.J., Blandin, G., Berriman, M., Crabtree, J.,
              Aggarwal, G., Caler, E., Renaud, H., Worthey, E.A., Hertz-Fowler, C.,
              Ghedin, E., Peacock, C., Bartholomeu, D.C., Haas, B.J., Tran, A.N.,
              Wortman, J.R., Alamek, D.C., Angioli, S., Anupama, A., Badger, J.,
              Brington, F., Cadag, E., Carlson, J.M., Cerqueira, G.C., Cressy, T.,
              Delcher, A.L., Djikeng, A., Embley, T.M., Hauser, C., Ivens, A.C.,
              Kummerfeld, S.K., Pereira-Leal, J.B., Nilsson, D., Peterson, J.,
              Salzberg, S.L., Shalom, J., Silva, J.C., Sundaram, J.,
              Westenberger, S., White, O., Melville, S.E., Donelson, J.E.,
              Anderson, B., Stuart, K.D. and Hall, N.
TITLE        Comparative genomics of trypanosomatid parasitic protozoa
              science 309 (5733), 404-409 (2005)
JOURNAL      Science 309 (5733), 404-409 (2005)
PUBMED       16020724
REFERENCE    3 (bases 1 to 990720)
AUTHORS      NCBI Genome Project
TITLE        Direct Submission
JOURNAL      Submitted (02-JAN-2008) National Center for Biotechnology
              Information, NIH, Bethesda, MD 20894, USA
REFERENCE    4 (bases 1 to 990720)
AUTHORS      El-Sayed, H., Bartholomeu, D. and Haas, B.
TITLE        Direct Submission
JOURNAL      Submitted (06-JUL-2005) The Institute for Genomic Research, 9712
              Medical Center Dr, Rockville, MD 20850, USA
```

```
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
              NCBI review. The reference sequence was derived from CH473309.
FEATURES             source          1..990720
                     /organism="Trypanosoma cruzi strain CL Brener"
                     /mol_type="genomic DNA"
                     /strain="CL Brener"
                     /db_xref="taxon:353153"
                     misc_feature    1..39326
                                     /note="heterozygous, non-Esmeraldo-like haplotype"
                     gene            complement(<1..>769)
                                     /locus_tag="Tc00.1047053511261.4"
                                     /note="heterozygous, non-Esmeraldo-like haplotype;"
                                     allele of Tc00.1047053509809.39"
                                     /db_xref="GeneID:3546716"
                     mRNA            complement(<1..>769)
                                     /locus_tag="Tc00.1047053511261.4"
                                     /product="P-type H+ATPase, putative"
                                     /transcript_id="XM_809990.1"
                                     /db_xref="GeneID:3546716"
                     CDS              complement(<1..>769)
                                     /locus_tag="Tc00.1047053511261.4"
                                     /note="heterozygous, non-Esmeraldo-like haplotype;"
                                     allele of Tc00.1047053509809.39"
                                     /codon_start=1
                                     /product="P-type H+ATPase"
                                     /protein_id="XP_815083.1"
                                     /db_xref="GeneID:3546716"
                                     /translation="MNQINDRSVLNNSNGFNFEQHPQKPKRQGVLSKAISEHKED
DOVEVPHLPFSGKGLTAAEAEELLARYGNELPEKTPMLIFVNLGMPFALNVAI
IIEFLEINPQGLALIAIQLAKITIGWETIKAGDAVALANSILVAVVARDCAWQQ
LDNALVPGDILVKLAGSVAIPADCSINEGVINDDEALTGESLPTVMGTMRMPKMGSN
VVVGEDVTVQYTGQNTFFGKTVLTAQVESDLGHVH"
```

(ii) saida.fasta

```
>putative 60S ribosomal protein L32
MVKPFVQRIIVKRTKRTFRTHKCELFQLSSSWRKPRGSDSPVRRRYKGQKAMPNKGYS
DRKTKYITPSGFKNFPIHNVQDLYMLLMQNRKYAGVISHTVGAKSRAIVKKAHELDVRL
INGNAKLKRVSQ
>putative 60S ribosomal protein L36
MPARKKEVPKAEAPAPRTGIAGFNKGKHTRRARAPSSNDRYALPHKKLRVAKAIITDL
VGLSPMERVQELLRVGDKDRALKFKCKRLGNFKAAKRRRAKMEALRHQAQKK
>putative dynein-associated protein
MSEIEETFQRISQRPNVITGIIVVNNEGVPIRSTIEDTVTQINQYAHLITALAAKARHCVRD
LDPTNDLSFLIRSRKKNEMIVAPDKDFTLIVIQRFSDM
>putative 60S ribosomal protein L26
MASIKGCSRRAKRAHFOAPSHVRILMSAPLSKELAKYINVRAMPYRKDDDEVYKRGAY
KGREGKVTACYRLRWVHIIDKVNREKANGTTVPVGVHPSPINVEITKLKLNHNRKAILERKD
RSTKSDGKGKGVTAETAKAMQMD
>kinetoplastid membrane protein 11
MATTLEEFSAKLDRDLDAEFKAKMEEQNKFFADKPDESTLSPEMKEHYEKFEMIQEHTD
KFNKKMHSEHFKAFAELLEQQKNAQFPFG
>kinetoplastid membrane protein 11
MATTLEEFSAKLDRDLDAEFKAKMEEQNKFFADKPDESTLSPEMKEHYEKFEMIQEHTD
KFNKKMHSEHFKAFAELLEQQKNAQFPFG
>kinetoplastid membrane protein 11
MATTLEEFSAKLDRDLDAEFKAKMEEQNKFFADKPDESTLSPEMKEHYEKFEMIQEHTD
KFNKKMHSEHFKAFAELLEQQKNAQFPFG
>kinetoplastid membrane protein 11
MATTLEEFSAKLDRDLDAEFKAKMEEQNKFFADKPDESTLSPEMKEHYEKFEMIQEHTD
KFNKKMHSEHFKAFAELLEQQKNAQFPFG
>kinetoplastid membrane protein 11
MATTLEEFSAKLDRDLDAEFKAKMEEQNKFFADKPDESTLSPEMKEHYEKFEMIQEHTD
KFNKKMHSEHFKAFAELLEQQKNAQFPFG
>putative lactoylglutathione lyase-like protein
MSTRLMHTMIRVGDLDRSIKFYTEALGMRLNLRKMDCPEDKFTLVFLGYGTESETAVLEI
TNNYGGQSEYKHGDAGYHIAIGVDVNEIEARLKKMNVPIDVESEDGFMFIVDPDGYIE
LLNTERMLEKSRQEMNEQGTGA
>60S ribosomal protein L27a
MPTRFKTKTRQSGSTFCGVRGVKRRKHPSGRGNAGGQHHHRINFDFYHPGYFGKGMNH
YHLKNPLMKPTIINLNLTKIAKDEALKAKKGETLPVVDLLANGYSKLLGNGLHLQPCII
VHAKRWVSLADKRIKRGAGVVLQA
>putative small nuclear ribonucleoprotein Sm-F
MDANVPAFLASLVGSTVHVRSKWGPVYVGTLVSCDPYMNQLRDTVEKAKEETELGEML
LRCNNVLYIREVPQ
>putative lipoid acid containing carrier protein
MRRVFQLSSTLFLASARLYGTRYFTDSHEWVHDGGVATIGITAHAGESLGDVVYVALPN
VGDQLNAKEVGEVESVKATSCMYSPIINGVVDVANDRVKDEPALINRSPEDGWLKVKC
SELPQGLMTEEYKFKID
>putative ribosomal protein L37
MTKGTTSMGQRHGRHTILCCRGRNAYHVQWERCAACAYPRASRRRYNNSVKAIKRRRTG
TGRCRYLKEVNRIRIKNHKFTSLKA
>putative Lish domain-containing protein FOPNL
MHAMARQESLKEAMREVELTGVMDHVKAEALRAAIFHALQDSSAHDCASSNRPPPENL
LLNELIKEMYVFNMGHSLSVFRVSESGAKNTSAFVPRVLAQQLNMTGAPASVFLLYAM
LHESRAAADG
>putative ribosomal protein L37
MTKGTTSMGQRHGRHTILCCRGRNAYHVQWERCAACAYPRASRRRYNNSVKAIKRRRTG
TGRCRYLKEVNRIRIKNHKFTSLKA
```

(iii) enoveladoras.fasta

```
> putative heat shock protein DNAJ
MFVDVYAVLSLHSPSCSRDVRDAFRKLLALYHDPREPGEESTECFRIKEAYDVLSDPTTRYLDLGYAIEQW
RQRQEEAQLQQQQQRRRERMEDELMRNQVLQPOPTINGAYVPSLPSSESTRVSSGSSSRVLLTPASTSH
RTLAFTCYRREAVKESPEAAQAVPRQSGGVAVVRGRGLNAVYRSESDAGAGANTAGQYVTVDESSSDVRS
LPTFGSGSGSGSGSGLSKQVREQVQLNRQWRGHSVDRSSDGRKS TLVPTITTLQWGEKRLMPRSNAVSP
DFFRSVVKTRVFFHVATQGRK
> putative heat shock protein
MRRVYQRIKRDALSHSTPSGRAFAAASITLSAAVDGSRGNKISALATPMRFCTSSDAAKTKPVITDIED
VYIDFTPMACGTGTADAASPSGSSAKNHEDSRVVGESEMFXTETQGLLDIVACSLTEKEVPIRELV
NNSDALEKRHLLEISKPEYPREDEDAFFIAISCNQSKSRFVIRDTQYGMTRLEAEALGTIAGSGSKEFVR
ELQSAASGAQAEEKIIQQGVGVYASPMVAKHVKYVSRSAKSGSGYLNESDGTGTFKITECEGVDRGTIVL
DVKUTLSFTCTPQVCERVLYKNSVSVSEITLNGRVNITVEALMMKDMIAVSNHEHIDFYFISGAYDSMPFR
LHYVDAPLIRKALLYPQSHTEKYGGGMEGVNLYCHRVLTQKAKQVLPFWIRFKIAGVDSSEIPLNVR
EHTQDGSMMRLSLVLTKRIIRWLEEAQQRQRYERFIQYQYFLKEGVCTDQVHKMLAKLIRFETKSDI
DIPLYSLEYDRDLNQTTHIYIINAPSKEMALQSPYYEQKEHELVLCSTPIDDFVMQHLDTYAKHKLQV
IEMFASLDGSGVQHKKLEGEHDVKEVQKLTAEQVGLSDFIAKRLVGRVGVKTSRLRDSPAVIDHESA
QLRKIVYTCQKAGPPKYNIFNFKPIYKRLITLSISPTAAEEVETAGLVEQLDNAVISAGLEDFPSIV
SRLLNIMSVMVNEPTADK
> putative chaperone DNAJ protein
MESSRIDWIIANERHYIGLIPRADERDVRSSYLTLALQHPKNIINNRKAGEAFRAVGKAYAVLKDCKMR
MIDGGVORVHGLDAEULSTINLAALSKIVARVLCTVARRAHAVELIQRHGWDAFTSWGPAADFAQLQ
KSPREAAARVYKVFIVTMFLCVSYGHTFTQIKTKTINEGSSKEPYRLSRVNDGTATFTVIRGGGM
TDTVIRWPSISENDARALVLQIHEMKPTEKLSWASERYNRTATLRVGRKLRAAPSIRALRRKWAPEGM
RVKQFEFFITSADLVLPCLCEGVG
> putative chaperone DNAJ protein
MLRFTFFVLRWRKAPAAVAAJLPGILISLSSEKDYKILGVSRITASVSDIKKAYKRALETHPDQGGKKEE
FAEVAEAYECLSNEEKRRVYDYGSEAAANNAANGMGFGAHSANDIFAEFFRSRGMGFGDMRGPAQVQP
IEVLRMTLEIYKGVTKPRVNRVVKACDRGFGTKSQTKPKCAHCDGSGHVHVQHRMGPGMVQQTVTQCP
RCGSGGTMAKPDQCPCHCGMGVRLSGVNIIDIPGVPNSVTLVVRGEGSGMDEAFEGDLHVHVEVAPHKIF
TRGDOLLMKEIISLEALQTQGVSMRLDGRHVTVPHENVLRDVSVALVSGEMBADQGRDLVITHL
KMPAKLTAQQRREALIQAFGRPHKEKHTSKATTTARVMREGALEDMRDQKDAEGGGSGAGGHRGRQ
SSQQAECVHQ
> Co-chaperone protein P23
MAHIFPKWAERKIKLYLVTLQGSADVDVKTFTNTISITGKGIPTKASEPHELDNKTILLKEIPEKSSFEVL
GVAGVCAVKEEGYWNKLQNGSSSTANLSDVNLNMFDEDEDDGPAFGDYGDLNSENNGMGMGMDMG
GMGMDGSDDDDEDEAPQAADLQLDLAK
```

(iv) PDBxNCBI

BLASTP 2.10.1+

Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Reference for composition-based statistics: Alejandro A. Schaffer, L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and Stephen F. Altschul (2001), "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements", Nucleic Acids Res. 29:2994-3005.

Database: User specified sequence set (Input: C:\Python27\rcsb_pdb_20200811180304.fasta). 161 sequences; 50,696 total letters

Query= putative kinesin

Length=493

Score E
Sequences producing significant alignments:
(Bits) Value

50PT_27|Chain P|40S ribosomal protein S6|Trypanosoma cruzi (strain... 22.7 2.5
ST70_1|Chains A,B,C,D|Bifunctional dihydrofolate reductase-thymid... 21.6 6.4

> 50PT_27|Chain P|40S ribosomal protein S6|Trypanosoma cruzi (strain CL Brener) (353153)
Length=250

Score = 22.7 bits (47), Expect = 2.5, Method: Compositional matrix adjust.
Identities = 13/41 (32%), Positives = 22/41 (54%), Gaps = 0/41 (0%)

Query 277 KAKSMKERTSIENTGALRFASEERDAIRKAQQEEMQKE 317
K + S +ERR + G R A + +R+ R A + QK+
Sbjct 201 KVRKSAEERRAYLHLIGTRRRARQRNSARRHAHKVAAQKQ 241

> ST70_1|Chains A,B,C,D|Bifunctional dihydrofolate reductase-thymidylate synthase|Trypanosoma cruzi (strain CL Brener) (353153)
Length=521

Score = 21.6 bits (44), Expect = 6.4, Method: Compositional matrix adjust.
Identities = 13/72 (18%), Positives = 31/72 (43%), Gaps = 4/72 (6%)

Query 207 ANKVKAMKVTTTGAERLLFDYMGSKICOTLLADLHI----
FEAETQSKSAVIRNC 262
+++KA+ T T + R+LF + L H+ + + +
+R+C
Sbjct 366 VDIKAIIVETLKTNPDDRMLFTAWNPSALPRMALPPCHLLAQFYVSNGELSCMLYQRSC 425

Query 263 RQNDGLYYSTAT 274
G+ ++ A+
Sbjct 426 DMGLGVFFNIAS 437

Lambda	K	H	a	alpha
0.312	0.125	0.335	0.792	4.96

Gapped	K	H	a	alpha	sigma
Lambda	0.267	0.0410	0.140	1.90	42.6
				43.6	

Effective search space used: 17218868

Figuras

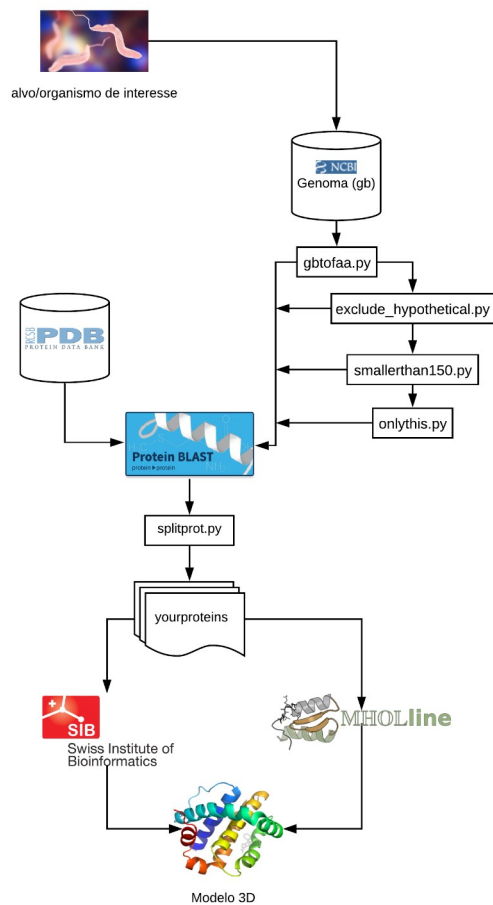


Figura 1|Fluxograma a partir do arquivo gb extraído do portal do NCBI.

NCBI Resources How To Sign in to NCBI

Genome Limits Advanced Search Help

Trypanosoma cruzi
Representative genome: Trypanosoma cruzi strain CL Brener (assembly ASM20906v1)
Download sequences in FASTA format for [genome](#), [transcript](#), [protein](#)
Download genome annotation in GFF: [GenBank](#) or [tabular](#) format
BLAST against Trypanosoma cruzi [genome](#)
All 34 genomes for species:
Browse the list
Download sequence and annotation from [RefSeq](#) or [GenBank](#)
NEW Try [NCBI Datasets](#) - a new way to download genome sequence and annotation we're testing in NCBI Labs

Tools
BLAST Genome

Related information
Assembly
BioProject
Components
Protein
PubMed
Taxonomy

Recent activity
[Turn Off](#) [Clear](#)

Trypanosoma cruzi
Genome

Trypanosoma Cruzii (4)
Genome

Trypanosoma cruzi strain CL Brener
1047053517087 genomic scaffold, v Nucleotide

Trypanosoma cruzi strain CL Brener
Genome

trypanosoma cruzi (4)
Genome

[See more...](#)

Display Settings: Overview Send to: ID: 25

Organism Overview : [Genome Assembly and Annotation report \[34\]](#) : [Organelle Annotation Report \[1\]](#)

Trypanosoma cruzi
This protozoan parasite causes American Trypanosomiasis or Chagas' disease

Lineage: Eukaryota[6063]; Discoba[63]; Euglenozoa[57]; Kinetoplastea[55]; Metakinetoplastina[54]; Trypanosomatida[53]; Trypanosomatidae[53]; Trypanosoma[9]; Schizotrypanum[1]; Trypanosoma cruzi[1]

Trypanosoma cruzi, a protozoan parasite, causes American Trypanosomiasis or Chagas' disease. The disease is transmitted to humans through the feces of infected bloodsucking insects in endemic areas of Central and South America, or occasionally by nonvectorial mechanisms, such as blood transfusion. Between 16 and 18 million persons are estimated [More...](#)

Summary

Sequence data: genome assemblies: 34; sequence reads: 2 (See [Genome Assembly and Annotation report](#))
Statistics:
median total length (Mb): 32.5392
median protein count: 13965
median GC%: 50.6

Publications (limited to 20 most recent records)

- Sequence of *Trypanosoma cruzi* reference strain SC43 nuclear genome and kinetoplast maxicircle confirms a strong genetic structure among closely related parasite discrete typing units. DeCuir J, et al. Genome 2021 May
- Nanopore Sequencing Significantly Improves Genome Assembly of the Protozoan Parasite *Trypanosoma cruzi*. Diaz-Viraqué F, et al. Genome Biol Evol 2019 Jul 1
- Expanding an expanded genome: long-read sequencing of *Trypanosoma cruzi*. Berná L, et al. Microb Genom 2018 May

Figura 2 | Portal NCBI. Banco de dados GenBank exibindo informações genômicas sobre o protozoário de interesse de estudo *Trypanosoma cruzi*

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Log in

Genome > **Genome Information by Organism**

Trypanosoma cruzi

Organism name (common or scientific) or Accession (Assembly, BioProject or replic [Q Search](#) Download Reports from FTP site

[Organism Overview](#), [Genome Assembly and Annotation report \(34\)](#), [Organelles \(1\)](#) [Filters](#) [Download](#)

#	Organism Name	Organism Groups	Strain	BioSample	BioProjec	Assembly	Levi	Size(Mb)	GC%	Replicons
1	Trypanosoma cruzi	Eukaryota; Protists; Kinetoplasts	231	SAMEA3359340	PRJEB9129	GCA_900252365.1	35.36	48.60		
2	Trypanosoma cruzi	Eukaryota; Protists; Kinetoplasts	Arequipa	SAMN03329996	PRJNA27444	GCA_003594685.1	19.05	-		
3	Trypanosoma cruzi	Eukaryota; Protists; Kinetoplasts	B.M. Lopez	SAMN13545730	PRJNA59507	GCA_010117175.1	18.51	48.30		
4	Trypanosoma cruzi marnickelii	Eukaryota; Protists; Kinetoplasts	B7	SAMN02953810	PRJNA77843	GCA_000300495.1	34.23	50.90		
5	Trypanosoma cruzi	Eukaryota; Protists; Kinetoplasts	Berenice	SAMN10335319	PRJNA49880	GCA_013358655.1	40.80	51.20		
6	Trypanosoma cruzi	Eukaryota; Protists; Kinetoplasts	Brazil clone A4	SAMN10689846	PRJNA51286	GCA_015033625.1	45.56	51.58	chromosome 1: CM026583.1 chromosome 2: CM026584.1 chromosome 3: CM026585.1 Show all 43 replicons	
7	Trypanosoma cruzi	Eukaryota; Protists; Kinetoplasts	Bug2148	SAMN07305944	PRJNA39309	GCA_002749415.1	55.16	51.30		
8	Trypanosoma cruzi	Eukaryota; Protists; Kinetoplasts	CL	SAMN04560603	PRJNA31539	GCA_003719155.1	65.00	39.80		
9	Trypanosoma cruzi strain CL Brener	Eukaryota; Protists; Kinetoplasts	CL Brener	SAMN02953627	PRJNA11755	GCA_000200065.1	89.94	51.70		
10	Trypanosoma cruzi	Eukaryota; Protists; Kinetoplasts	Colombiana	SAMN03339015	PRJNA27444	GCA_003594625.1	30.85	50.80		
11	Trypanosoma cruzi	Eukaryota; Protists; Kinetoplasts	Dm28c	SAMN08469708	PRJNA43304	GCA_003177105.1	53.27	51.60		
12	Trypanosoma cruzi cruzi	Eukaryota; Protists; Kinetoplasts	Dm28c	SAMN05432872	PRJNA33097	GCA_002219105.2	50.99	51.67	mitochondrion MT: CM008275.1	
13	Trypanosoma cruzi Dm28c	Eukaryota; Protists; Kinetoplasts	Dm28c	SAMN02953876	PRJNA22592	GCA_000496795.1	27.35	50.60		
14	Trypanosoma cruzi strain Esmeraldo	Eukaryota; Protists; Kinetoplasts	Esmeraldo cl. 3	SAMN00016463	PRJNA50493	GCA_000327425.1	38.08	50.60		
15	Trypanosoma cruzi	Eukaryota; Protists; Kinetoplasts	G	SAMN04566062	PRJNA31539	GCA_003719455.1	25.17	47.40		
16	Trypanosoma cruzi	Eukaryota; Protists; Kinetoplasts	Ikiakarora	SAMN13546110	PRJNA59509	GCA_010117215.1	18.49	48.70		
17	Trypanosoma cruzi JR cl. 4	Eukaryota; Protists; Kinetoplasts	JR cl. 4	SAMN02953827	PRJNA59941	GCA_000331405.1	41.48	51.20		

Figura 3 | Portal NCBI Página de exibição de projetos genômicos submetidos por organismo/cepa.

Organism Overview · Genome Assembly and Annotation report

Trypanosoma cruzi strain CL Brener

Reference genome sequence

Lineage: Eukaryota[6863]; Discoba[63]; Euglenozoa[57]; Kinetoplastea[54]; Trypanosomatida[53]; Trypanosomatidae[53]; Trypanosoma[9]; Schizotrypanum[1]; Trypanosoma cruzi[1]; Trypanosoma cruzi strain CL Brener[1]

Summary

Submitter: Trypanosoma cruzi consortium
 Assembly level: Scaffold
 Assembly: GCA_000209065.1_ASM20906v1 scaffolds: 29,495 contigs: 32,746 N50: 14,669 L50: 1,277
 BioProjects: PRJNA15540, PRJNA11755
 Whole Genome Shotgun (WGS): RefSeq: NZ_AAHK00000000.1; INSDC: AAHK00000000.1
 Statistics: total length (Mb): 89.9375
 protein count: 19607
 GC%: 51.7

Publications (limited to 20 most recent records)

1. Comparative genomics of trypanosomatid parasitic protozoa. El-Sayed NM, et al. Science 2005 Jul 15
2. The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. El-Sayed NM, et al. Science 2005 Jul 15
3. The Trypanosoma cruzi proteome. Atwood JA 3rd, et al. Science 2005 Jul 15

Replicon Info

Loc	Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
	Un	-	-	-	89.94	51.7	19,607	219	115	1,656	25,100	3,503

Genome Region

Trypanosoma cruzi strain CL Brener 1047053517087 genomic scaffold, whole genome shotgun sequence

Go to nucleotide: [Graphics](#) [FASTA](#) [GenBank](#)

taxonomy

Recent activity

Turn Off Clear

Trypanosoma cruzi strain CL Brener
Genome

Trypanosoma cruzi
Genome

Trypanosoma Cruzii (4)
Genome

Trypanosoma cruzi strain CL Brener 1047053517087 genomic scaffold, w Nucleotide

trypanosoma cruzi (4)
Genome

See more...

Figura 4|Portal NCBI Página de exibição do projeto genômico PRJNA15540 *Trypanosoma cruzi* strain CL Brener 1047053517087 genomic scaffold, whole genome shotgun sequence.

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Search

Advanced Help

GenBank

Trypanosoma cruzi strain CL Brener 1047053517087 genomic scaffold, whole genome shotgun sequence

NCBI Reference Sequence: NW_001849447.1

[FASTA](#) [Graphics](#)

Go to:

LOCUS NI_001849447 990720 bp DNA linear CON 28-NOV-2009

DEFINITION Trypanosoma cruzi strain CL Brener 1047053517087 genomic scaffold, whole genome shotgun sequence.

ACCESSION NI_001849447

VERSION NI_001849447.1

DBLINK Project: 15540

BioProject: PRJNA15540

KEYWORDS WGS; RefSeq.

SOURCE Trypanosoma cruzi strain CL Brener

ORGANISM Trypanosoma cruzi strain CL Brener

Eukaryota; Discoba; Euglenozoa; Kinetoplastea; Metakinetoplastina; Trypanosomatida; Trypanosomatidae; Trypanosoma; Schizotrypanum.

REFERENCE 1 (bases 1 to 990720)

AUTHORS El-Sayed, M.H., Myler, P.J., Bartholomew, D.C., Nilsson, D., Aggarwal, G., Tran, A.N., Ghedin, E., Worthey, E.A., Delcher, A.L., Blandin, G., Westenberg, S.J., Coler, E., Cerqueira, G.C., Branche, C., Haas, B., Anupama, A., Arner, E., Aslund, L., Attipoe, P., Bontempi, E., Brinkmann, F., Burton, P., Cadag, E., Campbell, D.A., Carrington, M., Crabtree, J., Darban, H., da Silva, J.F., de Jong, P., Edwards, K., Englund, P.T., Fazelina, G., Feldblyum, T., Ferella, M., Frasch, A.C., Gull, K., Horn, D., Hou, L., Huang, Y., Kindlund, E., Klingbeil, M., Kluge, S., Koo, H., Lacerda, D., Levin, M.J., Lorenzi, H., Louie, T., Machado, C.R., McCulloch, R., McKenna, A., Mizuno, Y., Mottram, J.C., Nelson, S., Ochaya, S., Osoegawa, K., Pai, G., Parsons, M., Pentony, H., Pettersson, U., Pop, M., Ramirez, J.L., Rinta, J., Robertson, L., Salzberg, S.L., Sanchez, D.O., Seyler, A., Sharma, R., Shetty, J., Simpson, A.J., Sisk, E., Tammi, H.T., Tarleton, R., Teixeira, S., Van Aken, S., Vogt, C., Ward, P.N., Wickstead, B., Wortman, J., White, O., Fraser, C.H., Stuart, K.D. and

Send to: Complete Record Coding Sequences Gene Features

Choose Destination File Clipboard Collections Analysis Tool

Download 1 item.

Format GenBank

Show GI

Create File

Assembly

BioProject

Protein

PubMed

Taxonomy

Components (Core)

Gene

Identical GenBank Sequence

PubMed (Weighted)

RNA

Recent activity

Turn Off Clear

Figura 5|Portal NCBI Página de extração do relatório em formato de arquivo gb.

```

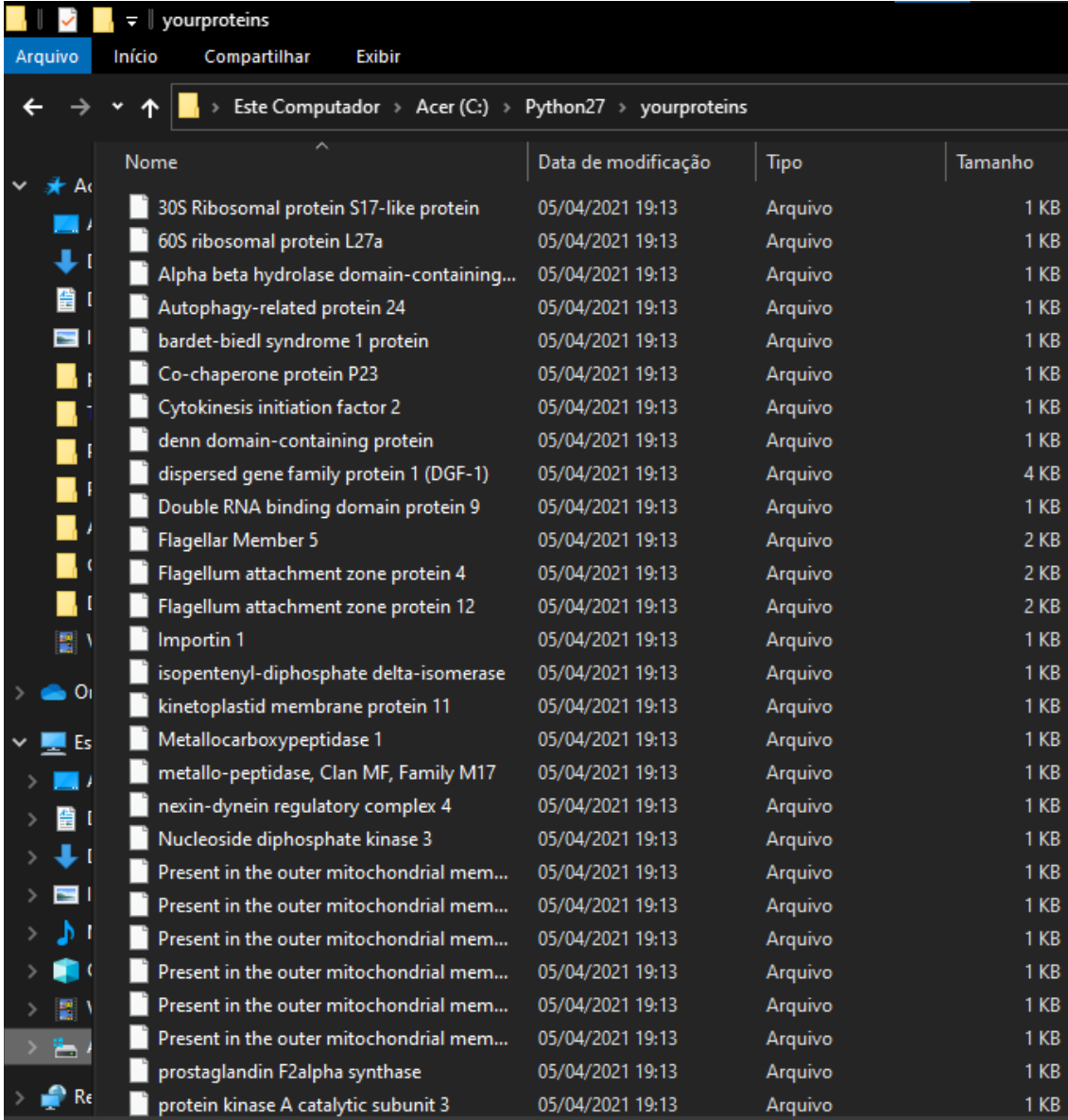
CDS      complement(<1..769)
          /locus_tag="Tc00.1047053511261.4"
          /note="heterozygous, non-Esmeraldo-like haplotype;
          allele of Tc00.1047053509809.39"
          /codon_start=1
          /product="P-type H+-ATPase"
          /protein_id="XP_815083.1"
          /db_xref="GeneID:3546716"
          /translation="MNQKNDRSVLNNNSNGNFNEQHPPQKPQKRQSVLSKAISEHKED
          DVDEVPMLPPSKGLTTAEAEELLAKYGRNELPEKKTPSWLI FVRNLWGPMPPFALWVAI
          IIEFALENWPDGAILLAIQLANATIGWYETIKAGDAVAALKNSLKPVATVHRDGAWQQ
          LDAALLVPGDLVKLASGSAPADCSINEGVIDVDEAALTGESLPVTMGTDHMPKMGSN
          VVRGEVDGTVQYTGQNTFFGKTAVLLQSVESDLGNIHV"

```

Figura 6|Portal NCBI Trecho do arquivo sequence.gb extraído do portal NCBI/Genbank, contendo uma determinada região de codificação de proteína do organismo.

The screenshot shows the RSCB PDB website interface. At the top, there's a navigation bar with links like 'RCSB PDB', 'Deposit', 'Search', 'Visualize', 'Analyze', 'Download', 'Learn', 'More', and 'Documentation'. Below this is the PDB logo and a search bar. The main content area shows search results for 'Trypanosoma cruzi'. A list of structures is displayed on the left, with '7AOR' selected. To the right of the list is a 3D model of the protein complex. Below the model, detailed information about the entry is provided, including the title 'mt-SSU from Trypanosoma cruzi in complex with mt-IF-3', the authors 'Soufari, H., Waltz, F., Parrot, C., Bochler, A., Hashem, Y.', the release date '2021-04-28', the method 'ELECTRON MICROSCOPY 3.5 Å', and the organisms 'Leishmania arabica' and 'Trypanosoma cruzi'.

Figura 7|Portal RSCB PDB. Banco de estruturas moleculares exibindo proteínas modeladas do protozoário de interesse *Trypanosoma cruzi*.



Nome	Data de modificação	Tipo	Tamanho
30S Ribosomal protein S17-like protein	05/04/2021 19:13	Arquivo	1 KB
60S ribosomal protein L27a	05/04/2021 19:13	Arquivo	1 KB
Alpha beta hydrolase domain-containing...	05/04/2021 19:13	Arquivo	1 KB
Autophagy-related protein 24	05/04/2021 19:13	Arquivo	1 KB
bardet-biedl syndrome 1 protein	05/04/2021 19:13	Arquivo	1 KB
Co-chaperone protein P23	05/04/2021 19:13	Arquivo	1 KB
Cytokinesis initiation factor 2	05/04/2021 19:13	Arquivo	1 KB
denn domain-containing protein	05/04/2021 19:13	Arquivo	1 KB
dispersed gene family protein 1 (DGF-1)	05/04/2021 19:13	Arquivo	4 KB
Double RNA binding domain protein 9	05/04/2021 19:13	Arquivo	1 KB
Flagellar Member 5	05/04/2021 19:13	Arquivo	2 KB
Flagellum attachment zone protein 4	05/04/2021 19:13	Arquivo	2 KB
Flagellum attachment zone protein 12	05/04/2021 19:13	Arquivo	2 KB
Importin 1	05/04/2021 19:13	Arquivo	1 KB
isopentenyl-diphosphate delta-isomerase	05/04/2021 19:13	Arquivo	1 KB
kinetoplastid membrane protein 11	05/04/2021 19:13	Arquivo	1 KB
Metallo-carboxypeptidase 1	05/04/2021 19:13	Arquivo	1 KB
metallo-peptidase, Clan MF, Family M17	05/04/2021 19:13	Arquivo	1 KB
nexin-dynein regulatory complex 4	05/04/2021 19:13	Arquivo	1 KB
Nucleoside diphosphate kinase 3	05/04/2021 19:13	Arquivo	1 KB
Present in the outer mitochondrial mem...	05/04/2021 19:13	Arquivo	1 KB
Present in the outer mitochondrial mem...	05/04/2021 19:13	Arquivo	1 KB
Present in the outer mitochondrial mem...	05/04/2021 19:13	Arquivo	1 KB
Present in the outer mitochondrial mem...	05/04/2021 19:13	Arquivo	1 KB
Present in the outer mitochondrial mem...	05/04/2021 19:13	Arquivo	1 KB
prostaglandin F2alpha synthase	05/04/2021 19:13	Arquivo	1 KB
protein kinase A catalytic subunit 3	05/04/2021 19:13	Arquivo	1 KB

Figura 10|Arquivos de saída das proteínas. Conteúdo do diretório “yourproteins”, contendo cada arquivo nomeado com a respectiva sequência, extraído do arquivo Fasta *sequencesnohyp.fasta*.

Model Results ⓘ

Order by: GMQE ▾



Template	Seq Identity	Coverage	Description
2ctp.1.A	43.75%	<div></div>	DnaJ homolog subfamily B member 12 Solution structure of J-domain from human DnaJ subfamily B member 12

Model-Template Alignment

Model_01	MFV DYYAVLDSLHPSCSSRDVRDAFKRLALLYHPDRPE EGSTECFREIKEAYDVLSDPTR	59
2ctp.1.A	---DYYEIDGVSRGASQEDLKKAYRRLLAKFHDPKNHAFGATEAFKAIGTAYAVLSDNPEK	64
Model_01	RYLYDLGYAEIQWVRQRQQEEAQLQQQQQRRREREMRMDLMRNQVLQPQPTRNGAYVSP	119
2ctp.1.A	RKQYDQFG-----	72
Model_01	SLPSESTRRVSSGSSSRVLLTPASTSHRTLAPTCVRREAVEKSREAAQAVPRQSGGFVA	179
2ctp.1.A	-----	
Model_01	VRRGRLAVVERSDAGIGAQNTAHGQYRQTVDEFSSSDVRS LPVTGSQSSQSGGENSLSGK	239
2ctp.1.A	-----	
Model_01	QPSEQEQLQRWGRHSVKDKSNDGKSTLVPKITTLQWGEKKLMRRSNVAVPSPDFFKRSV	299
2ctp.1.A	-----	
Model_01	VKTIWRVFFHVPATGQRDK	317
2ctp.1.A	-----	

Figura 11|Página de resultados do Swiss-Model. Resultados do modelo, estimativas de qualidades local e global, o molde utilizado, o alinhamento, dentre outros dados.

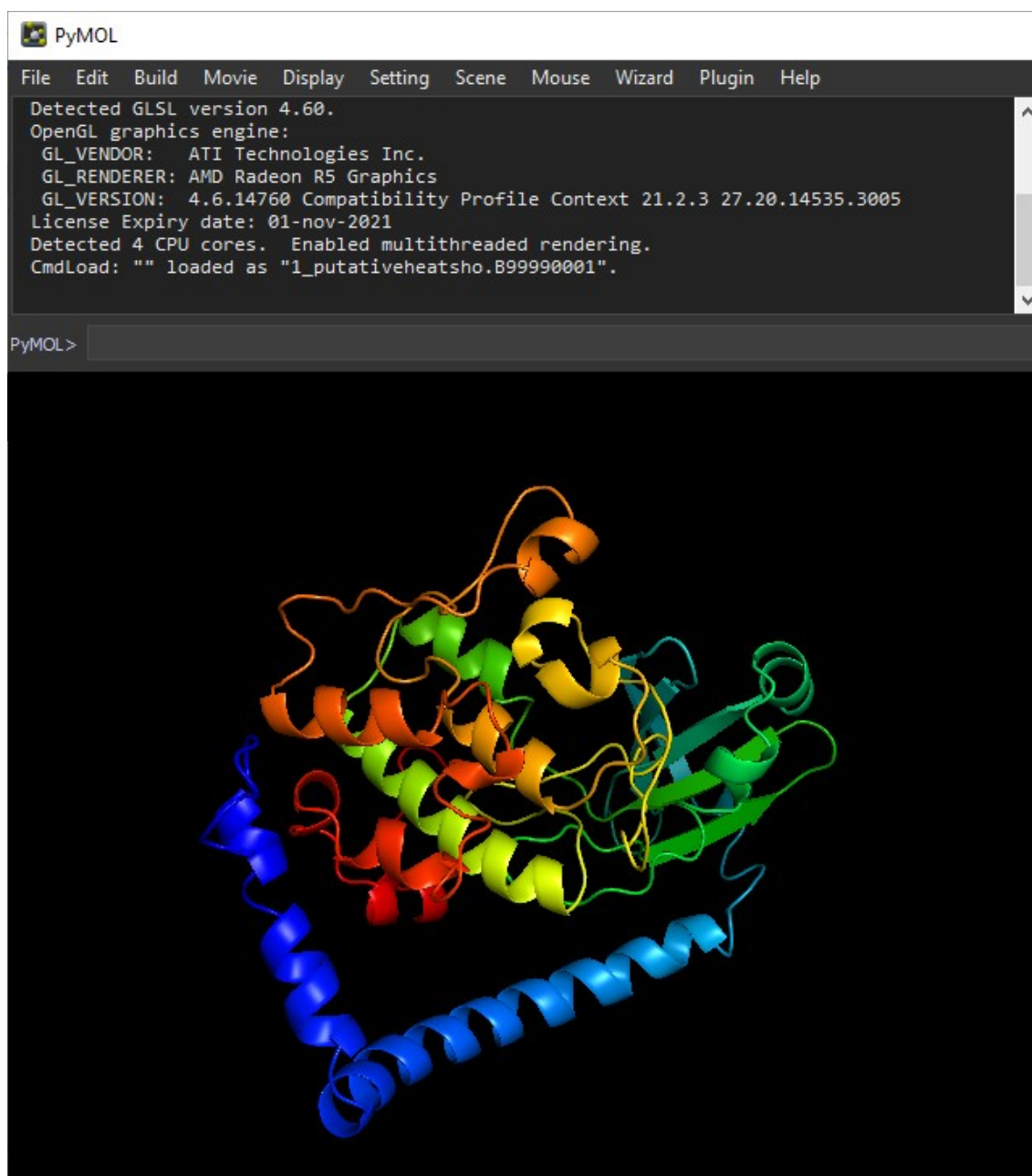


Figura 12|Representação do modelo gerado. Arquivo PDB gerado pelo Modeller a partir do MHOLline e visualizado a partir do aplicativo PyMol.