

Princípios FAIR na Bioinformática

Nicole Scherer

25-09-2023

Escola Gaúcha de Bioinformática

00 : 01 : 00

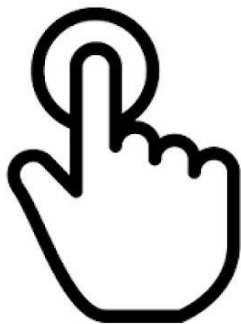
Um minuto para ouvir a palavra

FAIR

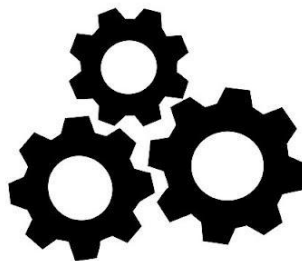
F_{indable}



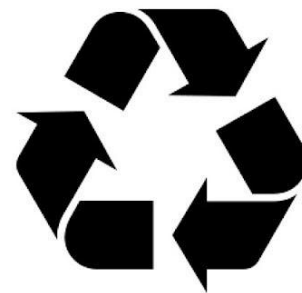
A_{ccessible}



I_{nteroperable}



R_{eusable}



Contextualizando

- FAIR é só um acrônimo poético?
- Por que a bioinformática precisa disto?
- Objeto digital de pesquisa
- Metaverso, metafísica, metadado, metamorfose ambulante...
- Para humanos e máquinas... e a Skynet?
- Com licença, por favor,
- é pessoal e sensível
- Ontologiaaaa ... eu quero uma pra viveeer

Localizável



AIR

Metadados e dados
devem ser encontrados
tanto por humanos
quanto por máquinas

Identificador único,
global e persistente
(PID)

Metadados
descritivos
suficientemente
detalhados



F_{indable}



Acessível



Armazenados num
repositório confiável

compreensíveis
tanto para os seres
humanos como para
as máquinas

Recuperáveis por seu
identificador único

PRJNA992786
SRR25233779
10.3233/DS-190026

Os metadados são
acessíveis, mesmo
quando os dados
não estão mais
disponíveis

Protocolo de acesso
padronizado,
aberto, globalmente
implementado

HTTP, FTP, SMTP, ...

Autenticação e
autorização, quando
necessário

A_{ccessible}



Interoperável

FA



R

metadados devem utilizar uma linguagem formal, acessível, partilhada e amplamente aplicável

utiliza vocabulário controlado, estruturado e padronizado na área de conhecimento

Dados armazenados em arquivos de formatos abertos e amplamente utilizados

Referências a outros dados e ou dependências de software documentadas nos metadados



Reutilizável

R_{eusable}



FAI

R

Descritos com uma pluralidade de atributos precisos e relevantes

Disponibilizados com uma licença de uso clara e acessível

Metadados incluem proveniência detalhada

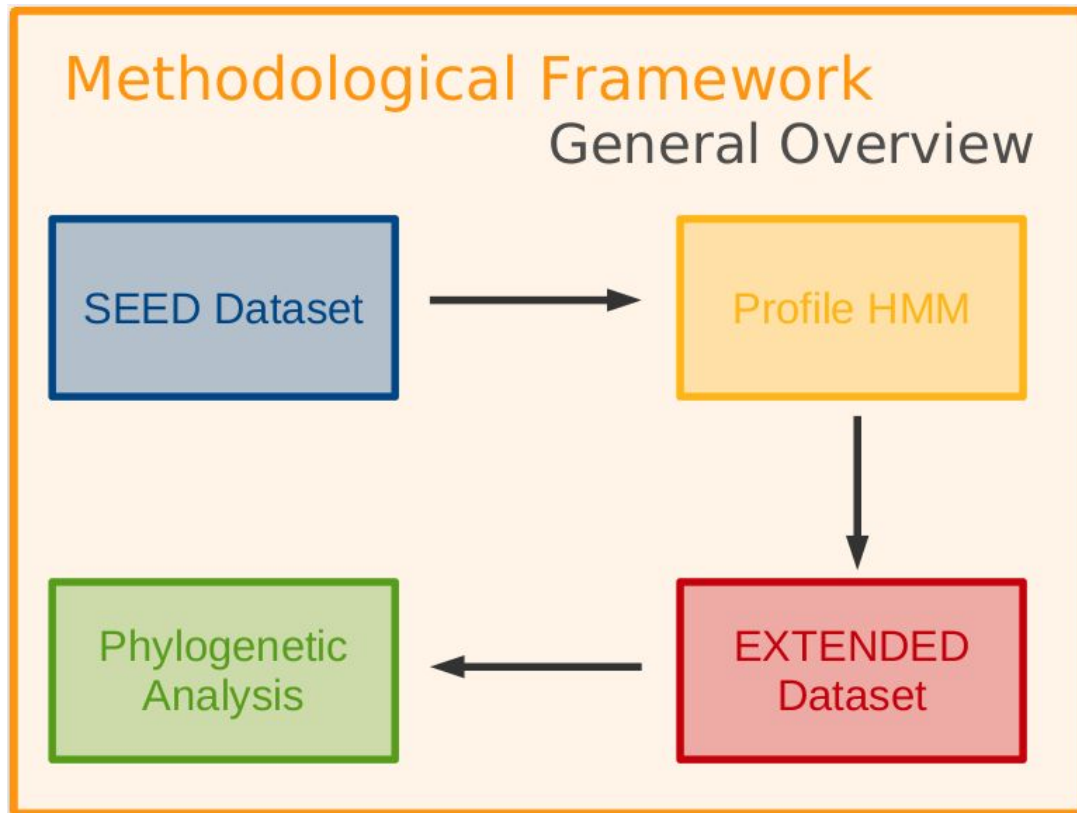


Por que precisamos de
FAIR?

História pessoal

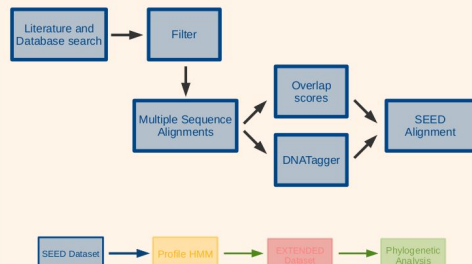
Pipeline desenvolvido
em **Perl** e **R**

Análise evolutiva em
proteínas relacionadas
à patogênese em
plantas

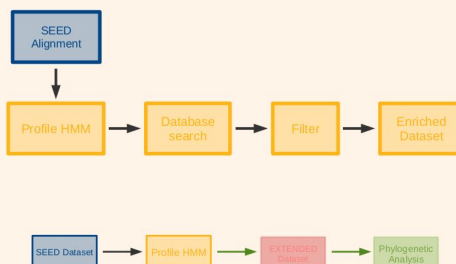


Conjunto de scripts e datasets

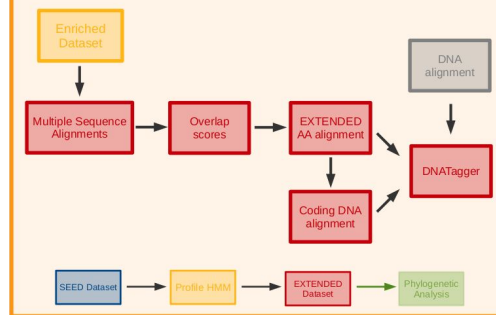
SEED Dataset



Profile Hidden Markov Models



EXTENDED Dataset



Alignment Visualization

DNATagger

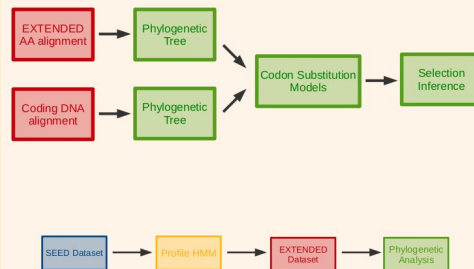
```

ADFLG4_MUSPR  A A N N I L S S A S L G N G S G
022317_MUSAC  A A N N I L S S A S L G N G S G
A7U707_MUSAC  A A N N I L S S A S L G N G S G
00Q3Y6_ZINOF  A A N N I L S S A S L G N G S G

ADFLG4_MUSPR  G C C A C G A A G A T C A A A T G G T T T C T C C G G T G G C T G C A A A C C A G A T A A A G T G T C A C A C G G G T C
022317_MUSAC  G C C A C G A A G A T C A A A T G G T T T C T C C G G T G G C T G C A A A C C A G A T A A A G T G T C A C A C G G G T C
A7U707_MUSAC  G C C A C G A A G A T C A A A T G G T T T C T C C G G T G G C T G C A A A C C A G A T A A A G T G T C A C A C G G G T C
00Q3Y6_ZINOF  G C C A C G A A G A T C A A A T G G T T T C T C C G G T G G C T G C A A A C C A G A T A A A G T G T C A C A C G G G T C
    
```

N. M. Scherer and D. M. Basso. (2008) DNATagger, colors for codons.
Genet. Mol. Res., 7(3):853–860.

Phylogenetic Analysis



Phylogenetic Analysis

Tree reconstruction



Codon Substitution Models

ATG	TCT	AAC	CTG	TGT	TCA	TAG
Met	Ser	Asn	Leu	Cys	Ser	stop
<div style="display: flex; justify-content: space-around;"> ↓ ↓ </div>						
ATG	TCC	AAC	CTG	TGG	TCA	TAG
Met	Ser	Asn	Leu	Trp	Ser	stop
<div style="display: flex; justify-content: space-around;"> synonymous non-synonymous </div>						

- Selection inference
- Positive selection

Meus resultados seguem os princípios FAIR?

Localizável ?

Identificador único e persistente?

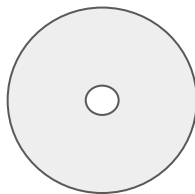
Protocolo de comunicação padronizado?

Metadados?

Acessível?

Interoperável?

Reutilizável?



URN (NBN):

urn:nbn:de:hbz:061-20100621-134521-6

<https://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=15271>

hhu Heinrich Heine Universität Düsseldorf

Universitäts- und Landesbibliothek Publikationsservice

HHU Düsseldorf > Publikationsservice > Startseite > Dokument 15271

[Zurück zur Universität](#)

Dokument: Pathogenesis-Related Proteins: Phylogenetic Characterization

Startseite	Title: Pathogenesis-Related Proteins: Phylogenetic Characterization
Mein Dokumentenkorb	Weiterer Titel: An der Pathogenese beteiligte Proteine: Phylogenetische Charakterisierung
Publizieren	URL für Lesezeichen: https://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=15271
Suche im Volltextbestand	URN (NBN): urn:nbn:de:hbz:061-20100621-134521-6
Spezielle Sammlungen	Kollektion: Dissertationen
Autorenbetreuung	Sprache: Englisch
Server Policy	Dokumententyp: Wissenschaftliche Abschlussarbeiten » Dissertation
	Medientyp: Text
	Author: Scherer, Nicole de Miranda [Autor]
	Daten: [Dateien anzeigen] Adobe PDF [Details] 23,72 MB in einer Datei [ZIP-Datei erzeugen] Dateien vom 04.06.2010 / geändert 04.06.2010

Bioinformática FAIR

Onde aplicar os princípios FAIR na Bioinformática

dados

software

protocolos

tutoriais

Onde aplicar os princípios FAIR na Bioinformática

dados



software

protocolos

tutoriais

Onde aplicar os princípios FAIR na Bioinformática

dados



protocolos

software



tutoriais

Onde aplicar os princípios FAIR na Bioinformática

dados



protocolos



software



tutoriais

Onde aplicar os princípios FAIR na Bioinformática

dados



protocolos



software



tutoriais



Onde aplicar os princípios FAIR na Bioinformática

dados



protocolos



software

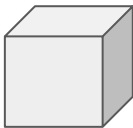


tutoriais



Meta

-dados



Metadados

Dados que descrevem dados

Metadados

Exemplo do disco de vinil

Dados: músicas registradas nos sulcos do vinil

Metadados: registrados no encarte do disco

- autores / músicos
- engenheiros de som
- letras das músicas
- estúdio / palco
- data da gravação
- licenças



Com licença

Licenças

Licenças são necessárias para definir como os dados podem ser reutilizados

Tipos de licenças públicas:

Creative Commons - protocolos, dados, tutoriais

LICENÇA PÚBLICA GERAL GNU - software



Attribution
CC BY



Attribution — ShareAlike
CC BY-SA



Attribution — NoDerivs
CC BY-ND



Attribution — NonCommercial
CC BY-NC



Attribution — NonCommercial — ShareAlike
CC BY-NC-SA



Attribution — NonCommercial — NoDerivs
CC BY-NC-ND

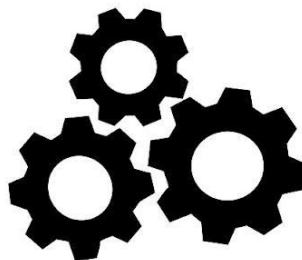
F_{indable}



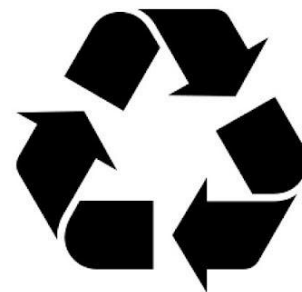
A_{ccessible}



I_{nteroperable}

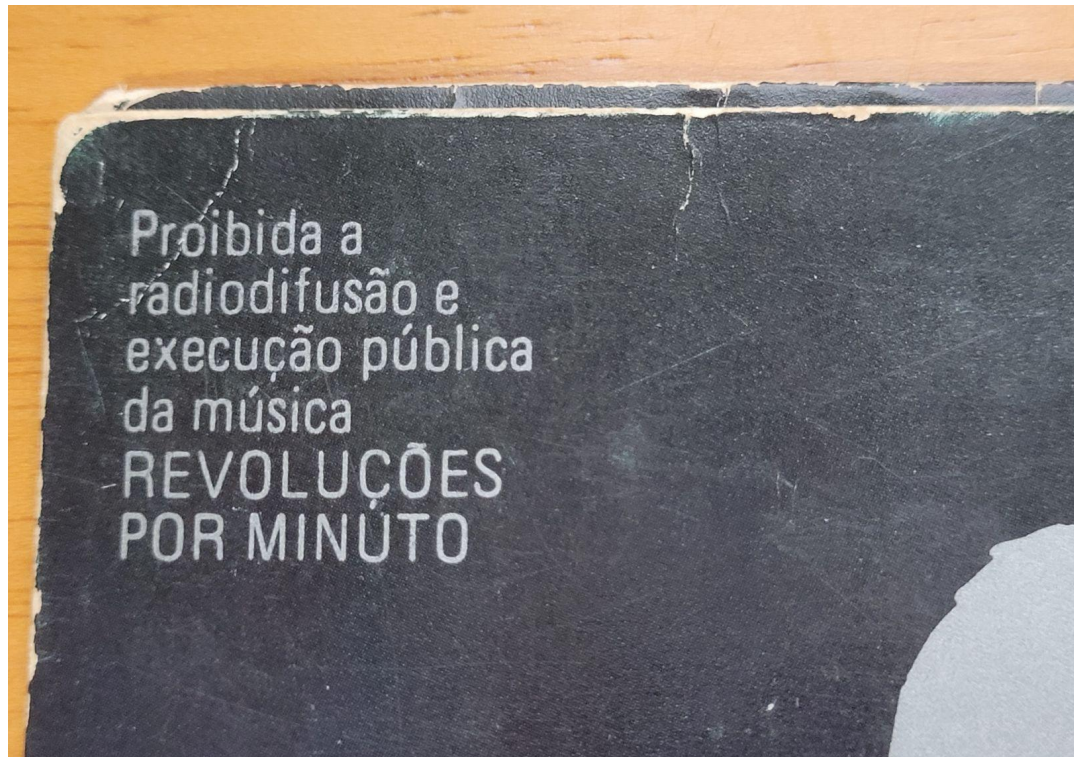


R_{eusable}



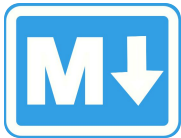
Permissões

Em quais contextos o dado /
software /protocolo pode ser
utilizado



Formatos

Formatos de archivos



FASTA

BAM

HDF5

FASTQ

VCF

CRAM

GTF

JSON

PDB

BED

Formatos de archivos



FASTA

BAM

HDF5

FASTQ

VCF

CRAM

GTF

JSON

PDB

BED

To Share or
Not to Share

Compartilhar ou Não compartilhar - eis a questão

Quando compartilhar?

pesquisa financiada com
recursos públicos

dados necessários para
reprodutibilidade

resultados negativos para uma
pergunta podem ser dados
valiosos para outras perguntas

Quando não compartilhar?

Compartilhar ou Não compartilhar - eis a questão

Quando compartilhar?

pesquisa financiada com
recursos públicos

dados necessários para
reprodutibilidade

resultados negativos para uma
pergunta podem ser dados
valiosos para outras perguntas

Quando não compartilhar?

Dados sensíveis

Dados pessoais

Dados que dão suporte a um
pedido de patente ainda não
finalizado

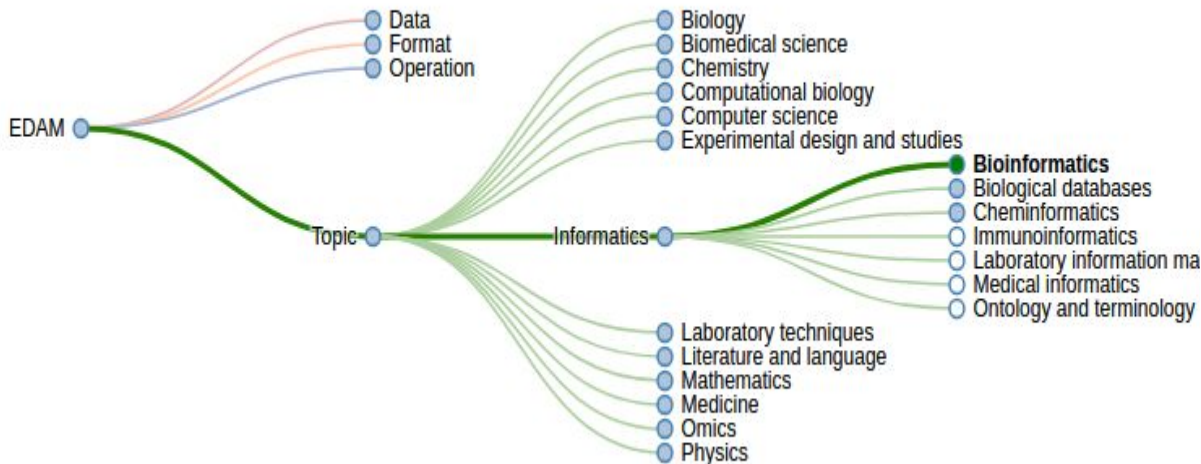
Aberto o possível

Fechado o necessário

Ontologias

e interoperabilidade

EDAM ontology



Details of term "Bioinformatics"



Term	Bioinformatics
Definition	The archival, curation, processing and analysis of complex biological data.
Comment	This includes data processing in general, including basic handling of files and databases, datatypes, workflows and annotation.
Exact synonyms	
Narrow synonyms	
URI	topic_0091
Parents	Informatics

Vocabulários, ontologias, esquemas



Práticas de FAIRificação

- ★ Documentação
- ★ Formatos de arquivos
- ★ Metadados
- ★ Acesso
- ★ Identificadores Persistentes
- ★ Licenças de dados e software

Iniciativas FAIR



Obrigada!

Organização da EGB

AB3C

ELIXIR-PT

ELIXIR-BE

SIB

Todos vocês que
ficaram até tarde para
ouvir a palavra dos
princípios FAIR

<https://github.com/BioNica/aulas-principios-FAIR>