

Universidade Federal do Pará Programa de Pós-Graduação em Ciência da Computação

Disciplina: Metodologia Científica **Discente:** Edwin Jahir Rueda Rojas

Pesquisa

A pesquisa tem como prioridade identificar em uma base de dados de genes, quais são os candidatos os quais podem ser genes *HouseKeeping*, para isso, no começo, a gente tem que gerar mais dados os quais são considerados como *HouseKeeping* pela literatura, a gente faz isso com as GAN's (*Generative adversarial network*), depois disso, mediante aprendizado de maquina a gente consegui dizer se o gen novo pode ser o não um bom candidato, isso já que se precisa reduzir o número de "candidatos".

Protocolo do mapeamento sistemático

Foco da pesquisa: Identificar se um gene é o não candidato à ser *Housekeeping* baseado na utilização de redes *GAN's*.

Questão de pesquisa: Será que o gene é um gene candidato à ser *HouseKeeping?*

String de busca:

- **Passo 1 (termos-chave):** gene, candidato, *HouseKeeping*
- Passo 2 (sinônimos):
 - **gene**: DNA, genético, gênico
 - **candidato**: competidor, pretendente, postulante, aspirante, pretendedor, solicitante, proponente.
 - **HouseKeeping:** constitutivo, integrante, característico, peculiar, distintivo, típico.
- Passo 3 (operador OR):
 - gen **OR** DNA **OR** genético **OR** gênico
 - competidor OR pretendente OR postulante OR aspirante OR pretendedor OR solicitante OR proponente.
 - Housekeeping OR constitutivo OR integrante OR característico OR peculiar OR distintivo OR típico.
- Passo 4 (operador AND):

(gen **OR** DNA **OR** genético **OR** gênico **OR** gen*) **AND**

(competidor OR pretendente OR postulante OR aspirante OR pretendedor OR solicitante OR proponente) AND

(Housekeeping **OR** constitutivo **OR** integrante **OR** característico **OR** peculiar **OR** distintivo **OR** típico)

(gen* OR DNA) AND (candidate OR competitor OR suitor OR postulant OR aspirant OR proponent OR identify*) AND (housekeeping OR HKG) AND (in-silico)

Bases de dados: As seguintes bases de dados são escolhidas já que são as bases nas quais são publicados mais artigos que tentam com inteligencia artificial.

- IEEE
- Springer Link
- NIPS (Nueral Information Processing Systems)

Critérios de inclusão e exclusão:

Inclusão:

- Tem que estar escritos em português, espanhol ou inglês.
- Tem que ter resultados comparáveis.
- Tem que ter as palavras chaves.

Exclusão:

- Trabalhos publicados antes do ano 2014.
- Trabalhos que não empreguem métodos de inteligencia artificial.

String: (gen* **OR** DNA) **AND** (candidate **OR** competitor **OR** suitor **OR** postulant **OR** aspirant **OR** proponent **OR** identify*) **AND** (housekeeping **OR** HKG) **AND** (in-silico)

Resultados da busca

Base de dados	Passo 1	Passo 2	Passo 3
IEEE	11	1	1
Springer Link	1.754	953	1
Busca Manual	-	-	1
Total	1.765	954	3

Tabela 1. Quantidade de trabalhos selecionados.

Passo 1: Pesquisa inicial. Passo 2: Critérios de inclusão. Passo 3: Critérios de exclusão.

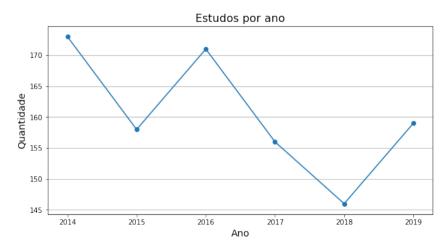


Figura 1. Estudos por ano sobre estimação de genes housekeeping.

Extração dos dados

Título: A Computational Approach Using Ratio Statistics for Identifying Housekeeping Genes from

cDNA Microarray Data

Fonte: IEEE Resumo:

Os autores do artigo tentam predizer se um gene é ou não *housekeeping*, para isso eles empregam *Ratio Statistics Based Normalization Strategy*, jogam os dados que conhecem que não são genes *housekeeping* e fazem um modelo gaussiano para cada *feature* do gene para assim ajustar o limite da gaussiana com os genes que são *housekeeping* mediante uma validação cruzada. O limite é ajustado com o produtório das probabilidades gaussianas de cada *feature*. Ao final os autores conseguem dizer se o gene é ou não *housekeeping* com um acerto razoável.

Resposta da questão de pesquisa:

Os autores conseguem dizer se um gene é ou não *HKG*, com um acerto considerável, sendo que **p(x)** < **épsilon** quer dizer que o gene é candidato a gene *HKG*, eles conseguem encontrar esse valor de **épsilon** que da o resultado mais favorável.

Título: RNA-sequence data normalization through in silico prediction of reference genes: the

bacterial response to DNA damage as case study

Fonte: Springer

Resumo:

Os autores conseguem fazer predições de possíveis novos *genes housekeeping* através de programação dinâmica (DP) e mediante de uma normalização quadrática. Fazendo *clustering* hierárquicos dos respetivos genes para ao final jogar os possíveis genes *housekepping* para assim poder ver que tão semelhante aos *cluster* é o gene, e assim dizer se o gene é o não um possível gene *housekepping*.

Resposta da questão de pesquisa:

Os autores conseguem dizer se um novo gene é ou não candidato a gene *HKG*, baseados nos *clusters* hierárquicos feitos com os genes *HKG* da literatura.

Título: *Elucidating tissue specific genes using the Benford distribution*

Fonte: Springer

Resumo:

Os autores conseguem modelar os genes mediante uma distribuição de *Benford*, eles conseguem ver que para certos genes *HKG*, a distribuição desses genes consegui ser semelhante a uma de *Benford*, modelando esses genes dessa forma, eles conseguem fazer um classificador baseado na distancia euclidiana (KNN), sendo o melhor k igual a sete. Assim, eles conseguem ter um classificador com um erro baixo e assim poder dizer se um novo gene é o não candidato a ser gene *housekeeping*.

Resposta da questão de pesquisa:

Os autores conseguem dizer com certo acerto (o acerto do classificador) se o novo gene \acute{e} ou não um gene HKG.