

Extração de características e classificação de sequências de RNAs

Fabrício M. Lopes e Matheus H. Pimenta-Zanon

fabricao@utfpr.edu.br

UTFPR-CP

Grupo de Pesquisa em Bioinformática e
Reconhecimento de Padrões

bioinfo-cp@utfpr.edu.br



EPB 2022

III ESCOLA PARANAENSE
DE BIOINFORMÁTICA



Escola Paranaense de Bioinformática (EPB)

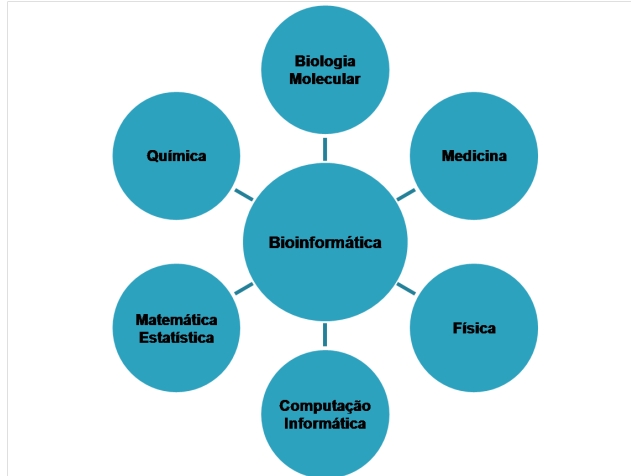
08-AGO-2022

Organização

- 1 Bioinformática
- 2 Reconhecimento de Padrões
- 3 Redes Complexas
- 4 Aplicação: BASiNET

Bioinformática

Introdução - Bioinformática



Introdução - Definição

- A bioinformática trata da **compreensão de como a vida funciona**, é uma ciência guiada por hipóteses;
- Está na **interseção entre a ciência da computação e as ciências da vida**;
- Trata da **integração de temas biológicos com a ajuda de ferramentas informáticas**, bases de dados biológicos e obtenção de novos conhecimentos a partir dessas informações.

Introdução - Características

- É uma área emergente de investigação altamente **interdisciplinar**;
- Trata da **análise computacional de informações biológicas**: genes, genomas, proteínas, células, sistemas ecológicos, informação médica, etc.;
- Trata do **desenvolvimento e uso de sistemas computacionais** para a análise, interpretação, simulação e predição de sistemas biológicos.

Surgimento da Bioinformática

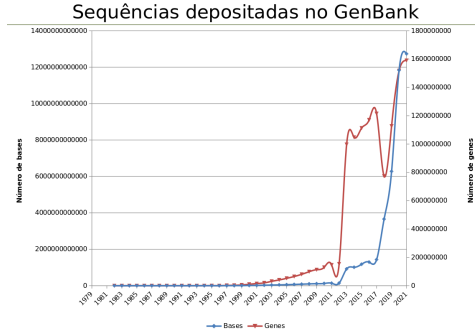
- A **bioinformática** surgiu com o propósito do estudo da biologia molecular e bioquímica, em larga escala;
- Em termos globais, em 1988 foi lançado o primeiro banco de dados público contendo sequências de DNA, o ***NCBI - National Center for Biotechnology Information***;
- A bioinformática chegou ao Brasil em 1999, com o **sequenciamento completo do DNA** da bactéria *Xylella fastidiosa*, patógeno que gera prejuízos à cultura de cítricos.

Surgimento da Bioinformática

- Desde então, o **surgimento de enormes bancos de dados** internacionais com dados biológicos/moleculares tornaram os dados acessíveis e comparáveis;
- Desenvolvimento de **algoritmos** para o alinhamento de sequências gênicas, para a predição de genes e proteínas, entre outras metodologias estatísticas e computacionais.

Crescimento da Bioinformática

From 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months.¹



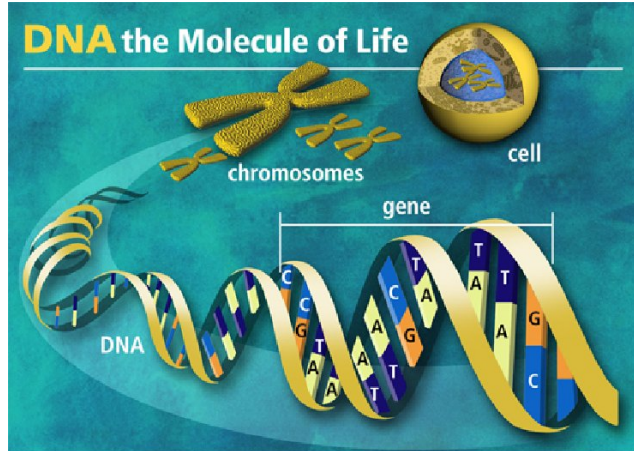
¹ <https://www.ncbi.nlm.nih.gov/genbank/release/current/>

Biologia Sistêmica

- Uma aplicação recente que tem recebido muita atenção pela comunidade de pesquisa é a **biologia sistêmica** (*Systems Biology*);
- Trata do **desenvolvimento e uso de sistemas computacionais** para a análise, interpretação, simulação e predição de sistemas biológicos;
- **Desafio**: investigar, identificar e descrever de que forma as componentes interagem como um sistema e, como esse sistema funciona.

Reconhecimento de Padrões

Origem dos dados



Como extrair informação dos dados ?

```
ATGTATCCAGGTAGTGGACGTTACACCTACAACAACGCTGGTGGTAATAATGGCTACCAA  
CGGCCCATGGCTCCTCCACCTAACAGCAGTATGGACAGCAATATGGTCAGCAATATGAA  
CAGCAGTATGGACAGCAATATGGGCAACAAAATGATCAGCAATTCAGTCAGCAATATGCT  
CCACCACCAGGTCCCTCCCCCTATGGCTTATAACAGGCCTGTGTATCCCCCCCCCTCAATTC  
CAGCAGGAACAGGCAAGGCACAATTAAGCAACGGCTACAACAATCCTAATGTAAACGCA  
TCCAATATGTACGGTCCACCCAGAAATATGTCATTACCTCCACCTCAAACACAAACTATT  
CAAGGTACAGACCAACCTTATCAGTATTCTCAATGTACTGGGCGTAGAAAGGCTTTGATT  
ATCGGTATAAACTACATAGGTTCAAAAAATCAACTGCGTGGTTGTATCAATGATGCTCAT  
AACATCTTCAACTTTTTGACTAATGGGTACGGTTACAGTTCAGATGACATTGTATATTA  
ACTGATGATCAGAACGATTTGGTCAGGGTTCCTACTAGGGCTAATATGATTAGGGCCATG  
CAATGGTTGGTCAAGGATGCGCAACCCAATGATTCTTTGTTCTTCATTATTCTGGACAT  
GGTGGCCAACTGAAGATTTGGATGGGGACGAAGAAGATGGGATGGATGATGTTATATAT  
CCGGTCGATTTTCAAACTCAAGGGCCAATTATCGACGATGAAATGCACGATATAATGGTG  
AAGCCCTTACAACAAGGTGTTAGACTAACAGCATTGTTTGA CTCTTGTCATTTCGGGTACA  
GTGTTGGATCTTCCATATACCTATTCTACTAAGGGTATTATTAAGGAGCCCAATATTTGG  
AAGGATGTTGGCCAAGATGGCCTGCAAGCAGCTATTTTCATATGCCACAGGAAACAGGGCT  
GCTTTGATTGGTTCTTTAGGTTCTATATTCAAGACCGTTAAGGGAGGTATGGGCAATAAT  
GTGGATAGAGAACGCGTGAGACAGATCAAATTCAGCAGCAGATGTTGTTATGTTATCA  
GGTTCGAAGGATAAATCAAACCTTCTGCAGATGCTGTGCAAGATGGGCAAAAATACAGGTGCA  
ATGTCCCACGCCTTCATCAAGGTTATGACTTTTACAACCACAGCAATCATATTTATCTCTT  
TTACAGAACATGAGGAAAGAATTGGCTGGTAAGTATTCTCAAAAACCACAATTATCATCG  
TCACACCCTATTGACGTAAATCTGCAATTTATTATGTAG
```

Conhecimento sobre o domínio dos dados

		Second letter				
		U	C	A	G	
First letter	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Stop UGG Trp	U C A G
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	U C A G
	A	AUU } AUC } Ile AUA } AUG Met	ACU } ACC } Thr ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	U C A G
	G	GUU } GUC } Val GUA } GUG }	GCU } GCC } Ala GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } Gly GGA } GGG }	U C A G

2

²<https://courses.lumenlearning.com/wm-biology1/chapter/reading-codons/>

Conhecimento sobre o domínio da solução

Definição de Reconhecimento de Padrões:

- *“É uma área de pesquisa que tem por objetivo a classificação de objetos (padrões) em um número de categorias ou classes”, Theodoridis e Koutroumbas [Theodoridis and Koutroumbas, 2008].*
- *“O ato de observar os dados brutos e tomar uma ação baseada na categoria de um padrão”, Duda et al. [Duda et al., 2001].*

Classificadores

- **Classificadores:** utilizados para classificar ou descrever padrões ou objetos a partir de um conjunto de propriedades ou características.
- Existem essencialmente dois casos particulares de reconhecimento de padrões:
 - **Classificação supervisionada.**
 - **Classificação não supervisionada.**

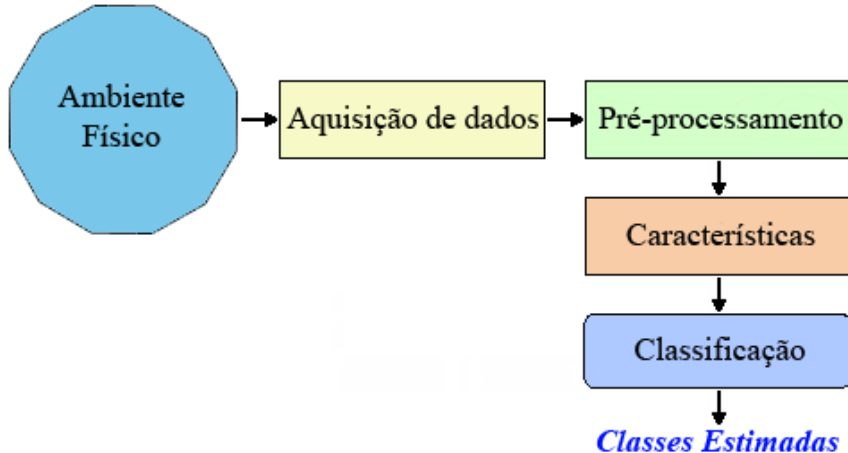
Classificação supervisionada

- Selecionam-se amostras representativas para cada uma das classes que se deseja classificar.
- Conhecemos o padrão e classes que estamos procurando.
- Também conhecido como Aprendizado supervisionado.

Classificação não supervisionada

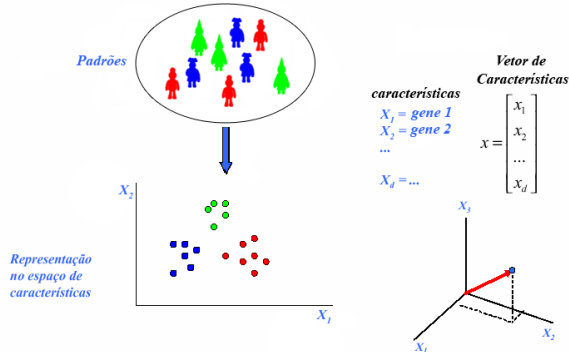
- Não conhecemos o padrão, nem o número total de classes a serem encontradas durante a classificação.
- Também conhecido como aprendizado não supervisionado ou análise de agrupamentos (**clusters**).
- O conjunto de dados é particionado em grupos, baseados em características específicas, tais que os pontos dentro de um grupo (cluster) sejam mais similares do que os pontos de outros grupos.
- Pode ajudar compreender funções de muitos genes para os quais não há informações disponíveis, Jiang et al. [Jiang et al., 2004].

Etapas do Reconhecimento de Padrões

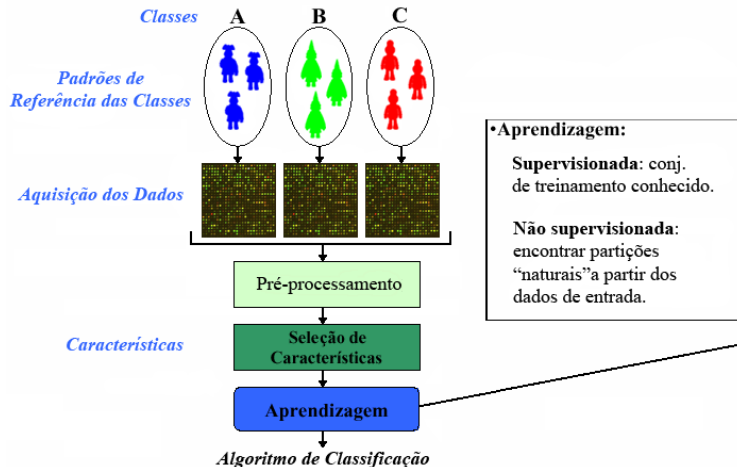


Características

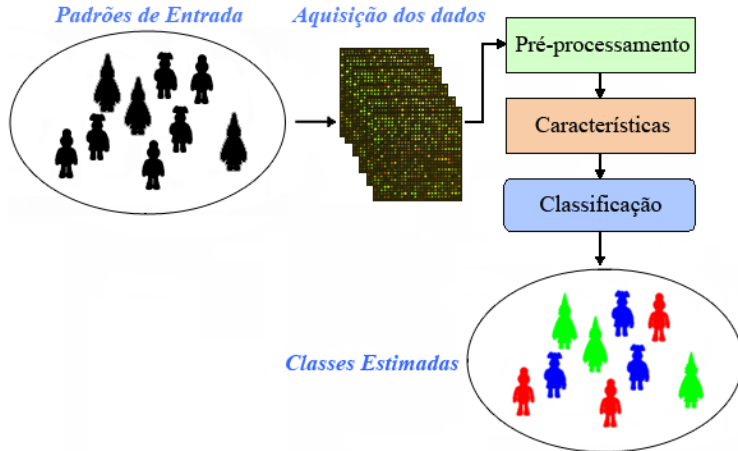
- **Característica** ou **Atributo**: dado extraído de uma amostra por meio de medida e/ou processamento.



Introdução - Treinamento



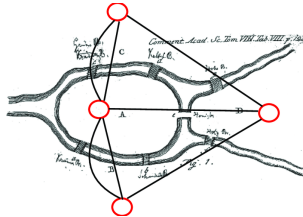
Classificação (Generalização)



Redes Complexas

Redes Complexas

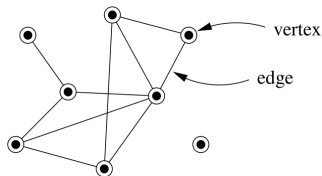
- “Pode-se descobrir se é ou não possível atravessar cada ponte exatamente uma vez”?
- Euler provou que o problema não tem solução, “Königsberg Bridge Problem” é considerada como o início da teoria das redes (1735) [Paoletti, 2011].



[Karimi, 2015]

Redes Complexas

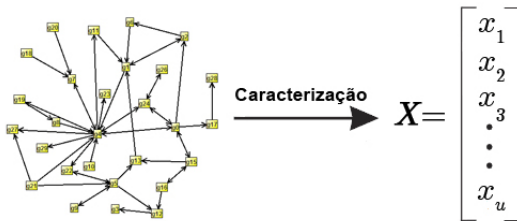
- Uma **rede** é um grafo, com um conjunto de **vértices** interligados através de **arestas**.
- A teoria de redes complexas pode ser entendida como o estudo das redes (estrutura e função), i.e. das relações de seus vértices [Newman, 2003].



[Newman, 2003]

Redes Complexas

- Redes complexas possuem **topologias** distintas e **propriedades** bem definidas [Boccaletti et al., 2006, Costa et al., 2007].
- As redes complexas podem ser caracterizadas em termos de **medidas específicas**.

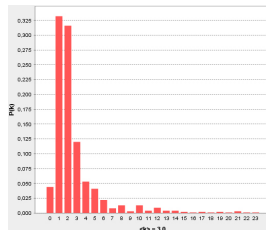
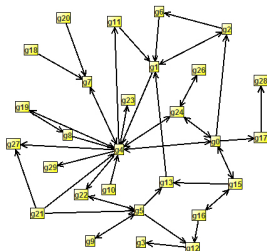


Medidas de Redes Complexas

- É possível extrair medidas que caracterizem a topologia da rede [Boccaletti et al., 2006, Costa et al., 2007], tais como:
 - Average Betweenness Centrality
 - Cluster Coefficient
 - Average Path Length
 - Assortativity
 - Average, maximum, minimum degree
 - Frequency of motifs with size 3 and 4, etc.

Redes Complexas

- Modelos teóricos de redes complexas podem ser considerados para a definição da topologia de redes biológicas [Lopes et al., 2011, Costa et al., 2008]:
 - small-world** propostas por Watts e Strogatz (**WS**) [Watts and Strogatz, 1998];
 - scale-free** propostas por Barabási-Albert (**BA**) [Barabási and Albert, 1999].



Redes BA - Aplicações

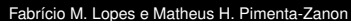
- O modelo de redes *scale-free* e suas propriedades têm sido utilizado para simular e descrever o comportamento de redes biológicas [Barabási, 2009].
- Muitas das redes biológicas conhecidas apresentam uma estrutura *scale-free* [Albert, 2005, Khanin and Wit, 2006, Costa et al., 2008, Lopes et al., 2014], implicando que a distribuição das relações entre os genes (k , grau) é irregular.

BASiNET - BiologicAI Sequences NETwork: a case study on coding and non-coding RNAs identification

BASiNET

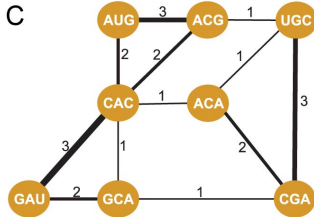
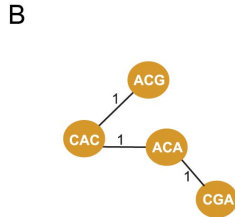
- É um método [Ito et al., 2018] de extração de características para a classificação de sequências biológicas com base nas medidas de redes complexas:
 - Considera a vizinhança dos códons para a construção de uma rede que mapeia os padrões de vizinhança
 - Extrai medidas topológicas para a caracterização da rede e composição de vetor de características
 - O vetor de características é usado na classificação das sequências

- Explora padrões existentes nas sequências de entrada considerando três etapas principais



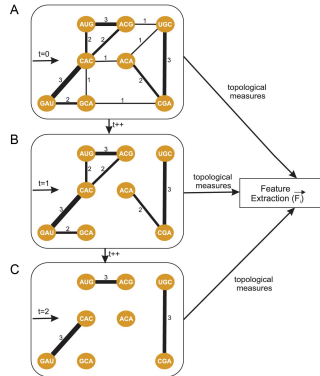
Mapping

- Mapeamento da sequência em uma rede (grafo) com arestas ponderadas



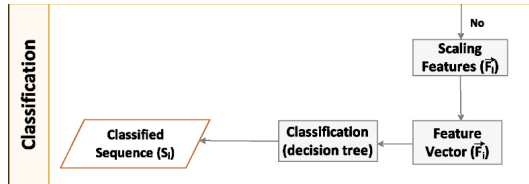
Feature extraction

- Extração de características iterativo, considerando os pesos das arestas para a extração de medidas topológicas



Classification

- A classificação é realizada de forma independente a partir do vetor de características reescalado



Pacote R

- O BASiNET foi implementado em R e está disponível livremente para uso³

BASiNET: Classification of RNA Sequences using Complex Network Theory

It makes the creation of networks from sequences of RNA, with this is done the abstraction of characteristics of these networks with a methodology of threshold for the purpose of making a classification between the classes of the sequences. There are four data present in the "BASiNET" package, "sequences", "sequences2", "sequences-predict" and "sequences2-predict" with 11, 10, 11 and 11 sequences respectively. These sequences were taken from the data set used in the article (LI, Aimin; ZHANG, Junying; ZHOU, Zhongyin, 2014) <[doi:10.1186/1471-2105-15-311](https://doi.org/10.1186/1471-2105-15-311)>, these sequences are used to run examples. The BASiNET was published on Nucleic Acids Research, (ITO, Eric; KATAHIRA, Isaque; VICENTE, Fábio; PEREIRA, Felipe; LOPES, Fabricio, 2018) <[doi:10.1093/nar/gky462](https://doi.org/10.1093/nar/gky462)>.

Version: 0.0.4
Depends: R (≥ 3.4.0)
Imports: [igraph](#), [BiocSings](#), [RWeka](#), [randomForest](#), [rmcfs](#), grDevices, graphics, stats, [rJava](#)
Suggests: [knitr](#), [rmarkdown](#)
Published: 2018-10-02
Author: Eric Augusto Ito
Maintainer: Eric Augusto Ito <ericaugustoito@hotmail.com>
License: [GPL-3](#)
NeedsCompilation: no
CRAN checks: [BASiNET results](#)

Downloads:

Reference manual: [BASiNET.pdf](#)
Vignettes: [Classification of mRNA and lncRNA sequences](#)
Package source: [BASiNET_0.0.4.tar.gz](#)
Windows binaries: r-devel: [BASiNET_0.0.4.zip](#), r-release: [BASiNET_0.0.4.zip](#), r-oldrel: [BASiNET_0.0.4.zip](#)
macOS binaries: r-release (arm64): [BASiNET_0.0.4.igz](#), r-release (x86_64): [BASiNET_0.0.4.igz](#), r-oldrel: [BASiNET_0.0.4.igz](#)
Old sources: [BASiNET archive](#)

Linking:

Please use the canonical form <https://CRAN.R-project.org/package=BASiNET> to link to this page.

³<https://cran.r-project.org/web/packages/BASiNET/>

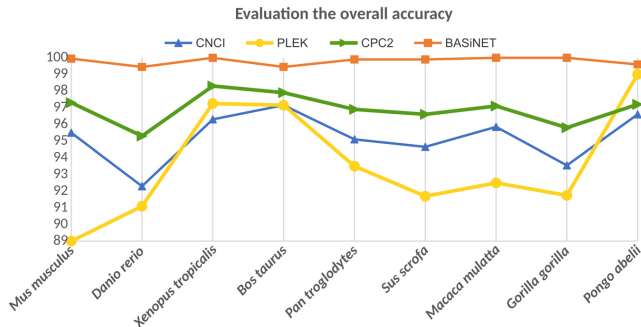
Sequências adotadas

Dois datasets foram adotados com diferentes espécies:

- Um composto por nove espécies de vertebrados, com duas classes mRNA e ncRNA, adotado pelo método PLEK [Li et al., 2014]
- Outro composto por seis espécies (quatro vertebrados, uma planta e um nematóide), com três classes: mRNA, sncRNA e lncRNA adotado pelo método CPC2 [Kang et al., 2017]

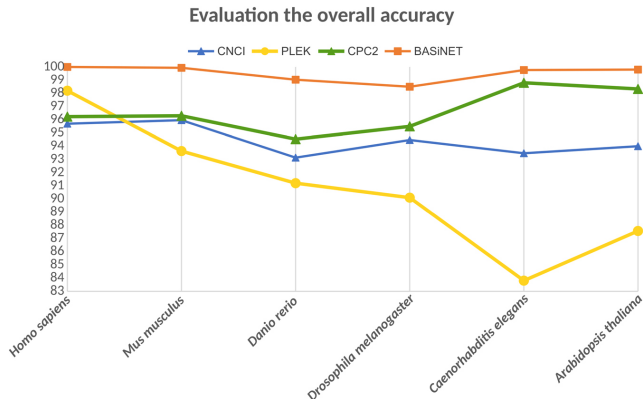
Resultados - Dataset PLEK

- Considerando o dataset (PLEK), os resultados foram superiores aos métodos concorrentes e com menor variação, mostrando assertividade e robustez do BASiNET



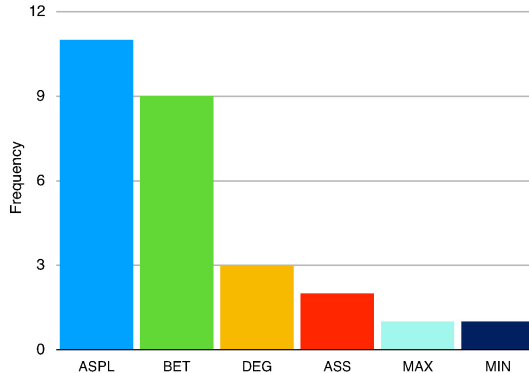
Resultados - Dataset CPC2

- Considerando o dataset (CPC2), os resultados reforçam a superioridade com relação aos métodos concorrentes e a robustez do BASiNET



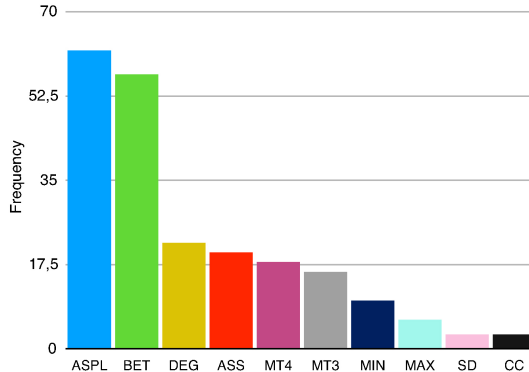
Discussão

- As características extraídas que foram consideradas na classificação, foram analisadas considerando o dataset (PLEK).



Discussão

- As características extraídas que foram consideradas na classificação, foram analisadas considerando o dataset (CPC2).



Discussão

- Considerando os dois experimentos, algumas medidas topológicas se destacaram: ASPL, BET, DEG e ASS.
- Um caminho mínimo médio (ASPL) é a média dos caminhos mais curtos entre todos os pares de vértices (codons) da rede.

Discussão

- O betweenness (BET) quantifica a relevância de um vértice em relação a todos os caminhos da rede, i.e. um vértice com uma conectividade.
- O DEG quantifica do grau médio da rede (quantidade média de conexões).
- O ASS quantifica a tendência dos vértice se ligarem a outros vértices semelhantes de alguma forma, tais como o grau do nó.

Discussão

- ASPL e DEG estão diretamente associadas à conectividade (grau) e à distância (ASPL) entre os vértices da rede.
- BET e ASS estão associadas à presença de subestruturas, como a ligação de vértices semelhantes (ASS) e centralidade (BET).

Conclusão

- BASiNET é um método de extração de características para classificação de sequências biológicas (RNAs) baseado em redes complexas e suas medidas topológicas.
- As sequências são mapeadas e representadas por meio de redes complexas.
- As medidas formam um vetor de características que é utilizado para classificar as sequências.

Conclusão

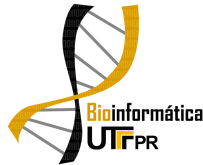
- O método foi aplicado em dois conjuntos de dados apresentados nos métodos PLEK e CPC2. Os resultados do BASiNET foram comparados com os métodos CNCI[Sun et al., 2013], PLEK[Li et al., 2014] e CPC2[Kang et al., 2017].
- Os resultados de acurácia do BASiNET comparados aos outros métodos, mostraram que o BASiNET superou os outros em todos os conjuntos de dados e com menor variação (robustez).
- O BASiNET foi implementado em código aberto (linguagem R) e o programa está disponível livremente em <https://cran.r-project.org/package=BASiNET>.

<https://groups.google.com/g/bioinfo-news>

943 participantes!

Participe!

PPGBIOINFO - Mestrado



EDITAL DE SELEÇÃO 2022-2 - Mestrado

Inscrições abertas até 15/08/2022

[http:
//www.utfpr.edu.br/cursos/coordenacoes/stricto-sensu/
ppgbioinfo/editais/edital-de-selecao-2022-2](http://www.utfpr.edu.br/cursos/coordenacoes/stricto-sensu/ppgbioinfo/editais/edital-de-selecao-2022-2)

PPGAB - Doutorado



EDITAL DE SELEÇÃO 2022 - Doutorado




Inscrições em fluxo contínuo!

[http://www.utfpr.edu.br/cursos/coordenacoes/
stricto-sensu/ppgab/editais](http://www.utfpr.edu.br/cursos/coordenacoes/stricto-sensu/ppgab/editais)




Obrigado!

fabricao@utfpr.edu.br




Referências Bibliográficas I

-  Albert, R. (2005).
Scale-free networks in cell biology.
J Cell Sci 118, 4947–4957.
-  Barabási, A.-L. (2009).
Scale-Free Networks: A Decade and Beyond.
Science 325, 412–413.
-  Barabási, A.-L. and Albert, R. (1999).
Emergence of scaling in random networks.
Science 286, 509–512.




Referências Bibliográficas II

-  Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. and Hwang, D. U. (2006).
Complex networks: Structure and dynamics.
Physics Reports 424, 175–308.
-  Costa, L. d. F., Rodrigues, F. A. and Cristino, A. S. (2008).
Complex networks: the key to systems biology.
Genetics and Molecular Biology 31, 591–601.
-  Costa, L. d. F., Rodrigues, F. A., Travieso, G. and Villas-Boas, P. R. (2007).
Characterization of complex networks: a survey of measurements.
Advances in Physics 56, 167–242.




Referências Bibliográficas III

-  Duda, R., Hart, P. and Stork, D. (2001).
Pattern classification.
Pattern Classification and Scene Analysis: Pattern Classification, 2nd edition,
Wiley.
-  Ito, E. A., Katahira, I., Vicente, F. F., Pereira, L. P. and Lopes, F. M. (2018).
BASiNET - Biological Sequences NETWORK: a case study on coding and
non-coding RNAs identification.
Nucleic Acids Research , gky462.
-  Jiang, D., Tang, C. and Zhang, A. (2004).
Cluster Analysis for Gene Expression Data: A Survey.
IEEE TKDE 16, 1370–1386.




Referências Bibliográficas IV

-  Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L. and Gao, G. (2017).
CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features.
Nucleic Acids Research 45, W12–W16.
-  Karimi, F. (2015).
Tightly Knit - Spreading processes in empirical temporal networks.
PhD thesis, Umeå University.
-  Khanin, R. and Wit, E. (2006).
How Scale-Free Are Biological Networks.
Journal of Computational Biology 13, 810–818.



Referências Bibliográficas V

-  Li, A., Zhang, J. and Zhou, Z. (2014).
PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme.
BMC Bioinformatics 15, 311.
-  Lopes, F. M., Cesar-Jr, R. M. and Costa, L. d. F. (2011).
Gene Expression Complex Networks: Synthesis, Identification, and Analysis.
Journal of Computational Biology 18, 1353–1367.
-  Lopes, F. M., Jr., D. C. M., Barrera, J. and Jr., R. M. C. (2014).
A feature selection technique for inference of graphs from their known topological properties: Revealing scale-free gene regulatory networks.
Information Sciences 272, 1–15.

Referências Bibliográficas VI

-  Newman, M. E. J. (2003).
The Structure and Function of Complex Networks.
SIAM Review 45, 167–256.
-  Paoletti, T. (2011).
Leonard Euler's solution to the Königsberg bridge problem.
Convergence 1.
-  Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., Liu, Y., Chen, R. and Zhao, Y. (2013).
Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts.
Nucleic Acids Research 41, e166–e166.

Referências Bibliográficas VII

-  Theodoridis, S. and Koutroumbas, K. (2008).
Pattern Recognition.
4th edition, Academic Press, USA.
-  Watts, D. J. and Strogatz, S. H. (1998).
Collective dynamics of small-world networks.
Nature 393, 440–442.