Minicurso: Ferramenta rSWeeP – método para representação vetorial de sequências biológicas.

Camila P Perico
Escola Paranaense de Bioinformática – 2024

Universidade Federal do Paraná

6–8 de Novembro de 2024

Organização do Minicurso

1. O que são métodos livres de alinhamento (AF)
   - problemas dos métodos baseados em alinhamento
   - tipos de métodos AF
2. SWeeP
   - descrição do método
   - pacote rSWeeP
3. **Aplicações e Prática com rSWeeP**
   - Aplicação em proteoma mitocondrial e quickStart
   - Aplicação em SARS-CoV-2 e Prática 1
   - Parametrização
   - Aplicação em proteomas bacterianos
4. Conclusão

Métodos livres de Alinhamento

Qual o grande desafio atual da bioinformática (e outras áreas)?

Grande volume de dados (**big data**)

Novas tecnologias $\rightarrow$ mais dados coletados

- Grande número de sequências biológicas
- Sequências biológicas longas (genomas, metagenomas etc)
- Transcriptomas de milhares de células (scRNAseq)
- entre outros

Desafios na análise de sequências biológicas

Desafios:

- Dificuldade atual em analisar grandes quantidades de dados moleculares
- **Alto custo computacional** para realizar **alinhamento** (método tradicional)

Soluções:

- Métodos **livres de alinhamento** vêm se mostrando vantajosos para Big data[1]
- Particularmente métodos de **representação vetorial** de sequências biológicas possibilitam usar técnicas de **ML** e aplicações mais efetivas em bioinformática[2]

---

[1] Asgari and Mofrad (2015) Continuous distributed representation of biological sequences for deep proteomics and genomics. PloS one, 10(11).

[2] Randhawa, et al. (2020) Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: Covid-19 case study. Plos one, 15(4)
Leimeister, et al. (2019). Prot-spam: Fast alignment-free phylogeny reconstruction based on whole-proteome sequences. GigaScience, 8(3)
Zhang, et al. (2017) Viral phylogenomics using an alignment-free method: A three-step approach to determine optimal length of k-mer. Scientific reports, 7(1)

Porque usar métodos livres de alinhamento (AF)?

5 casos em que métodos de alinhamento falham (e AF não)[3]:

1. Alinhamentos pressupõem segmentos homólogos linearmente arranjados.
   Mundo real: genomas são **mosaicos**.
   Alinhamentos são sensíveis a alta taxa mutacional, recombinação, duplicações, transferência horizontal de genes, ganho/perda de genes

2. acurácia dos alinhamentos cai rapidamente conforme **diminui a identidade**

3. alinhamentos são muito **custosos em RAM** e tempo de processamento

4. limitação de escala
   múltiplos genomas $\rightarrow$ impossível de resolver em tempo realista

5. Pressupostos na **parametrização**: escolha da matriz de substituição, penalidade de gaps, limites estatísticos etc. Diferentes parâmetros $\rightarrow$ **resultados diferentes**

[3] Zielezinski, et al. (2017). Alignment-free sequence comparison: benefits, applications, and tools. Genome biology, 18

Métodos livres de alinhamento (AF)

Definição *Métodos AF podem ser definidos como qualquer método para quantificar a similaridade/dissimilaridade de sequências que não use ou produza alinhamento em nenhuma etapa da aplicação do algoritmo* [4]

Métodos AF **não** enfrentam os mesmos desafios dos métodos baseados em alinhamentos
Há 2 classes principais:

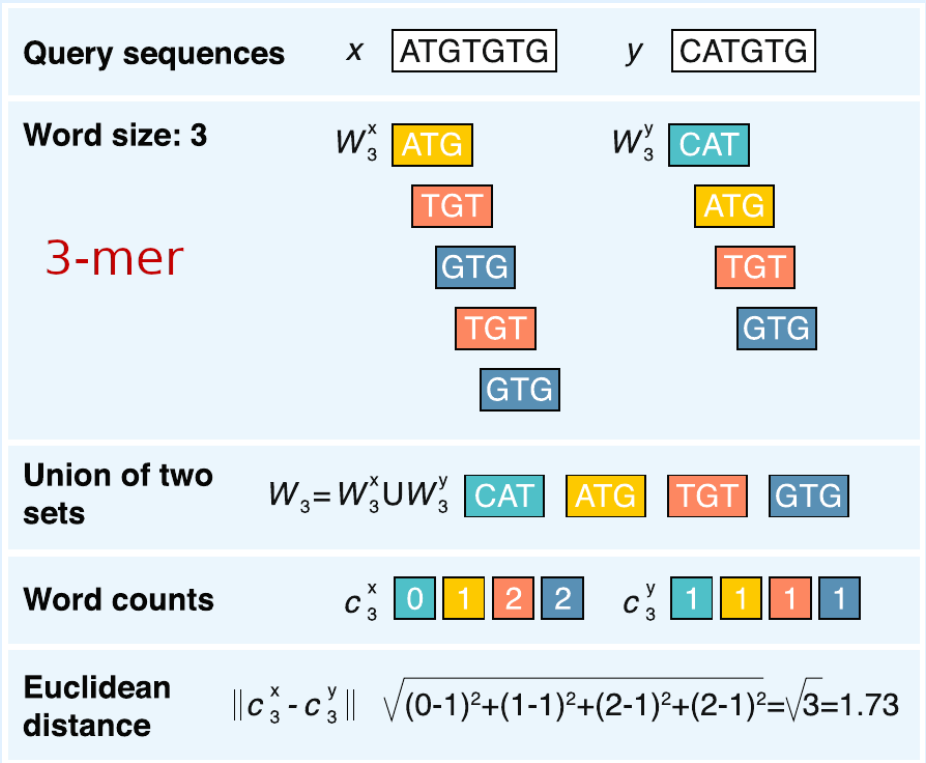
1. *word-based* ou frequência de palavras (k-mers): calculam similaridade/dissimilaridade baseada na frequência de subsequências
2. baseados em teoria da informação: avaliam o conteúdo informacional completo entre duas sequências

******word-based* se mostram superiores na detecção de padrões complexos

[4] Zielezinski, et al. (2017). Alignment-free sequence comparison: benefits, applications, and tools. Genome biology, 18

Métodos livres de alinhamento (AF)

Os métodos AF estão se tornando cada vez mais presentes na literatura[5]



word-based ou frequência de palavras (k-mers)

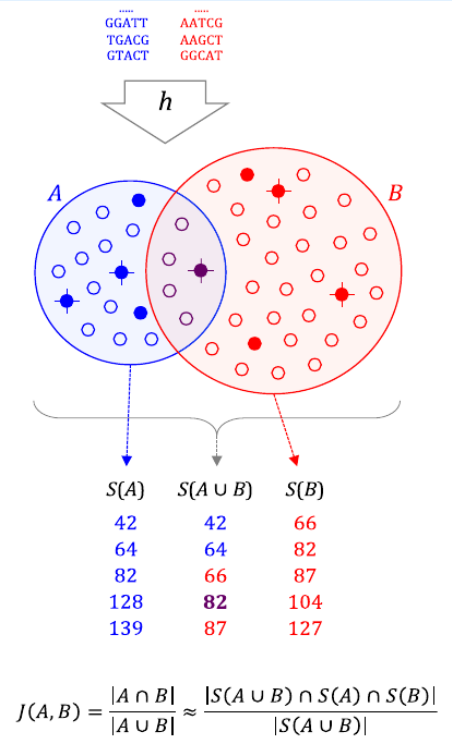
Considere duas sequências de DNA:
`ATGTGTG` e `CATGTG`
A distância euclidiana entre elas é 1.73

Adaptado de: Zielezinski, et al (2017)

[5] Forsdyke (2019). Success of alignment-free oligonucleotide (k-mer) analysis confirms relative importance of genomes not genes in speciation and phylogeny. Biological Journal of the Linnean Society, 128(2)
Zielezinski, et al. (2017). Alignment-free sequence comparison: benefits, applications, and tools. Genome biology, 18

Alguns importantes métodos AF

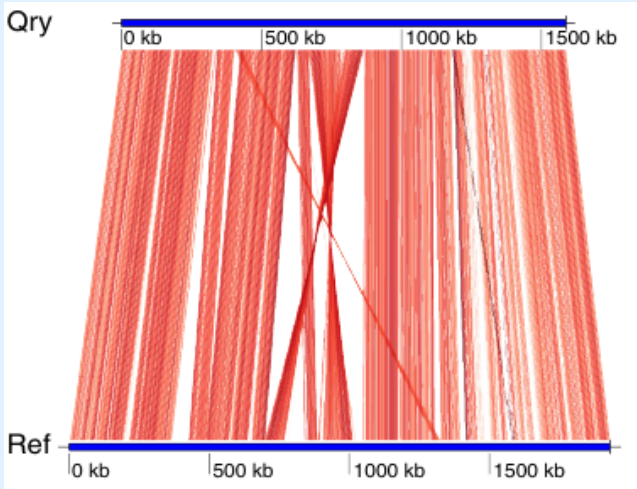
Mash → técnica baseada em estatística de conjuntos (índice de Jaccard) de k-mers



Ondov, et al (2016). Mash: fast genome and metagenome distance estimation using minhash. Genome biology, 17:1–14.

fastANI (Average Nucleotide Identity) → **NCBI**

fragmentação de sequências de genoma, seguida de pesquisa de sequência de nucleotídeos, alinhamento e cálculo de identidade.



Jain, et al (2018). High throughput ani analysis of 90k prokaryotic genomes reveals clear species boundaries. Nature communications,9(1):5114.

Alguns importantes métodos AF

Mash → técnica baseada em estatística de conjuntos (índice de Jaccard) de k-mers

Ondov, et al (2016). Mash: fast genome and metagenome distance estimation using minhash. Genome biology, 17:1–14.

fastANI (Average Nucleotide Identity) → **NCBI**

fragmentação de sequências de genoma, seguida de pesquisa de sequência de nucleotídeos, alinhamento e cálculo de identidade.

Jain, et al (2018). High throughput ani analysis of 90k prokaryotic genomes reveals clear species boundaries. Nature communications,9(1):5114.

** ambos baseados em nucleotídeos
** ambos retornam um índice de similaridade (matriz de distâncias), a partir de FASTAS fornecidos
** ambos podem analisar genomas completos

O Método SWeeP

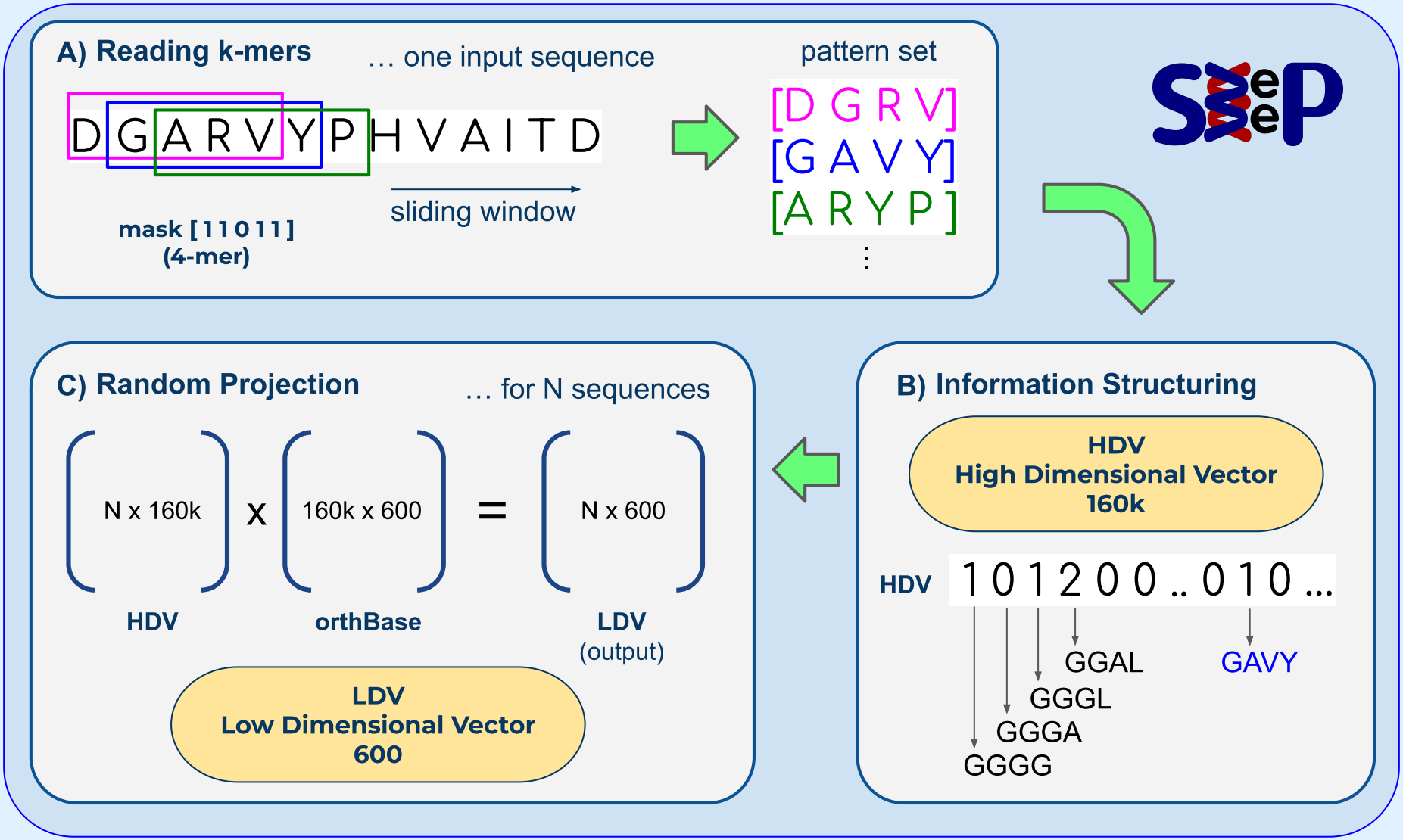
O que é SWeeP?

SWeeP (Spaced Words Projection)
método de representação vetorial de sequências biológicas[6]

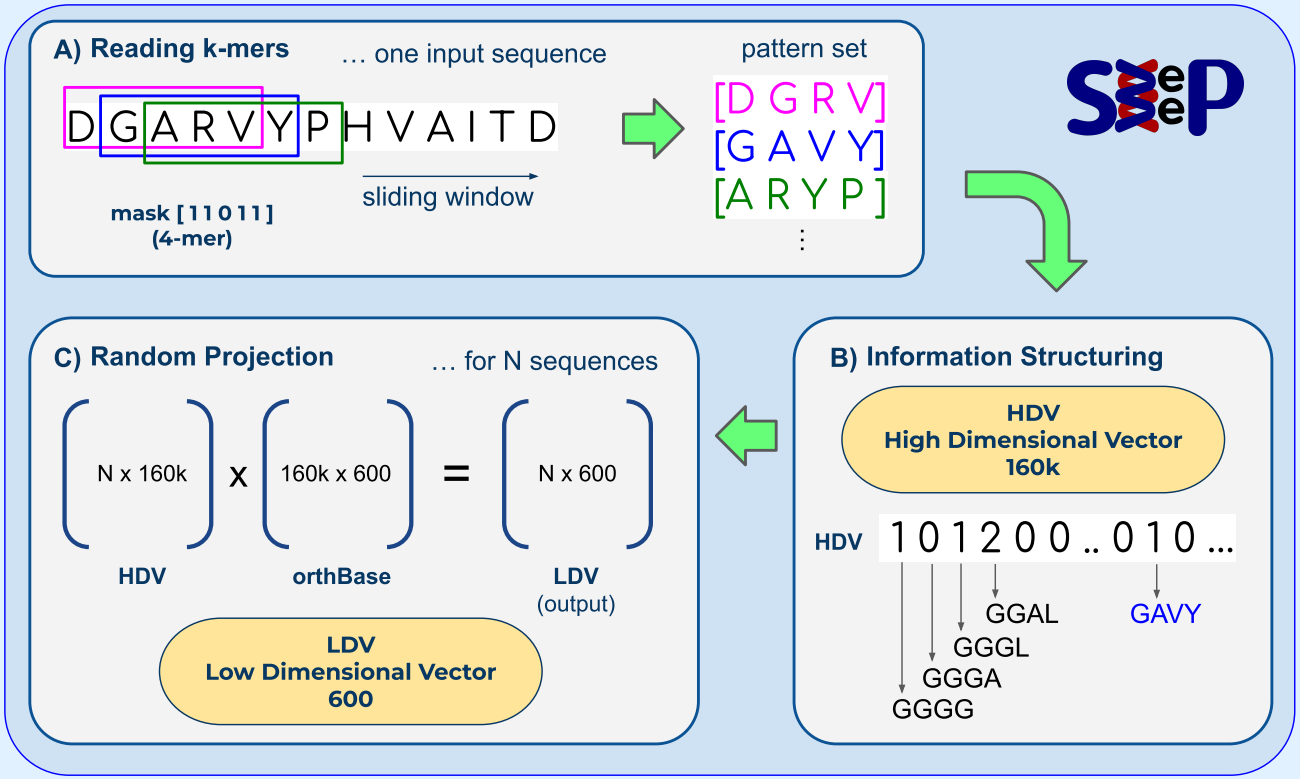
Ferramenta **AF** *word-based* que representa as sequências (nucleotídicas ou proteicas) em vetores compactos por meio de **Random Projections** (apoiada pelo lema de Johnson-Linderstrauss JL[7]).

[6] De Pierri,et al.(2020)Sweep: representing large biological sequences datasets in compact vectors. Scientific reports,10(1)

[7] Johnson, Lindenstrauss (1984) Extensions of Lipschitz mappings into a Hilbert space, Contemp. Matemática



Processamento: embedding por projeção randômica da frequência de k-mers de sequências biológicas.



Propriedades:

- preserva distâncias entre as amostras no espaço de menor dimensão
- vetor numérico estruturado e diretamente comparável
- vetor de dimensões reduzidas

Terminologia

embedding	conversão de um vetor de alta dimensionalidade em outro de baixa dimensionalidade (ex. projeção randômica, RP, ou PCA)
alta dimensão	o número de características é maior que o número de observações (amostras)
HDV	vetor estruturado de alta dimensão (frequência de k-mers, vetor contagem RNAseq)
LDV	projeção do HDV em vetor compacto
k-mer	oligonucleotídeo
máscara	[1111] e [11011] são duas máscaras possíveis para 4-mer

k-mers e significado biológico

O uso de *k-mers* é válido? Traz informações relevantes da evolução das espécies?

Segundo Forsdyke (2019), *k-mers* possuem **justificativa biológica**[8]:

- Variações na frequência de k-mers está fortemente relacionado à **especiação** (redução na capacidade de hibridização das fitas de DNA)
  espécies possuem frequência de k-mers bem determinado
- **códons preferenciais** $\rightarrow$ tendem a ser preservados na espécie
- pode revelar relações que métodos de alinhamento não conseguiriam

Em DNA:

- 1-mer - (%GC) permite diferenciar principais grupos de eubacterias
- 3-mer - genomas específicos
- 4-mer - padrão ouro para vírus e procariotos
- N-mer - maior capacidade de discriminação até nível de indivíduo

[8] Forsdyke, D. R. (2019). Success of alignment-free oligonucleotide (k-mer) analysis confirms relative importance of genomes not genes in speciation and phylogeny. Biological Journal of the Linnean Society, 128(2)

Desafios na alta dimensionalidade

Quanto maior o k-mer, maior o vetor HDV (vetor contagem)

- 1-mer = 4
- 2-mer = 16
- 4-mer = 256
- 7-mer = 16384
- 15-mer = 268 435 456
- 25-mer = 1.1259e+15 (ref[9])

maior k	$\rightarrow$	maior o vetor HDV	$\rightarrow$	maior consumo de RAM	$\rightarrow$	maior a dificuldade em analisar os vetores

Métodos de ML têm dificuldade de lidar com a alta dimensionalidade (tempo computacional)

Necessário o uso de métodos para a **redução da dimensionalidade**

[9] Chan CX, Bernard G, Poirion O, Hogan JM, Ragan MA. Inferring phylogenies of evolving sequences without multiple sequence alignment. Sci Rep. 2014;4:6504.

Redução de Dimensionalidade

3 técnicas básicas de redução de dimensão linear: **RP, PCA, FS**

RP – projeções randômicas ; **PCA** – Análise de Componentes Principais; **FS** – seleção de características

[10] Grellmann, et al. (2016). Random projection for fast and efficient multivariate correlation analysis of high-dimensional data: A new approach. Frontiers in Genetics, 7:102.
Ray, et al. (2021) Various dimension reduction techniques for high dimensional data analysis: a review. Artificial Intelligence Review, 54

Redução de Dimensionalidade

3 técnicas básicas de redução de dimensão linear: **RP, PCA, FS**

Limitações:[10]

- **FS, PCA** – depende explicitamente do conjunto de dados de entrada
- **PCA** – computacionalmente custoso com tempo de processamento quadrático em relação ao número de dimensões (características)
- **FS** – perda de informação potencialmente relevante
- **FS** – problema dependente
- **RP, PCA** – dificuldade em alocar altíssima dimensionalidade em RAM

Não foi encontrado na literatura sintetize do gap na redução de dimensionalidade.

Então, introduzimos **nosso conceito de generalidade** e propomos uma **solução** que satisfaz as limitações dos métodos atuais.

---

[10] Grellmann, et al. (2016). Random projection for fast and efficient multivariate correlation analysis of high-dimensional data: A new approach. Frontiers in Genetics, 7:102.
Ray, et al. (2021) Various dimension reduction techniques for high dimensional data analysis: a review. Artificial Intelligence Review, 54

GAPS nos métodos AF e Perspectiva de Generalidade

Na ausência de um termo específico, vamos utilizar a designação de **"generalidade"** como um método que incorpora os seguintes conceitos:

a) **usabilidade** $\rightarrow$ capaz de realizar *embedding* de alta dimensionalidade [11].

b) **conservação das distâncias** (local e global) $\rightarrow$ preserva informação[12]

c) método **conjunto independente** $\rightarrow$ *embedding* **reprodutível** a qualquer conjunto subsequente, dada a mesma parametrização

Impomos tais exigências para obter de um método:
geral, implementável, representação vetorial de sequências universal

Nenhum método da literatura satisfaz tais propriedades (exceto SWeeP)

[11] Grellmann, et al. (2016). Random projection for fast and efficient multivariate correlation analysis of high-dimensional data: A new approach. Frontiers in Genetics, 7:102.

[12] Heiser e Lau (2020). A quantitative framework for evaluating single-cell data structure preservation by dimensionality reduction techniques. Cell reports, 31(5)

Motivação

Principal motivação generalização da ferramenta $\rightarrow$ nova versão: SWeePlite

Não se limita a sequências biológicas (genes, genomas, proteínas e proteomas)[13] mas possui **potencial** para:

`metagenomas, espectrometria de massa, expressão gênica, análise multiômica, textos etc`

SWeeP já foi implementado em **mineração de textos** [14] de forma bem sucedida.

---

[13] da Silva Filho, et al. (2021). Prediction and analysis in silico of genomic islands in aeromonas hydrophila. Frontiers in microbiology, 12.
Raittz, et al. (2021). Comparative genomics provides insights into the taxonomy of azoarcus and reveals separate origins of nif genes in the proposed azoarcus and aromatoleum genera. Genes, 12(1)

[14] Ieger Raittz, et al. (2022). What are we learning with yoga: a text mining approach to literature. medRxiv

Consumo memória do vetor HDV

quero projetar uma sequência de **aminoácidos** em um vetor de comprimento **1000**

SWeeP Original

máscara	k-mer	# combinações	tamanho da matriz	memória (float)
11011	4	20^4	$1.6 \cdot 10^5$ x 1000	640 Mb
110111	5	20^5	$3.2 \cdot 10^6$ x 1000	12.8 Gb
1110111	6	20^6	$6.4 \cdot 10^7$ x 1000	**256 Gb**

SWeePlite

uso de memória: 80 - 800 Mb (parametrizável)

**permite alocar a matriz completa se quiser*

O Pacote rSWeeP

https://bioconductor.org/packages/release/bioc/html/rSWeeP.html
https://bioconductor.org/packages/devel/bioc/html/rSWeeP.html

Package 'rSWeeP'

October 5, 2024

Type Package

Title Spaced Words Projection (SWeeP)

Version 1.17.3

Date 2024-04-28

Maintainer Camila P Perico <camilapp94@gmail.com>

Description ``Spaced Words Projection (SWeeP)" is a method for representing biological sequences using vectors
preserving inter-sequence comparability.

Depends foreach, doParallel, parallel, Biostrings, methods, utils

Imports tools, stringi,

Suggests Rtsne, ape, Seurat, knitr, rmarkdown, tictoc, BiocStyle, testthat (>= 3.0.0)

Encoding UTF-8

RoxygenNote 7.3.2

VignetteBuilder knitr

Config/testthat/edition 3

O Pacote rSWeeP

https://bioconductor.org/packages/release/bioc/html/rSWeeP.html
https://bioconductor.org/packages/devel/bioc/html/rSWeeP.html

Contents

2 funções principais e 4 auxiliares:

- SWeeP – função SWeeP original, necessita que forneça a matriz de projeção
- SWeePlite – versão lite com função FGOP para gerar internamente matriz proj.
- orthBase – gera matriz de projeção
- extractHDV – extrai o HDV das sequências biológicas
- PMPG – Percentage of Mono or Paraphyletic Groups
- PCCI – PhyloTaxonomic Consistency Cophenetic Index. Estimativa do grau de agrupamento das amostras do mesmo taxon na árvore filogenética.

Função original:

- SWeeP – função SWeeP original, necessita que forneça a matriz de projeção
- orthBase – gera matriz de projeção

```
library(rSWeeP)

# mascara
mask = c(2,1,2)
# tamanho da projecao
psz = 600
# sequencia tipo aminoacido
seqtype="AA"
# endereco da pasta com FASTAS
path = "pasta/FASTAS/"

# crio a base ortonormal de projecao - 160000 x 600
base160k = orthBase(mask=mask,col=psz,seqtype='AA')

sw = SWeeP(path,base160k,seqtype=seqtype,mask=mask)
```

Função lite:

- SWeePlite – versão lite com função FGOP para gerar internamente matriz proj.

```
library(rSWeeP)

# mascara
mask = c(2,1,2)
# tamanho da projecao
psz = 600
# sequencia tipo aminoacido
seqtype="AA"
# endereco da pasta com FASTAS
path = "pasta/FASTAS/"

sw = SWeePlite(path,seqtype=seqtype,mask=mask,psz=psz)
```

Saída das funções `SWeeP` e `SWeePlite`

- `output$proj` uma matriz numérica com psz colunas e uma linha por sequência, correspondendo cada linha a um vetor compacto
- `output$info` informação adicional do processo. Este objeto subdivide-se em:
  `$ProjectionSize`: um número inteiro correspondente a psz (o comprimento do vetor de saída)
  `$bin`: um booleano que contém se é binário (TRUE) ou de contagem (FALSE)
  `$mask`: um vetor que contém a máscara utilizada
  `$SequenceType`: um carácter que contém o tipo da sequência (aminoácido: 'AA', ou nucleótido: 'NT')
  `$concatenate` : um booleano que corresponde à concatenação de sequências
  `$version` : um carácter correspondente à versão do pacote
  `$norm` : um carácter que contém a normalização utilizada
  `$extension`: um carácter que contém a lista de extensões consideradas
  `$timeElapsed`: um double que contém o tempo decorrido em segundos
  `$headers`: lista de cabeçalhos para cada sequência analisada

Funções adicionais

- extractHDV – extrai o HDV das sequências biológicas

```
library(rSWeeP)

# mascara
mask = c(2,1,2)
# sequencia tipo aminoacido
seqtype="AA"
# endereco da pasta com FASTAS
path = "pasta/FASTAS/"

output = extractHDV(path,mask=mask,seqtype=seqtype)
```

- output$HDV High Dimensional Vectors (vectores de alta dimensão) do FASTAS fornecido. A dimensão do vetor corresponde a 20^k para os aminoácidos e 4^k para os nucleótidos. k corresponde ao k-mer da máscara utilizada.
- output$info informação adicional do processo.

Funções adicionais

- PMPG – Percentage of Mono or Paraphyletic Groups
- PCCI – PhyloTaxonomic Consistency Cophenetic Index. Estimativa do grau de agrupamento das amostras do mesmo taxon na árvore filogenética.

```
library(rSWeeP)
[...]

mdist = dist(sw$proj,method='euclidean')
tree = ape::nj(mdist)

PMPG(tree,taxons)
PCCI(tree,taxons)
```

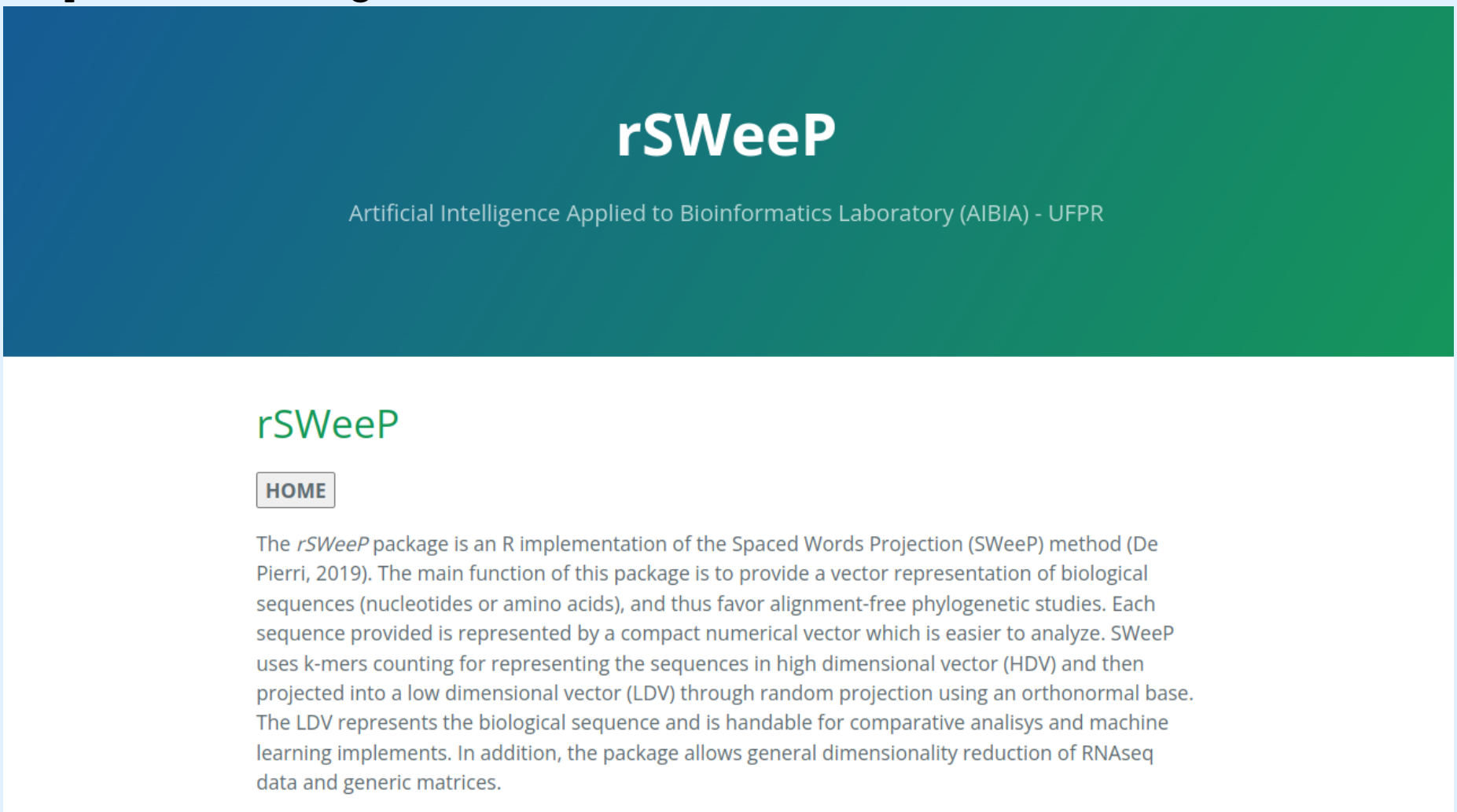
- PCCI$mean retorna o valor médio do agrupamento
- PMPG$percMono retorna o percentual de táxons monofiléticos
- PMPG$percPara retorna o percentual de táxons parafiléticos
- PMPG$mean retorna o percentual somado entre mono e paraiflético

Prática

Prática

Tutoriais

`https://aibialab.github.io/rSWeeP`



Instalação pacotes R necessários

Instalação do rSWeeP:

```
# instalo o manager do Bioconductor
install.packages("BiocManager")

# instalo o rSWeeP depositado no Bioconductor
BiocManager::install("rSWeeP")
```

Pacotes necessários para a prática

```
install.packages("ape")
install.packages("umap")
install.packages("Rtsne")
BiocManager::install("ggtree")
```

Manual do pacote

```
?rSWeeP
```

Dúvidas

Dúvidas e dificuldades instalação R