PROPOSTA DE FORMATO DE ARQUIVO PARA ANÁLISE DE CONTEÚDO

[1]Emanuel Diego dos Santos Penha, [1]Luana Maria Castelo Branco, [1]Markênia Kelia Santos Alves Martins, [1]Natássia Albuquerque Pinheiro, [1]Ana Paula Moreira Bezerra

1. DeVry | Fanor

emanuel.penha@fanor.edu.br

1 INTRODUÇÃO

A grande quantidade de dados gerados em campos do conhecimento e pela internet possui um volume muito grande. A tal ponto que a dificuldade de sua leitura e interpretação é um desafio, um assunto recorrente. Mas isso não importa somente aos que trabalham em genômica ou astronomia (BELL; HEY; SZALAY, 2009). Todas as vezes que recorremos ao Google para completar frases com palavras que ainda digitaremos, usufruímos do resultado da exploração de *big data,* desses dados que são tão grandes que impõem grandes obstáculos logísticos em seu uso.

Mas nem tudo precisa ou deve ter essa escala. Um pesquisador que transcreve entrevistas para um trabalho específico, dificilmente terá tempo e recursos para obter uma amostra grande o suficiente para que se compense falta de estruturação.

Registros realizados e mantidos de maneira ineficiente prejudicam a reprodutibilidade dos resultados. Várias vezes o pesquisador recorre a ferramentas que ele já conhece (MICROSOFT, 2016), mas que não são a melhor escolha. Existem implementações de algumas soluções em análise de sequências de genes e proteínas, mas elas são voltadas prioritariamente para ensino (ANZALDI; MUÑOZ-FERNÁNDEZ; ERILL, 2012).

Ter transcrições de várias respostas à perguntas de várias amostras em um arquivo do Microsoft WORD não é uma boa opção, pois o arquivo só poderá ser aberto por uma

quantidade limitada de programas e a interoperabilidade pode ser prejudicada (SHAH; KESAN, 2009).

Tomando como exemplo a análise de conteúdo (BARDIN, 2008), no qual definimos para uma parte do texto uma categoria, um conceito, fica claro a limitação. Se partimos de um único arquivo adicionando essas *tags* com marcadores, caixas de texto, comentários, a contagem dessas será feita de maneira manual. Nesse momento existe a possibilidade de erro na quantificação, podendo comprometer o resultado e consequentemente as conclusões do estudo. Embora seja claro que a interpretação é a parte mais importante em pesquisa qualitativa, o peso que a frequência das categorias possui na análise de conteúdo não pode ser ignorado.

O formato FASTA é um padrão para armazenamento de sequências de DNA, RNA e peptídeos (LIPMAN; PEARSON, 1985). Resumidamente, ele é um arquivo de texto semi estruturado. A primeira linha do arquivo possui um cabeçalho, cujo primeiro caractere é "$>$". Além do nome da sequência em si, nessa mesma linha podemos encontrar outros campos que são separados pelo caractere "|". A sequência de caracteres da sequência segue na próxima linha. O objetivo do presente trabalho é adaptar essa estrutura à Análise de Conteúdo.

2 METODOLOGIA

Esse é um trabalho de cunho metodológico, composto por três etapas. A primeira consiste em identificar os padrões da estrutura do FASTA que podem ser usados para organizar texto. A segunda parte é criar um algoritmo para a contagem das categorias. A terceira é, como prova de conceito, adaptar um *script* em *perl*, linguagem bastante usada em processamento de texto, para implementação do mesmo.

Esse deve ler todos os arquivos de texto dentro de um diretório e criar uma tabela com a contagem de categorias para o combinado dos arquivos. O *script* foi desenvolvido originalmente para um trabalho de conclusão de Residência em Saúde da Família (CAMPOS, 2015)

.

3 RESULTADOS

3.1 ESTRUTURA DO FASTA

Como exemplos menos estruturado de um FASTA, temos o cabeçalho do gene do Fator de Necrose Tumoral no *Homo Sapiens:* >NC_000006.12:31575567-31578336 Homo sapiens chromosome 6, GRCh38.p7 Primary Assembly

E como um exemplo mais estruturado, podemos citar o cabeçalho da isoforma p170 do receptor do fator de crescimento epidérmico: >sp|P00533|EGFR_HUMAN Epidermal growth factor receptor OS=Homo sapiens GN=EGFR PE=1 SV=2

Depois do cabeçalho, na linha seguinte, vêm a sequência em si. Um exemplo de uma sequência curta de peptídeo, com o cabeçalho mais estruturado pode ser encontrado abaixo:

>gi|31563518|ref|NP_852610.1| microtubule-associated proteins 1A/1B light chain 3A isoform b [Homo sapiens]
MKMRFFSSPCGKAAVDPADRCKEVQQIRDQHPSKIPVIIERYKGEKQLPVLDKTKFLV
PDHVNMSELVKI
IRRRLQLNPTQAFFLLVNQHSMVSVSTPIADIYEQEKDEDGFLYMVYASQETFGFIRE
N

3.2 USO DA ESTRUTURA DO FASTA PARA ANÁLISE DE CONTEÚDO

A idéia central é que o texto coletado em entrevistas ou algo que o valha, seja armazenado em arquivos de texto. O cabeçalho teria os dados da questão ou pergunta, sendo adicionadas as informações e ou metadados para análise posterior. Assim ferramentas para processar essa informação poderia ser criadas com diferentes linguagens de programação, sistemas operacionais. Seria muito mais fácil disponibilizar os dados de pesquisa qualitativa para que outros possam avaliar a classificação e contagem das categorias.

A definição de como o FASTA pode ser usado para armazenamento de texto para análise de conteúdo é trivial. A título de exemplo, suponhamos uma resposta a uma questão hipotética abaixo. Em um primeiro momento, o pesquisador transcreveria as questões em um arquivo de texto. Pelo cabeçalho podemos identificar que essa é a primeira questão, respondida pela amostra 1 e ela possui uma data específica de coleta. Dessa maneira, esse exemplo corresponderia à:

>questão 1 | amostra 1 | data de coleta
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

A definição das categorias e inserção delas no meio do texto é feita pelo próprio pesquisador, depois da transcrição. Identificando as categorias as quais cada pedaço do texto pertence. Assim, o texto “ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore” pertenceria à categoria B. Seguindo a mesma lógica, cada pedaço de texto ficaria associado a uma categoria, sendo limitado pela identificação de uma categoria anterior, como pode ser visualizado na figura abaixo.

Figura 1. Identificação visual dos trechos de texto associados à categorias..

>questão 1 | amostra 1 | data de coleta
Lorem ipsum dolor sit amet, consectetur adipiscing elit, <categoria A> sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation <categoria A> ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore <categoria B> eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt <categoria C> mollit anim id est laborum <Categoria B>.

3.3 DESCRIÇÃO DO ALGORITMO

Os programas a serem desenvolvidos devem encontrar a linha cujo primeiro caractere seja “>”, verificar a existência de “|” e em caso afirmativo, dividir o conteúdo do cabeçalho usando “|” como separador e armazenar essa informação em um *array* ou estrutura que o valha.

Em seguida, deve identificar no texto abaixo do cabeçalho, todas as palavras que estão entre “<” e “>”, e contar sua frequência. O resultado desse processo para a Figura 1 seria:

Tabela 1. Frequência de categorias em texto de análise de conteúdo.

Categorias	Frequência
A	2
B	2
C	1

Os dados para análise podem ser oriundos de formulários online, textos de livros, transcrições de entrevistas. Enfim, qualquer coisa que possa ser expressa em texto e tenha a necessidade de definição de categorias internas pode usar dessa metodologia.

3.4 IMPLEMENTAÇÃO DO ALGORITMO

O *script* original foi adaptado para atender a casos mais gerais. Pode ser encontrado sob controle de versão em https://github.com/diegopenhanut/flf/blob/master/perl/cont.cat.pl e pode ser executado em vários sistemas operacionais.

4 DISCUSSÃO

Foi possível verificar que várias das características do arquivo FASTA podem ser utilizados para organização de dados para Análise de Conteúdo. A principal seria a presença do cabeçalho com metainformação de uma maneira simples.

Embora existam vários programas que possam fazer a análise de conteúdo (FRIESE, 2014) e que de fato auxiliem bastante na organização e análise dos conteúdos, eles acabam por limitar o pesquisador quanto à reprodutibilidade. A nossa proposta de formato de arquivo resolve esse problema, ao definir uma estrutura comum que pode ser usada, lida, avaliada e reavaliada por qualquer programa que entenda a sua especificação.

5 CONCLUSÕES

Foi possível adaptar um formato de texto usado para armazenar sequências biológicas para análise de conteúdo com sucesso. Além disso, foi criada também ferramenta

que mostra a implementação do algoritmo e contagem das categorias. Como desdobramento desse trabalho, é possível planejar softwares com interface gráfica que usem essa estrutura para armazenamento e manipulação de textos e categorias.

REFERÊNCIAS

ANZALDI, L. J.; MUÑOZ-FERNÁNDEZ, D.; ERILL, I. BioWord: A sequence manipulation suite for Microsoft Word. **BMC bioinformatics**, v. 13, n. 1, p. 124, 7 jun. 2012.

BARDIN, L. Análise de conteúdo. **Lisboa: edições**, v. 70, p. 225, 1977.

BELL, G.; HEY, T.; SZALAY, A. Beyond the data deluge. **Science**, v. 323, n. 5919, p. 1297–1298, 6 mar. 2009.

CAMPOS, K. S. **Percepção dos profissionais e gestores de saúde sobre a estratégia eSUS atenção básica e sua relação com a vigilância alimentar e nutricional**. Residência Integrada em Saúde - Escola de Saúde Pública do Ceará, 2015.

FRIESE, S. **Qualitative Data Analysis with ATLAS.ti**. SAGE, 2014.

LIPMAN, D. J.; PEARSON, W. R. Rapid and sensitive protein similarity searches. **Science**, v. 227, n. 4693, p. 1435–1441, 22 mar. 1985.

MICROSOFT. Word 2016. Microsoft Corporation, 2017.

SHAH, R; KESAN, J. Interoperability challenges for open standards: ODF and OOXML as examples. In: **Proceedings of the 10th Annual International Conference on Digital Government Research: Social Networks: Making Connections between Citizens, Data and Government.** Digital Government Society of North America, 2009.