Trabalho apresentado no XXXIX CNMAC, Uberlândia - MG, 2019.

**Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**

# Desenvolvimento de Bactérias Artificiais Mutantes de *S. agalactiae* Híbridas entre Humano e Tilápia Usando Algoritmo Evolucionário e Lógica *Fuzzy*

Edgar L. Aguiar[1]
Departamento de Modelagem Matemática e Computacional, CEFET-MG, Belo Horizonte, MG
Gustavo H. M. Mendonça
Departamento de Modelagem Matemática e Computacional, CEFET-MG, Belo Horizonte, MG
Claudia B. Assunção
Núcleo de Pós-Graduação e Pesquisa da Santa Casa de Belo Horizonte - Belo Horizonte, MG
Sandro R. Dias
Departamento de Modelagem Matemática e Computacional, CEFET-MG, Belo Horizonte, MG
Thiago. S. Rodrigues
Departamento de Modelagem Matemática e Computacional, CEFET-MG, Belo Horizonte, MG

**Resumo**. Os microrganismos podem ser classificados e observados em diversos reinos biológicos, sendo amplamente distribuídos nos mais diferentes ambientes do planeta, podendo ser detectados nos lugares mais comuns aos mais hostis, causando impactos diretamente no padrão ambiental e em diversos outros organismos. Um dos microrganismos de elevado potencial biotecnológico e patogênico são as bactérias. O objetivo desse trabalho é gerar bactérias artificiais mutantes híbridas de humano com peixe da espécie *Streptococcus agalactiae* por meio de algoritmo evolucionário, combinando os genes de bactérias encontradas em peixes e humanos, e sua posterior classificação em uma máquina de inferência *Fuzzy*. Para auxiliar o trabalho foram realizadas análises comparativas entre os perfis genômicos das bactérias, para classificá-las e tratá-las de forma preventiva, evitando possíveis surtos.

**Palavras-chave**. *Streptoccus Agalactiae*. Bactéria Híbrida. Algoritmo Genético. Inteligência Computacional. Lógica *Fuzzy*.

## 1 Introdução

Os microrganismos podem ser classificados e observados em diferentes reinos biológicos, sendo amplamente distribuídos na natureza. A alta diversidade de microrganismos contribui para um alto volume de informações que impulsionam estudos em diversas áreas: agrária, medicina e indústria [12]. Os microrganismos podem ser detectados nos lugares mais hostis, afetando diretamente o padrão ambiental e outros organismos. Levando-se em consideração essas informações, é possível apontar que tais características possam refletir em impactos significativos na economia, e saúde de forma positiva ou negativa [8].

[1]edgarlaguiar@gmail.com

 

Dentre os microrganismos mais estudados estão as bactérias. Elas são de grande importância medicinal, biotecnológica, veterinária, ambiental e um dos mais antigos organismos da Terra, com amostras localizadas em rochas de 3,8 milhões de anos [8]. Existem vários fatores que tornam as bactérias bons organismos modelos para diversos estudos genômicos como: conservação de funções celulares em comum com organismos mais complexos, a possibilidade de se obter rapidamente condições de cultivo adequadas com uma elevada quantidade de material biológico, e possuir um custo de desenvolvimento significativamente menor quando comparado a organismos eucariotos [5], [8].

Entre as inúmeras características das bactérias, uma das mais importantes é a reprodução, pois ela permite a geração de mutações, transferências de genes entre espécies, trocas e ganhos de fatores de virulência e aquisição de novas resistências [8]. Os novos fatores de virulência e resistência aumentam significativamente o risco para os hospedeiros, pois a cada nova resistência adquirida, aumenta-se a necessidade de se desenvolver novos antibacterianos e bactericidas para combatê-los, sendo que a cada dia, antigos fármacos se tornam menos eficazes e, assim, aumentam o custo de pesquisas para se combater as novas resistências obtidas pelas novas gerações [11].

Tendo em vista as informações que são categoricamente discutidas pela literatura, nota-se que os microrganismos, em especial as bactérias, despertam grande interesse em pesquisas direcionadas às áreas abordadas neste trabalho. O total de projetos de genomas depositados na base de dados (BD) Gold desde da sua criação em 2007 as bactérias eram o maior grupo. No ano de 2011, o BD armazenava aproximadamente 10 mil genomas bacterianos, e já em 2015, o valor chegou a aproximadamente 50 mil genomas, em 2018 teve um salto significativo para aproximadamente 140 mil genomas [3].

Dentro do filo das *Firmicutes* encontra-se o gênero *Streptococcus*. Esse gênero atualmente tem 121 espécies, 23 subespécies identificadas e, aproximadamente, 100 mil artigos relacionados [7], [10]. Dentre as espécies mais estudadas, destaca-se a *Streptococcus agalactiae*, uma bactéria patogênica que afeta diversos hospedeiros.

Atualmente existem aproximadamente 40 genomas completos de *S. agalactiae* depositados no banco de dados públicos, tornando-se necessário um aumento significativo no número de linhagens, pois essa quantidade baixa é um grande dificultador das análises comparativas. Outros fatores dificultadores são: o custo elevado de sequenciamento com plataformas NGS (*Next-Generation Sequencing*) e a elevada dificuldade técnica na montagem de genomas completos. Por tais motivos o necessário desenvolvimento de novas abordagens *in sílico* para simular novas linhagens [9].

Neste trabalho, foi implementado um algoritmo bioevolutivo, de modo a criar computacionalmente bactérias mutantes que contenham os genes de peixe e os de humano para análises futuras. Essas bactérias serão geradas considerando o perfil genômico das espécies, sua probabilidade de sobrevivência e sua capacidade adaptativa mais próximos à realidade. Serão dados maiores pesos aos genes de peixes originários da Ásia e genes de humanos da América, pois o maior objetivo é estudar e prevenir mutações de bactérias com características destes peixes e com capacidade de afetarem *Homo sapiens* na América. Em seguida, será criada um classificador de inferência *Fuzzy* onde as bactérias serão classificadas considerando as informações obtidas pelo algoritmo genético, e resultando em uma nota para cada bactéria referente à sua probabilidade de sobrevivência e sua capacidade de

DOI: 10.5540/03.2020.007.01.0368  

adaptação. Uma das formas de realizar essa classificação é por meio da utilização de Inteligência Computacional (IC). Pois através do conhecimento das especialistas Biomedicina, Medicina da Santa Casa BH, e Microbiologia pela UFMG, em conjunto com desenvolvimento da IC foi possível avaliar a eficiência da máquina *Fuzzy*, considerando os aspectos de sobrevivência e adaptação, para uma melhor comparação e análise dos resultados.

## 2 Algoritmo Genético

Observando a evolução tecnológica principalmente na inteligência artificial, fica claro que a natureza serve como fonte de inspiração para os cientistas. Um dos exemplos mais difundidos é a computação evolucionária, baseada no princípio da evolução das espécies e na genética. Neste contexto, existem algoritmos evolucionários, criados a partir das ideias de evolução dos indivíduos na natureza, inspiradas pelo Darwinismo e Lamarckismo. Os algoritmos de computação evolutiva trabalham com um conjunto (população) de soluções candidatas que interagem entre si e competem pela permanência na população. A evolução é alcançada basicamente pelos processos de reprodução dos indivíduos com herança genética, variação em uma população de indivíduos (mutação, que ajuda a criar diversidade na população) e aplicação da "seleção natural"para produção da próxima geração, onde indivíduos mais bem adaptados tem maiores chances de sobreviverem [6].

Segundo Goldberg e Holland(1988), um algoritmo genético (AG) é uma técnica de busca que localiza uma sequência ótima, através do processamento de uma população de sequências inicializadas aleatoriamente, usando técnicas inspiradas na biologia evolutiva, como hereditariedade, mutação, seleção natural e recombinação (*Crossover*) [4]. O primeiro passo do AG é iniciar com uma população de indivíduos (que pode ser feita de forma aleatória ou determinística), representando um conjunto de soluções candidatas para o problema. Em seguida, a população é avaliada, de maneira que cada indivíduo recebe um valor de aptidão (*Fitness*), indicando quão boa aquela solução é para o problema. Este valor é utilizado para comparação entre os indivíduos e dá maior probabilidade aos indivíduos com aptidão maiores de serem selecionados para a próxima geração e para gerar descendentes no cruzamento, porém não descartando a possibilidade de indivíduos menos adaptados de também se reproduzirem ou serem selecionados. Esta característica torna o processo de seleção probabilístico, pois cada indivíduo possui uma determinada probabilidade de ser selecionado de acordo com o valor de aptidão [6].

## 3 Lógica *Fuzzy*

Ao descrevermos certos fenômenos ou características relacionados ao mundo é comum a utilização de graus que representam qualidades ou verdades parciais. Como exemplo, podemos considerar o grupo "pessoas altas", e uma abordagem para caracterizar este grupo é dada de maneira que os indivíduos sejam considerados altos com maior ou menor grau, ou seja, existem elementos que pertenceriam mais à classe dos altos que outros. Por consequência, quanto menor (ou maior) for a medida da altura do indivíduo, menor (ou maior) será seu grau de pertinência a esta classe. Desse modo, podemos dizer que

 

os indivíduos pertencem à classe das pessoas altas, com maior ou menor intensidade. Partindo deste tipo de questões, onde a propriedade que define o conjunto é incerta, que surgiu a teoria dos conjuntos *Fuzzy*, que tem crescido consideravelmente em nossos dias, tanto do ponto de vista teórico como nas aplicações em diversas áreas de estudo, como computação e matemática [2].

Na lógica *Fuzzy* a ambiguidade semântica da inteligência humana pode ser representada por meio de variáveis linguísticas e seus termos primários. Uma variável linguística é uma entidade utilizada para se representar de modo impreciso e, portanto, linguístico, um conceito ou variável de um dado problema. Ela admite como valores as expressões linguísticas (chamadas de termos primários), em contraste com uma variável numérica que assume apenas valores precisos. Os termos primários de uma variável linguística formam a sua estrutura de conhecimento. Por exemplo, a variável linguística "altura"poderia admitir os termos primários "muito alto"e "razoavelmente baixo" [1].

A estrutura adotada nos métodos de inferência *Fuzzy* de Mamdani define regras de produção que possuem relações *Fuzzy* tanto em seus antecedentes quanto em seus consequentes [1]. Neste método, o módulo de interface (entrada) recebe valores numéricos e os converte em conjuntos *Fuzzy* equivalentes, ocorrendo uma conversão escalar $\rightarrow$ *Fuzzy* (Fuzzificação). A máquina de inferência então busca em seu banco de conhecimento e processa as regras disparadas pela entrada, fazendo uma composição pelo método Max-Min. Então, o módulo de interface de saída recebe um conjunto *Fuzzy* (processado anteriormente) para cada variável de entrada e o converte em um valor escalar correspondente, gerando saídas compatíveis com os demais sistemas [1].

## 4 Análise Comparativa Genômica

A análise comparativa genômica consiste em comparar e selecionar quais os genes que são encontrados em determinadas linhagens do *Cluster* sobre uma determinada condição [13]. Essas análises inicialmente podem ser realizadas pelos identificadores de genes, porém o mesmo gene é anotado com diferentes termos em diversas linhagens tal fato dificulta uma análise comparativa por termos e ontologias. Foi necessário realizar alinhamento de sequências locais entre os genes de cada linhagem do Cluster, os parâmetros foram de no mínimo 90% de identidade e 90% de cobertura. Após obtenção das linhagens no formato gbff no Site do NCBI, as linhagens foram separadas em diversos clusters gênicos, sendo realizadas diversas análises: 1° - *Cluster Homo sapiens*: *Core, Parcial e Unique*; 2° - *Cluster Homo sapiens* América X *Homo sapiens* Ásia: *Core, Parcial e Unique*; 3° - *Cluster Oreochromis sp*: *Core, Parcial e Unique*; 4° - *Cluster Oreochromis sp* América X *Oreochromis sp* Ásia: *Core, Parcial e Unique*; 5°- *Cluster Homo sapiens X Oreochromis sp*: *Core, Parcial e Unique*. A análise comparativa de Core consiste em selecionar quais genes são encontradas em todas as linhagens do *cluster*. Já a Parcial são todos os genes que existem em um determinado grupo e não existem no outro e o *Unique* são os genes que existem exclusividade em um sub grupo refinado de um determinado grupo. As saídas das análises comparativas foram usadas para as próximas etapas do trabalho.

DOI: 10.5540/03.2020.007.01.0368  

## 5 Implementação do Algoritmo Genético

O AG foi implementado em linguagem Java 6, na IDE NetBeans. As informações referentes ao genes foram obtidas no NCBI. Todos os genes do *Core* Genoma foram definidos com peso 1 (pois são essenciais), os genes do *Parcial Core* foram definidos com peso 2, os genes *Exclusive* de peixe América foi dado peso 1,5. Já para genes *Exclusive* de peixe da Ásia foi dado peso 3, pois são genes que tem maior importância para este trabalho, uma vez que os peixes da Ásia afetando humanos da América seriam o caso mais grave devido às diferentes possibilidades de mutações e o atual despreparo para esta situação hipotética. Pelo mesmo motivo, os genes *Exclusive* de humano da América possuem peso 3. Finalmente, aos genes *Exclusive* de humano da Ásia foi dado peso 2. Para uma bactéria ser factível, e mais fiel a uma possível bactéria real, foi considerado que a mesma deve ter tamanho entre 2000kb e 2400kb, deve possuir todo o core genoma do hospedeiro (humano), deve possuir entre 40% e 60% do *Parcial Core* (peixe) e deve possuir entre 40% e 70% do *Unique Core*. Para a seleção foram utilizadas a roleta simples e com ranking. O número de pareamentos para o cruzamento foi definido como a metade do número de indivíduos da população. Se a população possui 100 bactérias, ocorrerão 50 pareamentos, por exemplo. O *crossover* foi implementado considerando os genes *Parcial* e *Unique* das bactérias, uma vez que o core genoma do hospedeiro deve estar totalmente presente na bactéria (sendo inclusive critério de factibilidade), não fazendo sentido utilizar *crossover* neste grupo de genes. O *Crossover* é realizado com 1 ponto de corte definido aleatoriamente para o *Parcial Core* e outro para o *Unique core* da bactéria, podendo ser pontos diferentes. Os parâmetros utilizados nas execuções foram os seguintes: Tamanho da População: 250; N. de Gerações: 400; Taxa de Mutação: 0,3; Taxa de Cruzamento: 0,8; Critério de parada: N. de gerações atingido.

## 6 Máquina de Inferência *Fuzzy*

As bactérias resultantes das execuções do algoritmo completo de AG foram classificadas em um método de inferência *Fuzzy* de Mamdani. O sistema possui 3 variáveis linguísticas de entrada e 2 de saída. As variáveis linguísticas de entrada são o tamanho da bactéria, sua pontuação (valor) de *Parcial Core* e sua pontuação (valor) de *Unique Core*. A entrada (tamanho) que admite os termos primários pequeno, ideal e grande. Os 3 termos foram representados por funções triangulares, sendo que o termo ideal abrange de 1800kb a 2600kb, possuindo pertinência 1 em 2200kb. As variáveis de entrada *Score Parcial* e *Score Unique* admitem os termos primários baixo, médio e alto, todas sendo representadas por funções triangulares. Para estas variáveis, quanto maior seu valor, melhor a bactéria.

As variáveis de saída foram adaptabilidade e probabilidade de sobrevivência. Ambas admitem os termos primários muito baixa, baixa, média, alta e muito alta. Procuramos a bactéria com a melhor adaptabilidade e melhor probabilidade de sobrevivência possível. Os termos muito baixa e muito alta foram representados por funções trapezoidais, enquanto os termos baixa, média e alta foram representados por funções triangulares. As partições *Fuzzy* de saída admitem valores entre 0 e 1, indicando qual bactéria que possui melhor valor de adaptabilidade e probabilidade de sobrevivência.

DOI: 10.5540/03.2020.007.01.0368  

# 7 Resultados Obtidos

O AG foi executado 10 vezes em sequência devido ao tempo de execução e a limitação computacional da máquina utilizada. Após o fim de cada execução foi selecionado a melhor bactéria gerada (elitismo). Em seguida as características de cada uma foram inseridas na máquina de inferência *Fuzzy*, criada no Matlab, e as bactérias foram classificadas considerando sua adaptabilidade e probabilidade de sobrevivência.

Após a classificação das bactérias pelo método de inferência *Fuzzy*, as especialistas Microbiologia e Medicina também classificaram as bactérias considerando os mesmos fatores. Os resultados da classificação pelo método *Fuzzy* e pelo especialistas podem ser vistos na Tabela 1, onde é mostrado para cada uma das execuções o tamanho da bactéria, sua pontuação do *Parcial Core*, pontuação do *Unique Core*, classificação de 0 a 1 referente à adaptabilidade da bactéria pelo método *Fuzzy*, classificação 0 a 1 referente à adaptabilidade da bactéria pelo método *Fuzzy*, classificação quanto adaptação e sobrevivência pelas especialista. Os resultados demonstram que as bactérias geradas pelo AG possuem bons atributos tanto em adaptabilidade quanto em probabilidade de sobrevivência, sendo bons padrões para um estudo aprofundado de suas cargas genéticas e características, além das consequências de sua existência. Isso fica evidenciado ao perceber que todas as 10 bactérias alcançaram pontuação superior a 65% para ambos os atributos considerados tanto pela classificação do método de inferência *Fuzzy* e 60% pelas especialistas.

Tabela 1: Resultado das classificações das bactérias pelo Fuzzy e as Especialistas

| | | | | Nota Fuzzy | Nota Fuzzy | Nota Esp 1 | Nota Esp 2 |
|---|---|---|---|---|---|---|---|
| Exec. | Tamanho | Score P. | Score U. | Adapt. | Sobreviv. | Adapt. | Sobreviv. |
| 1 | 2395 | 3809 | 16,47 | 0,656 | 0,661 | 0,611 | 0,67 |
| 2 | 2397 | 3809 | 18,50 | 0,657 | 0,714 | 0,62 | 0,73 |
| 3 | 2400 | 3829 | 13,52 | 0,671 | 0,653 | 0,70 | 0,66 |
| 4 | 2397 | 3799 | 17,72 | 0,662 | 0,685 | 0,60 | 0,69 |
| 5 | 2400 | 3842 | 15,96 | 0,814 | 0,655 | 0,76 | 0,665 |
| 6 | 2399 | 3821 | 16,05 | 0,653 | 0,654 | 0,605 | 0,665 |
| 7 | 2399 | 3817 | 18,10 | 0,655 | 0,696 | 0,61 | 0,71 |
| 8 | 2398 | 3825 | 17,42 | 0,656 | 0,677 | 0,611 | 0,68 |
| 9 | 2398 | 3812 | 15,12 | 0,651 | 0,651 | 0,609 | 0,66 |
| 10 | 2400 | 3816 | 17,51 | 0,654 | 0,678 | 0,61 | 0,685 |

# 8 Conclusões

Aplicação do método *Fuzzy* se mostrou muito útil para auxiliar na tomada de decisões e na classificação dos diversos perfis das bactérias, pois através da mesma foi possível condensar um alto volume regras, para criar um classificador funcional de microrganismos híbridos. Observarmos que após execuções dos algoritmos foi possível gerar bactérias mutantes híbridas contendo os genes das linhagens de humano e peixe factíveis, e que elas tiveram bons resultados de adaptabilidade e sobrevivência variando entre 65 a 80%.

DOI: 10.5540/03.2020.007.01.0368

As opiniões das especialistas na classificações das bactérias foram bem próxima da classificação realizada pela máquina *Fuzzy*, e isso demonstra que a mesma está bem calibrada para esta função. Dependendo do perfil desejado da bactéria e do objetivo, podem ser selecionadas diferentes bactérias. Por exemplo no caso da melhor bactéria com maior capacidade adaptativa, seria a bactéria 5 que possui valor de 0,814 de acordo com a máquina *Fuzzy* e 0,760 na opinião do especialista. Já no caso de melhor bactéria com maior capacidade de sobrevivência seria a bactéria 2 com valor de 0,714 atribuído pela máquina *Fuzzy* e 0,730 de acordo com o especialista.

## Referências

[1] P. E. M. Almeida, and G. E. Alexandre, *Sistemas fuzzy*, SO Sistemas Inteligentes: fundamentos e aplicações, Cap 7, Manole,(2003): 169-201.

[2] M. Amendola, and L. C. Barros. *Manual do uso da teoria dos conjuntos Fuzzy no MATLAB 6.5*, Feagri e Imecc/Unicamp (2005): 1-44.

[3] Gold, *GOLD Statistics 2019.*, Disponível em:<https://gold.jgi-psf.org/statistics>. Acesso em: 13 de Março de 2019.

[4] D. E. Goldberg, and J. H. Holland., *Genetic algorithms and machine learning*, Machine learning 3.2 (1988): 95-99.

[5] M. J. Pelczar Jr, E.C.S. Chan, N. R. Kreig, *Microbiologia: conceitos e aplicações*, Vol. 1., São Paulo, Makron Books, 1996.

[6] R. Linden, *Algoritmos genéticos*, 2a edição, Brasport, 2008.

[7] Lpsn, *List of Prokaryotic names*, Disponível em: http://www.bacterio.net/streptococcus.html. Acesso em: 13 de Março de 2019.

[8] M. T. Madigan et al., *Microbiologia de Brock* , 14a Edição, Artmed Editora, 2016.

[9] D. C. B. Mariano, F. L. Pereira, E. L. Aguiar et al., *SIMBA: a web tool for managing bacterial genome assembly generated by Ion PGM sequencing technology.* BMC Bioinformatics , v. 17, p. 65-72, 2016.

[10] NCBI, *PubMed Search Streptococcus.* Disponível em: https://www.ncbi.nlm.nih.gov/pubmed/?term=Streptococcus , Acesso em: 13 de Março de 2019.

[11] P. Quinn et al., *Microbiologia veterinária e doenças infecciosas*, Artmed Editora, 2005.

[12] G. J. Tortora; C. L. Case; B. R. Funke, *Microbiologia* 12a, Artmed Editora, 2016.

[13] J.C. Venter, M.D. Adams, E.W. Myres et al., *The Sequence of the Human Genome*, Science, 291(5507): 1304-1351.

 