Anotações de Funções de Proteínas utilizando Processamento de Linguagem Natural e Alinhamento Local de Proteínas

Gabriel Bianchin de Oliveira

Orientador: Prof. Dr. Zanoni Dias
Coorientador: Prof. Dr. Hélio Pedrini

21 de Março de 2023

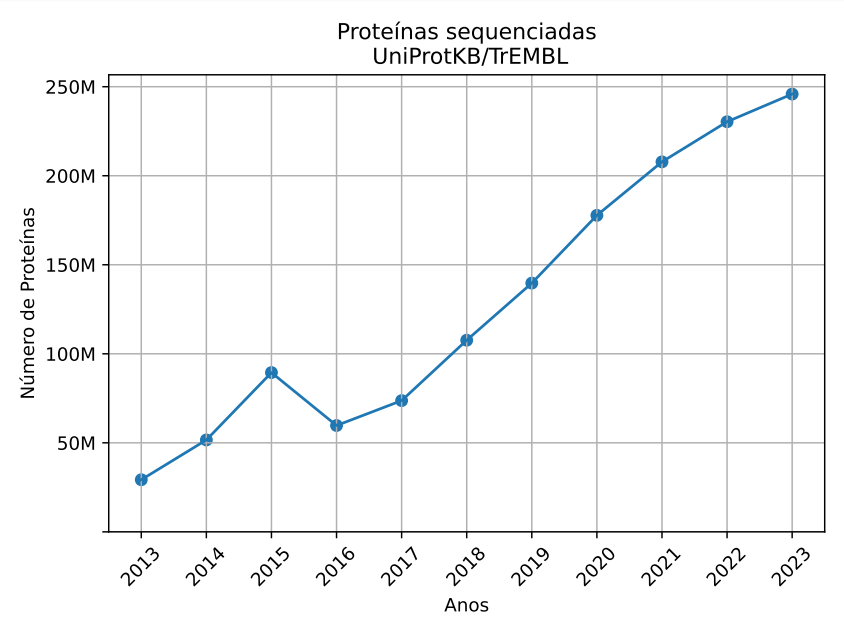
Instituto de Computação
Universidade Estadual de Campinas

Sumário

Sumário

Introdução

- Proteínas, aminoácidos e funções
- Sequenciamento de proteínas nos últimos anos
- Custo para definição de funções



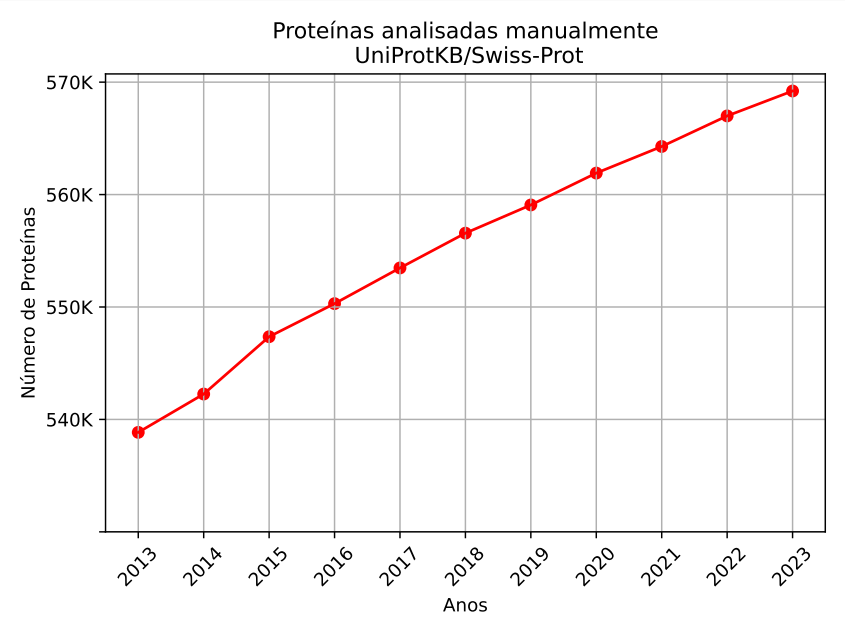


Figura 1: Número de proteínas sequenciadas e analisadas manualmente na base UniProtKB.

Introdução

- Métodos computacionais para classificação utilizando a sequência
- Aprendizado de Máquina e Processamento de Linguagem Natural (PLN)
- Alinhamento local de proteínas
- Problemas na classificação:
  - Limite de entrada dos modelos
  - Pouca eficiência de modelos de PLN
  - Dependência de proteínas homólogas anotadas
  - Dificuldade na classificação de termos raros

Objetivos

Desenvolver uma abordagem capaz de predizer funções de proteínas utilizando técnicas de PLN e alinhamento local de proteínas

- Estudar as abordagens disponíveis na literatura para a tarefa em questão
- Avaliar as bases de dados disponíveis na literatura
- Investigar técnicas de processamento de linguagem natural e alinhamento local no problema de predição de funções de proteínas
- Propor um método para realizar a predição de funções de proteínas utilizando técnicas de processamento de linguagem natural e alinhamento local de proteínas
- Realizar experimentos com o método proposto
- Avaliar e comparar o método com outras abordagens da literatura
- Documentar e publicar os resultados obtidos

Questões de Pesquisa

1. É possível obter resultados competitivos utilizando redes Transformers de processamento de linguagem natural na classificação de funções de proteínas?
2. A utilização de métodos baseados em alinhamento local de proteínas pode auxiliar na predição quando combinados com o modelo baseado em processamento de linguagem natural?
3. Como tratar funções que estão presentes em poucas amostras e adaptar o método proposto para estes casos?
4. Qual é o impacto do uso de aumentação de dados durante o treinamento de modelos baseados em Transformers?

Questões de Pesquisa

5. Considerando técnicas baseadas em regras e métodos generativos, qual é a melhor forma de aplicar aumentação de dados para a tarefa sob investigação?
6. A adaptação de um modelo com limitação de entrada para um modelo que aceite uma entrada maior pode melhorar os resultados na tarefa de predição de funções de proteínas?
7. É possível transferir o conhecimento de um modelo maior (professor) para um modelo menor (estudante) sem perder a eficácia e aumentar a eficiência?

Sumário

Função de Proteína

- Anotação via Ontologia Genética (GO) [1]
- Ontologia de Componente Celular (CC), Função Molecular (FM) e Processo Biológico (PB)
- Grafo acíclico direcionado
- Classificação multirrótulo

Função de Proteína

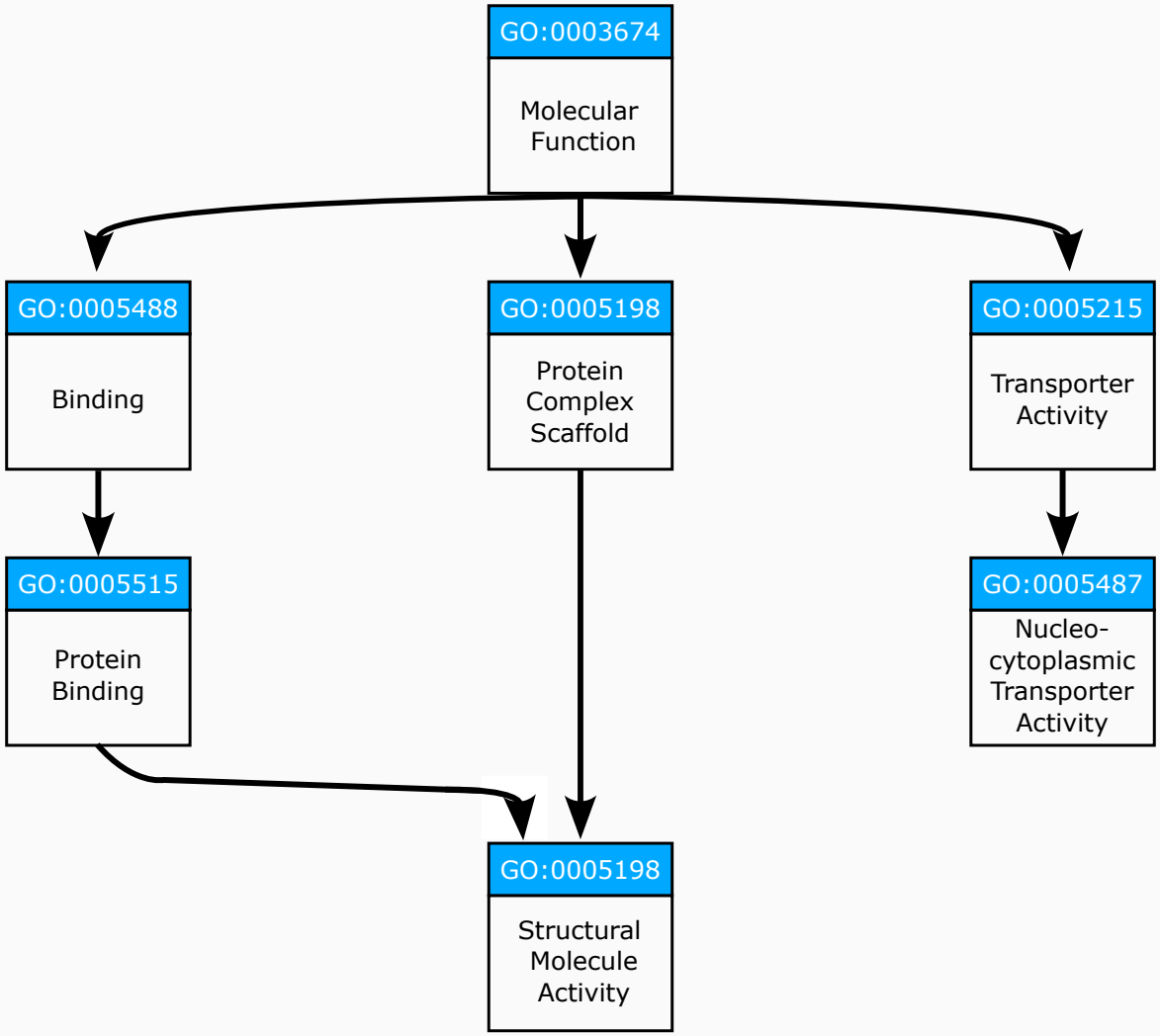


Figura 2: Parte da ontologia de Função Molecular.

Função de Proteína

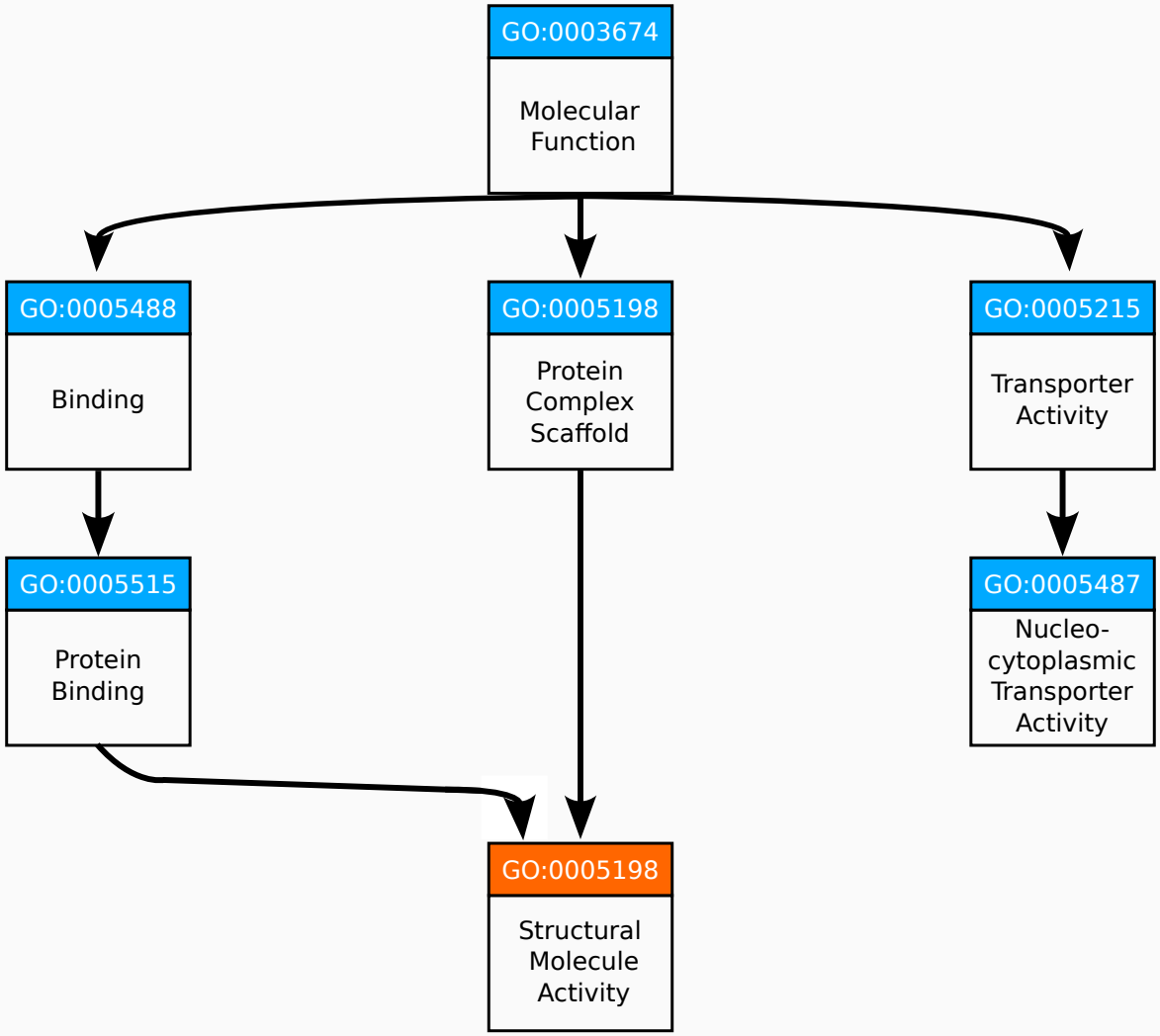


Figura 2: Parte da ontologia de Função Molecular.

Função de Proteína

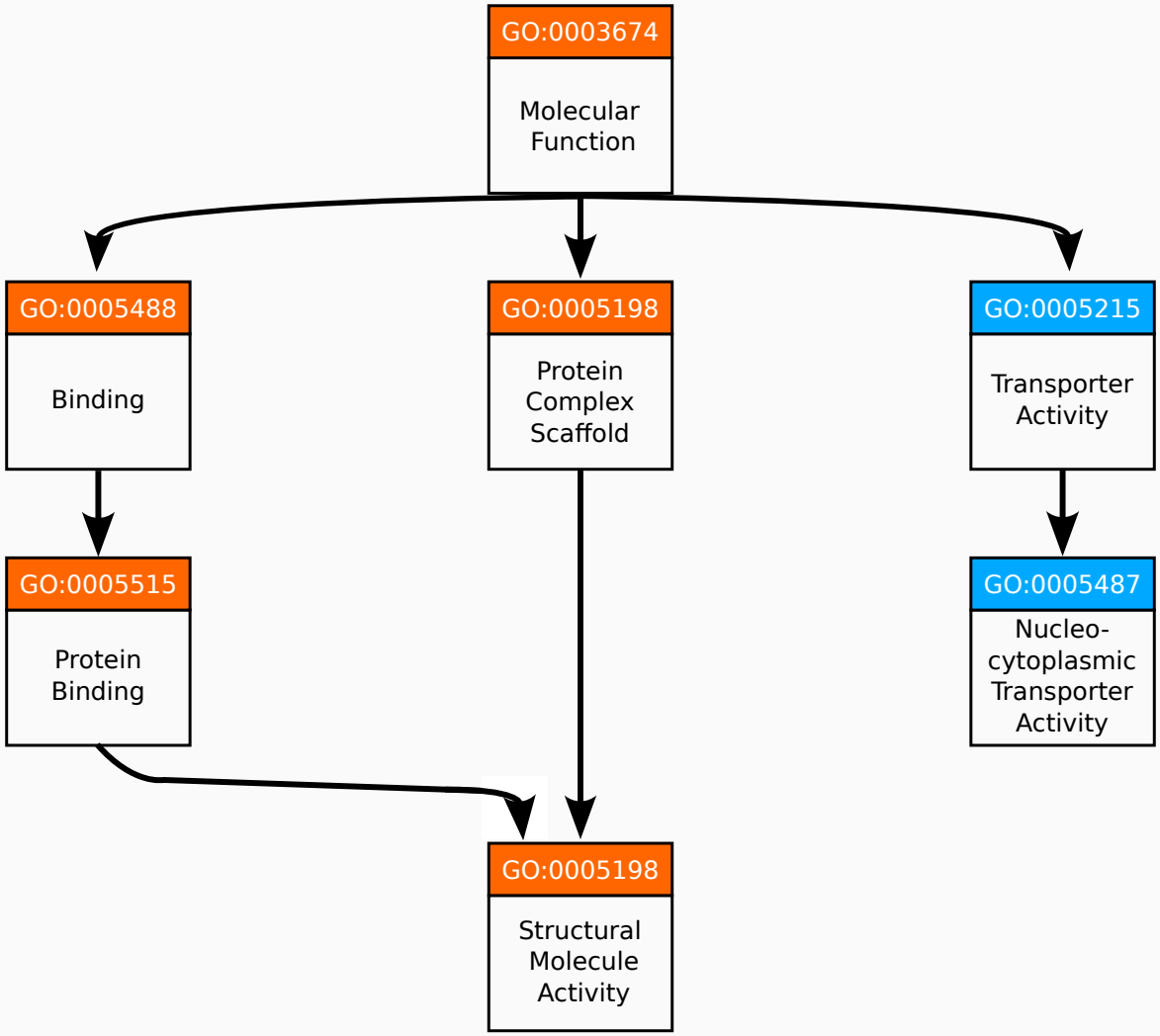


Figura 2: Parte da ontologia de Função Molecular.

Alinhamento Local de Proteínas

- Subsequências de duas proteínas com alta similaridade
- *Matches*, *mismatches* e *gaps*

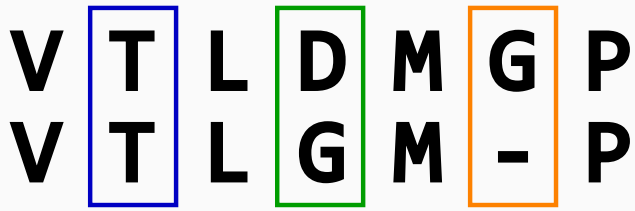


Figura 3: Exemplo de *matches*, *mismatches* e *gaps*.

- Matriz de pontuação
- *bitscore* e *e-value*
- BLAST [2] e DIAMOND [3]

Transformers

- Arquiteturas com módulos atencionais utilizadas em PLN

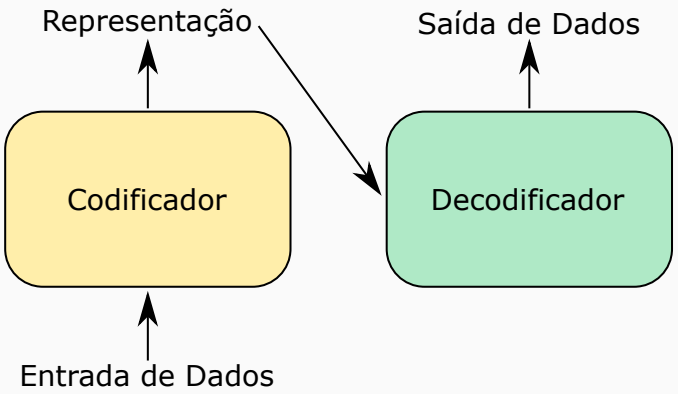


Figura 4: Arquitetura Transformer de maneira simplificada.

- Arquiteturas codificadora-decodificadora
  - Transformer Vanilla [4]
- Arquiteturas codificadoras
  - BERT [5], RoBERTa [6] e LongFormer [7]
  - ProtBERT, ProtBERT-BFD [8], TAPE [9] e ESM [10]
- Arquiteturas decodificadoras
  - Modelos GPT [11]
  - ProtGPT2 [12]

Aumentação de Dados

- Aumentar a variabilidade dos dados
- Transformações nas amostras originais
- Aumentações em PLN:
  - Regras manuais
  - Interpolação de dados ou rótulos
  - Métodos generativos
- Técnicas pouco aplicadas no domínio de proteínas

Zero-Shot Learning e Few-Shot Learning

- Classes não presentes ou esparsamente representadas por poucas amostras no treinamento
- *Zero-Shot Learning*
- *Few-Shot Learning*
- Aplicações na predição de funções de proteínas [13, 14]

Destilação de Conhecimento

- Modelos robustos e pouco eficientes
- Aplicação de destilação de conhecimento:
  - Arquitetura já treinada (professor)
  - Arquitetura que recebe o conhecimento (estudante)
  - Estudante imita o comportamento do professor

Sumário

Abordagens

- Baseada em homologia
- Baseada em redes
- Baseada em informação
- Baseada em mineração de texto
- Baseada em sequência

Métodos Baseados em Homologia

- Busca por proteínas homólogas com funções definidas
- Ferramentas de alinhamento local de proteínas:
  - BLAST [2]
  - DIAMOND [3]
- Presença em métodos que usam junto com aprendizado de máquina [15, 16, 17]
- Problemas:
  - Dependência de proteínas com funções definidas significativamente similares
  - Dificuldade em homologia remota

Métodos Baseados em Redes

- Análise de redes de interações de proteínas
- Utilização como extratora de características [18, 19]
- Presença em métodos com outros tipos de características [20]
- Problemas:
  - Dependência de redes de interação de proteínas
  - Custo para a análise laboratorial

Métodos Baseados em Informação

- Características biológicas para a classificação:
  - Características estruturais [21]
  - Características biofísicas [21]
  - Super-famílias [22]
  - Domínio [22]
- Presença em métodos que usam homologia, sequência e redes de interação de proteínas [23, 24]
- Problemas:
  - Dependência de dados biológicos
  - Custo para análise laboratorial

Métodos Baseados em Mineração de Texto

- Análise de textos científicos
- Utilização de processamento de linguagem natural [25]
- Problemas:
  - Dependência de textos científicos que relatem as funções
  - Dificuldade para predizer funções de proteínas ainda não estudadas

Métodos Baseados em Sequência

- Classificação usando a sequência de aminoácidos
- Utilização de aprendizado de máquina e processamento de linguagem natural:
  - Redes convolucionais [15, 20]
  - Módulos atencionais [26]
  - TF-IDF [27]
  - Transformers [16, 17]

Sumário

Metodologia

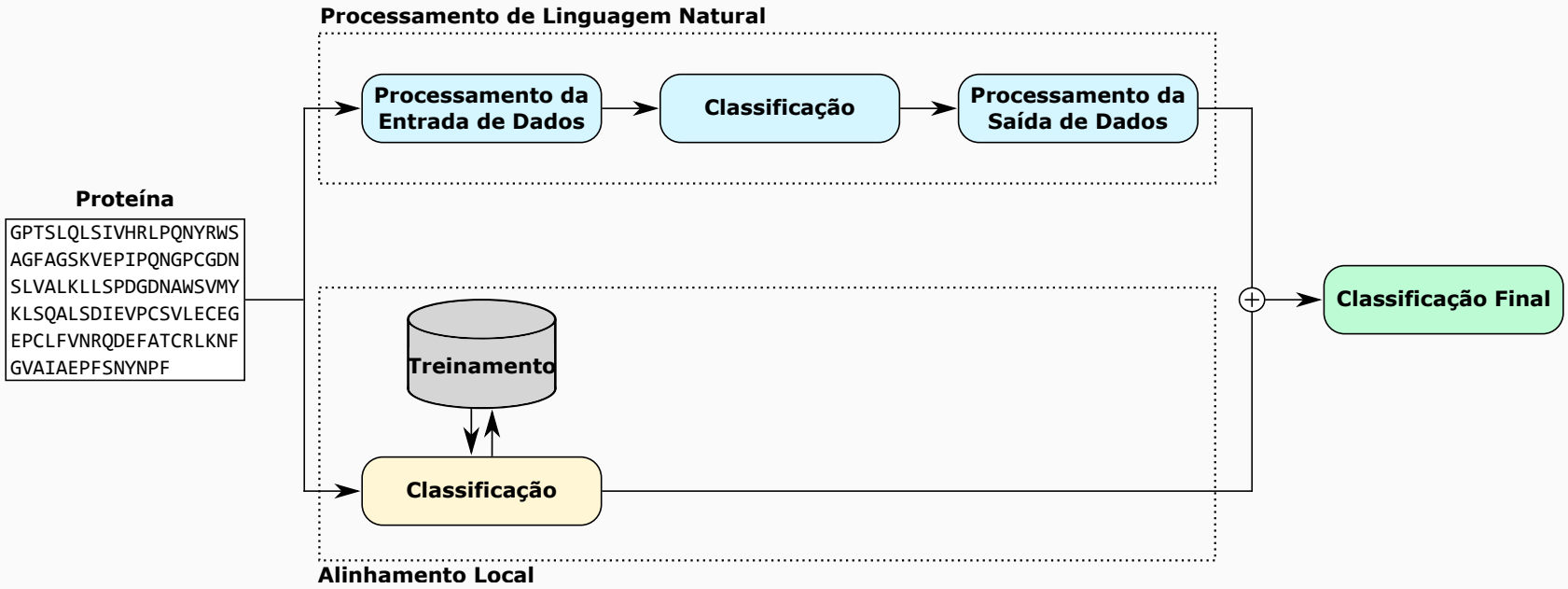


Figura 5: Visão geral do método proposto.

Metodologia

- Aumentação de dados
  - Troca de aminoácidos com a matriz PAM [28]
  - Inserção e exclusão aleatória
  - Métodos generativos
- Classificação de termos raros
  - *Zero-Shot Learning*
  - *Few-Shot Learning*
- Destilação de conhecimento
  - Arquitetura Transformers mais enxuta
  - Redes recorrentes bidirecionais

Base de Dados

- Bases do *Critical Assessment of protein Function Annotation* (CAFA)
- Base derivada do CAFA3 [29], criada por Kulmanov e Hoehndorf [15]
- CAFA3, CAFA4 e CAFA5

Bases de Dados

Tabela 1: Quantidade de proteínas no treinamento, validação e teste, e quantidade de rótulos em cada uma das ontologias.

	CC	FM	PB
Treinamento	45.309	32.421	47.691
Validação	4.985	3.587	5.252
Teste	1.265	1.137	2.392
Rótulos	551	677	3.992

Métricas de Avaliação

- F_{max}
- AuPRC
- IAuPRC
- S_{min}

$$\text{pr}(\tau) = \frac{1}{m(\tau)} \sum_{i=1}^{m(\tau)} \frac{\sum_f I(f \in P_i(\tau) \wedge f \in T_i)}{\sum_f I(f \in P_i(\tau))} \tag{1}$$

$$\text{rc}(\tau) = \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_f I(f \in P_i(\tau) \wedge f \in T_i)}{\sum_f I(f \in T_i)} \tag{2}$$

$$F_{\text{max}} = \max_{\tau} \left\{ \frac{2 \times \text{pr}(\tau) \times \text{rc}(\tau)}{\text{pr}(\tau) + \text{rc}(\tau)} \right\} \tag{3}$$

AuPRC e IAuPRC

AuPRC:

- Gerada a partir do cálculo do $F_{\max}$

IAuPRC:

- Métrica proposta durante a pesquisa
- Interpolação da AuPRC

$$\text{pr}(\text{rc}(\tau)) = \max\{\text{pr}(\text{rc}(\tau')) : \tau \leq \tau' \leq 1\} \tag{4}$$

$$\text{IC}(f) = -\log(Pb(f|Pr(f)) \tag{5}$$

$$\text{ru}(\tau) = \frac{1}{n}\sum_{i=1}^{n}\sum_{f \in T_i - P_i(\tau)} \text{IC}(f) \tag{6}$$

$$\text{mi}(\tau) = \frac{1}{n}\sum_{i=1}^{n}\sum_{f \in P_i(\tau) - T_i} \text{IC}(f) \tag{7}$$

$$S_{\text{min}} = \min_{\tau} \sqrt{\text{ru}(\tau)^2 + \text{mi}(\tau)^2} \tag{8}$$

Recursos Computacionais

- Bibliotecas
  - `NumPy`
  - `scikit-learn`
  - `TensorFlow`
  - `Hugging Face`
  - `Matplotlib`
  - `ktrain`
- Ambiente virtual
  - Google Colaboratory

Sumário

Atividades

Tabela 2: Cronograma de atividades dividido em trimestres.

Atividades	1º ano				2º ano				3º ano				4º ano			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	•	•	•	•												
2			•	•	•	•	•	•	•	•						
3									•							
4	•	•	•	•	•	•	•	•	•	•	•	•				
5			•	•	•	•	•	•	•							
6					•	•										
7						•	•	•	•							
8						•	•	•	•							
9								•	•	•						
10										•	•	•				
11											•	•	•	•		
12													•	•		
13															•	•
14																•

1. Obtenção dos créditos obrigatórios em disciplinas do programa de doutorado.
2. Participação no Programa de Estágio Docente (PED).
3. Exame de Qualificação Específico (EQE).

Atividades

Tabela 2: Cronograma de atividades dividido em trimestres.

Atividades	1º ano				2º ano				3º ano				4º ano			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	●	●	●	●												
2			●	●	●	●	●	●	●	●						
3									●							
4	●	●	●	●	●	●	●	●	●	●	●	●				
5			●	●	●	●	●	●	●							
6					●	●										
7						●	●	●	●							
8						●	●	●	●							
9								●	●	●						
10										●	●	●				
11											●	●	●	●		
12													●	●		
13															●	●
14																●

4. Revisão da Literatura.
5. Construção do método baseado em PLN utilizando Transformers.
6. Construção do método baseado em alinhamento local.
7. Agregação dos métodos baseados em PLN e alinhamento local.

Atividades

Tabela 2: Cronograma de atividades dividido em trimestres.

Atividades	1º ano				2º ano				3º ano				4º ano			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	●	●	●	●												
2			●	●	●	●	●	●	●	●						
3									●							
4	●	●	●	●	●	●	●	●	●	●	●	●				
5			●	●	●	●	●	●	●							
6					●	●										
7						●	●	●	●							
8						●	●	●	●							
9								●	●	●						
10										●	●	●				
11											●	●	●	●		
12													●	●		
13															●	●
14																●

8. Avaliação das técnicas de aumentação de dados.
9. Adaptação do modelo para Transformers longos.
10. Adaptação do modelo final para *Zero-Shot Learning* e *Few-Shot Learning*.
11. Aplicação de destilação de conhecimento no modelo final.

Atividades

Tabela 2: Cronograma de atividades dividido em trimestres.

Atividades	1º ano				2º ano				3º ano				4º ano			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	●	●	●	●												
2			●	●	●	●	●	●	●	●						
3									●							
4	●	●	●	●	●	●	●	●	●	●	●	●				
5			●	●	●	●	●	●	●							
6					●	●										
7						●	●	●	●							
8						●	●	●	●							
9								●	●	●						
10										●	●	●				
11											●	●	●	●		
12													●	●		
13															●	●
14																●

12. Escrita da tese.
13. Revisão da tese.
14. Defesa da tese.

Sumário

Métodos Comparados

- Ingênuo
- BLASTp [2]
- DIAMOND Score (DS) [3]
- DeepGO [20]
- DeepGOPlusCNN e DeepGOPlus [15]
- TALE+Transformers e TALE+ [16]
- ATGO e ATGO+ [17]
- Nossos métodos:
  - TEMPROT
  - DS-TEMPROT
  - TEMPROT+

TEMPROT

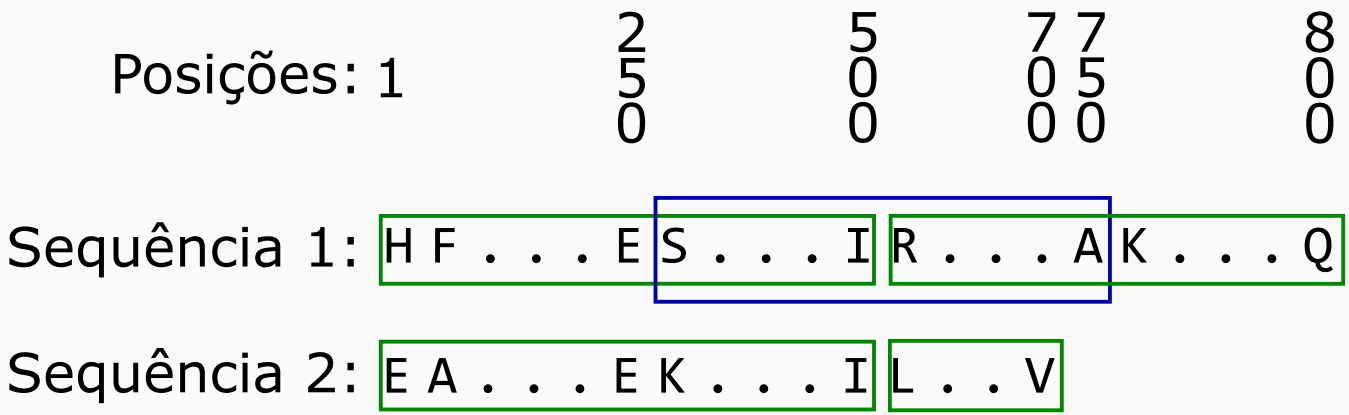


Figura 6: Exemplos das janelas criadas para o TEMPROT.

- Aumentação de dados via substituição
- Ajuste fino do ProtBERT-BFD
- Extração de características do *token* `CLS`
- Treinamento do meta-classificador

Avaliação nas Ontologias

Tabela 3: F_{max} e S_{min} dos nossos métodos e dos métodos da literatura no conjunto de teste das três ontologias (CC, FM e PB).

	F_{max}			S_{min}		
Método	CC	FM	PB	CC	FM	PB
Ingênuo	0,611	0,446	0,402	10,268	9,349	25,423
DeepGO	0,379	0,489	0,337	11,880	8,821	27,414
DeepGOPlusCNN	0,664	0,531	0,498	9,783	8,240	23,799
TALE+Transformers	0,661	0,550	0,491	9,682	8,115	23,929
ATGO	0,684	0,616	**0,547**	9,437	7,228	**22,228**
TEMPROT	**0,689**	**0,643**	0,499	**9,209**	**6,973**	23,652
DS	0,593	0,572	0,519	9,957	7,164	23,066
BLASTp	0,637	0,620	0,561	9,795	6,805	22,183
DeepGOPlus	0,677	0,619	0,553	9,515	7,090	22,648
TALE+	0,681	0,631	0,555	9,363	6,949	22,615
ATGO+	0,690	0,652	**0,589**	9,286	**6,617**	**21,233**
DS-TEMPROT	**0,692**	0,658	0,562	9,187	6,761	22,504
TEMPROT+	**0,692**	**0,662**	0,581	**9,169**	6,662	21,892

Avaliação nas Ontologias

Tabela 4: AuPRC e IAuPRC dos nossos métodos e dos métodos da literatura no conjunto de teste das três ontologias (CC, FM e PB).

	AuPRC			IAuPRC		
Método	CC	FM	PB	CC	FM	PB
Ingênuo	0,521	0,228	0,266	0,634	0,370	0,345
DeepGO	0,257	0,309	0,247	0,382	0,465	0,304
DeepGOPlusCNN	0,637	0,460	0,444	0,634	0,528	0,465
TALE+Transformers	0,613	0,444	0,477	0,706	0,549	0,469
ATGO	**0,667**	**0,623**	**0,506**	**0,724**	0,632	**0,524**
TEMPROT	0,639	0,561	0,459	0,719	**0,664**	0,483
DS	0,237	0,320	0,286	0,483	0,462	0,417
BLASTp	0,380	0,360	0,402	0,586	0,562	0,502
DeepGOPlus	0,638	0,559	0,514	0,717	0,635	0,536
TALE+	0,643	0,621	0,547	0,724	0,643	0,540
ATGO+	**0,660**	**0,650**	**0,550**	**0,731**	**0,689**	**0,571**
DS-TEMPROT	0,648	0,584	0,510	0,724	0,683	0,540
TEMPROT+	0,641	0,595	0,529	0,728	**0,689**	0,558

Resultados Principais

- "Predição de Funções Moleculares de Proteínas utilizando Aprendizado de Máquina" (Felipe Lopes de Mello, Gabriel Bianchin de Oliveira e Zanoni Dias) [30], apresentado no relatório técnico (2021).
- "Prediction of Protein Molecular Functions Using Transformers" (Felipe Lopes de Mello, Gabriel Bianchin de Oliveira, Hélio Pedrini e Zanoni Dias) [31], apresentado na *21st International Conference on Artificial Intelligence and Soft Computing* (ICAISC 2022).

Resultados Principais

- "Protein Molecular Function Annotation Based on Transformer Embeddings" (Gabriel Bianchin de Oliveira, Hélio Pedrini e Zanoni Dias) [32], apresentado na *11th Brazilian Conference on Intelligent Systems* (BRACIS 2022).
- "TEMPROT: Protein Function Annotation using Transformers Embeddings and Homology Search" (Gabriel Bianchin de Oliveira, Hélio Pedrini e Zanoni Dias), submetido para um periódico internacional (2023).

Outros Resultados

- "Deep Learning-based COVID-19 diagnostics of low-quality CT images" (Daniel Ferber, Felipe Vieira, João Dalben, Mariana Ferraz, Nicholas Sato, Gabriel Oliveira, Rafael Padilha e Zanoni Dias) [36], apresentado no *Brazilian Symposium on Bioinformatics* (BSB 2021).
- "Ensemble of Patches for COVID-19 X-ray Image Classification" (Thiago Dong Chen, Gabriel Bianchin de Oliveira e Zanoni Dias) [35], apresentado na *14th International Conference on Agents and Artificial Intelligence* (ICAART 2022).

Outros Resultados

- "Bias Assessment in Medical Imaging Analysis: a Case Study on Retinal OCT Image Classification" (Gabriel Oliveira, Lucas David, Rafael Padilha, Ana Paula da Silva, Francine de Paula, Lucas Infante, Lucio Jorge, Patricia Xavier e Zanoni Dias) [34], apresentado na *14th International Conference on Agents and Artificial Intelligence* (ICAART 2022).
- "Algorithmic Fairness Applied to the Multi-Label Classification Problem" (Ana Paula dos Santos Dantas, Gabriel Bianchin de Oliveira, Daiane Mendes de Oliveira, Hélio Pedrini, Cid Carvalho de Souza e Zanoni Dias) [33], apresentado na *18th International Conference on Computer Vision Theory and Applications* (VISAPP 2023).

Sumário

Próximas Etapas

- Aumentação de Dados:
  - Inclusão e exclusão aleatória de aminoácidos
  - Métodos generativos
  - Combinação de aumentações
- Transformers longos:
  - Adaptação do método para Transformers longos
- Classificação de termos raros:
  - *Zero-Shot Learning*
  - *Few-Shot Learning*
- Destilação de conhecimento:
  - Adaptação do método para arquiteturas mais eficientes

Referências

[1] Gene Ontology Consortium.
The Gene Ontology (GO) database and informatics resource.
Nucleic Acids Research, 32(suppl_1):D258–D261, 2004.

[2] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman.
Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.
Nucleic Acids Research, 25(17):3389–3402, 1997.

[3] Benjamin Buchfink, Klaus Reuter, and Hajk-Georg Drost.
Sensitive protein alignments at tree-of-life scale using DIAMOND.
Nature Methods, 18(4):366–368, 2021.

Referências

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin.
Attention is all you need.
In *30th Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
BERT: Pre-training of deep bidirectional transformers for language understanding.
arXiv:1810.04805, pages 1–16, 2018.

Referências

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov.
RoBERTa: A robustly optimized bert pretraining approach.
arXiv:1907.11692, pages 1–13, 2019.

[7] Iz Beltagy, Matthew E Peters, and Arman Cohan.
Longformer: The long-document transformer.
arXiv:2004.05150, pages 1–17, 2020.

[8] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost.

Referências

ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing.
arXiv:2007.06225, pages 1–29, 2021.

[9] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song.
Evaluating protein transfer learning with TAPE.
In *32nd Advances in Neural Information Processing Systems (NeurIPS)*, page 9689. NIH Public Access, 2019.

Referências

[10] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives.
Evolutionary-scale prediction of atomic level protein structure with a language model.
bioRxiv, pages 1–28, 2022.

[11] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott

Referências

Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei.
Language models are few-shot learners.
arXiv:2005.14165, pages 1–75, 2020.

[12] Noelia Ferruz, Steffen Schmidt, and Birte Höcker.
ProtGPT2 is a deep unsupervised language model for protein design.
Nature Communications, 13(1):1–10, 2022.

[13] Hanwen Xu and Sheng Wang.
ProTranslator: zero-shot protein function prediction using textual description.
In *26th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, pages 279–294, 2022.

Referências

[14] Maxat Kulmanov and Robert Hoehndorf.
DeepGOZero: Improving protein function prediction from sequence and zero-shot learning based on ontology axioms.
bioRxiv, pages 1–9, 2022.

[15] Maxat Kulmanov and Robert Hoehndorf.
DeepGOPlus: improved protein function prediction from sequence.
Bioinformatics, 36(2):422–429, 2019.

[16] Yue Cao and Yang Shen.
TALE: Transformer-based protein function Annotation with joint sequence–Label Embedding.
Bioinformatics, 37(18):2825–2833, 2021.

Referências

[17] Yi-Heng Zhu et al.
Integrating unsupervised language model with triplet neural networks for protein gene ontology prediction.
PLoS Computational Biology, 18(12):e1010793, 2022.

[18] Vladimir Gligorijević, Meet Barot, and Richard Bonneau.
deepNF: deep network fusion for protein function prediction.
Bioinformatics, 34(22):3873–3881, 2018.

[19] Jiajie Peng, Hansheng Xue, Zhongyu Wei, Idil Tuncali, Jianye Hao, and Xuequn Shang.
Integrating multi-network topology for gene function prediction using deep neural networks.
Briefings in Bioinformatics, 22(2):2096–2105, 04 2020.

Referências

[20] Maxat Kulmanov, Mohammed Asif Khan, and Robert Hoehndorf.
DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier.
Bioinformatics, 34(4):660–668, 2018.

[21] Domenico Cozzetto, Federico Minneci, Hannah Currant, and David T. Jones.
FFPred 3: feature-based function prediction for all Gene Ontology domains.
Scientific Reports, 6(1):1–11, 2016.

[22] Sayoni Das, Ian Sillitoe, David Lee, Jonathan G. Lees, Natalie L. Dawson, John Ward, and Christine A. Orengo.
CATH FunFHMMer web server: protein functional annotations using functional family assignments.
Nucleic Acids Research, 43(W1):W148–W153, 2015.

Referências

[23] Damiano Piovesan and Silvio C. E. Tosatto.
INGA 2.0: improving protein function prediction for the dark proteome.
Nucleic Acids Research, 47(W1):W373–W378, 2019.

[24] Chengxin Zhang, Peter L. Freddolino, and Yang Zhang.
COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information.
Nucleic Acids Research, 45(W1):W291–W299, 2017.

[25] Samah Fodeh, Aditya Tiwari, and Hong Yu.
Exploiting PubMed for protein molecular function prediction via NMF based multi-label classification.
In *IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 446–451. IEEE, 2017.

Referências

[26] Kyle Hippe, Sola Gbenro, and Renzhi Cao.
ProLanGO2: protein function prediction with ensemble of encoder-decoder networks.
In *11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB)*, pages 1–6. ACM, 2020.

[27] Ashish Ranjan, David Fernandez-Baca, Sudhakar Tripathi, and Akshay Deepak.
An ensemble tf-idf based approach to protein function prediction via sequence segmentation.
IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14(8):1–12, 2021.

Referências

[28] Margaret O. Dayhoff.
Atlas of Protein Sequence and Structure.
National Biomedical Research Foundation, 1972.

[29] Naihui Zhou, Yuxiang Jiang, Timothy R. Bergquist, Alexandra J. Lee, Balint Z. Kacsoh, Alex W. Crocker, Kimberley A. Lewis, George Georghiou, Huy N. Nguyen, Md Nafiz Hamid, Larry Davis, Tunca Dogan, Volkan Atalay, Ahmet S. Rifaioglu, Alperen Dalkıran, Rengul Cetin Atalay, Chengxin Zhang, Rebecca L. Hurto, Peter L. Freddolino, Yang Zhang, Prajwal Bhat, Fran Supek, José M. Fernández, Branislava Gemovic, Vladimir R. Perovic, Radoslav S. Davidović, Neven Sumonja, Nevena Veljkovic, Ehsaneddin Asgari, Mohammad R.K. Mofrad, Giuseppe Profiti, Castrense Savojardo, Pier Luigi Martelli, Rita Casadio, Florian Boecker, Heiko Schoof, Indika Kahanda, Natalie Thurlby, Alice C. McHardy, Alexandre

Referências

Renaux, Rabie Saidi, Julian Gough, Alex A. Freitas, Magdalena Antczak, Fabio Fabris, Mark N. Wass, Jie Hou, Jianlin Cheng, Zheng Wang, Alfonso E. Romero, Alberto Paccanaro, Haixuan Yang, Tatyana Goldberg, Chenguang Zhao, Liisa Holm, Petri Törönen, Alan J. Medlar, Elaine Zosa, Itamar Borukhov, Ilya Novikov, Angela Wilkins, Olivier Lichtarge, Po-Han Chi, Wei-Cheng Tseng, Michal Linial, Peter W. Rose, Christophe Dessimoz, Vedrana Vidulin, Saso Dzeroski, Ian Sillitoe, Sayoni Das, Jonathan Gill Lees, David T. Jones, Cen Wan, Domenico Cozzetto, Rui Fa, Mateo Torres, Alex Warwick Vesztrocy, Jose Manuel Rodriguez, Michael L. Tress, Marco Frasca, Marco Notaro, Giuliano Grossi, Alessandro Petrini, Matteo Re, Giorgio Valentini, Marco Mesiti, Daniel B. Roche, Jonas Reeb, David W. Ritchie, Sabeur Aridhi, Seyed Ziaeddin Alborzi, Marie-Dominique Devignes, Da Chen Emily Koo, Richard Bonneau, Vladimir Gligorijević, Meet Barot, Hai Fang,

Referências

Stefano Toppo, Enrico Lavezzo, Marco Falda, Michele Berselli, Silvio C.E. Tosatto, Marco Carraro, Damiano Piovesan, Hafeez Ur Rehman, Qizhong Mao, Shanshan Zhang, Slobodan Vucetic, Gage S. Black, Dane Jo, Erica Suh, Jonathan B. Dayton, Dallas J. Larsen, Ashton R. Omdahl, Liam J. McGuffin, Danielle A. Brackenridge, Patricia C. Babbitt, Jeffrey M. Yunes, Paolo Fontana, Feng Zhang, Shanfeng Zhu, Ronghui You, Zihan Zhang, Suyang Dai, Shuwei Yao, Weidong Tian, Renzhi Cao, Caleb Chandler, Miguel Amezola, Devon Johnson, Jia-Ming Chang, Wen-Hung Liao, Yi-Wei Liu, Stefano Pascarelli, Yotam Frank, Robert Hoehndorf, Maxat Kulmanov, Imane Boudellioua, Gianfranco Politano, Stefano Di Carlo, Alfredo Benso, Kai Hakala, Filip Ginter, Farrokh Mehryary, Suwisa Kaewphan, Jari Björne, Hans Moen, Martti E.E. Tolvanen, Tapio Salakoski, Daisuke Kihara, Aashish Jain, Tomislav Šmuc, Adrian Altenhoff, Asa Ben-Hur, Burkhard Rost, Steven E. Brenner,

Referências

Christine A. Orengo, Constance J. Jeffery, Giovanni Bosco, Deborah A. Hogan, Maria J. Martin, Claire O'Donovan, Sean D. Mooney, Casey S. Greene, Predrag Radivojac, and Iddo Friedberg. **The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens.** *Genome Biology*, 20(1):244, 2019.

[30] Felipe Lopes Mello, Gabriel Bianchin Oliveira, and Zanoni Dias. **Predição de Funções Moleculares de Proteínas utilizando Aprendizado de Máquina.** Technical Report IC-PFG-21-44, Institute of Computing, University of Campinas, December 2021.

Referências

[31] Felipe Lopes Mello, Gabriel Bianchin Oliveira, Helio Pedrini, and Zanoni Dias.
Prediction of Protein Molecular Functions Using Transformers.

In *21st International Conference on Artificial Intelligence and Soft Computing (ICAISC)*, pages 379–387. Springer, 2022.

[32] Gabriel Bianchin Oliveira, Helio Pedrini, and Zanoni Dias.
Protein Molecular Function Annotation Based on Transformer Embeddings.
In *11th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 210–220. Springer, 2022.

Referências

[33] Ana Paula Santos Dantas, Gabriel Bianchin Oliveira, Daiane Mendes Oliveira, Helio Pedrini, Cid Carvalho Souza, and Zanoni Dias.
Algorithmic Fairness Applied to the Multi-Label Classification Problem.
In *18th International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 1–8. SciTePress, 2023.

[34] Gabriel Oliveira, Lucas David, Rafael Padilha, Ana Paula da Silva, Francine de Paula, Lucas Infante, Lucio Jorge, Patricia Xavier, and Zanoni Dias.
Bias Assessment in Medical Imaging Analysis: a Case Study on Retinal OCT Image Classification.
In *14th International Conference on Agents and Artificial Intelligence (ICAART)*, pages 574–580. SciTePress, 2022.

Referências

[35] Thiago Dong Chen, Gabriel Bianchin Oliveira, and Zanoni Dias. **Ensemble of Patches for COVID-19 X-ray Image Classification.** In *14th International Conference on Agents and Artificial Intelligence (ICAART)*, pages 561–567. SciTePress, 2022.

[36] Daniel Ferber, Felipe Vieira, João Dalben, Mariana Ferraz, Nicholas Sato, Gabriel Oliveira, Rafael Padilha, and Zanoni Dias. **Deep Learning-based COVID-19 diagnostics of low-quality CT images.** In *Brazilian Symposium on Bioinformatics (BSB)*, pages 69–80. Springer, 2021.

Anotações de Funções de Proteínas utilizando Processamento de Linguagem Natural e Alinhamento Local de Proteínas

Gabriel Bianchin de Oliveira

Orientador: Prof. Dr. Zanoni Dias
Coorientador: Prof. Dr. Hélio Pedrini

21 de Março de 2023

Instituto de Computação
Universidade Estadual de Campinas