https://eventos.utfpr.edu.br//sicite/sicite2019

Busca, Indexação e definição de base de dados integrada de Escherichia Coli

Search, Indexing and integrated database definition of Escherichia Coli

Douglas Ribeiro Violante
douglasviolante@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná,
Cornélio Procópio, Paraná, Brasil

Fabrício Martins Lopes
fabricio@utfpr.edu.br
Universidade Tecnológica Federal do Paraná,
Cornélio Procópio, Paraná, Brasil

RESUMO

Este trabalho adota dados provenientes da base de dados da KEGG (*Kyoto Encyclopedia of Genes and Genomes*), devido a sua divisão hierárquica em vias metabólicas e descrições interespécies para calcular a similaridade semântica entre quaisquer genes G1 e G2 de organismos modelos, como a Escherichia Coli, Drosophila Melanogaster, Saccharomyces Cerevisiae e Arabdopsis Thaliana. Foi desenvolvido um script que pode ser descrito em duas partes, a primeira parte trata da geração do grafo acíclico direcionado, e a segunda parte trata do cálculo da similaridade baseada em similaridade semântica. Os scripts permitem que futuramente poderá ser levada em conta informações como os termos GO (*Gene Ontology*) e KO (*KEGG Orthology*), atualmente introduzidos e considerados, podendo assim elevar a precisão do cálculo.

PALAVRAS-CHAVE: Bioinformática. Integração de Dados. Escherichia Coli.

Recebido: 19 ago. 2019.
Aprovado: 01 out. 2019.

ABSTRACT

This paper adopts data from KEGG (*Kyoto Encyclopedia of Genes and Genomes*), due to its hierarchical division into metabolic pathways and the inter-species descriptions to calculate the semantic similarity between any G1 and G2 genes of model organisms, such as Escherichia Coli, Drosophila Melanogaster, Saccharomyces Cerevisiae and Arabdopsis Thaliana. Has been developed a script that can be described in two parts, the first part deals with the generation of a directed acyclic graph, and the second part deals with the calculation of similarity based in semantic similarity. The scripts in the future can taken into account informations like GO (*Gene Ontology*) terms and KO (*KEGG Orthology*), currently introduced and considered, being able to increase precision of the calculation.

KEYWORDS: Bioinformatics. Data integration. Escherichia Coli.

INTRODUÇÃO

Recentemente, mudanças relacionadas a custos de obtenção e disponibilidade de dados, conforme Juan; Norberto; Francisco (2016, p.1) descrevem que avanços nas áreas da indústria de biotecnologia levaram ao decréscimo nos custos de obtenção, e consequentemente o aumento da quantidade de dados, entretanto se extraindo pouco deste montante de informação.

Com esse montante de informação, este trabalho aborda os já conhecidos chamados termos GO (*Gene Ontology*), que segundo Norberto; Jesús (2011, p.1) se trata de um vocabulário controlado, usado na descrição das funções dos genes e determinação de coerência entre *sets* de genes, além dos termos KO (*KEGG Orthology*), no qual nas palavras de Minoru; Yoko; Kanae (2015, p.1) são termos de uma base de dados de funções moleculares organizados hierarquicamente, e divididos em uma forma semelhante a um grafo acíclico direcionado. E estes termos KO são obtidos observando não apenas no contexto do mapa de uma via metabólica KEGG, mas também em hierarquias BRITE, redes moleculares, nome e módulos KEGG, com ferramentas como KOALA e BlastKOALA, e assim então atribuindo um identificador a cada KO chamado K-número.

Então utilizando-se a base de dados do KEGG, e mais especificamente bases de ortólogos funcionais KO, desenvolveu-se um algoritmo portado atualmente na linguagem de programação Python na versão 3.x. O algoritmo desenvolvido constrói um grafo direcionado acíclico com todos os genes e suas hierarquias associadas ao organismo, e a partir do grafo, funções de análise e cálculo de similaridade foram aplicadas, e demonstraram resultados adequados para organismos testados, como *Arabidopsis Thaliana*, *E. Coli*, *Drosophila Melanogaster* e *Saccharomyces Cerevisiae.*

MATERIAIS E MÉTODOS

O método GFD-Net, utiliza-se de uma análise topológica do grafo acíclico direcionado, usando apenas os termos GO e o número de arrestas contidas no caminho entre os termos, para o cálculo de similaridade semântica. O GFD-Net não faz uso das propriedades de quantidade de informação, conforme desenvolvida, testada e explicada por Juan; Norberto; Francisco (2016, p.1) onde o cálculo em si, consiste basicamente do cálculo da distância entre dois pares de termos GO, como é descrito na Eq. (1).

$$\text{Distância}\left(t_\alpha, t_\beta\right) = \frac{\text{Distância}\left(t_\alpha, t_\beta\right)}{2 * \text{Profundidade}\left(\text{LCA}\left(t_\alpha, t_\beta\right)\right) + \text{Distância}\left(t_\alpha, t_\beta\right)} \quad (1)$$

Onde sejam t_α e t_β termos GO, em que a distância é o número de arrestas ligando os pares t_α e t_β , e profundidade como o número de arrestas entre a raiz do grafo e o ancestral comum mais profundo entre os pares de termos t_α e t_β.

Com isso por meio da topologia e produto cartesiano, tal como os genes a que pertencem os termos GO, a análise feita pelo autor, conseguiu determinar que para quaisquer valores retornados pela equação, pode se aferir como similaridade semântica.

Neste trabalho foram adotados quatro organismos modelos da biologia: a *Arabdopsis Thaliana* e *Drosophila Melanogaster* citadas como modelo por Bob; Nicole (2016, p.1), sendo a *Escherichia Coli* citada como modelo por Ingrid et al (2005, p.1), e *Saccharomyces Cerevisiae* citada por Michael et al (2000, p.1). Esses organismos retratam modelos da biologia para serem testados pela metodologia desenvolvida neste trabalho, cuja suas descrições estão no quadro (1).

Quadro 1 - Informações dos organismos utilizados.

Organismo	Número de Genes	Número de Termos KO	Tipo do Organismo	Arquivo .KEG
Escherichia Coli	5433	5631	Bactéria	eco00001
Arabdopsis Thaliana	15873	15940	Planta	ath00001
Drosophila Melanogaster	12994	13027	Inseto	dme00001
Saccharomyces Cerevisiae	7814	7875	Fungi	sce00001

Fonte: Autoria própria (2019).

Para a realização dos objetivos almejados neste trabalho, utilizou-se um computador com uma CPU Intel i5 3340m @ 2.6 GHz e memória RAM 8Gb DDR3, além da linguagem de programação Python em sua versão 3.x, na qual é uma linguagem interpretada, orientada a objetos e interativa segundo a Python Software Foundation (2019). Além de que se importou funções de outras bibliotecas, que foram necessárias ao desenvolvimento do trabalho, tais como a da *iGraph*, uma biblioteca externa de trabalho em grafos, e outra biblioteca própria da linguagem conhecida como RE, usada para promover o tratamento e uso de expressões regulares no código.

Como materiais, foi adotada a base de dados que segundo Minoru; Yoko; Kanae (2015, pp. 1-5) contém as funções moleculares em termos de ortólogos funcionais, de cada organismo em análise neste trabalho, obtida da KEGG (*Kyoto Encyclopedia of Genes and Genomes*). Esta base de dados disponibiliza um arquivo hierárquico e estruturado de extensão .KEG, no qual cada linha com nível D demonstra informações sobre um gene no organismo e ao mesmo tempo indica o nível mais específico, tal como a indicação de termos KO se existir. Um trecho deste arquivo é apresentado na Figura 1.

Figura 1 - Trecho de um arquivo .KEG da organismo E. Coli MG, usado no trabalho, demonstrando a hierarquização em níveis específicos.

```
C    00983 Drug metabolism - other enzymes
B
B  09112 Not included in regular maps
C    09113 Global maps only
#
A09120 Genetic Information Processing
B
B  09121 Transcription
C    03020 RNA polymerase [PATH:eco03020]
D      b3295 rpoA; RNA polymerase subunit alpha K03040 rpoA; DNA-directed RNA polymerase subunit alpha [EC:2.7.7.6]
D      b3987 rpoB; RNA polymerase subunit beta  K03043 rpoB; DNA-directed RNA polymerase subunit beta [EC:2.7.7.6]
D      b3988 rpoC; RNA polymerase subunit beta' K03046 rpoC; DNA-directed RNA polymerase subunit beta' [EC:2.7.7.6]
D      b3649 rpoZ; RNA polymerase subunit omega K03060 rpoZ; DNA-directed RNA polymerase subunit omega [EC:2.7.7.6]
C    03022 Basal transcription factors
C    03040 Spliceosome
B
B  09122 Translation
C    03010 Ribosome [PATH:eco03010]
D      b0911 rpsA; 30S ribosomal subunit protein S1 K02945 RP-S1; small subunit ribosomal protein S1
D      b0169 rpsB; 30S ribosomal subunit protein S2 K02967 RP-S2; small subunit ribosomal protein S2
```

Fonte: Autoria própria (2019).

Por fim, para uma melhor visualização dos grafos gerados neste trabalho, estes foram exportados para extensão .graphml referente a cada organismo. Foi adotado o *software open-source*, *Cytoscape* na versão 3.7, utilizado largamente na integração e visualização de grafos complexos, segundo a Cytoscape Consortium (2019). Um exemplo é mostrado na Figura 2.

Figura 2 - Grafo de uma pequena parte da ortologia do Escherichia Coli, gerado utilizando o script desenvolvido no trabalho e exibido pelo Cytoscape, onde a raiz-*OBJECT* e níveis hierárquicos do arquivo .KEG estão visíveis.

Fonte: Autoria própria (2019).

Na execução deste trabalho foi desenvolvido um único script universal em Python para a geração do grafo de cada organismo, e o cálculo de similaridade entre genes. É possível descrever o script em duas partes, um de geração do grafo e o outro de cálculo de similaridade.

A parte da geração do grafo de cada organismo, é realizado mediante a leitura do arquivo .KEG do organismo, lendo linha por linha e identificando as hierarquias presentes em cada uma delas. Desta forma faz-se uso de variáveis chamadas de *placeholders* para cada nível A, B, C e D, com a finalidade de indicar onde cada nível pertence, e para a ligação de arestas direcionadas entre os vértices.

Na parte de cálculo de similaridade, o grafo do organismo gerado anteriormente, é percorrido, identificando o menor caminho entre os genes de entrada G1 e G2. Dessa forma são recuperados seus ancestrais em comum, sendo que os mesmos são utilizados para o cálculo de similaridade utilizando a Eq. (1), juntamente com os cálculos de profundidade do ancestral comum mais especifico.

RESULTADOS

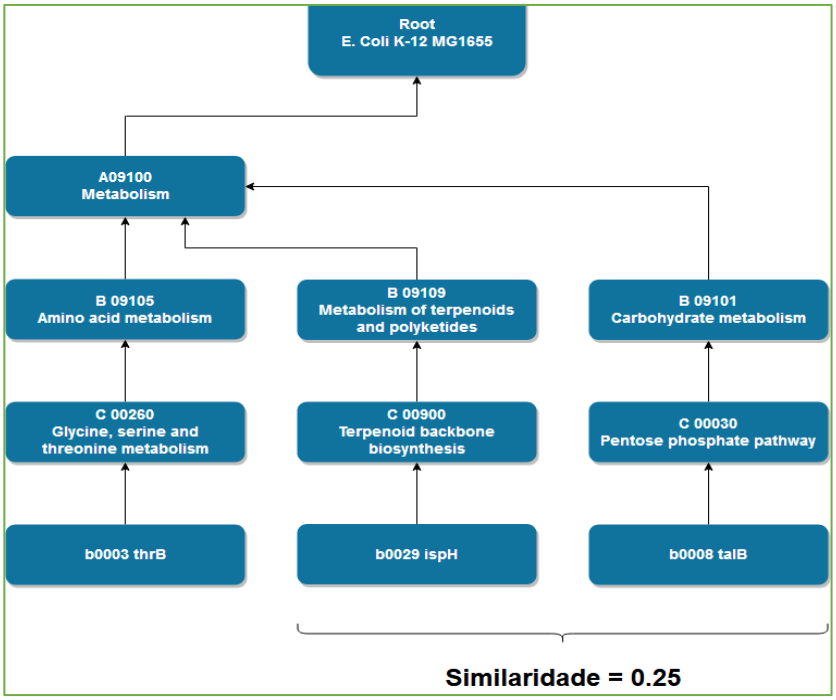
Na execução das partes descritas, obteve-se um único grafo, para cada organismo baseado no arquivo .KEG como pode ser visto na Figura 3.

Figura 3 - Grafos gerados com os scripts descritos, para cada organismo.

Fonte: Autoria própria (2019).

A partir dos grafos gerados, torna-se possível o cálculo da similaridade entre genes nos organismos, atingindo uma faixa de valores de 0, 0.25, 0.5, 0.75 e 1.00, relacionados aos níveis hierárquicos presentes. Um exemplo calculado entre dois genes da *Escherichia Coli* pode ser verificado na Figura 4.

Figura 4 - Exemplo calculado na Escherichia Coli entre os genes b0029 e b0008, indicando uma similaridade de 0.25.



Fonte: Autoria própria (2019).

CONCLUSÃO

Com os resultados apresentados, consegue-se perceber que os scripts desenvolvidos tanto para a geração de grafos, quanto para cálculo de similaridade, estão em pleno funcionamento e prontos, os quais serão utilizados em processos e trabalhos futuros. O desenvolvimento deste trabalho permite que pesquisas com os termos KO, e levantamento da similaridade semântica entre eles. Essa informação gerada contribui com a descoberta de relações entre diferentes genes, a qual será explorada em trabalhos futuros vinculados ao projeto de pesquisas deste trabalho.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico CNPq – Brasil.

REFERÊNCIAS

DÍAZ-DÍAZ, Norberto; AGUILAR-RUIZ, Jesús S. GO-based Functional Dissimilarity of Gene Sets. **Bmc Bioinformatics**, [s.l.], v. 12, n. 1, p.1-9, 1 set. 2011. Springer Nature. http://dx.doi.org/10.1186/1471-2105-12-360.

DÍAZ-MONTAÑA, Juan J.; DÍAZ-DÍAZ, Norberto; GÓMEZ-VELA, Francisco. GFD-Net: A novel semantic similarity methodology for the analysis of gene networks. **Journal Of Biomedical Informatics**, [s.l.], v. 68, p.71-82, abr. 2017. Elsevier BV. http://dx.doi.org/10.1016/j.jbi.2017.02.013.

KANEHISA, Minoru et al. KEGG as a reference resource for gene and protein annotation. **Nucleic Acids Research**, [s.l.], v. 44, n. 1, p.457-462, 17 out. 2015. Oxford University Press (OUP). http://dx.doi.org/10.1093/nar/gkv1070.

GOLDSTEIN, Bob; KING, Nicole. The Future of Cell Biology: Emerging Model Organisms. **Trends In Cell Biology**, [s.l.], v. 26, n. 11, p.818-824, nov. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.tcb.2016.08.005.

KESELER, I. M.. EcoCyc: a comprehensive database resource for Escherichia coli. **Nucleic Acids Research**, [s.l.], v. 33, n. 0, p.334-337, 17 dez. 2004. Oxford University Press (OUP). http://dx.doi.org/10.1093/nar/gki108.

ASHBURNER, Michael et al. Gene Ontology: tool for the unification of biology. **Nature Genetics**, [s.l.], v. 25, n. 1, p.25-29, maio 2000. Springer Nature. http://dx.doi.org/10.1038/75556.

KANEHISA, Minoru; SATO, Yoko; MORISHIMA, Kanae. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. **Journal Of Molecular Biology**, [s.l.], v. 428, n. 4, p.726-731, fev. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.jmb.2015.11.006.

CYTOSCAPE CONSORTIUM. **What is Cytoscape?** Disponível em: <https://cytoscape.org/what_is_cytoscape.html>. Acesso em: 01 ago. 2019.

PYTHON SOFTWARE FOUNDATION. **General Python FAQ.** Disponível em: <https://docs.python.org/3/faq/general.html>. Acesso em: 27 jul. 2019.