10th Brazilian Congress on Computational Intelligence (CBIC'2011), November 8 to 11, 2011, Fortaleza, Ceará Brazil

GENEBINGO: IDENTIFICAÇÃO DE GENES UTILIZANDO REDE NEURAL ARTIFICIAL

Dieval Guizelini, Fábio de Oliveira Pedrosa, Jeroniza Nunes Marchaukoski, Lucas Martins Ferreira, Maria Berenice Reynaud Steffens, Giselle Munhoz Alves, Michelly Alves C. Gehlen, Roberto Tadeu Raittz

Programa de Pós-Graduação em Bioinformática, Universidade Federal do Paraná, 81520-260, Brasil
dievalg@gmail.com, fpedrosa@ufpr.br, jeroniza@ufpr.br, lucas@lmferreira.com, berenice.steffens@gmail.com, gmunhoz@gmail.com, cegempac.michelly@gmail.com, raittz@ufpr.br

Resumo – Este artigo descreve uma nova aplicação computacional, o GeneBingo, para encontrar genes nos genomas microbianos. Os testes realizados apresentam um desempenho médio de 36,57% superior aos resultados obtidos pelo GLIMMER (aplicação mais utilizada para este fim). O GeneBingo explora a aplicação de Rede Neural Artificial para classificação das "Open reading frames" candidatas. Para desenvolvimento da rede foram identificadas características baseadas em métodos conhecidos como RBSFinder, FramePlot, percentual de Guanina e Citosina (GC) e dicionários com regiões promotoras ou reguladoras.

Palavras-chave – Rede Neural Artificial, *Open Reading Frames* (ORF), Reconhecimento de Padrões, anotação de genomas.

1 Introdução

O desenvolvimento de novos equipamentos e técnicas de sequenciamento de DNA tem possibilitado o crescimento exponencial da quantidade de organismos com genomas completamente seqüenciados. A análise desses genomas tem apresentado uma grande densidade de regiões que codificam proteínas, geralmente com 90% ou mais da sequências do DNA [19].

Após o sequenciamento, em geral, os pesquisadores utilizam os programas GLIMMER [19], Gene Locator and Interpolated Markov ModelER do Institute for Genomic Research (TIGR) [4] e o RBS Finder [20] para identificar as regiões que codificam proteínas. Tais aplicações utilizam modelos de Markov[5], e o programa mais utilizado, o GLIMMER, classifica em média 57% dos genes[4]. A manipulação destes programas exige grande experiência e sensibilidade do pesquisador na área de genômica e uma expressiva necessidade de realizar várias iterações.

Este trabalho propõe a utilização de Rede Neural Artificial (ANN) [8] como método alternativo aos Modelos Ocultos de Markov [5]. Segundo Lancashire *et al* [8] as principais vantagens no uso de Redes Neurais Artificiais são: a tolerância a falhas, a capacidade de generalização e a possibilidade de trabalhar com dados que apresentem relações e inter-relações complexas.

O método Free Associative Neurons (FAN) [3, 21] foi criado em 1998 e aperfeiçoado em 2002. Tal método é uma ANN híbrida, que apresenta uma abordagem neuro-fuzzy para reconhecimento supervisionado de padrões.

Este artigo apresenta um programa que usa ANN para a busca de regiões codificadoras de proteínas em sequências de DNA. Os estudos foram realizados nos domínios Bactéria e Archaea, porém a metodologia pode ser aplicada aos outros domínios.

2 Metodologia

Este trabalho utiliza as abordagens qualitativas e quantitativas. Para isso, foi adotada uma metodologia própria, composta pelas etapa a seguir:

- Obtenção de genomas seqüenciados e anotados;
- Obtenção de um conjunto conhecido de ORFs (Open Reading Frames) verdadeiras;
- Obtenção de um conjunto de falsas ORFs;
- Identificação de características;
- Desenvolvimento de aplicação para extração das características;
- Treinamento de rede neural artificial para classificação das ORFs;
- Desenvolvimento de aplicação para extração de características e classificação das ORFs;

- Escolha de genomas previamente anotados para o processo de validação;
- Identificação das ORFs com a aplicação GLIMMER para fins de comparação de resultados;
- Desenvolvimento de aplicação para comparação dos resultados;
- Comparação dos resultados.

Para validação do método, os resultados foram comparados com os resultados apresentados por outros programas de identificação de ORFs.

3 Discussão e Análise de Resultados

3.1. Obtenção de genomas seqüenciados e anotados

Os genomas completamente sequenciados e anotados, disponíveis em formato GenBank, foram obtidos na pasta /Bactéria do serviço de FTP do GenBank do NCBI (no endereço ftp://ftp.ncbi.nih.gov/genomes/Bacteria). Dos arquivos disponíveis, foram separados os arquivos dos 24 organismos que são citados na literatura do GLIMMER (Tabela 1), para fins de comparação. Os demais foram agrupados em 5 grupos, organizados segundo os percentuais de GC (Tabela 2). Em cada grupos foram escolhidos aleatoriamente 10 arquivos para serem utilizados no estudo das características e no treinamento da rede (listados na Tabela 2). Nos artigos de Marin *et al*. [11] , Mackiewicz *et al*. [10], Berger *et al*. [1], Li *et al*. [9], bem como em outros artigos, foi observada a relevância do percentual de GC em relação a análise genômica. As informações dos organismos, quantidade de genes anotados foram obtidos do NCBI GenBank revisão 178.

Tabela 1 - Lista dos organismos utilizados nas etapas de validação e comparação

Organismo	**Genes anotados**	**%GC**
Archaeoglobus fulgidus DSM 4304	2486	48.58
Bacillus subtilis subsp. subtilis str. 168	4423	43.51
Campylobacter jejuni subsp. doylei 269.97	2037	30.57
Carboxydothermus hydrogenoformans Z-2901	2707	42.05
Caulobacter crescentus CB15	3819	67.21
Chlorobium tepidum TLS	2337	56.53
Clostridium perfringens ATCC 13124	3017	28.37
Colwellia psychrerythraea 34H	5054	38.00
Dehalococcoides ethenogenes 195	1642	48.85
Escherichia coli str. K-12 substr. MG1655	4493	50.79
Geobacter sulfurreducens PCA	3523	60.93
Haemophilus influenzae 86-028NP	1899	38.15
Helicobacter pylori P12	1624	38.80
Methylococcus capsulatus str. Bath	3052	63.58
Mycobacterium tuberculosis H37Ra	4084	65.61
Neisseria meningitidis MC58	2225	51.52
Porphyromonas gingivalis ATCC 33277	2155	48.35
Pseudomonas fluorescens Pf-5	6233	63.30
Ralstonia solanacearum GMI1000	3503	67.03
Thermotoga maritima MSB8	1928	46.24
Treponema denticola ATCC 35405	2838	37.87
Treponema pallidum subsp. pallidum str. Nichols	1095	52.77
Streptococcus agalactiae NEM316	2235	35.62
Streptococcus pneumoniae Hungary19A-6	2402	39.63

Tabela 2 - Listas dos organismos por grupo de %GC utilizados para extração de características

Grupo (faixa de %GC)	**Organismo**	**Genes Anotados**	**%GC**
I	*Buchnera aphidicola* str. APS (Acyrthosiphon pisum)	607	26,31
(16.6,32.51)	*Campylobacter hominis* ATCC BAA-381	1799	31,73
	Campylobacter jejuni RM1221	1940	30,31
	Campylobacter jejuni subsp. jejuni 81116	1681	30,54
	Candidatus Carsonella ruddii PV	213	16,56
	Candidatus Phytoplasma Mali	536	21,39
	Clostridium botulinum A str. ATCC 3502	3776	28,24
	Clostridium botulinum B str. Eklund 17B	3586	27,51
	Francisella tularensis subsp. tularensis SCHU S4	1852	32,26
	Mycoplasma mycoides subsp. mycoides SC str. PG1	1052	23,97

II	*Aliivibrio salmonicida* LFI1238	1115	38,24
(32.51,44.17)	*Haemophilus somnus* 2336	2065	37,38
	Lactobacillus johnsonii NCC 533	1918	34,61
	Leptospira borgpetersenii serovar Hardjo-bovis L550	270	40,16
	Marinomonas sp. MWYL1	4598	42,59
	Methanosarcina mazei Go1	3436	41,48
	Staphylococcus aureus subsp. aureus USA300_TCH1516	2802	32,76
	Streptococcus equi subsp. equi 4047	2243	41,28
	Streptococcus pyogenes str. Manfredo	1907	38,63
	Streptococcus thermophilus LMG 18311	1974	39,09
III	*Coprothermobacter proteolyticus* DSM 5265	1541	44,77
(44.17, 55.83)	*Corynebacterium glutamicum* R	3128	54,13
	Dickeya zeae Ech1591	4367	54,52
	Escherichia coli UMN026	5096	50,72
	Robiginitalea biformata HTCC2501	3259	55,29
	Salmonella enterica subsp. enterica serovar Paratyphi C strain	4830	52,16
	Shewanella amazonensis SB2B	3785	53,59
	Shewanella baltica OS155	4521	46,28
	Shewanella loihica PV-4	3993	53,67
	Yersinia pestis Antiqua	4274	47,70
IV	*Agrobacterium tumefaciens* str. C58	2819	59,38
(55.83,67.49)	*Bordetella avium* 197N	3510	61,58
	Brucella canis ATCC 23365	2200	57,21
	Burkholderia multivorans ATCC 17616	2166	67,13
	Deinococcus deserti VCD115	2650	63,39
	Desulfococcus oleovorans Hxd3	3323	56,17
	Edwardsiella ictaluri 93-146	3894	57,44
	Gluconobacter oxydans 621H	2499	61,07
	Parvibaculum lavamentivorans DS-1	3707	62,33
	Xanthomonas oryzae pv. oryzae KACC10331	4279	63,69
V	*Anaeromyxobacter dehalogenans* 2CP-1	4573	74,72
(67.49,74.9)	*Anaeromyxobacter* sp. K	4557	74,84
	Burkholderia mallei ATCC 23344	3393	68,15
	Methylobacterium extorquens AM1	1177	67,65
	Myxococcus xanthus DK 1622	7456	68,89
	Nakamurella multipartita DSM 44233	5471	70,92
	Nocardioides sp. JS614	4755	71,65
	Rhodobacter sphaeroides KD131	3143	69,18
	Streptomyces coelicolor A3(2)	7912	72,12
	Streptomyces griseus subsp. griseus NBRC 13350	7224	72,23

3.2. Obtenção de um conjunto conhecido de ORFs verdadeiras

Para fins desse estudo, são consideradas "ORFs verdadeiras" as ORFs que codificam genes que expressam proteínas, segundo descrição contida nas anotações dos genomas completamente seqüenciados, montados e anotados. Para produção do arquivo de ORFs verdadeiras, no formato Fasta [18], foram sorteadas até 1000 genes anotados de cada um dos 5 grupos de %GC, totalizando aproximadamente 5000 seqüências de nucleotídeos de regiões anotadas e codificantes das 158872 existentes nos 50 organismos descritos na Tabela 3. Esses números são aproximados, em função de alguns dos organismos possuírem menos de 1000 genes anotados.

3.3. Obtenção de um conjunto de falsas ORFs

Neste trabalho, foram consideradas "ORFs falsas", as seqüências de nucleotídeos contidas em genomas que não são anotadas como genes, ou seja, possuem os códons de *start* e *stop*, porém não possuem necessariamente região regulatória, não codificam genes ou simplesmente não se encontram anotados. Para produção do arquivo de ORFs falsas, no formato Fasta, foram selecionadas aleatoriamente 5000 seqüências delimitadas por um *start* e *stop* códon, porém de áreas intergênicas ou de áreas internas a genes conhecidos dos genomas listados na Tabela 3. Esses números são aproximados, em função de alguns organismos possuírem menos de 1000 regiões delimitadas por *stop codons* e que possuíssem um *start codon* para ser considerado com um provável gene falso.

3.4. Identificação de características

O processo de extração e identificação de características foi completamente experimental e realizado com análise e treinamento sucessivos das redes. Esse processo consiste em: identificar uma unidade de medida que possa ser extraída a partir das seqüências de nucleotídeos, treinar a rede com todas as características e validar os resultados da rede em um conjunto conhecido e não utilizado na fase de treinamento. No processo de teste e validação da característica, a mesma era mantida se obtivesse melhores resultados.

As características foram concebidas a partir de quatro critérios:

1. Características conhecidas das informações gênica/genômica (*start códon*, *stop códon*, *frame* etc).
2. Metodologia descrita por outros autores ou aplicações de predileção de genes.
3. Análises *in silico* realizadas e que apresentaram algum ganho de qualidade nos testes realizados nos conjuntos de treinamento.
4. As informações precisam ser obtidas única e exclusivamente das seqüências de nucleotídeos a serem examinados e anotados.

Tais critérios, as características observadas, desenvolvidas e em uso na versão atual são:

1. Identificação do *start códon* do gene (informações disponibilizadas pelo NCBI – The Genetic Codes), codificados de acordo com a freqüência observada nos genes: 1-ATG (77,9%), 2-GTG (15,7%), 3-TTG (6,3%) e 4-CTG (menor que 0,1%);
2. Identificação do *stop códon* do gene (NCBI – The Genetic Codes), codificados de acordo com a freqüência observada nos genes: 1- TGA (46,5%), 2-TAA (36%) e 3-TAG (17,5%);
3. Frame em que o gene foi anotado;
4. Relação de (G-C)/(G+C) anterior ao start códon, calculado com uma janela de 30pb a 60pb;
5. Relação de (G-C)/(G+C) no início do gene, incluindo o start códon, calculado com uma janela de 30pb a 60pb, mantendo o mesmo tamanho da janela utilizada na característica anterior;
6. Relação de (G+C)/(A+C+G+T) anterior ao start códon, calculado com uma janela mínima de 30pb e máxima de 60pb.
7. Relação de (G+C)/(A+C+G+T) observada na janela de 30pb a 60pb do início do gene.
8. Identificação em qual dos frames ocorreu a maior variação de GCs, baseada no FramePlot [6].
9. Pribnow box (TATA-BOX) [17], busca na região -10 pela seqüência TATAAT ou por uma dessas 3 expressões regulares: "[tc]a[acgt][atg][acgt]t", "[tg]a[acgt][atg][acgt]t" e "[ta][at][acgt][atg][acgt]t".
10. Gilbert Box, busca na região -30 pela seqüencia consenso TTGAC ou por uma dessas expressões regulares: "[at]ttg[atcg][cat]" e "[cg]ttg[atcg][cat]".
11. Procura nos 20pb anteriores e posteriores ao start códon a presença de uma das seqüências consensos (TAAGGAG, CAGGAG, AGGAG e AGGA) para o RBS [20].
12. Busca por uma das 159 hexâmero variações possíveis para o TATABOX [17], classificadas por rede previamente treinada para esse fim.
13. Distância entre os stops códons contidos no mesmo frame.
14. Freqüência simples da ocorrência do nucleotídeo citosina (C) na 2ª base no gene.
15. Freqüência simples da ocorrência do nucleotídeo timina (T) na 2ª base no gene.
16. Freqüência simples da ocorrência do nucleotídeo adenina (A) na 3ª base no gene.
17. Freqüência simples da ocorrência do nucleotídeo guanina (G) na 3ª base no gene.

3.5. Desenvolvimento de Aplicação para Extração das Características

Foi desenvolvida uma aplicação em Java que carrega e interpreta o arquivo no formato GenBank. Com os arquivos carregados e através da análise das seqüências de nucleotídeos e das *features* de anotações do GenBank, foram obtidas as características das informações de quais regiões codificam e não codificam genes nos organismos listados na tabela 3.. Um extrato do arquivo de saída da aplicação é apresentado na Figura 01.

```
3 2 -1 -0.267 -0.600 0.500 0.222 3 0 0 0 0 0.524 4 4 7 2 2
1 2 2 0.000 0.765 0.033 0.283 7 0 0 1 0 0.837 3 12 11 7 2
1 1 -2 -0.059 1.000 0.283 0.250 5 0 0 3 0 0.058 0 1 0 1 2
1 2 -2 0.429 0.200 0.233 0.250 6 0 0 1 1 0.958 19 43 84 12 1
1 2 -1 0.294 0.429 0.283 0.350 4 0 0 0 1 0.992 43 48 56 29 1
1 2 -3 -0.636 0.167 0.183 0.200 4 0 0 1 0 0.977 68 130 170 43 1
```

Figura 01 – Extrato do arquivo de treinamento.

No arquivo de saída, cada linha é um padrão formado pelas 17 características descritas anteriormente, delimitadas por um espaço em branco, e pelas duas classes (1 gene, 2 não gene), presentes na última coluna.

3.6. Treinamento de rede neural artificial para classificação das ORFs

Para treinamento da rede e validação foi utilizado a aplicação EasyFAN [7], versão 5.6.2, que permitiu a leitura do arquivo texto, descrito anteriormente e treinamento de uma rede no modelo FAN. A técnica FAN de Rede Neural Artificial foi escolhida por apresentar as seguintes características:

- Associa características das Redes Neurais (aprendizado automático) e dos modelos difusos (representação da informação);
- Tem como base os neurônios independentes associados a cada classe de representação de um modelo de reconhecimento de padrões supervisionado;
- Graus de pertinência associam os padrões a cada neurônio representante de uma classe no domínio do problema;
- O treinamento é realizado por um algoritmo específico que usa reforço e penalização;
- Dispensa a necessidade de configuração entre problemas diferentes de reconhecimento de padrões;
- O resultado do treinamento pode ser representado graficamente.

As configurações utilizadas no modelo são as definidas como padrão do aplicativo, sendo o valor 6 para o raio difuso e 100 para o tamanho do suporte do conjunto difuso.

As estratégias de treinamento utilizadas foram:

- Embaralhar o conjunto de padrões a cada 1 época;
- Técnica de normalização utilizada: pelo mínimo/máximo;
- Têmpera utilizada: nenhuma;
- Épocas na ultima versão: 10.051.

Ao longo do processo de treinamento, a aplicação do EasyFAN preserva em memória as melhores redes, segundo as médias harmônica, aritmética e máximo do mínimo. Os resultados do treinamento são apresentados na Tabela 3. Nesta tabela as siglas TP, TN, FP e FN significam respectivamente verdadeiro positivo, verdadeiro falso, falso positivo e falso negativo.

Em função desses resultados, adotamos a rede com melhor média aritmética.

Tabela 3 - Resultado do treinamento

Resultados	Média Aritmética	Média Harmônica	Média Máximo/Mínimo
Geral	91,13	91,13	89,32
TP	987	987	988
TN	118	118	117
FP	1028	1028	1024
FN	78	78	82

3.7. Desenvolvimento de aplicação para extração de características e classificação das ORFs

As etapas do algoritmo são:

- Passo 1: Leitura do arquivo fasta - Leitura do arquivo que contém a seqüência a ser anotada, em formato fasta.
- Passo 2: Carrega a rede – Carrega para a memória a rede previamente treinada (rede3ari.enn).
- Passo 3: Busca pelas ORFs – O processo consiste em identificar todas as regiões delimitadas por dois *stop códons* (LongStop) e que sejam maiores ou igual a 1000 na primeira iteração, 500 na segunda iteração, 200 na terceira

iteração e 50 na última iteração. Exclui as regiões identificadas em iterações anteriores. Identifica-se os *start codon* em cada um dos LongStop e produz o conjunto das ORFs candidatas.

- Passo 4: Classificação das ORFs – As ORFs candidatadas são avaliadas pela rede, identificando as possíveis e descartando as desclassificadas. As ORFs que possuem um mesmo *stop codon* são classificadas juntas e a ORF que possuir a melhor avaliação é preservada, as demais são descartadas.
- Passo 5: Exclusão das ORFs inclusas – O conjunto de ORFs classificadas são ordenados em ordem de tamanho, a maior é copiada para o conjunto de resultado, as demais são comparadas com as ORFs contidas no arquivo de resultado eliminado as ORFs que apresentam o início e o final da região anotada interna a outra ORF.
- Passo 6: Avaliação dos overlaps – Cada ORF que possui alguma sobreposição é avaliada e todas que possuem uma sobreposição maior que 3 pares de bases são excluídas. Esse passo marca o final da iteração, voltando a execução para o passo 3 e depois de 4 iterações executa o passo final.
- Passo 7: Geração do arquivo GenBank – O arquivo de ORFs resultante desse processo e a seqüencia de nucleotídeos carregados no passo 1 são utilizados para produzir o arquivo no formato GenBank.

3.8. Escolha de Genomas previamente anotados para o processo de validação

Dos 30 organismos apresentados nas tabelas de resultado do sistema GLIMMER[1] (Center for Bioinformatics and Computational Biology [14]) foram selecionados 24 organismos que tem os genomas disponibilizados pelo NCBI. Os arquivos no formato GenBank foram obtidos no servidor de arquivos do GenBank do NCBI (NCBI GenBank, 2010). Este formato foi utilizado por conter as seqüências de nucleotídeos e as respectivas regiões anotadas. Foi desenvolvida uma aplicação para separar e gerar tabelas e planilhas para comparação dos elementos. Os genomas foram armazenados em arquivos no formato Fasta para serem utilizados pela aplicação GLIMMER e pela aplicação desenvolvida nesse trabalho.

3.9 Identificação das ORFs com a aplicação GLIMMER para fins de comparação de resultados

A versão utilizada do GLIMMER nos nossos estudos foi a 3.0.2. Os passos utilizados na execução deste programa são apresentados na figura 02

```
#!/bin/sh
long-orfs –o $1 –g $2 –A atg,ttg,ctg,att,atc,ata \
–Z tag,tga,taa $3 orfs.out
extract $3 orfs.out > orfs.fasta
build-icm icm.out < orfs.fasta
glimmer3 –o $1 –g $2 –A atg,ttg,ctg,att,atc,ata \
–Z tag,tga,taa $3 icm.out resultado.fas
```

Figura 02 – Script utilizado na execução do GLIMMER

Os parâmetros $1 para "–o" e $2 para "-g" foram obtidos na tabela disponível no site do GLIMMER1. O parâmetro $3 representa o nome do arquivo, no formato fasta, que contém a seqüência completa do genoma. Os valores para start códons (parâmetro –A) e stop códon (parâmetro –Z) foram obtidas na tabela 11 - "The Bacterial, Archaeal and Plant Plastid Code (transl_table=11)" do NCBI (The Genetic Codes).

3.10. Comparação dos resultados

O método de comparação desenvolvido utiliza as métricas abaixo:

- Quantidade de genes indicados / quantidade de genes anotados
- Quantidade de genes com sobreposição completa (coincidência da extremidade 5' e 3')
- Quantidade de genes que possuem a extremidade terminal (*stop codon*) indicada corretamente.

Os resultados preliminares são apresentados na tabela 4 e 5.

[1] Disponível em: http://www.cbcb.umd.edu/software/glimmer/g3.table3.jun01.shtml

Tabela 4 - Resultados preliminares I

Organismo	% genes indicados		% genes corretos	
	GLIMMER	Rede	GLIMMER	Rede
A. fulgidus	63,56	71,84	35,60	61,30
B. subtilis	64,82	80,74	40,76	63,08
C. jejuni	73,64	75,70	48,70	66,37
C. hydrogenoformans	70,19	74,84	44,92	62,10
C. crescentus	49,70	64,49	16,16	38,18
C. tepidum	58,71	66,54	25,55	47,50
C. perfringens	82,07	87,74	51,18	73,48
C. psychrerythraea	68,03	80,35	46,40	69,94
D. ethenogenes	55,12	76,80	35,38	63,89
E. coli	94,61	80,59	34,39	63,79
G. sulfurreducens	68,07	70,91	22,96	43,00
H. influenzae	72,30	85,36	48,87	76,46
H. pylori	74,45	82,14	51,05	72,23
M. capsulatus	75,26	62,81	19,66	31,16
M. tuberculosis	54,41	68,78	19,98	46,03
N. meningitidis	91,96	78,43	30,25	56,76
P. gingivalis	80,37	75,45	35,31	57,96
P. fluorescens	77,55	70,59	19,81	37,11
R. solanacearum	84,87	56,95	13,10	20,35
T. marítima	90,56	70,90	44,76	57,21
T. denticola	76,46	77,98	47,67	65,29
T. pallidum	72,60	69,68	38,26	57,63
S. agalactiae	72,53	82,91	25,91	31,72
S. pneumoniae	64,86	76,94	43,80	66,11
MÉDIA	**72,36**	**74,56**	**35,02**	**55,36**

Tabela 5 - Resultados preliminares II

Organismo	% dos *stop codon* indicados corretamente	
	GLIMMER	rede
Archaeoglobus fulgidus DSM 4304	55,03	67,98
Bacillus subtilis subsp. subtilis str. 168	61,16	78,68
Campylobacter jejuni subsp. doylei 269.97	70,10	72,95
Carboxydothermus hydrogenoformans	63,76	72,48
Caulobacter crescentus CB15	30,09	46,22
Chlorobium tepidum TLS	41,34	55,84
Clostridium perfringens ATCC 13124	77,66	86,11
Colwellia psychrerythraea 34H	67,02	79,01
Dehalococcoides ethenogenes 195	50,61	72,96
Escherichia coli str. K-12 (MG1655)	58,36	75,98
Geobacter sulfurreducens PCA	41,64	55,04
Haemophilus influenzae 86-028NP	69,67	82,99
Helicobacter pylori P12	71,55	80,30
Methylococcus capsulatus str. Bath	35,88	41,51
Mycobacterium tuberculosis H37Ra	35,48	59,45
Neisseria meningitidis MC58	50,47	66,16
Porphyromonas gingivalis	59,40	67,52
Pseudomonas fluorescens Pf-5	34,25	46,13
Ralstonia solanacearum GMI1000	24,18	28,95
Thermotoga maritima MSB8	23,39	65,82
Treponema denticola ATCC 35405	70,68	76,32
Treponema pallidum	59,00	67,03
Streptococcus agalactiae NEM316	52,93	60,54
Streptococcus pneumoniae Hungary	61,12	69,53
MÉDIA	**52,70**	**65,65**

Podemos observar um aumento de 20 pontos percentuais (Tabela 4) na indicação completa dos genes (considerando os *start codons* e *stop codons*). E apesar do acréscimo de 2,2% (Tabela 4) no quantitativo de genes indicados, ganhamos uma redução de 19,86% do GLIMMER para 8,91% da rede em falsos genes (Tabela 5), juntamente com um acréscimo de 12,95% de acertos na extremidade terminal.

4 Conclusões

Os resultados preliminares indicam uma real possibilidade de se utilizar a presente metodologia para a identificação e anotação automática de genes em genomas de *Bactérias* e *Archaeas*. Nos estudos tivemos um aumento de 58% de acertos em relação aos resultados do GLIMMER e uma redução de 13,20% na predileção incorreta dos genes, o que representa uma melhora de qualidade na ordem de 49,77%. Porém, precisamos avaliar melhor as regiões indicadas para termos uma quantidade de falsos positivos menores. Os estudos das características precisam ser ampliados, na expectativa de obter maiores percentuais. Outras estratégias podem ser avaliadas, como a criação de uma rede para cada grupo ao invés de uma rede para todos os genomas como foi realizado nesse trabalho.

5 Referências

[1] Berger, J. A.; Mitra, S. K.; Carli, M., Neri, A. Visualization and analysis of dna sequences using dna walks. **Journal of the Franklin Institute**, v. 341, p. 37–53, 2004. doi:10.1016/j.jfranklin.2003.12.002

[2] **Center for Bioinformatics and Computational Biology**. Glimmer 3. Disponível em <http://www.cbcb.umd.edu/software/glimmer/g3.table1.jun01.shtml>, consultado em 10/08/2010

[3] Raittz, R. T.; Dandslini, G. A.; Pacheco, R. C. S.; Martins, A.; Gauthier, F. A.; Barcia, R. M.; Souza, J. A. Fan: Learning by means of free associative neurons. **Congress On Computational Intelligence**, FUZZ-IEEE'98, 1998.

[4] Delcher, Arthur L.; Bratke, Kirsten A.; POWERS, Edwin C.; SALZBERG, Steven L. Identifying bacterial genes and endosymbiont DNA with Glimmer. **Bioinformatics**, v. 23, n. 6, p. 673-679, jan. 2007. DOI 10.1093/bioinformatics/btm009.

[5] Eddy, S. R. Profile hidden markov models. **Bioinformatics**, n. 14, p. 755–763, 1990.

[6] Ishikawa, J.; Hotta, K. Frameplot: a new implementation of the frame analysis for predicting protein-coding regions in bacterial dna with a high g+c content. **FEMS Microbiol Lett**, v. 174, p 251-253, mai. 1999.

[7] Kuster, C. V.; Ignacio, F. A.; Lenfers, F. P.; Garrett, L. F. V.; Zotto, S. EasyFan. 2006. Trabalho de Conclusão de Curso. (Graduação em Tecnólogo em Informática) - **Universidade Federal do Paraná**. Curitiba.

[8] Lancashire, L. J.; Lemetre C; Ball, G. R. An introduction to artificial neural networks in bioinformatics - application to complex microarray and mass spectrometry datasets in cancer studies. **Briefings in Bioinformatics**, n. 10, p. 315–329, 2009.

[9] Li, W.; Bernaola-Galvan, P.; Haghighi, F.; Grosse, I. Applications of recursive segmentation to the analysis of dna sequences. **Computers and Chemistry**. v. 26, n. 5, p. 491–510, jul. 2002.

[10] Mackiewicz, P.; Zakrzewska-Czerwinska, J.; Zawilak, A.; Dudek, M. R.; Cebrat, S. Where does bacterial replication start? rules for predicting the oric region. **Nucleic Acids Research**, v. 32, n. 13, p. 3781–3791, jul. 2004.

[11] Marin, A.; Xia, X. H. Gc skew in protein-coding genes between the leading and lagging strands in bacterial genomes: New substitution models incorporating strand bias. **Journal of Theoretical Biology**, v. 253, p. 508–513, 2008. doi:10.1016/j.jtbi.2008.04.004

[12] Mcculloch, W. S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. **BullMath Biol**, p. 99–115, 1943.

[13] Mcculloch, W. S.; Pitts, W. A grigoriev analyzing genomes with cumulative skew diagrams. **Nucleic Acids Research**, v. 26, n. 10, p. 2286–2290, 1998.

[14] **NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION**. GenBank Database (ftp). Disponível em <ftp://ftp.ncbi.nih.gov/genomes/Bacteria>, consultado em 10/08/2010.

[15] **NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION**. The DDBJ/EMBL/GenBank Feature Table: Definition – versão 8.3. Abril, 2010 disponível em <http://www.ncbi.nlm.nih.gov/collab/FT/>, consultado em 10/08/2010.

[16] **NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION**. The Genetic Codes. Disponível em <http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi>, consultado em 10/08/2010.

[17] Okamoto, T.; Sugimoto, K.; Sugisaki, H.; Takanami, M. Dna regions essential for the function of a bacteriophage fd promoter. **Nucleic Acids Research**, p. 2213–2222, jul 1977.

[18] Pearson, W. R. Rapid and sensitive sequence comparison with fastp and fasta. **Methods in enzymology**, n. 183, p. 63–98, 1990.

[19] Salzberg, S.; Delcher, A.; Kasif, S.; White, O. Microbial gene identification using interpolated markov models. **Nucleic Acids Research**, v. 26, n.2, p. 544–548, jan. 1998.

[20] Suzek B. E.; Ermolaeva, M. D.; Schreiber, M.; Salzberg, S. L. A probabilistic method for identifying start codons in bacterial genomes. **Bioinformatics**, n. 17, p. 1123–1130, 2001.

[21] Raittz, Roberto Tadeu. Fan 2002: Um modelo neuro-fuzzy para reconhecimento de padroes. 2002. **Tese (Doutorado em Engenharia de Produção)**, **Universidade Federal de Santa Catarina**, Florianópolis.