UNIVERSIDADE FEDERAL DO PARANÁ

RICARDO ASSUNÇÃO VIALLE

SILA - UM SISTEMA PARA ANOTAÇÃO AUTOMÁTICA DE GENOMAS UTILIZANDO TÉCNICAS INDEPENDENTES DE ALINHAMENTO

CURITIBA

2013

RICARDO ASSUNÇÃO VIALLE

SILA - UM SISTEMA PARA ANOTAÇÃO AUTOMÁTICA DE GENOMAS UTILIZANDO TÉCNICAS INDEPENDENTES DE ALINHAMENTO

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, área de concentração Bioinformática.

Orientador: Professor Dr. Roberto Tadeu Raittz

Coorientador: Professor Dr. Fábio de Oliveira Pedrosa

CURITIBA

2013

VIALLE, Ricardo Assunção
SILA - Um sistema para anotação automática de genomas utilizando técnicas independentes de alinhamento / Ricardo Assunção Vialle; Oorientador, Roberto Tadeu Raittz; coorientador, Fábio de Oliveira Pedrosa. - Curitiba, PR, 2013.
95 f.

Dissertação (Mestrado) - Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica. Programa de Pós-Graduação em Bioinformática.

Inclui referências

1. Anotação automática de genoma. 2. Alignment-free. 3. Bioinformática. 4. Ciência da computação. I. Raittz, Roberto Tadeu. II. Pedrosa, Fábio de Oliveira. III. Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica. IV. Título.

TERMO DE APROVAÇÃO

RICARDO ASSUNÇÃO VIALLE

Anotação rápida de genomas independente de alinhamento de sequências

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre em Bioinformática, pelo Programa de Pós-graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná, pela seguinte banca examinadora:

Orientador: Prof Dr Roberto Tadeu Raittz

Coorientador: Prof Dr Fábio de Oliveira Pedrosa

Prof Dr José Miguel Ortega
Universidade Federal de Minas Gerais - UFMG

Prof Dr Leonardo Magalhães Cruz
Universidade Federal do Paraná - UFPR

Curitiba, 03 de maio de 2013

Aos meus pais e irmão.

AGRADECIMENTOS

Agradeço primeiramente aos meus pais Solange Assunção Vialle e Ubiratan Vialle por todo amor, carinho e dedicação na minha criação e educação. Ao meu irmão Rodrigo Assunção Vialle por todo apoio e incentivo.

Ao professor Roberto Tadeu Raittz, meu orientador e amigo, pela oportunidade e pelo apoio em todos os momentos. Agradeço por tudo que aprendi nesses ótimos anos. Ao professor Fábio de Oliveira Pedrosa, meu coorientador, pelas dicas e discussões que ajudaram neste trabalho.

Ao professor Emanuel Maltempi de Souza por toda a ajuda na escrita dos trabalhos. Ao professor Dieval Guizelini pelas dicas, sugestões e ensinamentos. Ao Vinícius Weiss as importantes dicas que ajudaram a moldar o trabalho. À Bruna Miranda pelas contribuições ao desenvolvimento do trabalho.

Aos colegas de graduação Fábio Jugler, Fábio Santos e Jackson Prostt pela amizade e pelo trabalho desenvolvido em conjunto.

Aos professores Adriano Barbosa, Jeroniza Marchaukoski, Leonardo Cruz, Maria Berenice e Geraldo Picheth por todo o conhecimento passado.

A todos meus colegas e amigos do curso, especialmente ao Vitor Cedran Piro, meu irmão de orientação, pelas boas conversas e pela grande amizade. À Kátia de Paiva Lopes e ao Fabiano Gomes pela amizade e apoio. À Juliana e Vanely por terem me recebido no laboratório nos tempos de iniciação científica. Aos demais colegas e amigos da minha turma de mestrado Fausto, Rodnei, Eslei, Helba e Guilherme. Aos colegas e amigos das turmas anteriores, Sérgio, Lucas, Antonio, Leviston, Kelly, Rafael, Barbara, Jeovane, Nicolas e Rodrigo.

À Suzana por toda atenção e disponibilidade em ajudar e ao Programa de Pós-Graduação em Bioinformática.

Aos órgãos financiadores da bolsa de estudo e fomento para participação em eventos, sem os quais eu não poderia ter me dedicado integralmente a esse Mestrado: CNPq, CAPES, INCT-FBN, PRPPG, UFPR e SEPT.

E a todos que, direta ou indiretamente, contribuíram para eu concretizar este trabalho, sou, sinceramente,

Muito grato!

Mama told me when I was young
Come sit beside me, my only son
And listen closely to what I say
And if you do this it will help you some sunny day

Take your time, don't live too fast
Troubles will come and they will pass
Go find a woman and you'll find love
And don't forget, son there is someone up above

And be a simple kind of man
Be something you love and understand
Baby, be a simple kind of man
Won't you do this for me, son?
If you can?

Forget your lust for the rich man's gold
All that you need is in your soul
And you can do this if you try
All that I want for you my son?
Is to be satisfied

(...)

Boy, don't you worry, you'll find yourself
Follow your heart and nothing else
And you can do this if you try
All I want for you my son
Is to be satisfied

(...)

Baby, be a simple man
Be something you love and understand
Baby, be a simple man

Simple Man
Lynyrd Skynyrd

RESUMO

Na Bioinformática, a comparação de sequências é uma tarefa essencial para diversas atividades e é onde geralmente se tem o maior custo computacional. A anotação de genomas é uma atividade que envolve tarefas distintas, dentre elas a comparação de sequências contra grandes bancos de dados. Este trabalho apresenta um sistema para anotação automática de genomas de procariotos chamado SILA e uma nova abordagem independente de alinhamento para busca e comparação de sequências de proteínas chamado Rapid Alignment Free Tool for Sequences Similarity Search (RAFTS3). O SILA combina duas ferramentas para realizar a predições de genes. Os genes preditos são comparados com três bancos de sequências de proteínas (NR, Pfam e COG) utilizando o RAFTS3 para buscar informações que possam inferir, através da semelhança de sequências, as funções dos genes. Os testes realizados mostraram que o RAFTS3 obteve resultados comparáveis ao BLASTp com desempenho até 500 vezes superior em buscas por sequências com alta similaridade contra grandes bancos. O SILA apresenta resultados de anotação comparáveis ao estado da arte com altas taxas de acerto na predição e identificação dos genes. O serviço de anotação está disponível em ambiente Web.

Palavras-chave: Anotação automática de genomas, Comparação de sequências, Alignment-free.

ABSTRACT

In Bioinformatics, the comparison of sequences is an essential task for various activities and is where usually has the highest computational cost. The annotation of genomes is an activity that involves different tasks, among them, sequence comparison against large databases. This work presents a system for automatic annotation of genomes of prokaryotes called SILA and a new alignment-free approach to search and comparison of protein sequences called Rapid Alignment Free Tool for Sequences Similarity Search (RAFTS3). The SILA combines two tools to perform gene predictions. The predicted genes are compared using the RAFTS3 against three protein sequence databases (NR, Pfam and COG) to seek information to infer by the similarity of sequences, the functions of genes. The tests show that the RAFTS3 obtained results comparable to BLASTp with performance up to 500 times higher in searches for sequences with high similarity against large databases. The SILA presents results of annotation comparable to state of the art with high accuracy in predicting and identifying the genes. The annotation service is available in web.

Key-words: Automated genome annotation, Sequence comparison, Alignment-free.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AG	Algoritmos genéticos
ANN	Artificial Neural Network
BASys	Bacterial Annotation System
BBH	Bidirectional Best Hits
BHR	Bi-directional Hit Rate
BioGRID	Biological General Repository for Interaction Datasets
BLAST	Basic Local Alignment Search Tool
BLOSUM	BLOck SUbstitution Matrix
CDS	Coding DNA Sequence
COG	Clusters of Orthologous Groups
CRITICA	Coding Region Identification Tool Invoking Comparative Analysis
DDBJ	DNA Databank of Japan
DNA	Ácido desoxirribonucleico
EBI	European Bioinformatics Institute
EC	Enzyme Commission
ELPH	Estimated Locations of Pattern Hits
EMBL	European Molecular Biology Laboratory Nucleotide Sequence Database
EST	Expressed Sequence Tag
ExPASy	Expert Protein Analysis System
FASTA	Formato utilizado para armazenar sequências de bases e de aminoácidos em arquivo texto
GenBank	Banco de dados público do National Center for Biological Information
GI	Número identificador de sequências do NCBI
Glimmer	Gene Locator and Interpolated Markov ModelER
GO	Gene Ontology
GOPArc	Gene Ontology and Pathway Architecture
HGF	Hybrid Gene Finder
HMM	Hidden Markov Models
HSP	High Scoring Pairs
IMM	Interpolated Markov Models
INSDC	International Nucleotide Sequence Database Collaboration
IUBMB	Nomenclature Committee of the International Union of Biochemistry and Molecular Biology
KASS	KEEG Automatic Annotation Server
KEGG	Kyoto Encyclopedia of Genes and Genomes
K-mer	(Semente) menor fração de uma sequência com tamanho pré-definido “k”
KO	KEEG Orthology
MATLAB	Matrix Laboratory
MLP	Multi-Layer Perceptron
mRNA	RNA mensageiro

MSP	Maximal Segment Pair
NCBI	National Center for Biotechnology Information
NGS	Next generation sequencing
NIH	US National Institutes of Health
NLM	National Library of Medicine
NR	Banco de dados de proteínas não redundante do NCBI
ORF	Open Reading Frame
PAM	Point Accepted Mutation
pb	Pares de base
PDB	Protein Data Bank
PIR	Protein Information Resource
PRF	Protein Research Foundation
Prodigal	Prokaryotic Dynamic Programming Genefinding Algorithm
RAST	Rapid Annotation using Subsystem Techology
RBS	Ribosome Binding Site
RefSeq	Reference Sequence, banco de dados mantido pelo NCBI
RAFTS3	Rapid Alignment Free Tool for Sequences Similarity Search
RNA	Ácido ribonucleico
rRNA	RNA ribossômico
Sabia	System for Automated Bacterial Integrated Annotation
SIB	Swiss Institute of Bioinformatics
Swiss-Prot	Banco de dados de proteínas que faz parte da UniProt Knowledgebase
TCDB	Transport Classification Database
tRNA	RNA de transferência
WGS	Whole Genome Shotgun

SUMÁRIO

1 INTRODUÇÃO

1.1 CONSIDERAÇÕES INICIAIS

Para entender como um genoma de um organismo, com milhões ou bilhões de bases, contém toda a informação necessária para que a célula realize todos os processos necessários para a sua sobrevivência, informação essa que é passada de geração em geração, grandes quantidades de dados de sequências devem ser coletados e armazenados de forma que possam ser acessados e analisados facilmente. A história dos bancos de dados de sequências, assim como da Bioinformática, tem início com Margaret Dayhoff no início dos anos 60 com sua coleção de sequências de proteínas conhecida como Atlas de Sequências de Proteínas e Estruturas (BAXEVANIS; OUELLETTE, 2001).

Com o desenvolvimento das tecnologias de sequênciamento, a quantidade de dados biológicos produzidos alcançou níveis que tornaram as análises manuais absurdas para o estudo de dados de genomas e transcriptoma, permitindo o surgimento e o crescimento da área da ciência conhecida como Bioinformática (PROSDOCIMI, 2007).

A Bioinformática é a ciência que utiliza abordagens computacionais para responder questões biológicas. Para responder essas questões, é necessária a utilização de grandes e complexos conjuntos de dados de maneira rigorosa afim de obter conclusões biológicas (BAXEVANIS; OUELLETTE, 2001). Dessa forma, a Bioinformática utiliza programas de computadores para fazer inferências a partir de dados biológicos, podendo realizar conexões entre eles, e assim derivar predições úteis e interessantes (LESK, 2002).

Para Gibas, a Bioinformática é definida como um subconjunto da Biologia Computacional, sendo a ciência do uso da informação para entender a biologia. Ao fornecer, algoritmos, bancos de dados, interfaces e ferramentas estatísticas, a Bioinformática torna possível atribuir significados à dados brutos (GIBAS; JAMBECK, 2001).

Duas importantes atividades de larga escala que utilizam a Bioinformática são a genômica e a proteômica. A genômica se refere ao sequênciamento e análise de genomas, que pode ser vista como o conjunto completo de sequências de DNA que codificam o material hereditário passado de geração para geração. Essas sequências de

DNA incluem todos os genes e transcritos. Já a proteômica trata da análise do conjunto completo de proteínas, ou proteoma, que os genes expressam. Além da genômica e a proteômica, existem muitas outras áreas da biologia onde a Bioinformática é aplicada, como por exemplo a metabolômica e a transcriptômica (FOX, 2005).

1.2 JUSTIFICATIVA DO TRABALHO

No campo da Biologia Molecular, as pesquisas realizadas sobre a estrutura química e molecular das estruturas genéticas, permitiram o surgimento de projetos como o "Projeto Genoma Humano" (LANDER et al., 2001). Hoje, com o desenvolvimento de novas técnicas de sequênciamento de alto desempenho e baixo custo, bancos de dados de genomas como o GenBank ® têm crescido exponencialmente (BENSON et al., 2008, 2009, 2011). Esse crescimento gera diversos desafios computacionais para a Bioinformática, em especial para a anotação de genomas.

A anotação de genomas têm como objetivo obter informações estruturais e/ou funcionais sobre uma ou várias sequências relacionadas a um determinado genoma, sendo um processo complexo que envolve diferentes aspectos. As tarefas envolvidas na anotação incluem, a identificação de regiões codificadoras (CDS), busca por sequências semelhantes, busca por sítios funcionais dentro da sequência e inferência da estrutura espacial. Existem diversas ferramentas para auxiliar o especialista humano a realizar essas tarefas, poupando tempo e esforço (STEIN, 2001). No entanto, o processo automático de anotação somente é possível até certo ponto. Sendo necessária a interpretação manual de um especialista em certos casos (BRYSON et al., 2006). Por outro lado, o crescimento no volume de dados genômicos deve permitir a melhora na qualidade das anotações possibilitando a reanotação de genomas e por consequência a identificação e correção de erros nos bancos de dados (LESK, 2002).

Outro aspecto que deve ser levado em consideração durante a anotação são as buscas por similaridade. Um típico fluxo de trabalho (do inglês *workflow*) de anotação utiliza para busca de sequências similares o BLAST, acrônimo para Basic Local Alignment Search Tool (ALTSCHUL et al., 1990), contra diversos bancos de sequências de proteínas diferentes. Entretanto, realizar comparações contra grandes bancos, como o banco de proteínas não redundante (NR) do National Center for Biotechnology Information

(NCBI), é altamente custoso computacionalmente ocasionando longos tempos de execução.

Os algoritmos baseados em alinhamento de sequência são eficientes em detectar similaridade entre sequências de proteínas. Existem duas abordagens de técnicas de alinhamento, são elas, alinhamentos globais, que alinham sequências de ponta a ponta, e alinhamentos locais, que buscam alinhar regiões das sequências. Originalmente as técnicas de alinhamento utilizavam técnicas de programação dinâmica para produzir um alinhamento otimizado entre as sequências. Apesar de diversas implementações eficientes terem sido desenvolvidas, a carga computacional para comparar grandes quantidades de sequências torna os algoritmos dinâmicos inviáveis (MAHMOOD et al., 2012).

Como alternativa, abordagens heurísticas foram criadas para compensar o custo computacional. Em geral, esses métodos começam por gerar uma lista de subsequências de tamanho "k" determinado (k-mers). As sequências do banco de dados são pesquisadas para buscar somente as sequências que possuem k-mers comuns. Então, os k-mers encontrados são estendidos utilizando esquemas de pontuações de alinhamento, para maximizar o tamanho do trecho alinhado. Os métodos heurísticos são eficientes para realizar buscas em grandes bancos de dados (MAHMOOD et al., 2012). Entretanto, no contexto de anotação de genomas, em que são necessárias realizar alguns milhares de buscas, pode custar longos tempo de execução. Alem de não evitar que seja necessária a avaliação manual dos resultados ao final do processo.

Dessa forma, um *workflow* que economize tempo na busca de sequências similares proporciona ganhos significativos nas atividades de anotação de genomas e mineração de dados genômicos. Algumas técnicas que podem oferecer uma alternativa eficiente ao alinhamento tradicional, são as chamadas *alignment-free*. Os métodos *alignment-free* são baseados na hipótese que duas sequências similares compartilham certa porção de k-mers. Estes métodos geralmente trabalham calculando o número de k-mers compartilhados entre um par de sequências, seguido pelo cálculo de métodos estatísticos. Diversas técnicas de *alignment-free* foram propostas na literatura, com diversas utilizações bem sucedidas (COMIN; VERZOTTO, 2012; FENG; ZHAO; ZHANG, 2011; HUANG et al., 2011; JING; WILSON; BURDEN, 2011; KANTOROVITZ; ROBINSON; SINHA, 2007; KURTZ et al., 2008; MAHMOOD et al., 2012; SOARES; GOIOS; AMORIM, 2012; YU et al., 2011). Essa abordagem é interessante para fins de anotação de genes,

tendo em vista que somente é necessária a identificação de sequências que possuem alto grau de similaridade e não há necessidade do alinhamento em si.

1.3 OBJETIVOS

1.3.1 Objetivo Geral

Propor uma ferramenta para rápida comparação de sequências de proteínas contra um banco de dados, integrar a ferramentas de predição gênica e desenvolver um sistema de anotação automática de genomas.

1.3.2 Objetivos Específicos

- Testar e aprimorar a ferramenta Hybrid Gene Finder (HGF) para marcação de ORFs (*Open Reading Frames*) em genomas de procariotos.
- Integrar o HGF com a ferramenta Prodigal para obter resultados de anotação mais precisos.
- Desenvolver uma ferramenta para comparação de sequências de proteínas para busca em grandes bancos de dados.
- Integrar o anotador proposto com os principais bancos de proteínas para identificação de funções e produtos.
- Comparar o método proposto com outros métodos de anotação automática.
- Disponibilizar a ferramenta para anotação automática em ambiente web para uso público.
- Disponibilizar a biblioteca de anotação com propósito de mineração de dados.

2 REVISÃO DE LITERATURA

2.1 BIOLOGIA MOLECULAR

Nesta seção, apresentaremos alguns conceitos biológicos relevantes para a compreensão deste trabalho. Abordaremos conceitos de DNA e RNA bem como o processo bioquímico conhecido como o Dogma Central da Biologia Molecular.

2.1.1 DNA e RNA

O material genético de todo ser vivo está codificado em longas cadeias de uma molécula chamada DNA (Ácido desoxirribonucleico). A descoberta de que o DNA tem este papel central foi feita em meados da década de 1940. Esta descoberta, seguida pela elucidação da sua estrutura tridimensional em 1953, estabeleceu as bases para muitos dos progressos em bioquímica e em muitos outros campos (BERG; TYMOCZKO; STRYER, 2010).

O DNA tem a forma de uma dupla hélice e é constituído por cadeias de ácidos nucleicos chamados nucleotídeos. Os nucleotídeos possuem três componentes: uma base nitrogenada, uma pentose e um fosfato. No DNA são encontrados quatro tipos de nucleotídeos: Adenina (A), Citosina (C), Guanina (G) e Timina (T). Usualmente os nucleotídeos são chamados de bases, um segmento de uma molécula de DNA que contenha a informação requerida para a síntese de um produto biológico funcional, proteína ou RNA (Ácido Ribonucleico), é referido como um gene (NELSON; COX, 2006).

O RNA difere do DNA por uma discreta modificação no componente açúcar e pela presença da base Uracila (U) no lugar da Timina (T). A FIGURA 1 mostra os cinco nucleotídeos. Enquanto o principal papel do DNA é armazenar o código genético, o RNA apresenta diferentes tipos e funções. Os principais tipos de RNA são: o RNA mensageiro (mRNA), o RNA transportador (tRNA) e o RNA ribossômico (rRNA). O mRNA é transcrito a partir do DNA, e transporta informação até uma organela presente no citoplasma chamada ribossomo, responsável pela síntese de proteínas. O tRNA é responsável pelo transporte de moléculas de aminoácidos até o ribossomo. O rRNA é o principal

componente dos ribossomos, responsável por catalisar algumas etapas no processo de síntese de proteínas (NELSON; COX, 2006).

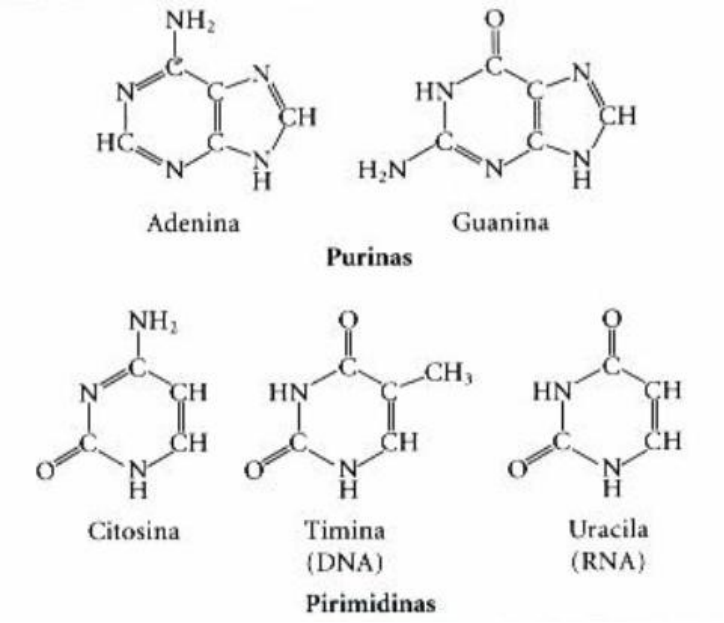


FIGURA 1 - AS BASES DOS ÁCIDOS NUCLÉICOS
FONTE: (NELSON; COX, 2006)

2.1.2 Proteínas

As proteínas são as macromoléculas mais versáteis nos sistemas vivos e servem para funções cruciais em essencialmente todos os processos biológicos. Elas desempenham diversos papéis nos organismos vivos, dentre eles: funcionam como catalisadores, transportam e armazenam outras moléculas, fornecem apoio mecânico e proteção imunitária, geram movimento, transmitem impulsos nervosos, e controlam o crescimento e a diferenciação (BERG; TYMOCZKO; STRYER, 2010).

As proteínas são polímeros lineares feitos de unidades monoméricas denominadas aminoácidos, que se unem ponta a ponta. Os aminoácidos são compostos de um grupo amina ($-NH_3$), um grupo carboxila (-COOH) e uma cadeia lateral específica para cada aminoácido. Cerca de 500 aminoácidos são conhecidos (WAGNER; MUSSO, 1983), no entanto vinte tipos de cadeias laterais são comumente encontrados em proteínas (FIGURA 2), variando em tamanho, forma, carga, capacidade de formação de pontes de hidrogênio, caráter hidrofóbico e reatividade química (BERG; TYMOCZKO; STRYER, 2010).

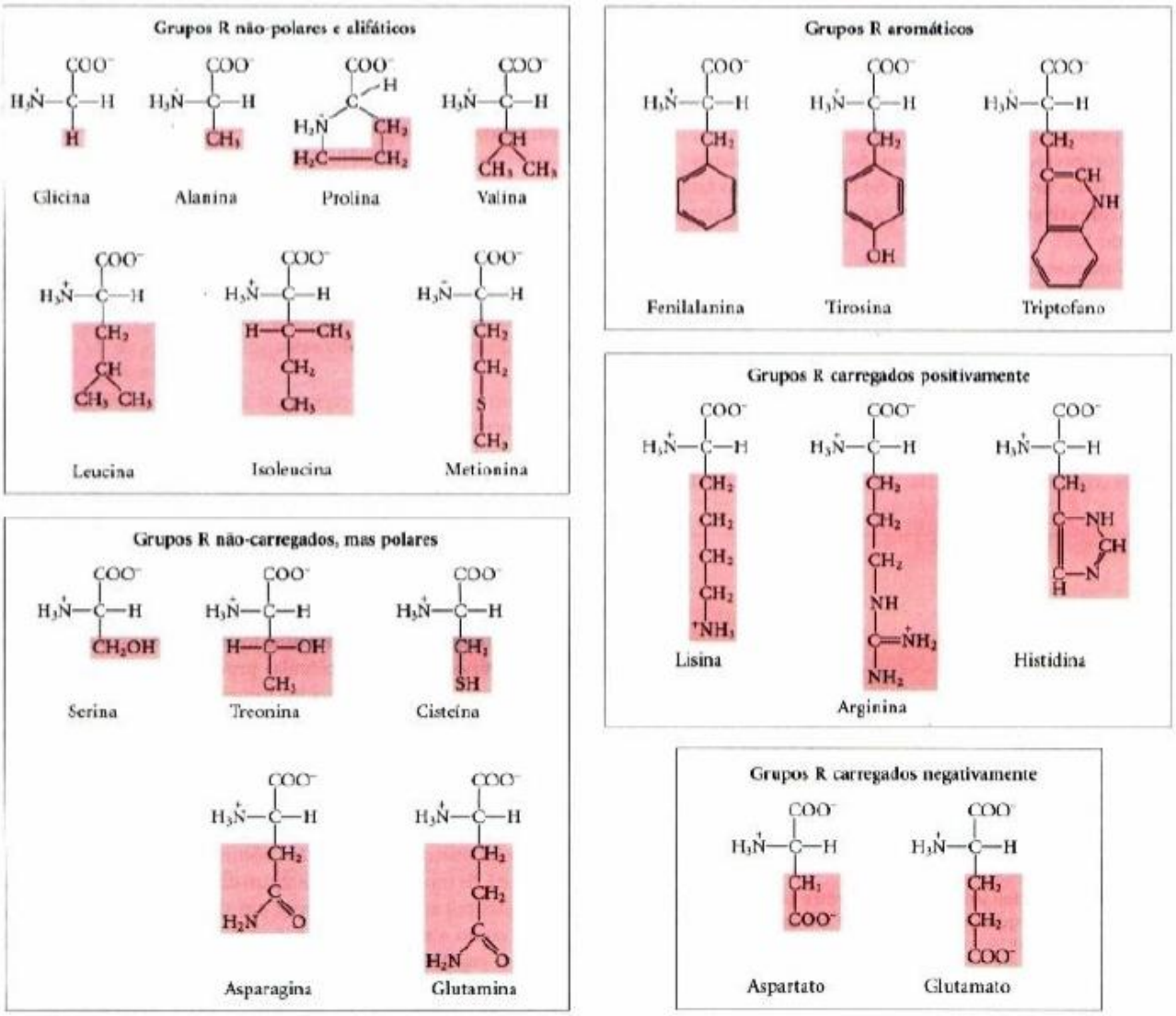


FIGURA 2 - OS 20 AMINOÁCIDOS PRIMÁRIOS DAS PROTEÍNAS
FONTE: (NELSON; COX, 2006)

As proteínas apresentam quatro níveis estruturais: a estrutura primária, é a sequência de aminoácidos; a estrutura secundária corresponde ao arranjo espacial de aminoácidos próximos entre si, por exemplo, as alfa-hélice e as folhas-beta; a estrutura terciária descreve a conformação da proteína inteira, onde as proteínas hidrossolúveis se enovelam em estruturas compactas com o interior apolar; e a estrutura quaternária são as conformações de duas ou mais cadeias polipeptídicas (BERG; TYMOCZKO; STRYER, 2010).

Tanto as sequências de proteínas quanto de DNA podem ser comparadas em termos de ancestrais comuns. Quando dois segmentos de DNA possuem ancestrais comuns são chamados homólogos, ocorrendo devido a um evento de especiação são

chamados ortologos, caso ocorra devido a um evento de duplicação são chamados parálogos. Usualmente, a homologia pode ser inferida a partir da similaridade das sequências, similaridade essa que pode ser obtida através de um alinhamento entre as sequências. No entanto há casos em que sequências homólogas apresentam baixa similaridade, porém suas estruturas são conservadas (KOONIN; GALPERIN, 2003).

2.1.3 Dogma Central da Biologia Molecular

O Dogma Central da Biologia Molecular é o conceito que descreve a maneira com que a informação é transferida através de um sistema biológico. Segundo (CRICK, 1970):

> *The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid.* (CRICK, 1970).

O conceito classifica em transferências gerais aquelas que representam o fluxo normal da informação biológica: o DNA pode ser copiado para DNA (Replicação), a informação do DNA pode ser copiada para um mRNA (Transcrição) e as proteínas podem ser sintetizadas a partir da informação no mRNA (Tradução). A **FIGURA 3** mostra as relações entre os três processos. Vale ressaltar que, apesar da transcrição ser apresentada de maneira unidirecional, algumas raras vezes as cadeias de RNA atuam como moldes para a síntese de cadeias de sequências complementares de DNA (WATSON et al., 2006).

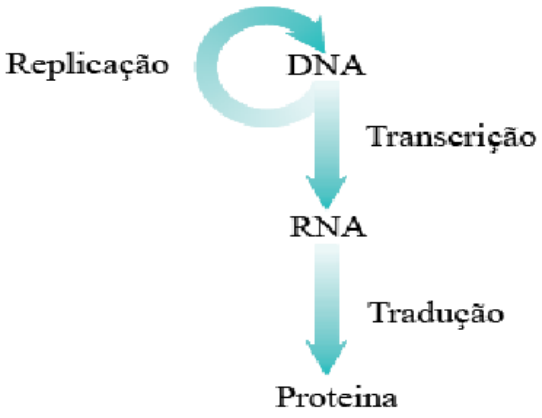


FIGURA 3 - DOGMA CENTRAL DA BIOLOGIA MOLECULAR
FONTE: (GUIZELINI, 2010)

Dada a existência de 20 aminoácidos e apenas quatro bases, o código genético é determinado por tríplices de nucleotídeos (códons) que codificam um determinado aminoácido. A finalização do código, em 1966, revelou que 61 dos 64 grupos permutáveis possíveis correspondiam a aminoácidos. Para cada códon presente no mRNA os tRNA associam um aminoácido gerando a sequência da proteína, sendo que a maioria dos aminoácidos é codificada por mais de um códon. Um códon presente no mRNA indica o inicio do processo de tradução no ribossomo. O final da tradução é indicado por um códon de parada (*Stop codon*) que separa o mRNA do ribossomo liberando a proteína (WATSON et al., 2006).

Assim, cada um dos códons, com exceção dos códons de parada, codificam apenas um aminoácido. No entanto, como alguns aminoácidos podem ser codificados por mais de um códon, o código genético é chamado degenerado. A TABELA 1 mostra os códons e os aminoácidos codificados.

TABELA 1 - CODIFICAÇÃO DE AMINOÁCIDOS

	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
	UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop	A
	UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

FONTE: (NASCIMENTO, 2005)

2.1.4 Expressão do genoma

A definição bioquímica para gene diz que: um gene é todo o DNA que codifica a sequência primária de algum produto gênico final, que pode ser um polipeptídio ou um RNA com função estrutural ou catalítica. Para determinar como um gene é expresso, é

preciso analisar como fluxo da informação determinada pelo dogma central funciona (NELSON; COX, 2006).

No processo de transcrição, uma máquina molecular composta por várias subunidades, conhecida como RNA-polimerase, cria uma “bolha” móvel na dupla hélice que desenrola trechos do DNA. A RNA-polimerase utiliza uma das fitas separadas de DNA como molde para a síntese progressiva de uma cópia complementar de RNA, através do pareamento de bases. O mRNA é produzido de modo semelhante em todas às células, no entanto nos eucariotos a maquinaria é mais complexa do que nos procariotos (WATSON et al., 2006).

Nos procariotos, o mRNA recém sintetizado já está pronto para a próxima etapa do fluxo de informação, em que servirá como molde para a síntese de proteínas. Nos eucariotos, o RNA produzido precisa sofrer uma série de eventos de maturação antes de estar pronto pra atuar como mRNA. Entre esses eventos, há a adição da estrutura chamada “quepe” (do inglês *cap*) à extremidade 5’ e da cauda poli-A à extremidade 3’. O evento mais crítico ocorre quando, nos eucariotos, os genes são interrompidos por um ou vários segmentos que não codificam proteínas, conhecidos como íntrons. Esses íntrons devem ser removidos para que os segmentos codificantes de proteínas, chamados éxons, possam ser unidos entre si, formando uma sequência contínua. Esse evento é conhecido como *splicing* (WATSON et al., 2006).

O DNA também contêm outros segmentos ou sequências que possuem função puramente reguladora. Sequências reguladoras fornecem sinais que podem denotar o início ou o fim dos genes, influenciar a transcrição ou funcionar como pontos de iniciação para replicação ou recombinação. Assim, alguns genes podem ser expressos de diferentes maneiras para gerar múltiplos produtos gênicos a partir de um segmento de DNA (NELSON; COX, 2006).

No processo de tradução, existem quatro principais participantes: a sequência codificante do mRNA; as moléculas adaptadoras (tRNA); as enzimas que ligam os aminoácidos nos tRNA; e a fábrica propriamente dita da síntese proteica, o ribossomo. As regiões que codificam uma proteína de um mRNA são compostas por uma sucessão contínua de códons não sobrepostos, chamada fase aberta de leitura ou ORF (*open-reading frame*). Cada ORF define uma única proteína e tem início e fim localizados internamente ao mRNA (WATSON et al., 2006).

A tradução inicia na extremidade 5’ da ORF e prossegue códon a códon, até a extremidade 3’. Nas bactérias, geralmente o códon de iniciação é o 5’-AUG-3’, mas

também são usados 5’-GUG-3’ e 5’-UUG-3’. As células eucarióticas sempre usam o 5’-AUG-3’ como códon de iniciação. Esse códon tem duas funções importantes: definir o primeiro aminoácido a ser incorporado à cadeia polipeptídica, e definir a fase de leitura para todos os códons imediatamente adjacentes entre si. Os códons de parada (5’-UAG-3’, 5’-UGA-3’ e 5’-UAA-3’) definem o final da ORF, sinalizando o término da síntese do polipeptídeo (WATSON et al., 2006). A FIGURA 4 exibe as possibilidades de tradução para cada frame de leitura possível.

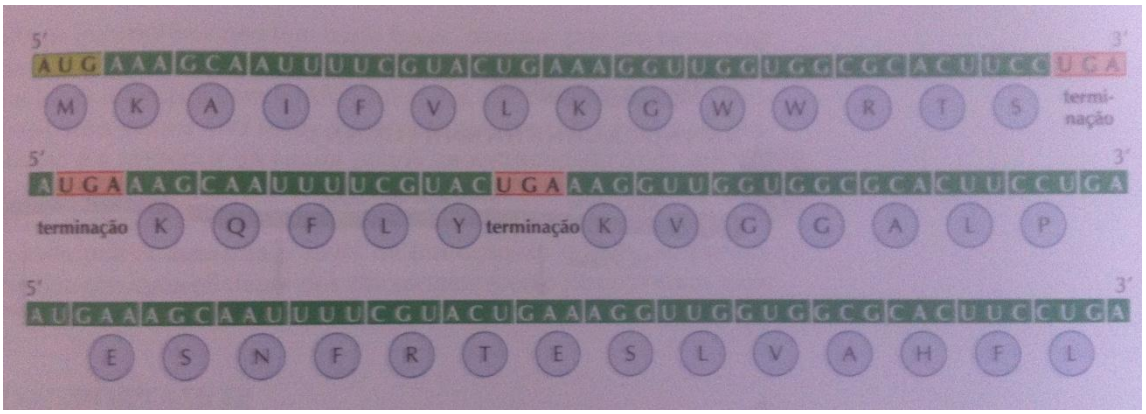


FIGURA 4 - TRÊS POSSÍVEIS FASES DE LEITURA
FONTE: (WATSON et al., 2006)

Os mRNA possuem pelo menos uma ORF, porém a quantidade pode variar entre eucariotos e procariotos. Nos eucariotos, na maior parte das vezes, o mRNA possui apenas uma ORF. Já nos procariotos, é frequente encontrar duas ou mais ORFs. Os mRNA que possuem mais de uma ORF são chamados mRNA policistrônicos enquanto os que contém uma só ORF são chamados mRNA monocistrônicos. Os mRNA policistrônicos frequentemente codificam proteínas que desempenham funções relacionadas (WATSON et al., 2006).

Para que a tradução ocorra, o ribossomo deve ser trazido para o mRNA para facilitar essa ligação. Várias fases de leitura de procariotos possuem uma sequência a montante do códon de iniciação, chamada sítio de ligação ao ribossomo (RBS, *Ribosome Binding Site*), também chamado de sequência de Shine-Dalgarno. O RBS geralmente está localizado de três a nove pares de base a extremidade 5’ do códon de iniciação, e é complementar à extremidade 3’ de um dos componentes do rRNA 16S (**FIGURA 5**). Dessa forma, nos procariotos, os sítios de ligação com o ribossomo são frequentemente sequências semelhantes à sequência 5’-AGGAGG-3’. O grau de complementaridade

entre o sítio de ligação ao ribossomo e o códon de início exerce forte influência sobre a atividade de tradução (WATSON et al., 2006).

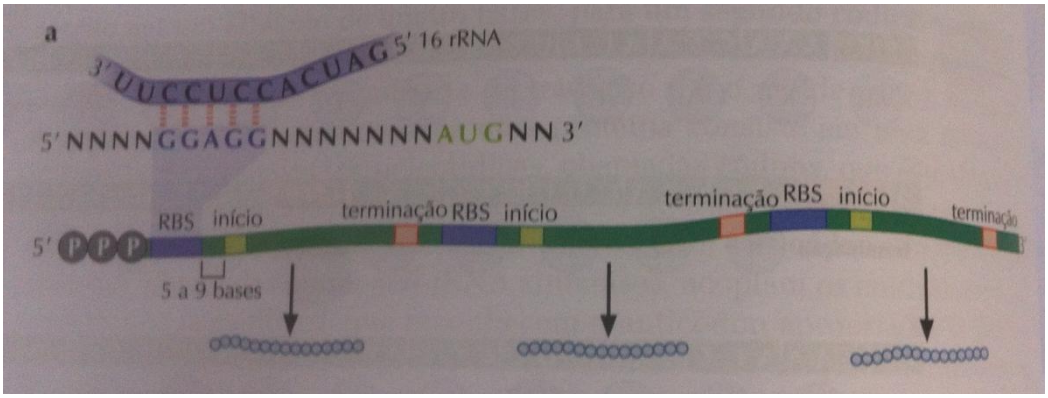


FIGURA 5 - ESTRUTURA DO mRNA
FONTE: (WATSON et al., 2006)

2.1.5 Sequênciamento

O processo bioquímico para descobrir as sequências das bases nitrogenadas presentes no DNA e no RNA é chamado de sequênciamento. Existem diversos métodos com diferenças no tamanho das sequências geradas, na qualidade das sequências e na cobertura.

O primeiro organismo não viral a ser sequênciado foi o da bactéria *Haemophilus influenzae*, em 1995 (FLEISCHMANN et al., 1995). Poucos anos depois, outras bactérias foram sequências como a *Escherichia coli* (BLATTNER et al., 1997) e a *Mycobacterium tuberculosis* (COLE et al., 1998). Junto com o sequênciamento do genoma humano (LANDER et al., 2001) houve um crescimento na quantidade de genomas sequênciados, em 2010 o NCBI comemorava a marca de mil genomas de procariotos completamente sequênciados.

O primeiro método de sequênciamento automático foi desenvolvido utilizando a metodologia de Sanger (SANGER; NICKLEN, 1977). No entanto o método era muito caro o que inviabilizava projetos maiores. A partir de 2004 novos sequênciadores surgiram no mercado, reduzindo o custo e o tempo. Essa nova tecnologia ficou conhecida como Sequênciamento de nova geração (*Next generation sequencing* - NGS) e revolucionou a pesquisa em genética e genômica (QUAIL et al., 2012). A FIGURA 6 mostra o barateamento do custo de sequênciamento ao longo dos anos.

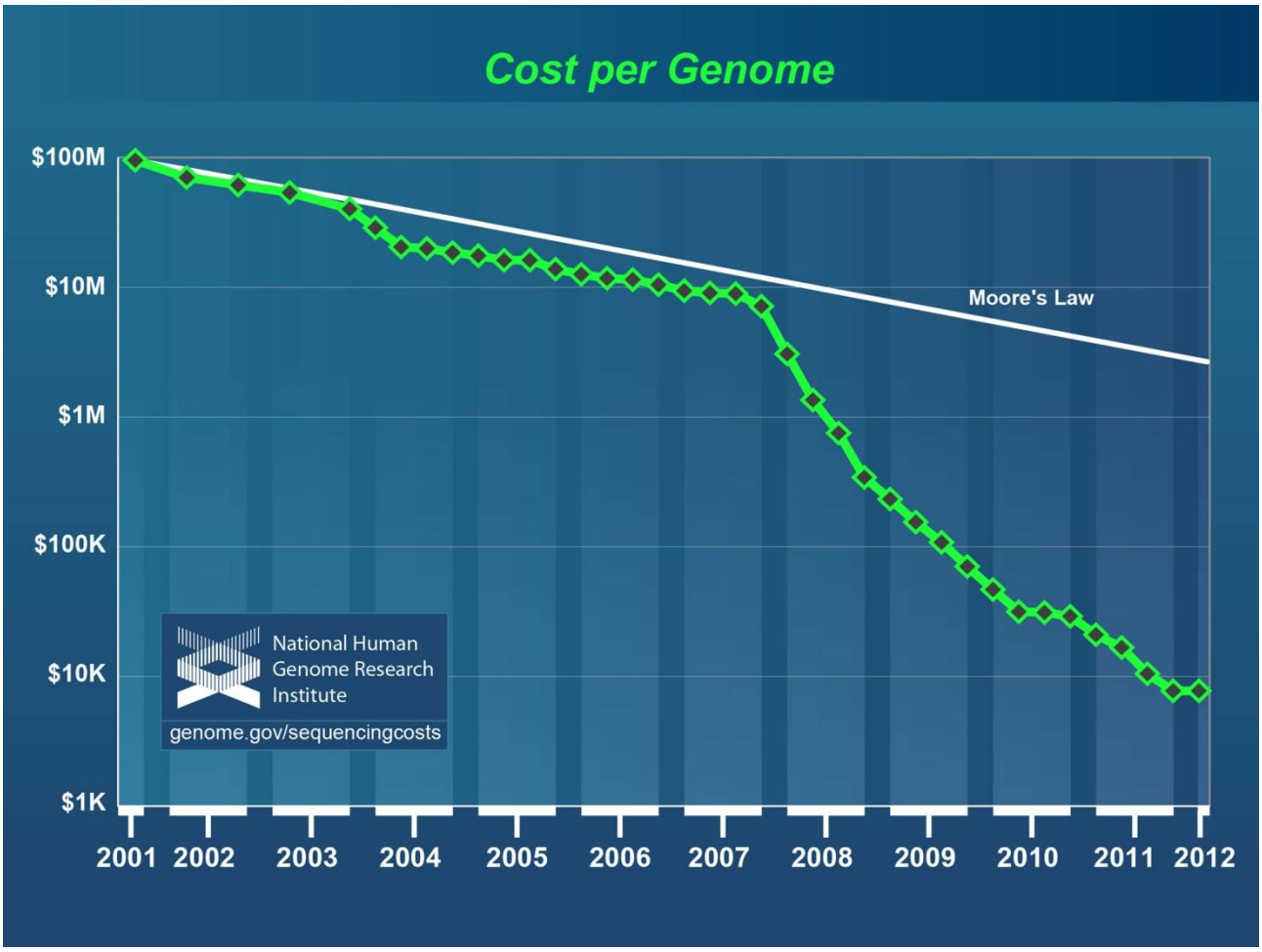


FIGURA 6 - CUSTO DO SEQUÊNCIAMENTO DE GENOMAS

Observa-se que a queda acentuada no custo do sequênciamento dos genomas (dólar) a partir de 2007 com a consolidação dos sequênciadores de nova geração.

FONTE: (genome.gov, 2013)

2.1.6 Montagem

A montagem de genomas refere-se ao processo de unir uma grande quantidade de sequências curtas de DNA, também chamadas de leituras (do inglês, *reads*), obtidas no processo de sequênciamento de forma a recriar a sequência do cromossomo original que gerou tais sequências.

Os algoritmos de montagem geralmente trabalham tentando unir os *reads* por sobreposição, estendendo as sequências o máximo possível gerando sequências consenso chamadas *contigs*. No entanto a montagem de genomas é um problema computacional muito complexo, especialmente devido a genomas que contém grandes

quantidades de sequências repetidas. Essas repetições podem ser de centenas de nucleotídeos e podem ocorrer em diferentes locais no genoma.

A união dos *contigs* através de um mapa físico que gere uma ordenação cria os *scaffolds*. O resultado gerado obtido pelos *contigs* e *scaffolds* é chamado de esboço (do inglês *draft*) do genoma. A etapa de acabamento (do inglês *finishing*) visa corrigir a saída fragmentada do montador e os erros nos *contigs*. O processo inclui fechamento de ausências (do inglês *gaps*) e validação da montagem. A FIGURA 7 ilustra os conceitos de *reads*, *contigs* e *scaffolds*.

FIGURA 7 - VIZUALIZAÇÃO DOS CONCEITOS DE READS, CONTIGS E SCAFFOLDS
FONTE: SOUZA (2012)

2.1.7 Anotação

A anotação de genomas consiste em obter informações estruturais e funcionais sobre uma ou várias sequências relacionadas a um determinado genoma (STEIN, 2001). Para isso, é preciso integrar análises computacionais, dados biológicos auxiliares e perícia biológica para obter a maior quantidade possível de dados úteis (LEWIS; ASHBURNER; REESE, 2000).

No processo de anotação genômica uma sequência desconhecida é documentada através da identificação de vários sítios e segmentos envolvidos na funcionalidade do genoma (ROUZÉ; PAVY; ROMBAUTS, 1999). A anotação consiste em duas etapas principais: predição de genes e agregação de informação. A primeira, também chamada estrutural, realiza a identificação de elementos como: ORFs, estrutura gênica, regiões codificantes e localização de motivos reguladores. A segunda também chamada funcional, identifica funções bioquímicas, biológicas, regulações, interações e expressões. Ambas podem envolver experimentos biológicos e análises *in silico*. A FIGURA 8 ilustra as etapas envolvidas na anotação de genes.

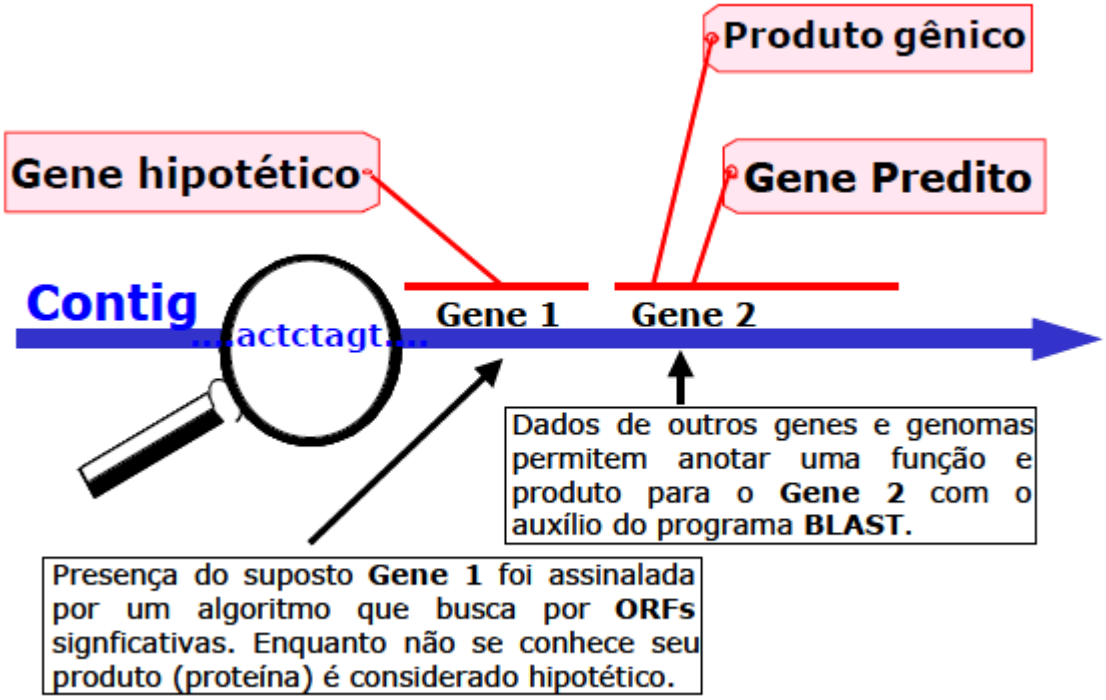


FIGURA 8 - ANOTAÇÃO DE GENES
FONTE: Adaptado de SANTOS e ORTEGA (2003)

Com o advento das NGS, foi possível produzir *drafts* em questão de semanas, gerando a necessidade da rápida anotação de genomas. Ferramentas de anotação automática tentam realizar todas as etapas através de analises por computador, enquanto a anotação manual necessita de uma curadoria realizada por um especialista humano. Dessa forma, anotações automáticas oferecem ganhos em tempo e custo em relação a anotação manual (PETTY, 2010). No entanto, segundo Bakke *et al.*, sistemas de anotação automática apresentam muitas diferenças e é preciso uma avaliação manual

sobre os resultados (BAKKE et al., 2009). Sendo assim, na prática, as duas abordagens devem trabalhar juntas se complementando.

O nível mais básico de anotação utiliza o BLAST, ou outra ferramenta de comparação, para buscar, por similaridades, informações contidas em outros genomas já anotados. No entanto, outras informações podem ser agregadas à anotação, quanto maior o volume de informação mais fácil se torna a tarefa do curador em resolver discrepâncias. Alguns bancos de dados baseiam-se no contexto do genoma, escores de similaridade, dados experimentais e integrações com outras fontes.

Segundo (GUIZELINI, 2010) os problemas mais comuns relatados no processo de anotação são: similaridade x homologia; genes depositados em bancos de dados sem função; genes hipotéticos; genes hipotéticos conservados; falta de padronização dos vocabulários de anotação; e erros/falta de anotação nos bancos de dados públicos. Sendo que espera-se que exista pelo menos um erro de anotação por cada genoma disponível no GenBank.

2.2 ARMAZENAMENTO E OBTENÇÃO DE INFORMAÇÕES BIOLÓGICAS

Muitas das informações obtidas a partir das pesquisas sobre a estrutura e função de moléculas de diversos organismos estão disponíveis digitalmente em bancos de dados. O maior banco, em termos de informações sobre sequências é o NCBI GenBank (SAYERS et al., 2011). Nessa seção abordaremos os principais bancos de dados biológicos relacionados a este trabalho.

2.2.1 GenBank

O GenBank é um banco de dados público criado e distribuído pelo *National Center for Biotechnology Information* (NCBI), sendo uma divisão da *National Library of Medicine* (NLM), localizado no campus do *US National Institutes of Health* (NIH) em Bethesda (BENSON et al., 2009). Juntamente com o *European Bioinformatics Institute* (EBI) do *European Molecular Biology Laboratory* (EMBL) e o *DNA Data Bank of Japan* (DDBJ) constituem o *International Nucleotide Sequence Database Collaboration* (INSDC), por meio da qual as informações são permutadas diariamente.

Segundo (BENSON et al., 2008) o GenBank dobra seu tamanho a cada 18 meses. Em 2011 o volume de dados era de aproximadamente 126.551.501.141 bases em 135.440.924 sequências nas divisões tradicionais e 191.401.393.188 bases em 62.715.288 sequências referentes a dados do projeto *Whole Genome Shotgun* (WGS) (BENSON et al., 2011).

Os dados do NCBI são disponibilizados gratuitamente através de arquivos de texto em formatos padronizados, além disso apresenta diversas ferramentas para análise dos dados. Cada entrada possui uma descrição concisa com: nome científico, taxonomia, referências bibliográficas, além de uma lista de características como regiões codificadoras, traduções em proteínas, unidades de transcrição, regiões repetitivas e sítios de mutação e modificação. Os arquivos seguem o padrão definido junto ao EMBL e recebe o nome de GenBank, esse formato de arquivo descreve com detalhamento as características das sequências, um exemplo de registro pode ser visualizado na FIGURA 9.

```
LOCUS       LISOD                   756 bp    DNA     linear   BCT 30-JUN-1993
DEFINITION  Listeria ivanovii sod gene for superoxide dismutase.
ACCESSION   X64011 S78972
VERSION     X64011.1  GI:44010
KEYWORDS    sod gene; superoxide dismutase.
SOURCE      Listeria ivanovii
  ORGANISM  Listeria ivanovii
            Bacteria; Firmicutes; Bacillales; Listeriaceae; Listeria.
REFERENCE   1  (bases 1 to 756)
  AUTHORS   Haas,A. and Goebel,W.
  TITLE     Cloning of a superoxide dismutase gene from Listeria ivanovii by
            functional complementation in Escherichia coli and characterization
            of the gene product
  JOURNAL   Mol. Gen. Genet. 231 (2), 313-322 (1992)
  MEDLINE   92140371
REFERENCE   2  (bases 1 to 756)
  AUTHORS   Kreft,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-APR-1992) J. Kreft, Institut f. Mikrobiologie,
            Universitaet Wuerzburg, Biozentrum Am Hubland, 8700 Wuerzburg, FRG
FEATURES             Location/Qualifiers
     source          1..756
                     /organism="Listeria ivanovii"
                     /strain="ATCC 19119"
                     /db_xref="taxon:1638"
                     /mol_type="genomic DNA"
     RBS             95..100
                     /gene="sod"
     gene            95..746
                     /gene="sod"
     CDS             109..717
                     /gene="sod"
                     /EC_number="1.15.1.1"
                     /codon_start=1
                     /transl_table=11
                     /product="superoxide dismutase"
                     /db_xref="GI:44011"
                     /db_xref="GOA:P28763"
                     /db_xref="InterPro:IPR001189"
                     /db_xref="UniProtKB/Swiss-Prot:P28763"
                     /protein_id="CAA45406.1"
                     /translation="MTYELPKLPYTYDALEPNFDKETMEIHYTKHHNIYVTKLNEAVS
                     GHAELASKPGEELVANLDSVPEEIRGAVRNHGGGHANHTLFWSSLSPNGGGAPTGNLK
                     AAIESEFGTFDEFKEKFNAAAAAARFGSGWAWLVVNNGKLEIVSTANQDSPLSEGKTPV
                     LGLDVWEHAYYLKFQNRRPEYIDTFWNVINWDERNKRFDAAK"
     terminator      723..746
                     /gene="sod"
ORIGIN
        1 cgttatttaa ggtgttacat agttctatgg aaatagggtc tatacctttc gccttacaat
       61 gtaatttctt .........
//
```

FIGURA 9 - EXEMPLO DE REGISTRO NO FORMATO GENBANK

FONTE: (NCBI, 2013)

Cada registro no GenBank possui um número de acesso estável e único, adicionalmente, para identificar sequências específicas o GenBank utiliza um identificador

único chamado GI. Cada GI refere-se a uma única sequência específica. Qualquer alteração na sequência gera um GI novo, mantendo o registro antigo.

Segundo (SAYERS et al., 2011) o maior banco de proteínas disponibilizado pelo GenBank é o NR, que é um banco não redundante que contem todas as traduções dos CDS contidos no GenBank, além das sequências dos bancos RefSeq (NCBI Reference Sequence (RefSeq), 2013), Swiss-Prot (SwissProt, 2013), PDB (PDB - Protein Data Bank, 2013), PIR (PIR - Protein Information Resource, 2013) e PRF (Protein Research Foundation, 2013).

2.2.2 FASTA

O formato de arquivos mais comum para representar sequências biológicas, sejam elas de nucleotídeos ou resíduos de aminoácidos, é o formato FASTA. Tendo origem junto ao software homônimo (LIPMAN; PEARSON, 1985), tornou-se um padrão em bioinformática. No formato FASTA cada sequência apresenta uma descrição indicada pelo símbolo "maior que" (>) ou por ponto e vírgula (;) no início da linha. A descrição não deve ultrapassar uma linha, caso ocorra, estas serão ignoradas pelo software servindo somente como comentário. Na linha seguinte à descrição tem-se a sequência propriamente dita, sendo que pode haver quebra de linha. Um arquivo FASTA pode apresentar varias sequências em um mesmo arquivo, sendo então chamado de Multi-FASTA onde cada sequência apresenta a mesma formatação definida para sequências únicas. A **FIGURA 10** ilustra um exemplo de sequência no formato FASTA.

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

FIGURA 10 - FORMATO FASTA
FONTE: Wikipédia (2013)

2.2.3 COG

O banco de dados COG (acrônimo para *Clusters of Orthologous Groups*) (TATUSOV et al., 2003) classifica filogeneticamente os genes presentes em genomas

bacterianos completos. A classificação se dá a níveis de categorias funcionais, como mostra a **TABELA 2**.

TABELA 2 - CATEGORIAS FUNCIONAIS DO COG

Código COG	Descrição
Processamento e armazenamento de informação	
J	Tradução, estrutura ribossomal e biogênese
A	Transformação e modificação de RNA
K	Transcrição
L	Replicação, recombinação e reparação
B	Estrutura e dinâmica da cromatina
Processos celulares e sinalização	
D	Controle do ciclo celular, divisão celular, e particionamento do cromossoma
Y	Estrutura nuclear
V	Mecanismos de defesa
T	Mecanismos de transdução de sinal
M	Biogênese da parede/membrana celular
N	Motilidade celular
Z	Citoesqueleto
W	Estruturas Extracelular
U	Tráfico intracelular, secreção e transporte vesicular
O	Modificação pós-traducional, renovação de proteínas, e chaperonas
Metabolismo	
C	Produção e conversão de energia
G	Transporte e metabolismo de carboidratos
E	Transporte e metabolismo de aminoácidos
F	Transporte e metabolismo de nucleotídeos
H	Transporte e metabolismo de Coenzima
I	Transporte e metabolismo de lipídios
P	Transporte e metabolismo de íons inorgânicos
Q	Biossíntese, transporte e catabolismo de metabólitos secundários
Não Caracterizadas	
R	Função geral predita
S	Função não conhecida

FONTE: (TIEPPO, 2011).

Cada COG consiste em proteínas individuais ou grupos de parálogos com pelo menos três linhagens que representam um domínio conservado. A ultima versão do banco apresenta 4.873 grupos, incluindo 136.711 proteínas relacionadas a 66 genomas. Apesar de ser um banco que não recebe atualizações a alguns anos ainda é utilizado para anotações.

2.2.4 Pfam

As proteínas são geralmente compostas por uma ou mais regiões funcionais, essas regiões são chamadas de domínios. Diferentes combinações de domínios dão origem as

diferentes proteínas encontradas na natureza. A identificação desses domínios pode ajudar na inferência das suas funções (FINN et al., 2010).

O Pfam é um banco de dados que contem uma vasta coleção de famílias de proteínas, cada uma representada por alinhamentos múltiplos e modelos escondidos de Markov (HMMs). Existem duas divisões nos bancos do Pfam: o Pfam-A, que possui registros de alta qualidade curados manualmente e o Pfam-B, que possui dados gerados automaticamente com dados do banco ADDA (HEGER; HOLM, 2003).

O Pfam também fornece agrupamentos de alto nível, chamados clãs. Cada clã é uma coleção de registros do Pfam-A que possuem relação de similaridade, estrutura ou perfil de HMM.

2.3 ALGORITMOS E SOFTWARES RELACIONADOS

2.3.1 Preditores de genes

Os softwares de predição de genes buscam identificar regiões codificadoras do genoma. Nessa seção apresentamos alguns softwares desenvolvidos com este propósito.

2.3.1.1 EasyGene

O EasyGene utiliza um método de predição de genes que estima a significância estatística do gene predito. É baseado em modelos escondidos de Markov (HMM). O conjunto de treinamento utiliza extensões de similaridades no banco de dados do Swiss-Prot para estimar o HMM. Os genes putativos são ranqueados com o HMM e baseado no comprimento da ORF é calculada a significância estatística. A medida de significância estatística de uma ORF é a quantidade esperada de ORFs em uma sequência aleatória de uma megabase com um nível de significância igual ou melhor, onde a sequência aleatória tem a mesma estatística do genoma em uma cadeia de Markov de terceira ordem (LARSEN; KROGH, 2003).

2.3.1.2 GeneMark

O GeneMark é uma família de programas de predição desenvolvida no *Georgia Institute of Technology* em Atlanta nos EUA. Para genomas procariotos ele disponibiliza o GeneMarkS que realiza o auto treinamento do genoma, suas características são:

> (...) *utilizes a non-supervised training procedure and can be used for a newly sequenced prokaryotic genome with no prior knowledge of any protein or rRNA genes. The GeneMarkS implementation uses an improved version of the gene finding program GeneMark.hmm, heuristic Markov models of coding and non-coding regions and the Gibbs sampling multiple alignment program. GeneMarkS predicted precisely 83.2% of the translation starts of GenBank annotated Bacillus subtilis genes and 94.4% of translation starts in an experimentally validated set of Escherichia coli genes. We have also observed that GeneMarkS detects prokaryotic genes, in terms of identifying open reading frames containing real genes, with an accuracy matching the level of the best currently used gene detection methods. Accurate translation start prediction, in addition to the refinement of protein sequence N-terminal data, provides the benefit of precise positioning of the sequence region situated upstream to a gene start. Therefore, sequence motifs related to transcription and translation regulatory sites can be revealed and analyzed with higher precision. These motifs were shown to possess a significant variability, the functional and evolutionary connections of which are discussed. (BESEMER; LOMSADZE; BORODOVSKY, 2001)*

2.3.1.3 Glimmer

O Glimmer (*Gene Locator and Interpolated Markov ModelER*) é um sistema que utiliza modelos interpolados de Markov (IMMs) para identificar regiões codificadoras. O método baseado em IMM realiza predições baseado variáveis de contexto, como por exemplo, o comprimento de um oligômero de DNA. Dessa forma, o Glimmer pode realizar mudanças dependendo da composição local da sequência, sendo dessa forma mais flexível que métodos que usam modelos de ordem fixa (SALZBERG et al., 1998). A abordagem do Glimmer usa a combinação de modelos de primeira à oitava ordem (DELCHER et al., 2007). As versões anteriores ao Glimmer3 utilizavam o software RBSfinder (SUZEK et al., 2001) para predição de sítios de ligação do ribossomos, no Glimmer3 a predição foi integrada no seu próprio sistema de pontuação com o software ELPH (ELPH : Estimated Locations of Pattern Hits, 2013).

2.3.1.4 HGF

O HGF (acrônimo para *Hybrid Gene Finder*) é um software desenvolvido no grupo de bioinformática da UFPR que utiliza técnicas de inteligência artificial para realizar predição de regiões codificadoras em procariotos (RAITTZ, 2010, não publicado).

Em técnicas tradicionais de reconhecimento de padrões (RP), duas atividades principais estão envolvidas: extração de características e classificação. Dentre várias técnicas de extração de características existentes os algoritmos genéticos (AG) apresentam uma alternativa interessante para aprimorar esse processo (GUO; ZHANG; NANDI, 2007; SAEYS; INZA; LARRAÑAGA, 2007; VAFAIE; DE JONG, [S.d.]). Os AG's são heurísticas que buscam imitar os processo de evolução natural para resolver problemas computacionais, geralmente sendo aplicados em problemas de otimização. Uma abordagem comum para classificação são as Redes Neurais Artificiais (ANN, acrônimo do inglês *Artificial Neural Network*). ANN são modelos para resolver problemas de inteligência artificial simulando propriedades de neurônios biológicos e suas conexões. A grande vantagem da utilização de ANN é a não necessidade da criação de regras para modelar um sistema, as redes são treinadas a partir de dados conhecidos e adquirem a habilidade de classificar dados futuros com características similares.

A fim de identificar a presença de algum padrão que possa representar uma informação biológica, o HGF representa cada códon com um valor numérico e associa cada códon a uma tabela que representa uma ordenação possível. A razão entre a média e o desvio padrão desta tabela é utilizada como uma característica de uma sequência.

As tabelas tem como objetivo encontrar transformações que possam ajudar na classificação de uma sequência em gene ou não. Para isso, várias podem ser geradas através de um AG que minimize para um conjunto de sequências codificantes e maximize para um conjunto de sequências não-codificantes.

Além das tabelas, o HGF utiliza como características o comprimento da sequência, o percentual da ocorrência de nucleotídeos G ou C, e a quantidade de códons "ATG" na sequência. Essas características extras foram escolhidas arbitrariamente por mostrarem melhora nos resultados dos testes realizados durante o desenvolvimento.

As características são então relacionadas a uma classe (codificante ou não-codificante) e uma ANN *Multi-Layer Perceptron* (MLP) é treinada usando um conjunto com: sequências anotadas presentes em organismos fechados, sequências falsas randômicas e ORFs que não possuem anotação (ORFs na fita complementar reversa de ORFs já anotadas e em regiões intergências).

2.3.1.5 Prodigal

O Prodigal (acrônimo para *Prokaryotic Dynamic Programming Genefinding Algorithm*) é um software desenvolvido no *Oak Ridge National Laboratory* da *University of Tennessee*, nos EUA.

O Prodigal utiliza uma abordagem de "tentativa e erro" no qual, um conjunto de genomas curados foi usado para o entendimento das regras gerais sobre os genes procariotos e como controle de qualidade das predições. Posteriormente 100 genomas do GenBank também foram incluídos (HYATT et al., 2010).

O Prodigal roda completamente de forma não supervisionada. Para isso determina automaticamente um conjunto de genes "reais" putativos, para aprender características como códons de início (ATG,GTG,TTG), sítios de ligação do Ribossomo, tendência de *frame plot* de GC (ISHIKAWA; HOTTA, 1999), estatísticas de hexameros codificantes, entre outros. O conjunto de treinamento examina o *frame plot* de GC, analisando a tendência de Gs e Cs para cada posição do códon em cada fase de leitura. A posição do códon com maior conteúdo de GC é considerada a vencedora e uma soma para a posição do códon é realizada. Após examinar todas as ORFs a soma oferece uma medida aproximada da preferência de G e C para cada posição do códon. Essa medida é então normalizada e dividida por 1/3, dessa forma se por exemplo 2/3 dos códons preferem G ou C na terceira posição o escore para essa posição é 2. Em seguida o escore para cada gene é calculado multiplicando o escore relativo a posição pelo número de códons em que aquela posição é o máximo do *frame plot* de GC com janela de 120 pares de base (HYATT et al., 2010).

Esses escores são calculados entre todos os códons de início e códons de parada com mais de 90pb. Então usando programação dinâmica, é determinado o caminho que maximize os genes para o treinamento. A programação dinâmica é utilizada tanto na etapa de treinamento quanto na fase final. Cada nó na matriz é um códon de início ou parada. A conexão de um nó de início com um nó de parada representa um gene cujo escore foi pré-calculado. Uma pontuação positiva ou negativa é atribuída de acordo com o tamanho da região intergênica (HYATT et al., 2010). A FIGURA 11 mostra as possíveis conexões entre os genes.

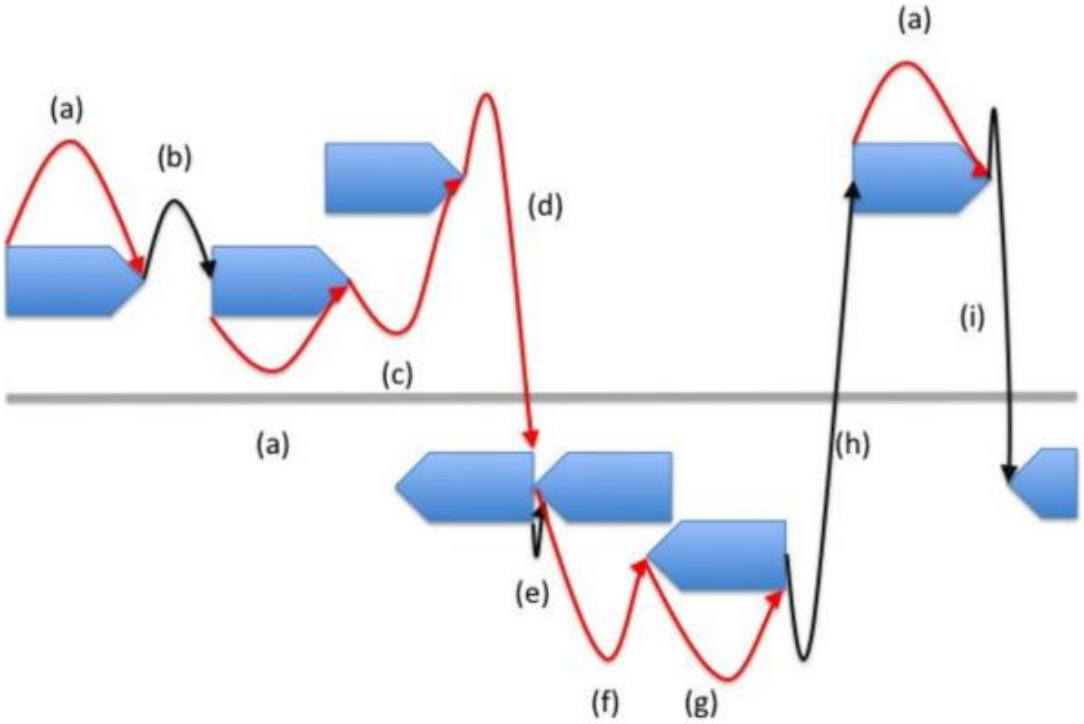


FIGURA 11 - ILUSTRAÇÃO DAS CONEXÕES ENVOLVIDAS NA PROGRAMAÇÃO DINÂMICA NO PRODIGAL

As setas vermelhas representam conexões de genes, e as setas pretas representam conexões intergênicas. (a) 5' para 3' na fita principal: Gene sobre a fita principal. (b) 3' para 5' na fita principal: Espaço Intergênico entre dois genes da fita principal. (c) 3'para 3' na fita principal : Sobreposição de genes na fita principal. (d) 5' na fita principal para 3' na fita reversa: As extremidades 3' dos genes se sobrepõem. (e) 3' para 5' na fita reversa: Espaço intergênico espaço entre dois genes de fita reversa. (f) 5' para 3' na fita reversa: Gene da fita reversa. (g) 3' para 3' na fita reversa: Sobreposição genes na fita reversa. (h) 5' na fita reversa para 5' na fita principal: Espaço intergênico entre dois genes em fitas opostas. (i) 3' na fita principal para 3' na fita reversa: Espaço intergênico espaço entre dois genes em fitas opostas.

FONTE: (HYATT et al., 2010)

Para lidar com o problema de genes em conflito, o Prodigal pré-calcula os melhores genes sobrepostos nos três *frames* para cada extremidade 3'. Dessa forma pode ser feita uma conexão entre a extremidade 3' de gene com outra extremidade de 3' de outro gene na mesma fita. O máximo de sobreposição permitido pelo Prodigal para dois genes na mesma fita é de 60pb e para fitas opostas é de 200pb, no entanto não é permitido a sobreposição de extremidades 5' (HYATT et al., 2010).

Após feita a seleção dos genes de treinamento, escores mais rigorosos são calculados. Para tal, são calculadas as frequências de hexameros relativos aos genes. O escore é calculado pelo logaritmo da relação entre a porcentagem de ocorrência de uma palavra no gene pela porcentagem no genoma inteiro. Assim, se uma palavra ocorre duas vezes mais em um gene do que no resto do genoma o escore é $log(2)$. Isso corresponde a um modelo de Markov de 5ª ordem. O escore final do gene é calculado pelo somatório de todas os hexameros. O Prodigal também altera os escores de genes vizinhos para

penalizar a escolha de genes truncados quando outro gene com maior escore pode ser escolhido (HYATT et al., 2010).

Para determinar os escores dos sítios de início de tradução, o Prodigal calcula as frequências dos códons de início e de *motifs* de sítios de ligação do ribossomo. O Prodigal inicialmente assume que será utilizado o Shine-Dalgarno (SD), caso contrario, utiliza um conjunto de *motifs* RBS e as distancias até os códons de início (o cálculo do escore segue a mesma lógica do escore do gene). Da mesma forma é calculado um escore para o códon de início, os valores são somados e multiplicados por uma constante para serem somados ao escore de codificação. Os valores globais são recalculados a cada novo conjunto de picos encontrados. No fim, tem-se um conjunto de pesos para os códons de início e para cada RBS. Caso não seja detectada a presença de *motifs* SD o Prodigal busca por *motifs* alternativos, buscando por trincas com ocorrência maior que 20% nos genes com maiores escores. O escore final do nó de códon de início relaciona o escore do RBS, o tipo de códon de início, o escore da região montante (*Upstream*) e o escore de codificação, junto com constantes determinadas experimentalmente. Após o cálculo de todos os escores, a programação dinâmica é novamente executada para calcular os escores das conexões entre os genes (HYATT et al., 2010).

O Prodigal apresenta alto desempenho em analisar genomas completos, sendo capaz de localizar todos os códons de parada e 96% dos códons de início dos genes verificados experimentalmente no Ecogene (ZHOU; RUDD, 2012), possui uma taxa de falsos positivos em média abaixo de 5%, e não apresenta problemas com genomas com alto percentual de GC (HYATT et al., 2010).

2.3.2 Comparadores de sequências

Nessa seção apresentaremos abordagens para comparação de sequências.

2.3.2.1 Smith-Waterman

O algoritmo de Smith-Waterman (SMITH; WATERMAN, 1981) foi proposto em 1981, sendo uma variação do algoritmo de Needleman-Wunsh (NEEDLEMAN; WUNSCH, 1970). Utilizando técnicas de programação dinâmica, é um algoritmo que garante que seja

alcançado o alinhamento local ótimo entre duas sequências, de acordo com o sistema de escores utilizado. Enquanto o algoritmo de Needleman-Wunsh realiza o alinhamento global das sequências, o Smith-Waterman, por não permitir escores acumulados negativos, realiza o alinhamento local.

2.3.2.2 BLAST

O BLAST (acrônimo para Basic *Local Alignment Search Tool*) (ALTSCHUL et al., 1990) é um conjunto de ferramentas para comparação de sequências tanto de proteínas quanto de DNA. O BLAST é um dos programas mais usados em bioinformática e apresenta um desempenho muito superior em comparação aos algoritmos que utilizam programação dinâmica.

O BLAST realiza alinhamentos locais, onde busca-se por regiões nas quais duas sequências possuam alto grau de similaridade, diferentemente do alinhamento global, que busca por sequências similares em toda a sua extensão. Sendo que é realizada a comparação entre uma sequência de entrada contra todas as sequências de um dado banco de dados. Seus resultados são confiáveis, tanto do ponto de vista estatístico quanto do ponto de vista de software, sendo uma ferramenta flexível que pode ser utilizada em diferentes cenários (BEDELL; KORF; YANDELL, 2003).

Disponibilizando diversos programas para comparação com todas as combinações entre sequências de nucleotídeos ou proteínas. O ferramental pode ser utilizado tanto através da web quanto *stand-alone,* onde o usuário pode utilizar bancos de sequências de sua preferência.

O BLAST utiliza um sistema de sementes para realizar uma pré-seleção de locais de possível alinhamento. Os alinhamentos são então estendidos seguindo um sistema de pontuação. O sistemas de pontuação podem ser alterados para tornar as comparações mais sensíveis as determinados níveis de similaridade. Para medir a similaridade local é utilizada a medida MSP (*Maximal Segment Pair*) de forma que são efetuadas buscas semi ótimas, encontrando as chamadas HSPs (*High Scoring Pairs*), que são pares de segmentos de sequências de alta pontuação (ALTSCHUL et al., 1990). Além dos alinhamentos, o BLAST fornece informações estatísticas como o valor esperado e a taxa de falso-positivos (YE; MCGINNIS; MADDEN, 2006).

Para comparação de sequências de aminoácidos são usadas matrizes de pontuação PAM (*Point Accepted Mutation*) ou BLOSUM (*BLOck SUbstitution Matrix*) que consideram semelhanças funcionais e evolutivas entre os aminoácidos.

O algoritmo básico do BLAST possui três etapas:

1 - Construção da lista de palavras candidatas - no caso dos nucleotídeos, são geradas todas as palavras de tamanho w presentes na sequência. Para aminoácidos são consideras as palavras de tamanho w fixo, que possuam pontuação mínima igual a um limite T, quando alinhadas, sem *gaps*, com alguma palavra também de tamanho w.

2 - Determinação dos hits no banco de dados - são encontradas todas as combinações exatas (hits) entre as palavras candidatas e as sequências do banco de dados.

3 - Extensão dos *hits* - cada equivalência é estendida até que sua pontuação atinja um limite mínimo x.

2.3.2.3 Técnicas independentes de alinhamento (alignment-free)

As técnicas *alignment-free* propõem uma forma de se obter uma medida de similaridade entre sequências sem realizar alinhamentos. Uma das primeiras propostas para uso de técnicas *alignment-free* para comparar sequências biológicas surgiu com Blaisdell em 1986, no qual o método proposto mostrou-se superior aos algoritmos de alinhamento na identificação de sequências com baixa similaridade (BLAISDELL, 1986).

Técnicas de *alignment-free* podem ser separadas em dois tipos: métodos baseados em palavras, aplicando métodos estatísticos e o métodos onde não é necessário utilizar palavras de tamanhos fixos, utilizando técnicas de compressão e complexidade de Kolmogorov. O primeiro inclui procedimentos baseados em métricas definidas em coordenadas no espaço, tais como distância Euclidiana e entropia relativa a distribuições de frequências. O segundo não necessita contar segmentos fixos, incluindo a teoria da complexidade de Kolmogorov e a representação independente de escala por mapas iterativos (VINGA; ALMEIDA, 2003).

Os métodos baseados em frequências de palavras partem do princípio que sequências similares irão compartilhar a mesma composição de palavras. Em contrapartida os métodos livres de palavras fixas seguem dois caminhos: o uso de ferramentas de compressão de sequências para medir a complexidade das sequências;

ou a representação da sequência usando funções iterativas como mapas bijetivos (VINGA; ALMEIDA, 2003).

Os métodos baseados na frequência de palavras são equivalentes a reconhecer alinhamentos locais entre segmentos idênticos. Sendo uma medida eficiente para filtrar sequências em métodos de alinhamento (PEVZNER, 1992). Sendo que essa medida é dependente do tamanho da palavra (resolução) e da medida (e.g. distância euclidiana).

2.3.3 Anotadores Automáticos

Nessa seção apresentaremos algumas ferramentas de anotação automática que possuem objetivos relacionados com a ferramenta proposta.

2.3.3.1 Sabia

O Sabia (acrônimo para *System for Automated Bacterial Integrated Annotation*), trabalha com montagem e anotação de genomas de procariotos. O sistema de montagem utiliza phred/phrap/consed para geração de *contigs*, gera *scaffolds* com um ordenador de *contigs*, e realiza finalização supervisionada com metodologia própria (ALMEIDA et al., 2004).

O sistema de anotação utiliza para predição o Glimmer e o GeneMark, junto com o tRNAscan (LOWE; EDDY, 1997). As ORFs são submetidas a diversos bancos. São levadas em consideração, a presença de sequências reguladoras de transcrição, sítios de ribossomos e presença de promotores. As buscas são feitas utilizando as sequências em nucleotídeos e aminoácidos utilizando o BLAST. São identificados os *motifs* com o InterPro (InterPro, 2013), é feita uma classificação funcional pelo *Kyoto Encyclopedia of Genes and Genomes* (KEGG) (KEGG - Kyoto Encyclopedia of Genes and Genomes, 2013) e COG, analise da localização com PSORT (PSORT, 2013), capacidade de possíveis transportadores de membrana com *Transport Classification Database* (TCDB) (TCDB, 2013) e ontologia com *Gene Ontology* (GO) (Gene Ontology, 2013). Alem disso, existem procedimentos para identificação de rRNA e *frameshifts* (ALMEIDA et al., 2004).

Uma interface gráfica oferece a opção de mudança no códon de início manual. Sendo possível acessar as vias metabólicas do KEGG durante o processo para auxiliar na

correção da montagem. Alem da identificação de sequências ortólogas usando BBH (*Bidirectional Best Hits*) (OVERBEEK et al., 1999).

2.3.3.2 GenDB

O GenDB é um sistema de anotação de genomas de procariotos que faz uma separação entre as informações das ferramentas e as suas interpretações. A ferramenta gera uma base SQL e permite acesso por interface gráfica e web. As ferramentas como BLAST e InterPro são acessadas via BioGRID (BioGRID - Biological General Repository for Interaction Datasets, 2013) no lado do servidor. Algumas funcionalidades podem ser acessadas via plug-in. A ferramenta importa arquivos nos formatos do GenBank, EMBL e FASTA. E oferece exportação nos formatos GenBank, EMBL, FASTA, GFF e PNG.

As ferramentas integradas são: Glimmer, CRITICA (BADGER; OLSEN, 1999) e tRNAscan para predição. BLAST para homologia, HMMER (EDDY, 2011) para *motifs*, BLOCKS (PIETROKOVSKI; HENIKOFF; HENIKOFF, 1996) e InterPro para classificar as sequências em diferentes tipos de *motifs*. TMHMM (TMHMM, 2013) para predição de alfa-hélice, SignalP (BENDTSEN et al., 2004) para predição de peptídeo sinal, CoBias (MCHARDY et al., 2004) para tendências de uso de códons e GOPArc (GOESMANN et al., 2003) para vias metabólicas (MEYER, 2003).

2.3.3.3 RAST

O RAST (acrônimo para *Rapid Annotation using Subsystem Techology*) fornece um serviço para anotação completa de genomas bacterianos e archaeas. O RAST utiliza subsistemas, um conjunto de papéis funcionais curados manualmente que relaciona grupos de funções com genomas, e FIGfams (MEYER; OVERBEEK; RODRIGUEZ, 2009), um conjunto de família de proteínas.

Os critérios utilizados para agrupar sequências no FIGfams são que ambas as sequências desempenhem o mesmo papel funcional (descrito pelos subsistemas) e tenham cobertura de pelo menos 70%. Para genomas próximos, onde sequências possuem mais de 90% de similaridade, o contexto no cromossomo pode ser utilizado para agrupar na mesma família (AZIZ et al., 2008).

O processo de anotação do RAST utiliza primeiramente o tRNAscan e o "search_for_rnas" (Niels Larsen, não publicado) para identificação de tRNA e rRNA. As anotações posteriores tentam não sobrepor estas. Em seguida, para marcar possíveis genes, são disponibilizadas duas opções, o software Glimmer3 ou uma abordagem própria que permite correções automáticas de *frameshifts* e preenchimento de *gaps*. Os genes marcados são buscados em um pequeno conjunto da FIGfam que agrega propriedades universais entre os procariotos, com os resultados obtidos são feitas correções nos starts e são realizadas buscas em FIGfams relacionados aos resultados já obtidos. Esse conjunto de genes identificados gera um conjunto de treinamento. Os demais são comparados contra os representantes do FIGfam. Por fim as indicações ainda não identificadas são comparadas contra o NR para resolver conflitos e starts. Após identificar as funções é construído o caminho metabólico baseado nos papeis dos subsistemas (AZIZ et al., 2008).

O RAST suporta tanto anotação de genomas de alta qualidade quanto de *drafts*. O serviço leva em torno de 12 à 24 horas para realizar a anotação de um genoma. Após completada a anotação, os genomas podem ser obtidos em diversos formatos ou visualizados *online*. A anotação inclui um mapa dos genes dos subsistemas e uma reconstrução metabólica.

2.3.3.4 BASys

O sistema BASys (acrônimo para *Bacterial Annotation System*) fornece aproximadamente 60 anotações por gene, utilizando mais de 30 programas. O sistema está dividido em 3 partes: a interface web, o sistema de anotação e o sistema de exibição dos dados.

Os dados podem ser submetidos anonimamente, ou atrelado a uma conta. O input deve ser um arquivo FASTA, sendo requerido as informações da topologia (circular ou linear), o tipo da técnica de Gram, e uma identificação (i.e. nome do cromossomo). Se nenhuma informação adicional for dada, o programa utiliza o Glimmer para realizar a predição dos genes. O BASys fornece também a opção de utilizar arquivos já marcados (VAN DOMSELAAR et al., 2005).

O tempo de processamento na época da publicação era de cerca de 24 horas para 3000 genes. As comparações de sequências utilizam o BLAST. O *pipeline* da anotação

utiliza comparação de bancos de dados e analise de sequências. Inicialmente as sequências são buscadas no UniProt (UniProt, 2013) e CyberCell (SUNDARARAJ et al., 2004), obtendo informações de função, papel metabólico, família estrutural e classificação de enzima. Para cada informação é usado um limiar diferente de similaridade. Em seguida, as sequências são comparadas com bancos de sequências modelos, um banco não redundante de bactérias, PDB e COG. Outras comparações são feitas sobre o Pfam e PROSITE, o PredictSPTM (J. Cruz, não publicado) é utilizado para predição de peptídeo sinal e PSIPRED (BRYSON et al., 2005) para predição de estrutura secundaria. Caso obtenha resultado do PDB é gerado o modelo com Homodeller (Homodeller, 2013) e gerado a analise com VADAR (WILLARD, 2003). Outras informações como peso molecular, ponto isoelétrico e *operons* são calculados. As informações descrevem escores para confiabilidade das informações. As informações são exibidas em forma circular com hiperlinks para cada gene. A visualização utiliza o CGView (STOTHARD; WISHART, 2005).

2.3.3.5 AGMIAL

O AGMIAL é um sistema integrado para anotação de genomas bacterianos que busca seguir as seguintes características: maximizar a automatização da anotação; possibilitar o trabalho com drafts de sequências; apresentar características modulares e extensíveis; utilizar padrões consolidados tanto da informática como da bioinformática; e distribuição em licença de código aberto.

O sistema é dividido em 2 subsistemas, CAM e PAM. O CAM trata da parte de predição, utiliza Markov no programa SHOW para identificar genes, tRNAscan, rRNAscan (K. Bryson, não publicado) e PETRIN (D'AUBENTON CARAFA; BRODY; THERMES, 1990). O CAM permite a inserção de novos contigs em projetos sem afetar anotações já realizadas. Para visualização o ele utiliza o Artemis (RUTHERFORD et al., 2000), MuGeN (HOEBEKE; NICOLAS; BESSIERES, 2003) para visualizar vários genomas e CGView para visualização circular (BRYSON et al., 2006).

O PAM utiliza ferramentas para detectar domínios integrado ao InterProScan (InterProScan, 2013). Também realiza buscas por domínios de família contra o Pfam ou TIGRFAMS (TIGRFAMS, 2013). Para as buscas é utilizado o PSI-BLAST (ALTSCHUL et al., 1997). Algumas buscas são genéricas como contra o Swiss-Prot, outras são

específicas contra organismos escolhidos pelo o usuário. Buscas são feitas também no COG, PROSITE e PRINTS (ATTWOOD, 2003). A localização é predita usando o PSORTb (GARDY et al., 2005).

2.3.3.6 KAAS

O KASS (*KEEG Automatic Annotation Server*) é um sistema para a anotação funcional de genes. O método utiliza o BLAST para comparar um gene contra um conjunto de sequências referencias do banco do KEGG. Os genes homólogos identificados são ranqueados pelo escore do BLAST e BHR (*Bi-directional Hit Rate*). O BHR realiza comparação entre os genes de 2 genomas, comparando todos contra todos em ambos os sentidos para gerar um escore. O melhor escore designa o identificador do grupo KEEG *Orthology* (KO). A ferramenta usa como input um FASTA com as ORFs ou *Expressed Sequence Tags* (ESTs). O banco de referência usa um conjunto pré-selecionado de organismos, podendo ser escolhido outros caso o usuário deseje (MORIYA et al., 2007).

3 MÉTODOS

3.1 BUSCA POR GENES

Para realizar a tarefa de predição de CDSs optamos por utilizar mais de uma ferramenta. Para isso, escolhemos uma ferramenta como sendo a principal, da qual utilizamos todas as suas indicações. E uma secundária, da qual utilizamos somente algumas indicações de forma a complementar a principal. A estratégia utilizada para união das duas ferramentas segue os seguintes passos:

1 - Geramos as ORFs utilizando as duas ferramentas separadamente.

2 - Todas as ORFs da ferramenta principal são incorporadas à anotação.

3 - Trechos de regiões intergênicas com mais de 200pb são estendidos até os códons de parada adjacentes.

4 - Caso haja indicações de novas ORFs nesses trechos indicados pela ferramenta secundária, elas são incorporadas à anotação.

Nos resultados testamos todas as combinações de ferramentas possíveis, e optamos por utilizar no SILA a combinação Prodigal e HGF, adotando o Prodigal como ferramenta principal e o HGF como ferramenta secundária. Os detalhes sobre a escolha são discutidos nos resultados.

3.2 RAFTS3

Devido ao alto custo computacional envolvido na comparação de sequências com as ferramentas do BLAST, que ocasionavam longos tempos de execução nas busca em grandes bancos, tornou-se necessário desenvolver uma nova abordagem para contornar esse problema que pudesse realizar a mesma tarefa com qualidade comparável e com um desempenho muito superior. Para isso, neste trabalho foi desenvolvido o RAFTS3 (acrônimo para *Rapid Alignment Free Tool for Sequences Similarity Search*), uma ferramenta para busca por similaridade em grandes bancos de dados que utiliza conceitos *alignment-free*. Os métodos da técnica e os seus resultados estão descritos no

APÊNDICE I - RAFTS3: A RAPID ALIGNMENT FREE TOOL FOR SEQUENCES SIMILARITY SEARCH.

3.3 SILA

A ferramenta de anotação automática, SILA, engloba as estratégias de *Gene-Finding* e de comparação de sequências com o RAFTS3 propostas anteriormente.

Após realizada a predição das CDSs, é preciso identificar quais são as proteínas produzidas por elas codificadas. Para isso são realizadas buscas por similaridade. As CDSs traduzidas são comparadas com sequências de proteínas já anotadas presentes nos bancos de dados. Para que possa ser inferida a homologia a partir da similaridade, foi utilizada a medida de escore relativo igual ou superior a 0.3 (BARBOSA-SILVA et al., 2008). O escore relativo é calculado através da razão do escore do alinhamento da sequencia buscada (*query)* com a sequencia resposta (*subject)* pelo escore do alinhamento da sequencia *query* com ela mesma. Consideramos esses resultados como *hits* significativos.

As informações que podem ser agregadas à anotação são obtidas em três consultas independentes a três bancos de dados. A anotação final é apresentada no formato de arquivo GenBank que possui campos específicos previstos para certas características. Na anotação, cada campo é preenchido caso obtenha um *hit* significativo no banco de dados. A FIGURA 12 exibe um exemplo de informações que podem ser associadas a um determinado gene.

- O campo Product é obtido pela descrição no banco NR do NCBI.
- O campo Function é obtido pelos grupos do COG.
- O campo EC Number é obtido pela descrição no banco Pfam.

```
/gene="0001.0008"
/function="[NOU]Cell motility|Posttranslational modification, protein turnover,
chaperones|Intracellular trafficking, secretion, and vesicular transport"
/product="Type 4 prepilin-like proteins leader peptide processing enzyme"
/EC_number="3.4.23.43"
/note="Organism:Methyloversatilis universalis FAM5 ScoreAlign:0.55301
ScoreRede:0.99837  ScoreCOG:0.52859 Query_coverage:91.32%  Subject_coverage:92.28%"
/color=5
```

FIGURA 12 - EXEMPLO DE INFORMAÇÕES ASSOCIADAS NA ANOTAÇÃO
FONTE: O AUTOR (2013).

Após a identificação das sequências e anotação de suas características, é realizado um ajuste de códons de início baseado nas anotações. O melhor *hit* significativo obtido no NR é usado como base para ajuste nas posições. As informações de escore e cobertura de alinhamento são utilizadas como critério na escolha. A FIGURA 13 ilustra um caso de mudança de códon baseada na anotação.

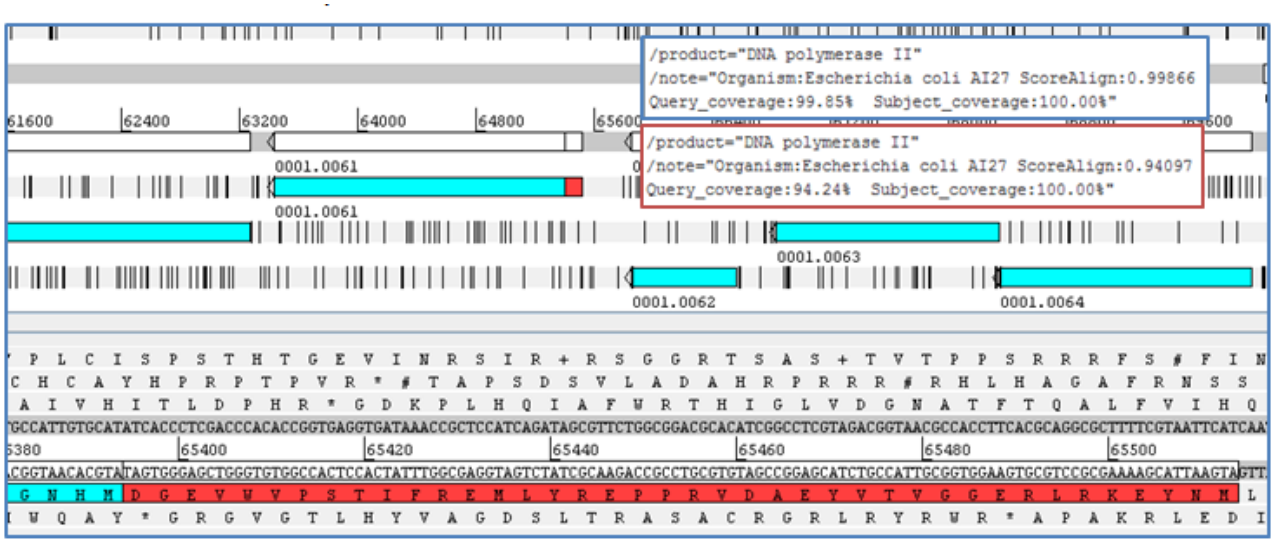


FIGURA 13 - CORREÇÃO DE CÓDONS DE INÍCIO APÓS A ANOTAÇÃO

Em vermelho está representada a indicação original, em azul a indicação com códon de início corrigido. Nota-se no exemplo que o escore de alinhamento antes da correção era de 0.94097 e após a correção é de 0.99866, indicando maior proximidade com a sequência presente no banco de dados.

FONTE: O AUTOR (2013).

3.3.1 SILA-WEB

A ferramenta SILA também foi disponibilizada como serviço na web. O serviço fornece uma interface no qual solicita ao usuário um arquivo no formato FASTA. Não é requerido nenhum parâmetro ou informação adicional, todo o serviço é realizado automaticamente e o resultado da anotação é disponibilizado para visualização na web, com diferentes formas de visualização, ou em arquivo no formato GenBank.

O sistema foi desenvolvido utilizando-se Java, HTML5 e MySQL e contou com a colaboração da equipe de TCC realizado em paralelo com o mestrado (VIALLE et al., 2013).

FIGURA 14 - SILA-WEB

FONTE: O AUTOR (2013).

4 RESULTADOS E DISCUSSÃO

4.1 RESULTADOS DO GENE-FINDING

Para definir a estratégia de *gene-finding* usada no SILA foram analisados os seguintes softwares de predição:

- EasyGene 1.2b Server[1];
- GeneMarkS 4.7a[2];
- Glimmer3[3];
- HGF[4];
- Prodigal v1.20[5];

Os softwares foram testados em 24 genomas de procariotos (TABELA 3), sendo observada a quantidade de genes preditos, o percentual de acerto de códons de parada, o percentual de acerto de códons de início, e o percentual de acerto de ambos (TABELA 4).

Os softwares EasyGene, GeneMarkS e Glimmer foram executados via serviço web, enquanto os softwares HGF e Prodigal foram executados localmente. O software EasyGene não disponibiliza a opção de treinamento para organismos via web, disponibilizando um conjunto de 138 organismo previamente treinados, a lista de organismos usados como referência encontra-se no APÊNDICE II. Os demais softwares foram utilizados com os parâmetros padrões. As ferramentas de análise foram desenvolvidas em MATLAB.

A análise das ferramentas separadamente (TABELA 4) mostrou que o Prodigal apresenta uma melhor taxa de acerto em relação ao número de indicações, além de obter os melhores resultados no acerto de códons de início. O Glimmer e o GeneMarkS obtiveram as melhores taxas de acerto de códons de parada. O HGF apresentou cerca de 64% de acerto dos códons de início, justamente por não realizar qualquer tipo de tratamento para esses casos, e resultados próximos aos demais em relação aos códons de parada. O EasyGene, devido ao conjunto de treinamento limitado, apresentou a menor quantidade de indicações obtendo uma taxa de acerto muito abaixo das outras ferramentas. Dessa forma decidimos por não utiliza-lo nos demais testes.

[1] Disponível em http://www.cbs.dtu.dk/services/EasyGene
[2] Disponível em http://exon.gatech.edu/genemarks.cgi
[3] Disponível em http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi
[4] Disponível em http://www.bioinfo.ufpr.br/hgf
[5] Disponível em http://prodigal.ornl.gov/server.html

TABELA 3 - LISTA DE ORGANISMOS USADOS PARA TESTES

ORGANISMO	ACCESSION	GENES ANOTADOS NO NCBI
Thermotoga maritima MSB8 chromosome, complete genome.	NC_000853	1854
Escherichia coli str. K-12 substr. MG1655 chromosome, complete	NC_000913	4268
Archaeoglobus fulgidus DSM 4304, complete genome.	NC_000917	2404
Treponema pallidum subsp. pallidum str. Nichols chromosome, complete genome.	NC_000919	1034
Bacillus subtilis subsp. subtilis str. 168 chromosome, complete genome.	NC_000964	4174
Caulobacter crescentus CB15 chromosome, complete genome.	NC_002696	3737
Chlorobium tepidum TLS chromosome, complete genome.	NC_002932	2245
Dehalococcoides ethenogenes 195, complete genome.	NC_002936	1579
Geobacter sulfurreducens PCA chromosome, complete genome.	NC_002939	3444
Treponema denticola ATCC 35405 chromosome, complete genome.	NC_002967	2767
Methylococcus capsulatus str. Bath chromosome, complete genome.	NC_002977	2955
Neisseria meningitidis MC58 chromosome, complete genome.	NC_003112	2063
Ralstonia solanacearum GMI1000 chromosome, complete genome.	NC_003295	3433
Colwellia psychrerythraea 34H chromosome, complete genome.	NC_003910	4909
Pseudomonas protegens Pf-5 chromosome, complete genome.	NC_004129	6137
Streptococcus agalactiae NEM316, complete genome.	NC_004368	2094
Haemophilus influenzae 86-028NP chromosome, complete genome.	NC_007146	1792
Carboxydothermus hydrogenoformans Z-2901 chromosome, complete genome.	NC_007503	2619
Clostridium perfringens ATCC 13124 chromosome, complete genome.	NC_008261	2875
Mycobacterium tuberculosis H37Ra chromosome, complete genome.	NC_009525	4034
Campylobacter jejuni subsp. doylei 269.97 chromosome, complete genome.	NC_009707	1731
Streptococcus pneumoniae Hungary19A-6, complete genome.	NC_010380	2155
Porphyromonas gingivalis ATCC 33277, complete genome.	NC_010729	2090
Helicobacter pylori P12 chromosome, complete genome.	NC_011498	1566
		Total: 67959

FONTE: O AUTOR (2013).

TABELA 4 - PERFORMANCE DAS FERRAMENTAS DE PREDIÇÃO

PREDITOR	START	STOP	START+STOP	INDICAÇÕES
EasyGene	0.638	0.778	0.637	54913
GeneMarkS	0.783	0.955	0.782	69755
Glimmer3	0.770	0.959	0.770	71013
HGF	0.644	0.938	0.644	73504
Prodigal	0.807	0.949	0.806	67757
RAST	0.767	0.958	0.767	70935

Os valores indicam o percentual de acerto de códons de início (START), códons de parada (STOP), e ambos (START+STOP) em relação à anotação disponível no NCBI.

FONTE: O AUTOR (2013).

Outra análise realizada foi em relação a utilização de códons de início alternativos ao códon ATG (TABELA 5). O HGF utiliza como critério de escolha de códon de início alternativo, somente casos onde não há a presença do códon ATG, dessa forma obteve os resultados mais distantes da anotação contida no NCBI. As ferramentas Prodigal e Glimmer apresentaram resultados semelhantes entre si, e próximos aos da anotação no NCBI. Enquanto o GeneMarkS foi o que mais se aproximou da anotação contida na anotação no NCBI.

TABELA 5 - CÓDONS DE INÍCIO ALTERNATIVOS

Códon de Início	GeneMarkS	Glimmer3	HGF	Prodigal	NCBI
ATG	33600	41448	65184	39344	34363
Outro	36155	29565	8320	28413	33596
Total	69755	71013	73504	67757	67959
Porcentagem de não ATG	51.83%	41.63%	11.31%	41.93%	49.43%

A tabela indica o numero total de códons de início ATG ou alternativos (Outro) nos 24 organismos analisados e a relação com a anotação disponível no NCBI

FONTE: O AUTOR (2013).

Testamos também as ferramentas separadamente em genomas gerados aleatoriamente (TABELA 6). Criamos 5 genomas aleatórios com 4Mb com diferentes percentuais de GC (30%, 40%, 50%, 60% e 70%). O HGF foi a ferramenta que apresentou menos indicações comparada as demais.

TABELA 6 - QUANTIDADE DE ORFs INDICADAS EM GENOMAS GERADOS ALEATORIAMENTE

Percentual de GC	GeneMarkS	Glimmer	HGF	Prodigal
30%	848	5778	601	2133
40%	930	7825	735	2741
50%	1622	10608	962	2681
60%	4004	10960	1324	5258
70%	4907	9326	1233	6762

Quantidade de indicação por ferramenta em genomas aleatórios de 4Mb com diferentes percentuais de CG.

FONTE: O AUTOR (2013).

A combinação das ferramentas seguiu a metodologia descrita nos métodos. Os testes mostraram resultados muito semelhantes nas combinações das três ferramentas. Sendo que a escolha da ferramenta principal na combinação afeta principalmente o acerto de códons de início, e a ferramenta secundária, utilizada nas intergênicas, serviu para elevar o acerto de códons de parada. A TABELA 7 mostra os resultados para todas as combinações.

TABELA 7 - PERFORMANCE DA COMBINAÇÃO DAS FERRAMENTAS DE PREDIÇÃO

COMBINAÇÃO	START	STOP	START+STOP	INDICAÇÕES
GeneMarkS+Glimmer3	0.789	0.964	0.788	73003
GeneMarkS+HGF	0.786	0.960	0.785	72765
GeneMarkS+Prodigal	0.785	0.958	0.784	70375
Glimmer3+GeneMarkS	0.776	0.966	0.775	72903
Glimmer3+HGF	0.775	0.967	0.774	74275
Glimmer3+Prodigal	0.776	0.965	0.775	72123
HGF+GeneMarkS	0.660	0.958	0.659	73893
HGF+Glimmer3	0.663	0.963	0.662	75538
HGF+Prodigal	0.658	0.954	0.657	72675

...continuação

COMBINAÇÃO	START	STOP	START+STOP	INDICAÇÕES
Prodigal+GeneMarkS	0.813	0.959	0.812	70055
Prodigal+Glimmer3	0.817	0.964	0.816	72183
Prodigal+HGF	0.811	0.957	0.810	70609

Percentual de acerto da combinação das ferramentas de códons de início (START), códons de parada (STOP), e ambos (START+STOP) em relação à anotação disponível no NCBI.

FONTE: O AUTOR (2013).

Como as combinações obtiveram resultados muito próximos, optamos por utilizar no SILA o Prodigal em conjunto com o HGF. O Prodigal mostrou-se superior a todas as demais ferramentas testadas em questão ao acerto de códons de início e à quantidade de falsos positivos. O HGF por sua vez, por realizar a predição com base apenas na sequência, sem analise de contexto, serviu como uma ótima ferramenta para complementar os resultados do Prodigal.

4.2 RESULTADOS DO RAFTS3

Os resultados da ferramenta RAFTS3 estão disponíveis no APÊNDICE I - RAFTS3: A RAPID ALIGNMENT FREE TOOL FOR SEQUENCES SIMILARITY SEARCH.

4.3 RESULTADOS DO SILA

Testamos o SILA realizando a anotação dos 24 organismos utilizados anteriormente nos testes com os preditores. Comparamos os resultados com a ferramenta RAST.

Notamos que as ferramentas apresentaram resultados semelhantes em todos os aspectos. O RAST indicou 70935 genes, enquanto o SILA (Prodigal+HGF) indicou 70609 genes. O percentual de acerto de códons de parada foi praticamente o mesmo para as duas ferramentas, e o SILA foi um pouco melhor em relação ao acerto de códons de início, com 81% contra 76% do RAST (TABELA 8).

TABELA 8 - PERFORMANCE DA COMBINAÇÃO DAS FERRAMENTAS DE ANOTAÇÃO

FERRAMENTA	START	STOP	START+STOP	INDICAÇÕES
RAST	0.767	0.958	0.767	70935
SILA	0.811	0.957	0.810	70609

Percentual de acerto de códons de início (START), códons de parada (STOP), e ambos (START+STOP) em relação à anotação disponível no NCBI.

FONTE: O AUTOR (2013).

Ao analisarmos as ocorrências de genes idênticos, considerando tanto códon de início como códon de parada, vimos que 49442 dos genes anotados no NCBI foram preditos tanto pelo SILA como pelo RAST (FIGURA 15). Vimos também, que a predição do SILA indicou mais 5634 genes anotados no NCBI, além dos genes anteriores, totalizando 55076, enquanto o RAST, obteve mais 2699, totalizando 52141 genes. As duas ferramentas indicaram 8081 genes comuns que não constam na anotação do NCBI. E como indicações únicas, o SILA apresentou 7452 e o RAST 10713.

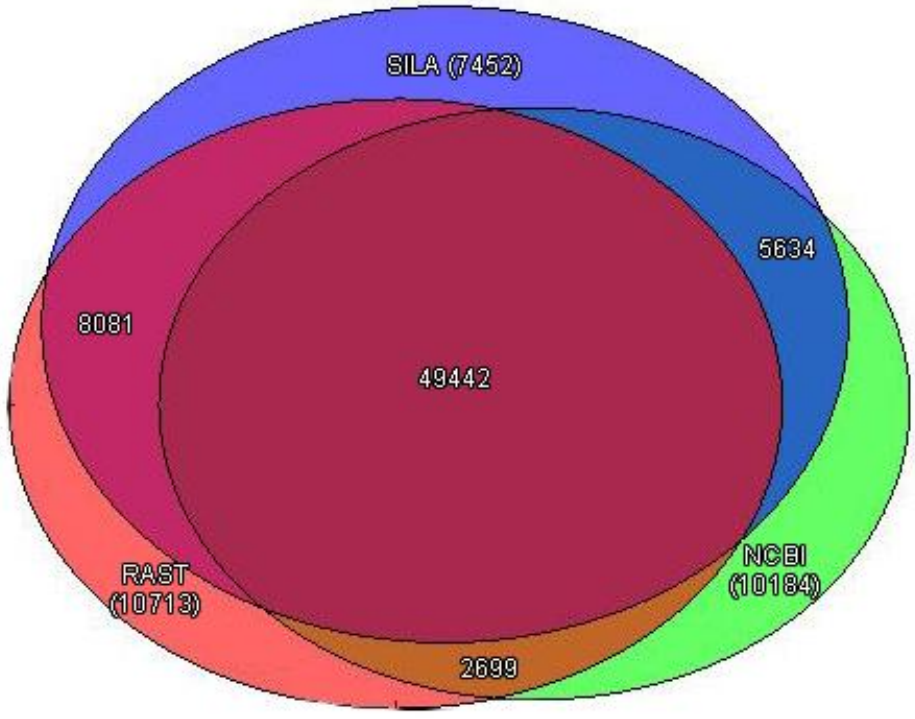


FIGURA 15 - QUANTIDADES DE GENES COMPARTILHADOS

O diagrama mostra as quantidades de indicações de genes (START+STOP) em comum entre as ferramentas SILA, RAST e anotação disponível no NCBI.

FONTE: O AUTOR (2013).

Realizando a análise de ocorrência de códons de parada comuns, notamos que o SILA apresentou 65092 ocorrências em comum com as anotações contidas no NCBI, e o

RAST apresentou 65113 ocorrências, sendo que 64243 foram comuns as duas ferramentas. De indicações novas o SILA apresentou 5517 e o RAST 5822 (FIGURA 16).

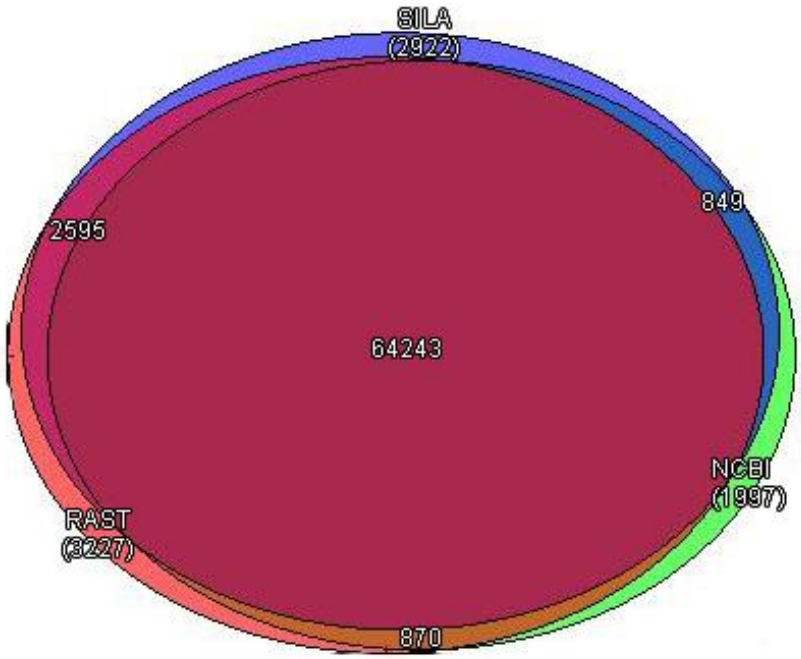


FIGURA 16 - QUANTIDADES DE CÓDONS DE PARADA COMPARTILHADOS

O diagrama mostra as quantidades de indicações de códons de parada (STOP) em comum entre as ferramentas SILA, RAST e anotação disponível no NCBI.

FONTE: O AUTOR (2013).

Na análise da utilização de códons de início (TABELA 9) vimos que tanto o SILA quanto o RAST apresentaram resultados praticamente idênticos, apresentando cerca de 88% das indicações de códons alternativos presentes nas anotações do NCBI.

TABELA 9 - CÓDONS DE INÍCIO ALTERNATIVOS ENTRE AS FERRAMENTAS DE ANOTAÇÃO

Códon de Início	SILA	RAST	NCBI
ATG	41018	41148	34363
Outro	29591	29787	33596
Total	70609	70935	67959
Porcentagem de não ATG	41.90%	41.99%	49.43%

FONTE: O AUTOR (2013).

Com a finalidade de verificar a capacidade de anotar produtos funcionais relevantes, outra análise comparativa foi realizada com relação aos produtos indicados pelas ferramentas (TABELA 10 e TABELA 11). Analisamos os 20 produtos que mais ocorreram nas anotações dos 24 organismos. Enquanto o SILA indicou 70609 produtos,

um para cada gene, o total de produtos indicados pelo RAST foi de 77483, devido a indicação de mais de um produto para alguns genes. O agrupamento dos produtos foi realizado na ocorrência idêntica das descrições. Como as ferramentas utilizam bancos de dados diferentes, com diferentes terminologias, é natural que os produtos indicados apresentem discrepâncias.

TABELA 10 - LISTA DOS 20 PRODUTOS COM MAIOR OCORRÊNCIA NO SILA

Produto	Quantidade
HYPOTHETICAL PROTEIN	6473
LIPOPROTEIN	459
ABC TRANSPORTER ATP-BINDING PROTEIN	429
CONSERVED HYPOTHETICAL PROTEIN	276
LYSR FAMILY TRANSCRIPTIONAL REGULATOR	221
SENSOR HISTIDINE KINASE	175
TRANSCRIPTIONAL REGULATOR	172
ACETYLTRANSFERASE	163
TRANSPOSASE	152
TETR FAMILY TRANSCRIPTIONAL REGULATOR	142
ABC TRANSPORTER PERMEASE	139
METHYL-ACCEPTING CHEMOTAXIS PROTEIN	138
OXIDOREDUCTASE	136
ARAC FAMILY TRANSCRIPTIONAL REGULATOR	131
TRANSMEMBRANE PROTEIN	129
SIGNAL PEPTIDE PROTEIN	122
TONB-DEPENDENT RECEPTOR	119
RESPONSE REGULATOR	114
CONSERVED PROTEIN	111
DNA-BINDING RESPONSE REGULATOR	110

FONTE: O AUTOR (2013).

TABELA 11 - LISTA DOS 20 PRODUTOS COM MAIOR OCORRÊNCIA NO RAST

Produto	Quantidade
HYPOTHETICAL PROTEIN	8482
MOBILE ELEMENT PROTEIN	756
CONSERVED HYPOTHETICAL PROTEIN	258
UNKNOWN	225
TRANSCRIPTIONAL REGULATOR, TETR FAMILY	196
LIPOPROTEIN, PUTATIVE	173
MEMBRANE PROTEIN, PUTATIVE	169
CONSERVED DOMAIN PROTEIN	161
ABC TRANSPORTER, ATP-BINDING PROTEIN	148
TRNA	144
TRANSCRIPTIONAL REGULATOR, LYSR FAMILY	142
LONG-CHAIN-FATTY-ACID--COA LIGASE	140
METHYL-ACCEPTING CHEMOTAXIS PROTEIN	140
PUTATIVE MEMBRANE PROTEIN	140
ACETYLTRANSFERASE, GNAT FAMILY	136
TRANSCRIPTIONAL REGULATOR, ARAC FAMILY	131
POSSIBLE MEMBRANE PROTEIN	126
NAD	119
PUTATIVE	115
PROBABLE TRANSMEMBRANE PROTEIN	113

FONTE: O AUTOR (2013).

5 CONCLUSÃO

Neste trabalho desenvolvemos uma ferramenta para realizar a anotação automática de genomas de procariotos com auto desempenho. Para tal, foi preciso desenvolver uma nova abordagem de comparação de sequências que pudesse oferecer uma alternativa as ferramentas do BLAST que apresentasse resultados semelhantes com baixíssimo custo computacional.

A ferramenta RAFTS3 mostrou-se eficiente para a anotação de genomas, possibilitando comparações de sequências contra grandes bancos de dados de proteínas com velocidades 500 vezes superiores em comparação à ferramenta BLASTp. Com perda de sensibilidade de cerca de 10% nas indicações com escore relativo superior a 0.3. Consideramos que o ganho no tempo justifica a perda de alguns resultados, sendo que caso seja necessário melhor qualidade, o BLAST pode ser utilizado para complementar os resultados.

Para desenvolver a ferramenta de anotação SILA analisamos diversas ferramentas de predição de genes. Nossa análise mostrou que as ferramentas apresentam diferenças em seus métodos gerando discrepâncias na identificação correta de regiões codificadoras. Dessa forma decidimos por utilizar a combinação de duas ferramentas, de maneira que uma complementasse os resultados da outra. Decidimos por utilizar o Prodigal como preditor principal devido as taxas de acerto elevadas e o baixo numero de indicações e o HGF como preditor secundário por apresentar uma abordagem diferente das demais ferramentas (sem análise de contexto) o que pode representar uma vantagem para complementar os resultados.

O SILA utilizando-se das predições feitas pela combinação das ferramentas Prodigal e HGF, realiza busca em bancos de sequências de proteínas com o RAFTS3 afim de identificar informações funcionais dos genes. Trabalhamos com três bancos de dados, o NR do NCBI como sendo o principal banco para a anotação, de onde obtemos as informações referentes ao produto gênico. O COG para identificação dos grupos funcionais dos genes, e o Pfam para classificação de enzimas.

Comparado o SILA com a ferramenta de anotação automática RAST, notamos que ambas apresentaram resultados bastante semelhantes, com vantagens para o SILA em alguns critérios. Como na identificação correta de códons de início, quantidade de indicações e informação relacionada ao produto gênico.

O sistema de anotação foi disponibilizado com serviço na web e oferece anotações de qualidade de maneira rápida e de fácil utilização. Dessa forma concluímos que esse trabalho apresenta uma alternativa vantajosa em relação a outras ferramentas de anotação disponíveis ao público.

RECOMENDAÇÕES E PROJETOS FUTUROS

A inclusão de novas informações que possam enriquecer as anotações fornecidas com o SILA podem ser realizadas com a inclusão ou integração de outros bancos de dados, bem como a utilização de outras ferramentas de predição.

A busca de rRNA e tRNA, predição de peptídeo sinal, papel metabólico, família estrutural e predição de estrutura secundária podem ser incluídos na anotação com a utilização de ferramentas já consolidadas com este propósito.

O desempenho da aplicação pode ser melhorada com a reescrita das bibliotecas em linguagem C/C++.

Os métodos aqui desenvolvidos podem ser utilizados para realização de mineração de dados para fins da descoberta de novos conhecimentos.

REFERÊNCIAS

ALMEIDA, L. G. P. et al. A System for Automated Bacterial (genome) Integrated Annotation--SABIA. **Bioinformatics (Oxford, England)**, v. 20, n. 16, p. 2832-3, 1 nov. 2004.

ALTSCHUL, S. F. et al. Basic local alignment search tool. **Journal of molecular biology**, v. 215, n. 3, p. 403-10, 5 out. 1990.

ALTSCHUL, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic acids research**, v. 25, n. 17, p. 3389-402, 1 set. 1997.

ATTWOOD, T. K. PRINTS and its automatic supplement, prePRINTS. **Nucleic Acids Research**, v. 31, n. 1, p. 400-402, 1 jan. 2003.

AZIZ, R. K. et al. The RAST Server: rapid annotations using subsystems technology. **BMC genomics**, v. 9, p. 75, jan. 2008.

BADGER, J. H.; OLSEN, G. J. CRITICA: coding region identification tool invoking comparative analysis. **Molecular biology and evolution**, v. 16, n. 4, p. 512-24, abr. 1999.

BAKKE, P. et al. Evaluation of three automated genome annotations for Halorhabdus utahensis. **PloS one**, v. 4, n. 7, p. e6291, jan. 2009.

BARBOSA-SILVA, A. et al. Clustering of cognate proteins among distinct proteomes derived from multiple links to a single seed sequence. **BMC bioinformatics**, v. 9, p. 141, jan. 2008.

BAXEVANIS, A.; OUELLETTE, B. **Bioinformatics**. New York, USA: John Wiley & Sons, Inc., 2001. v. 43

BEDELL, J.; KORF, I.; YANDELL, M. **BLAST**. [S.l.] O'Reilly, 2003.

BENDTSEN, J. D. et al. Improved prediction of signal peptides: SignalP 3.0. **Journal of molecular biology**, v. 340, n. 4, p. 783-95, 16 jul. 2004.

BENSON, D. A et al. GenBank. **Nucleic acids research**, v. 36, n. Database issue, p. D25-30, jan. 2008.

BENSON, D. A et al. GenBank. **Nucleic acids research**, v. 37, n. Database issue, p. D26-31, jan. 2009.

BENSON, D. A et al. GenBank. **Nucleic acids research**, v. 39, n. Database issue, p. D32-7, jan. 2011.

BERG, J. M.; TYMOCZKO, J. L.; STRYER, L. **Bioquímica**. 6. ed. Rio de Janeiro: Guanabara Koogan, 2010.

BESEMER, J.; LOMSADZE, A; BORODOVSKY, M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. **Nucleic acids research**, v. 29, n. 12, p. 2607-18, 15 jun. 2001.

BioGRID - Biological General Repository for Interaction Datasets. Disponível em: <http://thebiogrid.org/>. Acesso em: 1 fev. 2013.

BLAISDELL, B. E. A measure of the similarity of sets of sequences not requiring sequence alignment. **Proceedings of the National Academy of Sciences of the United States of America**, v. 83, n. 14, p. 5155-9, jul. 1986.

BLATTNER, F. R. et al. The complete genome sequence of Escherichia coli K-12. **Science (New York, N.Y.)**, v. 277, n. 5331, p. 1453-62, 5 set. 1997.

BRYSON, K. et al. Protein structure prediction servers at University College London. **Nucleic acids research**, v. 33, n. Web Server issue, p. W36-8, 1 jul. 2005.

BRYSON, K. et al. AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. **Nucleic acids research**, v. 34, n. 12, p. 3533-45, jan. 2006.

COLE, S. T. et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. **Nature**, v. 393, n. 6685, p. 537-44, 11 jun. 1998.

COMIN, M.; VERZOTTO, D. Alignment-free phylogeny of whole genomes using underlying subwords. **Algorithms for molecular biology : AMB**, v. 7, n. 1, p. 34, jan. 2012.

CRICK, F. Central dogma of molecular biology. **Nature**, 1970.

D'AUBENTON CARAFA, Y.; BRODY, E.; THERMES, C. Prediction of rho-independent Escherichia coli transcription terminators. A statistical analysis of their RNA stem-loop structures. **Journal of molecular biology**, v. 216, n. 4, p. 835-58, 20 dez. 1990.

DELCHER, A. L. et al. Identifying bacterial genes and endosymbiont DNA with Glimmer. **Bioinformatics (Oxford, England)**, v. 23, n. 6, p. 673-9, 15 mar. 2007.

EDDY, S. R. Accelerated Profile HMM Searches. **PLoS computational biology**, v. 7, n. 10, p. e1002195, out. 2011.

ELPH : Estimated Locations of Pattern Hits. Disponível em: <http://cbcb.umd.edu/software/ELPH/>. Acesso em: 1 fev. 2013.

FENG, J.; ZHAO, W.; ZHANG, H. Use of an Alignment-free Method to Compare Reduced Amino Acid Sequences. **Journal of Convergence Information Technology**, v. 6, n. 4, p. 213-221, 30 abr. 2011.

FINN, R. D. et al. The Pfam protein families database. **Nucleic acids research**, v. 38, n. Database issue, p. D211-22, jan. 2010.

FLEISCHMANN, R. D. et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. **Science (New York, N.Y.)**, v. 269, n. 5223, p. 496-512, 28 jul. 1995.

FOX, J. **WHAT IS BIOINFORMATICS?** Disponível em: <http://www.scq.ubc.ca/what-is-bioinformatics/>. Acesso em: 19 jan. 2013.

GARDY, J. L. et al. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. **Bioinformatics (Oxford, England)**, v. 21, n. 5, p. 617-23, 1 mar. 2005.

Gene Ontology. Disponível em: <http://www.geneontology.org/>. Acesso em: 1 fev. 2013.

genome.gov. Disponível em: <http://www.genome.gov/images/content/cost_per_genome.jpg>. Acesso em: 20 jan. 2013.

GIBAS, C.; JAMBECK, P. **Developing bioinformatics computer skills**. [S.l: s.n.].

GOESMANN, A. et al. Building a BRIDGE for the integration of heterogeneous data from functional genomics into a platform for systems biology. **Journal of Biotechnology**, v. 106, n. 2-3, p. 157-167, dez. 2003.

GUIZELINI, D. **BANCO DE DADOS BIOLÓGICO NO MODELO RELACIONAL PARA MINERAÇÃO DE DADOS EM GENOMAS COMPLETOS DE PROCARIOTOS DISPONIBILIZADOS PELO NCBI GENBANK**. [S.l.] Universidade Federal do Paraná, 2010.

GUO, H.; ZHANG, Q.; NANDI, A. K. FEATURE GENERATION USING GENETIC PROGRAMMING BASED ON FISHER CRITERION. p. 1867-1871, 2007.

HEGER, A.; HOLM, L. Exhaustive enumeration of protein domain families. **Journal of molecular biology**, v. 328, n. 3, p. 749-67, 2 maio. 2003.

HOEBEKE, M.; NICOLAS, P.; BESSIERES, P. MuGeN: simultaneous exploration of multiple genomes and computer analysis results. **Bioinformatics**, v. 19, n. 7, p. 859-864, 1 maio. 2003.

Homodeller. Disponível em: <http://wishartlab.com/homodeller/>. Acesso em: 1 fev. 2013.

HUANG, G. et al. Alignment-free comparison of genome sequences by a new numerical characterization. **Journal of theoretical biology**, v. 281, n. 1, p. 107-12, 21 jul. 2011.

HYATT, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. **BMC bioinformatics**, v. 11, p. 119, jan. 2010.

InterPro. Disponível em: <http://www.ebi.ac.uk/interpro/>. Acesso em: 1 fev. 2013.

InterProScan. Disponível em: <http://www.ebi.ac.uk/Tools/pfa/iprscan/>. Acesso em: 1 fev. 2013.

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. v. 174, p. 251-253, 1999.

JING, J.; WILSON, S.; BURDEN, C. Weighted k-word matches: a sequence comparison tool for proteins. **ANZIAM Journal**, v. 52, 2011.

KANTOROVITZ, M. R.; ROBINSON, G. E.; SINHA, S. A statistical method for alignment-free comparison of regulatory sequences. **Bioinformatics (Oxford, England)**, v. 23, n. 13, p. i249-55, 1 jul. 2007.

KEGG - Kyoto Encyclopedia of Genes and Genomes. Disponível em: <http://www.genome.jp/kegg/>. Acesso em: 1 fev. 2013.

KOONIN, E. V; GALPERIN, M. Y. **Sequence - Evolution - Function Computational Approaches in Comparative Genomics**. Boston: [s.n.].

KURTZ, S. et al. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. **BMC genomics**, v. 9, p. 517, jan. 2008.

LANDER, E. S. et al. Initial sequencing and analysis of the human genome. **Nature**, v. 409, n. 6822, p. 860-921, 15 fev. 2001.

LARSEN, T. S.; KROGH, A. EasyGene--a prokaryotic gene finder that ranks ORFs by statistical significance. **BMC bioinformatics**, v. 4, p. 21, 3 jun. 2003.

LESK, A. M. **Introduction to bioinformatics**. New York, USA: Oxford University Press, 2002.

LEWIS, S.; ASHBURNER, M.; REESE, M. G. Annotating eukaryote genomes. **Current opinion in structural biology**, v. 10, n. 3, p. 349-54, jun. 2000.

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

LOWE, T. M.; EDDY, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. **Nucleic acids research**, v. 25, n. 5, p. 955-64, 1 mar. 1997.

MAHMOOD, K. et al. Efficient large-scale protein sequence comparison and gene matching to identify orthologs and co-orthologs. **Nucleic acids research**, v. 40, n. 6, p. e44, mar. 2012.

MCHARDY, A. C. et al. Comparing expression level-dependent features in codon usage with protein abundance: an analysis of “predictive proteomics”. **Proteomics**, v. 4, n. 1, p. 46-58, jan. 2004.

MEYER, F. GenDB--an open source genome annotation system for prokaryote genomes. **Nucleic Acids Research**, v. 31, n. 8, p. 2187-2195, 15 abr. 2003.

MEYER, F.; OVERBEEK, R.; RODRIGUEZ, A. FIGfams: yet another set of protein families. **Nucleic acids research**, v. 37, n. 20, p. 6643-54, nov. 2009.

MORIYA, Y. et al. KAAS: an automatic genome annotation and pathway reconstruction server. **Nucleic acids research**, v. 35, n. Web Server issue, p. W182-5, jul. 2007.

NASCIMENTO, L. V. DO. **Um Sistema Baseado em Agentes para Re-anotação de Genomas**. [S.l.] Universidade Federal do Rio Grande do Sul, 2005.

NCBI. Disponível em: <http://www.ncbi.nlm.nih.gov/>. Acesso em: 18 jan. 2013.

NCBI Reference Sequence (RefSeq). Disponível em: <http://www.ncbi.nlm.nih.gov/RefSeq/>.

NEEDLEMAN, S. B.; WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **Journal of molecular biology**, v. 48, n. 3, p. 443-53, mar. 1970.

NELSON, D. L.; COX, M. M. **Lehninger princípios de bioquímica**. 4. ed. São Paulo: Sarvier, 2006.

OVERBEEK, R. et al. The use of gene clusters to infer functional coupling. **Proceedings of the National Academy of Sciences of the United States of America**, v. 96, n. 6, p. 2896-901, 16 mar. 1999.

PDB - Protein Data Bank. Disponível em: <http://www.rcsb.org/pdb/home/home.do>. Acesso em: 1 fev. 2013.

PETTY, N. K. Genome annotation: man versus machine. **Nature reviews. Microbiology**, v. 8, n. 11, p. 762, nov. 2010.

PEVZNER, P. A. Statistical distance between texts and filtration methods in sequence comparison. **Computer applications in the biosciences : CABIOS**, v. 8, n. 2, p. 121-7, abr. 1992.

PIETROKOVSKI, S.; HENIKOFF, J. G.; HENIKOFF, S. The Blocks database--a system for protein classification. **Nucleic acids research**, v. 24, n. 1, p. 197-200, 1 jan. 1996.

PIR - Protein Information Resource. Disponível em: <http://pir.georgetown.edu/>. Acesso em: 1 fev. 2013.

PROSDOCIMI, F. INTRODUÇÃO A BIOINFORMÁTICA. 2007.

Protein Research Foundation. Disponível em: <http://www.prf.or.jp/index-e.html>. Acesso em: 1 fev. 2013.

PSORT. Disponível em: <http://www.psort.org/>. Acesso em: 1 fev. 2013.

QUAIL, M. A et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. **BMC genomics**, v. 13, n. 1, p. 341, jan. 2012.

ROUZÉ, P.; PAVY, N.; ROMBAUTS, S. Genome annotation: which tools do we have for it? **Current opinion in plant biology**, v. 2, n. 2, p. 90-5, abr. 1999.

RUTHERFORD, K. et al. Artemis: sequence visualization and annotation. **Bioinformatics (Oxford, England)**, v. 16, n. 10, p. 944-5, out. 2000.

SAEYS, Y.; INZA, I.; LARRAÑAGA, P. A review of feature selection techniques in bioinformatics. **Bioinformatics (Oxford, England)**, v. 23, n. 19, p. 2507-17, 1 out. 2007.

SALZBERG, S. L. et al. Microbial gene identification using interpolated Markov models. **Nucleic acids research**, v. 26, n. 2, p. 544-8, 15 jan. 1998.

SANGER, F.; NICKLEN, S. DNA sequencing with chain-terminating. v. 74, n. 12, p. 5463-5467, 1977.

SAYERS, E. W. et al. Database resources of the National Center for Biotechnology Information. **Nucleic acids research**, v. 39, n. Database issue, p. D38-51, jan. 2011.

SMITH, T. F.; WATERMAN, M. S. Identification of common molecular subsequences. **Journal of molecular biology**, v. 147, n. 1, p. 195-7, 25 mar. 1981.

SOARES, I.; GOIOS, A.; AMORIM, A. Sequence comparison alignment-free approach based on suffix tree and L-words frequency. **TheScientificWorldJournal**, v. 2012, p. 450124, jan. 2012.

STEIN, L. Genome annotation: from sequence to biology. **Nature reviews genetics**, v. 2, n. July, 2001.

STOTHARD, P.; WISHART, D. S. Circular genome visualization and exploration using CGView. **Bioinformatics (Oxford, England)**, v. 21, n. 4, p. 537-9, 15 fev. 2005.

SUNDARARAJ, S. et al. The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate in silico modeling of Escherichia coli. **Nucleic acids research**, v. 32, n. Database issue, p. D293-5, 1 jan. 2004.

SUZEK, B. E. et al. A probabilistic method for identifying start codons in bacterial genomes. **Bioinformatics (Oxford, England)**, v. 17, n. 12, p. 1123-30, dez. 2001.

SwissProt. Disponível em: <http://web.expasy.org/docs/swiss-prot_guideline.html>. Acesso em: 1 fev. 2013.

TATUSOV, R. L. et al. The COG database: an updated version includes eukaryotes. **BMC bioinformatics**, v. 4, p. 41, 11 set. 2003.

TCDB. Disponível em: <http://www.tcdb.org/browse.php>. Acesso em: 1 fev. 2013.

TIEPPO, E. **MONTAGEM E ANÁLISE PRELIMINAR DO GENOMA DE Bradyrhizobium elkanii 587 UTILIZANDO LEITURAS DE SEQUÊNCIAS DE DNA CURTAS**. [S.l.] Universidade Federal do Paraná, 2011.

TIGRFAMS. Disponível em: <http://www.jcvi.org/cgi-bin/tigrfams/index.cgi>. Acesso em: 1 fev. 2013.

TMHMM. Disponível em: <http://www.cbs.dtu.dk/services/TMHMM/>. Acesso em: 1 fev. 2013.

UniProt. Disponível em: <http://www.uniprot.org/>. Acesso em: 1 fev. 2013.

VAFAIE, H.; DE JONG, K. Robust feature selection algorithms. **Proceedings of 1993 IEEE Conference on Tools with AI (TAI-93)**, p. 356-363, [S.d.].

VAN DOMSELAAR, G. H. et al. BASys: a web server for automated bacterial genome annotation. **Nucleic acids research**, v. 33, n. Web Server issue, p. W455-9, 1 jul. 2005.

VIALLE, R. A. et al. **SISTEMA INTEGRADO PARA ANOTAÇÃO AUTOMÁTICA DE GENOMAS**. [S.l.] Universidade Federal do Paraná, 2013.

VINGA, S.; ALMEIDA, J. Alignment-free sequence comparison--a review. **Bioinformatics**, v. 19, n. 4, p. 513-523, 1 mar. 2003.

WAGNER, I.; MUSSO, H. New Naturally Occurring Amino Acids. **Angewandte Chemie International Edition in English**, v. 22, n. 11, p. 816-828, 1983.

WATSON, J. D. et al. **Biologia molecular do gene**. 5. ed. Porto Alegre: Artmed, 2006. p. 760

WILLARD, L. VADAR: a web server for quantitative evaluation of protein structure quality. **Nucleic Acids Research**, v. 31, n. 13, p. 3316-3319, 1 jul. 2003.

YE, J.; MCGINNIS, S.; MADDEN, T. L. BLAST: improvements for better sequence analysis. **Nucleic acids research**, v. 34, n. Web Server issue, p. W6-9, 1 jul. 2006.

YU, C. et al. Protein map : An alignment-free sequence comparison method based on various properties of amino acids. **Gene**, v. 486, n. 1-2, p. 110-118, 2011.

ZHOU, J.; RUDD, K. E. EcoGene 3.0. **Nucleic acids research**, v. 41, n. November 2012, p. 613-624, 28 nov. 2012.

APÊNDICES

APÊNDICE I - RAFTS3: A RAPID ALIGNMENT FREE TOOL FOR SEQUENCES SIMILARITY SEARCH

RAFTS^3: Rapid Alignment Free Tool for Sequences Similarity Search

Ricardo A. Vialle[1, §], Fábio O. Pedrosa[2], Vinicius A. Weiss[2], Dieval Guizelini[1], Juliana H. Tibaes[1], Jeroniza N. Marchaukoski[1], Emanuel M. de Souza[2] and Roberto T. Raittz[1]

[1]Laboratory of Bioinformatics, Technological and Professional Education Sector, Federal University of Paraná, Curitiba, Paraná, Brazil.
[2]Department of Biochemistry and Molecular Biology, Federal University of Paraná, Curitiba, Paraná, Brazil.

[§]Corresponding author

Email addresses:
RAV: ricardovialle@gmail.com
FOP: fpedrosa@ufpr.br
VAW: viniciusweiss@gmail.com
DG: dievalg@gmail.com
JHT: tibaes.juliana@gmail.com
JNM: jeroniza@gmail.com
EMS: souzaem@ufpr.br
RTR: raittz@gmail.com

ABSTRACT

Background

A similarity search between a given protein sequence against a database is an essential task in genome analysis. Sequence alignment is the most frequent way to perform such analysis. Although this approach is efficient, the time required to perform searches against large databases is always a challenge. Alignment-free techniques offer alternatives to compare sequences without the need of alignment.

Results

We developed RAFTS^3, a fast protein similarity search tool that utilizes a filter step for candidate selection based on shared k-mers and a comparison measure using a binary matrix of co-occurrence of amino acid residues.

RAFTS^3 performed searches approximately 400-500 times faster than those with BLASTp against large protein databases such as NR, Pfam or UniRef, with a small loss of sensitivity depending on the similarity degree of the sequences.

Conclusions

RAFTS^3 is a new alternative for fast comparison of protein sequences, genome annotation and biological data mining.

KEYWORDS

Alignment-free, Protein sequence comparison, Biological data mining, Genome annotation.

BACKGROUND

Biological data mining deals with the discovery of patterns, trends, answers, or other meaningful information that is hidden in the data. Sequence comparison is the main component in the retrieval system from genomic databases. An efficient sequence comparison algorithm is critical for searching biologic data bases. Usually, bioinformatic workflows use algorithms based on sequence alignment such as BLAST [1] to search for similarity of DNA/RNA or protein sequences against large sequence databases. Comparisons involving large databases such as NCBI NR [2], however, are computationally costly and demand long running times. The development of new computationally faster algorithms may provide significant improvement in biological pattern search. A class of techniques that can speed up sequence comparison involves an alignment-free approach [3].

Algorithms based on sequence alignment are efficient in detecting similarities between protein sequences and there are two approaches for this technique. The first is the global alignment approach, where sequences are aligned from end to end; the second is the local alignment approach, that seeks to align regions within the sequences. Originally alignment techniques used dynamic programming to produce an optimized alignment between the

sequences. Though efficient implementations have been developed, the computational load to compare large amounts of sequences makes these algorithms very slow and demanding [3, 4].

To compensate the high computational cost of alignment, heuristic approaches were proposed. In general, these methods generate lists of subsequences of pre-determined length "k" (k-mers). The subject database is searched to find sequences that have common k-mers related to the query sequence. The k-mers are then extended using scores alignment schemes to maximize the size of the aligned regions. However, although heuristic methods are somewhat efficient to perform searches in large databases, they also have their limitations [4].

The alignment-free sequence comparison methods offer a way to obtain a similarity measure between sequences without the need to perform alignments. The very first proposal of an alignment-free method for biological sequence comparison was shown to be superior than sequence alignment algorithms in some respects such as the ability to compare sequences with low similarity [5].

Alignment-free methods are also based on the assumption that two similar sequences share a certain portion of k-mers that is equivalent to recognizing local alignments between identical segments. According to [6] this is an efficient technique for sequence filtering. Given a query sequence, the alignment-free methods generally work by selecting subject sequences with k-mers that are present in both query and subject sequences. The procedure then applies a statistical method to establish a similarity ranking for these sequences [7].

There are two classes of alignment-free techniques: a) methods based on words (sequences) with fixed sizes, followed by the explicit use of words for statistical analysis and includes procedures based on defined metrics in coordinate space, such as Euclidean distance and entropy of frequency distributions; b) methods where words of fixed sizes are not required for statistical analysis using data compression and/or Kolmogorov complexity scale independent representations by iterative maps. Reviews of those techniques are available [3, 7, 8].

Several alignment-free techniques have been proposed with different degrees of success. These have been applied in phylogenetic reconstruction [9–14], identification of homologous proteins [4], genome annotation [15], and identification of regulatory sequences [16].

There are some alignment-free approaches that proposes replace alignment-based approaches to search and compare sequences against large databases showing significant speed up results. PAUDA [17] is a alternative for BLASTx to search read sequences against protein databases in metagenomic context. And USEARCH [18] is a alternative for BLASTp that applies a k-mer approach to perform searches of protein sequences against a protein database. However, USEARCH require a paid license and only a limited free version is available for academic use.

In this paper we propose a fast, efficient alignment-free method, named RAFTS^3, to determine sequence similarity based on amino acid co-occurrence matrices, as well as a new heuristic approach for filtering sequences. The results show that RAFTS^3 was much faster than

BLASTp, with negligible loss of sensitivity when applied against large databases in all tests performed. RAFTS^3 has been successfully used in several biological data-mining tasks.

METHODS

Since RAFTS^3 deals with protein sequences comparison against protein databases, the first step to be considered is to set up the protein database into a specific RAFTS^3 format. The formatting consists of two steps to be applied to each protein sequence within a FASTA file: a) the sequences must be indexed by a hash function and b) a binary amino acid co-occurrence matrix (BCOM) that have to be assigned to each sequence to represent its contents.

When a formatted database is available, query searches can be performed. This process is also divided in two distinct steps. The filtering of candidates, that selects sequences whose indexed k-mers are shared with the query sequence, and the comparison of these candidates, that is done by means of the BCOM.

Format database

The formatting process takes a FASTA database as input and creates a structure comprising a hash table and BCOM matrices for all sequences in the database. Aiming to improve access to the sequences, RAFTS^3 also creates an index to allow direct access to each sequence in the FASTA file (Fig.1.A).

For each sequence in the database a set of k-mers is randomly selected and submitted to a hash function. The indexes are then stored into a hash table for fast selection of candidate sequences for comparison. These indexes will permit a further retrieval of any sequence in the database sharing a given k-mer (Fig.1.A). As standards, 10 k-mers with lengths of 6 amino acid residues are selected per sequence.

The formatting process also involves a BCOM assignment to each sequence. The BCOM was designed to represent the sequences with a low memory usage. Both the hash table and the BCOM matrices are stored in a common structure that is loaded in RAM during the application, aiming to minimize disk access in sequence comparison. The hash function and the BCOM structure will be detailed further.

Query sequence search

Searching is the main step in RAFTS^3. Its purpose is to retrieve sequences similar to a sequence of interest from a database. It is desirable that the recovered sequences are ranked by their similarity with the query sequence. Searching involves two main steps: filter and comparison.

In the filtering process, the search scope is reduced by selecting, through a hash table, only sequences containing common k-mers related to the query sequence. To perform a search based on a sequence of a given length n, hash indexes for all possible k-mers with length k are calculated by taking a sliding window that runs through the sequence from position 1 to $n - k + 1$

(Fig.1.B). The indexes generated for each k-mer are used to select the candidate sequences by searching in the hash table (Fig.1.B). The strategy of database indexing and candidate sequences selection is shown in Fig. 2.

The comparison is performed with the candidate sequences based on their BCOM. The details of the comparison method will be discussed later (Section - Binary co-occurrence matrix (BCOM)). Alignments of the best results can also be done to confirm the results or to assign them to a well established metric. The number of alignments can be customized by parameters; by default the alignment is performed only with the best stated result. As a measure of alignment quality, the relative score E (1) was used [19].

$$E = \frac{alignment\ score\ of\ S_1\ with\ S_2}{alignment\ score\ of\ S_1\ with\ S_1} \qquad (1)$$

Where S_1 and S_2 are protein sequences and the alignment algorithm is the Smith-Waterman [20] with the BLOSUM62 score table.

Hash function

The hash function of RAFTS^3 is an essential step in the filtering process and takes place either in the database format or in the query search. To perform the RAFTS^3 a specific hash function was developed.

The recursive indexing technique (INREC) [21] was used to assign a real number to a protein k-mer. INREC is a technique of dimensionality reduction and pattern recognition that uses a recursive process of a mathematical function to encapsulate, in a single number, the information that describes a pattern. Thereby, the indexes generated by similar sequences are equal or close to each other. The numbers generated by the INREC function are transformed in hash indexes H through the expression

$$H = \mathrm{mod}(\mathrm{INREC}(k\text{-}mer) \times largenumber, largeprime) \quad (2)$$

Where $largenumber$ is a value to express the decimal fraction of the INREC index as an integer number, and $largeprime$ defines size and spreading of the hash table.

To apply the INREC algorithm amino acid residues were converted to quaternary numeral system triplets by a two-way conversion table (Table 1). The codes are arbitrary but in correspondence to the possible codons representing a specific amino acid residue. The numerals 1, 2, 3 and 4 represent the nucleotide residues A, C, G and T/U.

Thus, $D = \{d_1, d_2, \ldots, d_m\}$ is a sequence of integers representing a sequence of length m, where $d_i \in \{1,2,3,4\}$. The INREC index I is generated from the recursion of the function f:

$$I = f(d_1 f(d_2 ... f(d_m))) \qquad (3)$$

where,

$$f(d_i) = tanh\left(\sqrt{\left(\frac{d_i}{4}\right)^{-1}}\right) \qquad (4)$$

Binary co-occurrence matrix (BCOM)

The binary co-occurrence matrix BCOM is a bi-dimensional fingerprint of an amino acid sequence. It not only represents an amino acid sequence but is a pattern of comparison with other sequences.

A BCOM is a binary matrix where each cell position (x, y) represents the occurrence of an amino acid pair "XY" in a sequence S. If the value within the cell is set to null the pair does not occur in S (Fig. 3). Thus for each sequence a 20x20 binary matrix is generated that represents the occurrence of all possible amino acid pairs within it. Any sequence can be represented by a matrix with 400 bits or 50 bytes. The small data volume needed and the uniform data structure of the BCOM allows databases with millions of sequences to be represented and stored in RAM. The whole NR database can be handle in a common laptop.

Let A and B be BCOMs corresponding to sequences S_1 and S_2 respectively. The sum of a binary operation and between the matrices A and B represents the occurrence of common amino acid residue pairs and reflects the sequences similarity. In a similar way, the binary operation xor is performed to calculate the degree of dissimilarity as a support for the comparison. Thus, the measure of difference e between A and B is given by the equation (5).

$$e = \frac{sum(xor(A,B))}{sum(and(A,B))} \qquad (5)$$

Each candidate sequence selected in the filter step will be related to a dissimilarity measure given by e and sorted in crescent order. Finally, a correlation coefficient r (6) between the matrices is calculated on the results of highest similarity obtained from e, and is used for reordering the results.

$$r = \frac{\sum_m \sum_n (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{\left(\sum_m \sum_n (A_{mn} - \bar{A})^2\right)\left(\sum_m \sum_n (B_{mn} - \bar{B})^2\right)}} \qquad (6)$$

Where, $\overline{A} = Mean\,(A)$ and $\overline{B} = Mean(B)$

Due the computational cost, the number of sequences compared with the correlation equation (6) was limited to 50.

IMPLEMENTATION AND DATASETS

The RAFTS^3 prototype was written in MATLAB [22] using its built-in functions, the Bioinformatics Toolbox [23], and an in-house library. Two protein databases were used, the NCBI NR with 19,689,576 sequences, PFAM [24] with 15,929,002 sequences and the UniRef50 [25] with 6,784,251 sequences. Comparisons were made with BLASTp.

RESULTS

Parameters selection

To determine the default parameters to be used by RAFTS^3 for the candidate selection step, sets of 1 to 20 k-mers of 4, 5, 6 or 7 amino acid residues were randomly extracted from each sequence of the NR database. A randomly selected subset of 1000 protein sequences from the NR database was used as query. Two criteria were considered to define the RAFTS^3 configuration settings: 1) The running time to search 1000 queries (Fig.4.A). 2) The number of queries with second best hit with relative score higher than 0.3 (Fig.4.B). The best hit was disregarded since that it always corresponds to the query sequence.

The purpose of this procedure was to find the number of k-mer per sequence and the k-mer size to be adopted as default parameters to carry out sequence searches by RAFTS^3. This analysis showed that the running times were lower using k-mer sizes of 6 and 7 residues and the number of hits with relative score higher than 0.3 reached a plateau with sets of 10 k-mers per sequence. Thereby, the following parameters were chosen as default: 10 k-mers of 6 amino acid residues per sequence.

Comparison of RAFTS^3 with BLASTp

The sensitivity and running time of the RAFTS^3 was compared with BLASTp version 2.2.26+ configured with its default parameters.

A data set of the bacteria *Herbaspirillum huttiense* subsp. *putei* IAM 15032 [26], available in NCBI (Genbank ANJR01000025.1), was used to perform the comparison. Contig 26 whose length is 1,354,643 bp with 1323 ORFs was analyzed. These ORFs were predicted by a combination of HGF (an in-house developed gene predictor) and Prodigal [27]. The NR and PFAM databases, frozen for the testing, did not contain any sequence of the referred organism. Thus the best hits from BLASTp and RAFTS^3 could be compared.

The performance was evaluated in terms of the number of sequences retrieved with relative score higher than a threshold, and the processing time spent in the search by both tools. The number of sequences retrieved by BLASTp was considered as the gold standard, representing 100% of the results.

RAFTS^3 showed from 87% to 98% of sensitivity compared with BLASTp when searching PFAM database and from 89% to 98% when searching the NR database, depending on the threshold of score (Table 2). RAFTS^3 showed to be almost 400 times faster than BLASTp when recovering PFAM database, and more than 500 times faster when recovering NR database.

Moreover, a data set of 1000 sequences randomly selected from a newer version of NR, absent in the database used for tests, was compared. This set represent sequences from more than 650 different organisms. For this tests RAFTS^3 showed from 81% to 92% of sensitivity compared with BLASTp when searching UniRef50 database and from 88% to 96% when searching the NR database, depending on the threshold of score (Table 3).

To illustrate some results in comparison with BLASTp, 3 different genes from the genome of the *Herbaspirillum seropedicae* SmR1 [28] were searched against the NR database and the top 4 similar sequences indicated by RAFTS^3 and BLASTp were selected. The E-value and the relative scores were calculated for both. The results showed that, despite some ordering differences, RAFTS^3 identified genes with close similarity as BLASTp did(Table 4).

A more extensive comparison is available as supplementary material on Table 5 showing results for 5 different genes randomly selected from the test set searched against the NR database and the Top 50 similar sequences indicated by RAFTS^3 and BLASTp. Also, to measure these ordering differences, the position where the RAFTS^3 best hits appears on BLASTp results were counted for the 1000 sequences used for tests. The results showed that 72% of the RAFTS^3 best hits occurred in the first 10 BLASTp results (Figure 5). A table with the Top 50 counts for both tools is available as supplementary material (Table 6).

Comparison of RAFTS^3 with USEARCH

We choose to use the small COG [29] database to test both software. For this case USEARCH was more fast and accurate than RAFTS^3. However, duo the limitations of the free version, we don't know how USEARCH behaves with large databases. We know that the memory consumption of USEARCH will be more than 40Gb for the NR database, while with RAFTS^3 it is 20 times less. Also we know that RAFTS^3 runtime don`t is much affected by the database growth and the sensibility tends to increase. Thereby, as RAFTS^3 aims to large databases, we were not able to do a fair comparison.

CONCLUSIONS

RAFTS^3 uses an aggressive filter approach with a fast comparison method based on BCOMs. Due the limitation of the free version of USEARCH, comparisons for searches against large databases cannot be performed. The comparison of RAFTS^3 with BLASTp, showed that RAFTS^3 could be used to achieve a fast protein similarity search with a small loss of sensitivity. The sensitivity compared to BLAST increases with the sequence similarity. RAFTS^3 also shows a minimal loss on performance when challenged with the increase of the database in comparison with BLASTp. We expect that RAFTS^3 can take advantage of the growth of the databases due the increase of *sequences of close relatives. As the database increases the filtering options can be more stringent avoiding the increase of the number of candidate sequences selected and thus of memory usage.*

The foregone demonstrates that the RAFTS^3 can perform high-speed protein search comparisons locally using a desktop computer or laptop. RAFTS^3 is being used in tasks as genome annotation by our Bioinformatics group at the Federal University of Parana with success, and presents a good solution for protein sequences data mining.

AVAILABILITY OF SUPPORTING DATA

The source code and the test files are available at: ftp://200.236.3.56/RAFTS3/

LIST OF ABBREVIATIONS

BCOM = Binary co-occurrence matrix

BLAST = Basic local alignment search tool

INREC = Recursive indexing technique

RAFTS^3 = Rapid alignment-free tool for sequences similarity search

COMPETING INTERESTS

All the authors declare that they have no competing interests.

ACKNOWLEDGEMENTS

Funding: National Institute of Science and Technologies of Biological Nitrogen Fixation, Fundação Araucária, CAPES and CNPq.

REFERENCES

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403–10.

2. Sayers EW, Barrett T, Benson D a, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Phan L,

Pruitt KD, Schuler GD, Sequeira E, et al.: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2011, **39**(Database issue):D38–51.

3. Vinga S, Almeida J: **Alignment-free sequence comparison--a review**. *Bioinformatics* 2003, **19**:513–523.

4. Mahmood K, Webb GI, Song J, Whisstock JC, Konagurthu AS: **Efficient large-scale protein sequence comparison and gene matching to identify orthologs and co-orthologs.** *Nucleic Acids Res* 2012, **40**:e44.

5. Blaisdell BE: **A measure of the similarity of sets of sequences not requiring sequence alignment.** *Proc Natl Acad Sci U S A* 1986, **83**:5155–9.

6. Pevzner P a: **Statistical distance between texts and filtration methods in sequence comparison.** *Comput Appl Biosci* 1992, **8**:121–7.

7. Mantaci S, Restivo A, Sciortino M: **Distance measures for biological sequences: Some recent approaches**. *Int J Approx Reason* 2008, **47**:109–124.

8. Giancarlo R, Scaturro D, Utro F: **Textual data compression in computational biology: a synopsis.** *Bioinformatics* 2009, **25**:1575–86.

9. Jing J, Wilson S, Burden C: **Weighted k-word matches: a sequence comparison tool for proteins**. *ANZIAM J* 2011, **52**.

10. Feng J, Zhao W, Zhang H: **Use of an Alignment-free Method to Compare Reduced Amino Acid Sequences**. *J Converg Inf Technol* 2011, **6**:213–221.

11. Huang G, Zhou H, Li Y, Xu L: **Alignment-free comparison of genome sequences by a new numerical characterization.** *J Theor Biol* 2011, **281**:107–12.

12. Yu C, Cheng S, He RL, Yau SS: **Protein map : An alignment-free sequence comparison method based on various properties of amino acids**. *Gene* 2011, **486**:110–118.

13. Comin M, Verzotto D: **Alignment-free phylogeny of whole genomes using underlying subwords.** *Algorithms Mol Biol* 2012, **7**:34.

14. Soares I, Goios A, Amorim A: **Sequence comparison alignment-free approach based on suffix tree and L-words frequency.** *ScientificWorldJournal* 2012, **2012**:450124.

15. Kurtz S, Narechania A, Stein JC, Ware D: **A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes.** *BMC Genomics* 2008, **9**:517.

16. Kantorovitz MR, Robinson GE, Sinha S: **A statistical method for alignment-free comparison of regulatory sequences.** *Bioinformatics* 2007, **23**:i249–55.

17. Huson D, Xie C: **A poor man's BLASTX-high-throughput metagenomic protein database search using PAUDA**. *Bioinformatics* 2013:1–2.

18. Edgar RC: **Search and clustering orders of magnitude faster than BLAST.** *Bioinformatics* 2010, **26**:2460–1.

19. Barbosa-Silva A, Satagopam VP, Schneider R, Ortega JM: **Clustering of cognate proteins among distinct proteomes derived from multiple links to a single seed sequence.** *BMC Bioinformatics* 2008, **9**:141.

20. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147**:195–7.

21. Souza JA de: **Reconhecimento de padrões usando indexação recursiva**. UNIVERSIDADE FEDERAL DE SANTA CATARINA; 1999.

22. MathWorks: **MATLAB**. .

23. MathWorks: **Bioinformatics Toolbox**. .

24. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR, Bateman A: **The Pfam protein families database.** *Nucleic Acids Res* 2010, **38**(Database issue):D211–22.

25. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH: **UniRef: comprehensive and non-redundant UniProt reference clusters.** *Bioinformatics* 2007, **23**:1282–8.

26. De Souza V, Piro VC, Faoro H, Tadra-Sfeir MZ, Chicora VK, Guizelini D, Weiss V, Vialle RA, Monteiro RA, Steffens MBR, Marchaukoski JN, Pedrosa FO, Cruz LM, Chubatsu LS, Raittz RT: **Draft Genome Sequence of Herbaspirillum huttiense subsp. putei IAM 15032, a Strain Isolated from Well Water**. *Genome Announc* 2013, **1**:e00252–12–e00252–12.

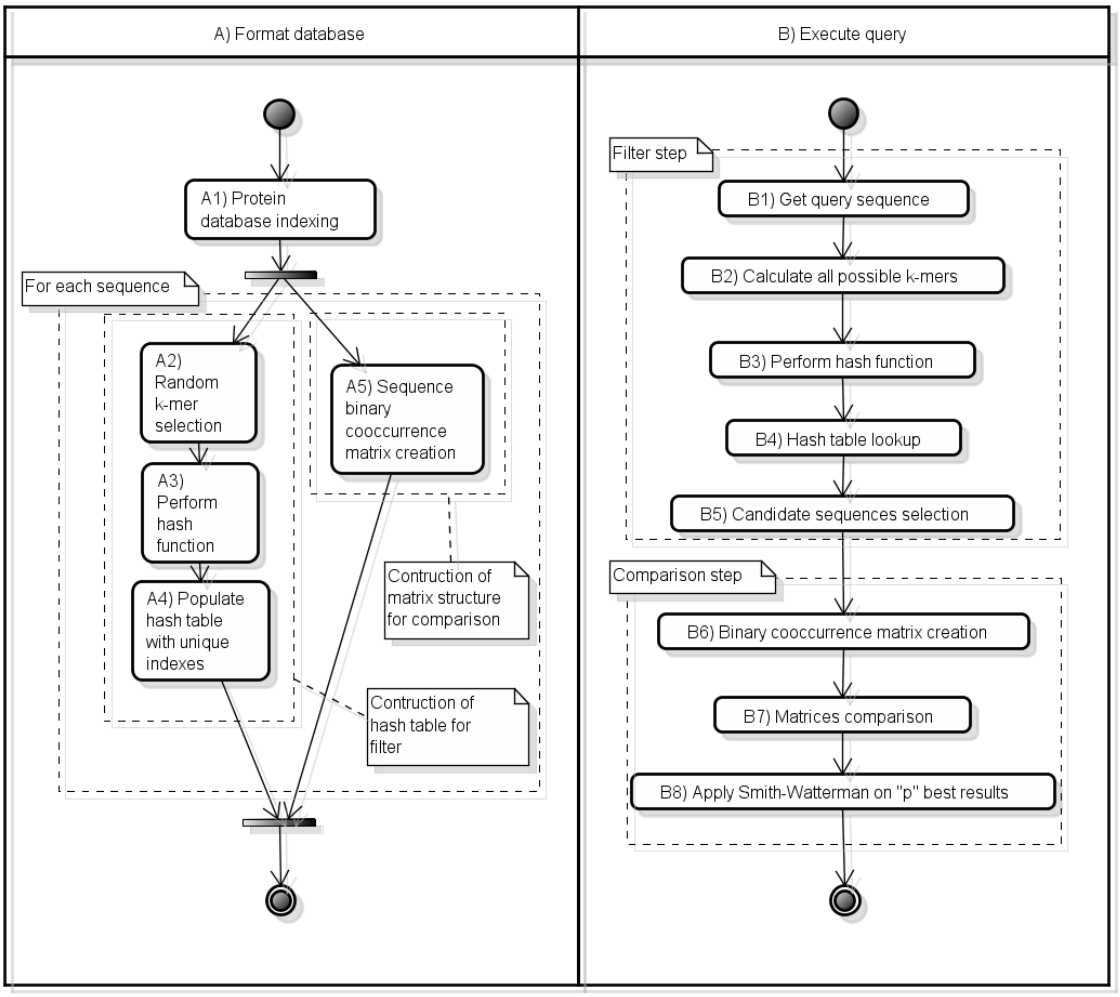
27. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: prokaryotic gene recognition and translation initiation site identification.** *BMC Bioinformatics* 2010, **11**:119.

28. Pedrosa FO, Monteiro RA, Wassem R, Cruz LM, Ayub R a, Colauto NB, Fernandez MA, Fungaro MHP, Grisard EC, Hungria M, Madeira HMF, Nodari RO, Osaku C a, Petzl-Erler ML, Terenzi H, Vieira LGE, Steffens MBR, Weiss V a, Pereira LFP, Almeida MIM, Alves LR, Marin A, Araujo LM, Balsanelli E, Baura V a, Chubatsu LS, Faoro H, Favetti A, Friedermann G, Glienke C, et al.: **Genome of Herbaspirillum seropedicae strain SmR1, a specialized diazotrophic endophyte of tropical grasses.** *PLoS Genet* 2011, **7**:e1002064.

29. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin E V, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov A V, Vasudevan S, Wolf YI, Yin JJ, Natale D a: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4**:41.

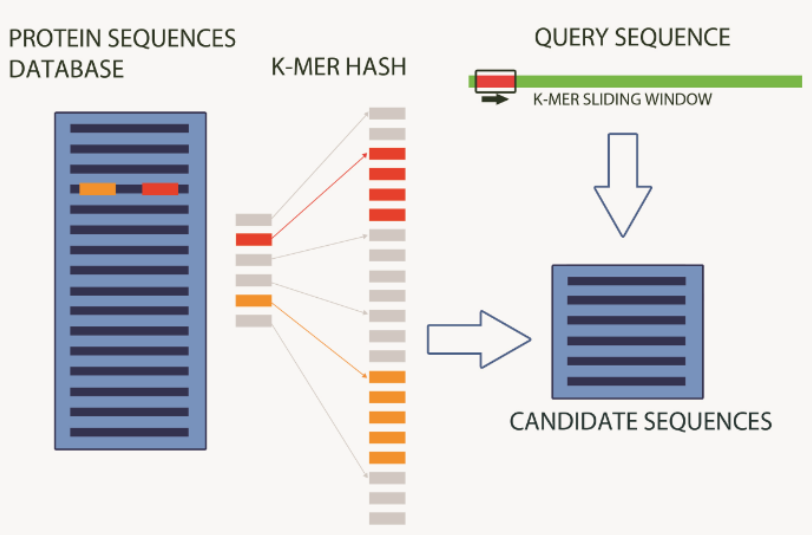
FIGURES

Figure 1 - RAFTS^3 activity diagram



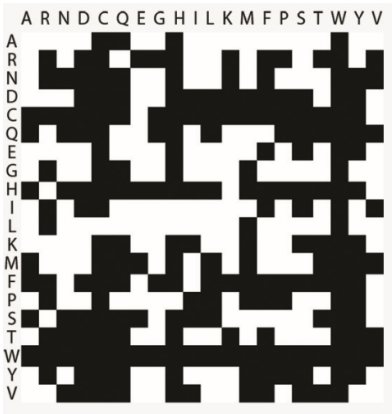
RAFTS^3 format database and query search overview. A) Shows the database formatting processes, which involves construction of two structures used in query sequence search, a hash table and a set of binary co-occurrence matrices. B) Shows the process for search and comparison of a query sequence, with filtering and comparison steps separated.

Figure 2 - Candidate selection



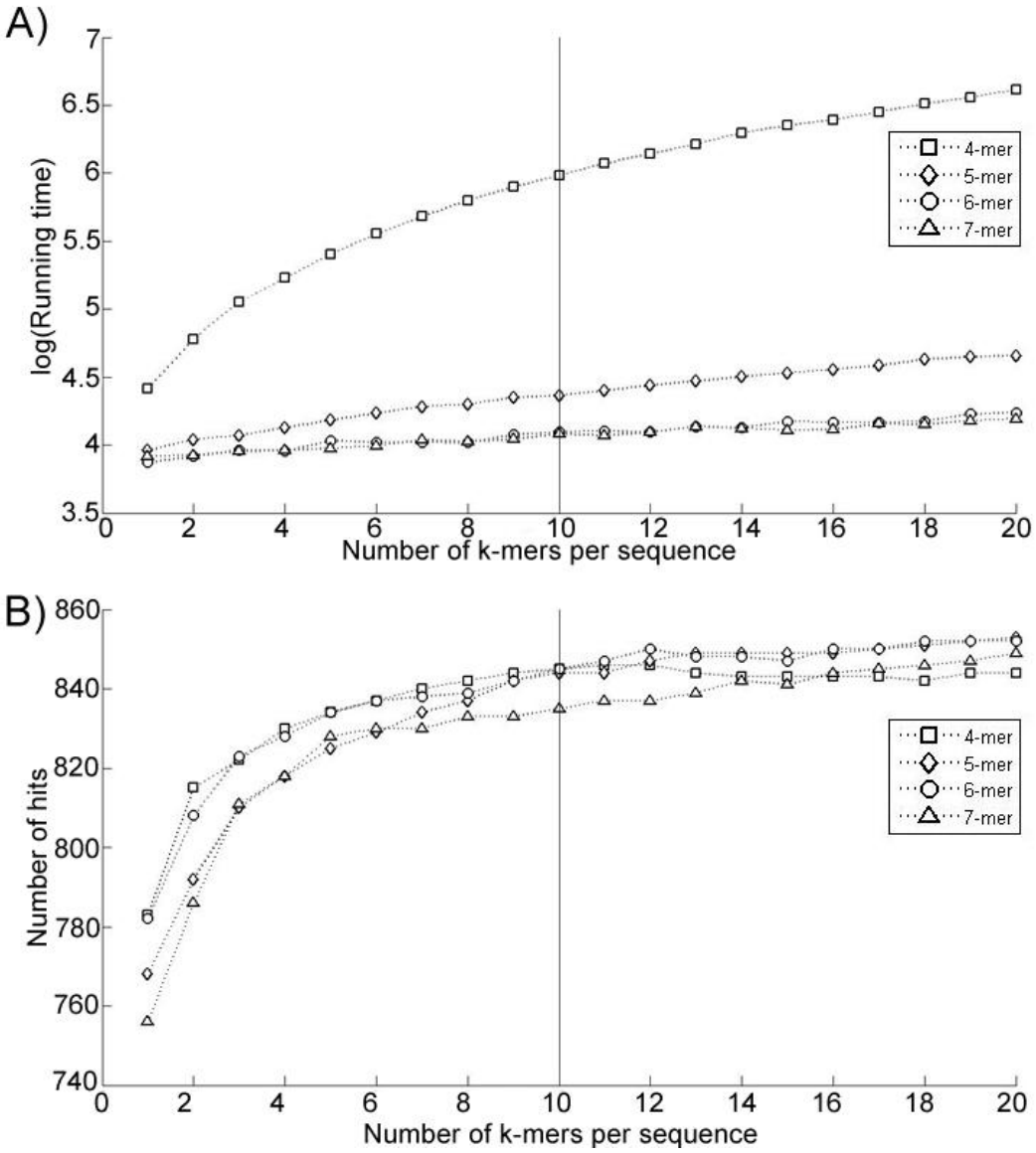
Hashing approach for candidate sequences selection.

Figure 3 - Binary co-occurrence matrix (BCOM)



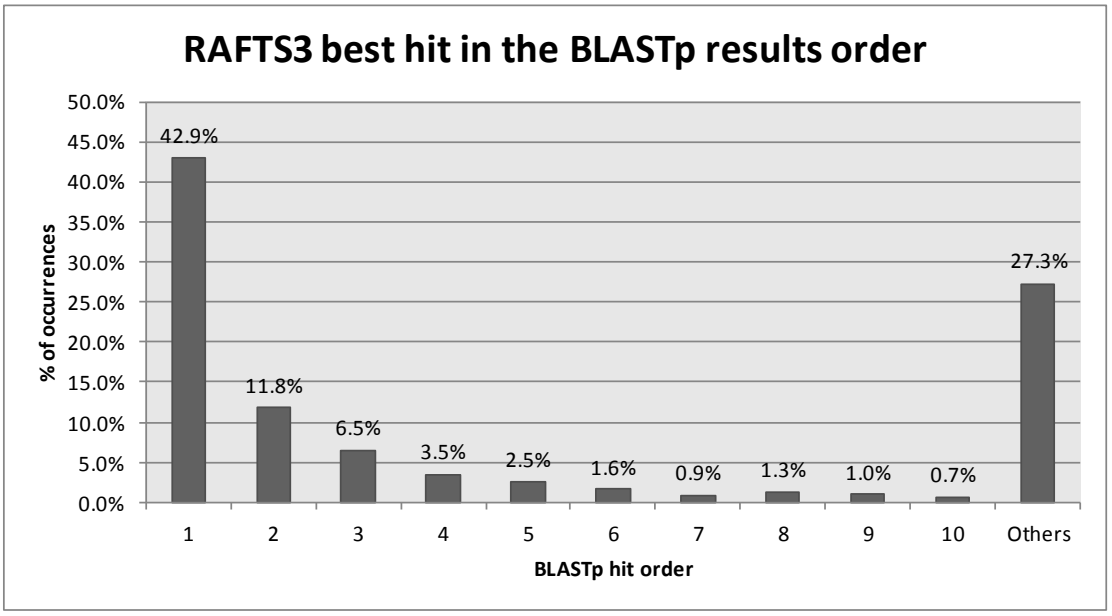
Co-occurrence matrix of a protein sequence. White dots represent the occurrence of the amino acid pair, the black dots represent the nonoccurrence.

Figure 4 - Parameters selection configuration tests



Comparison of different k-mer sets. The comparison is made analyzing the second best hit on the search of 1000 sequences randomly selected with sets of 1 to 20 k-mer lengths of 4, 5, 6 and 7 amino acid residues. A) Shows the logarithm of the running time to search 1000 queries for each configuration. B) Shows the number of queries with second best hit with relative score over 0.3.

Figure 5 - Percentage of occurrences of RAFTS3 best hits by BLASTp results position



The graphic shows the rate of occurrences of the RAFTS^3 best hits by the BLASTp results order position for 1000 sequences randomly selected from a newer NR.

TABLES

Table 1 - Amino acid numeric conversion

Amino acid	Code
A	3 2 1
R	1 3 1
N	1 1 2
D	3 1 2
C	4 3 2
Q	2 1 1
E	3 1 1
G	3 3 4
H	2 1 4
I	1 4 1
L	2 4 2
K	1 1 1
M	1 4 3
F	4 4 2
P	2 2 4
S	4 2 1
T	1 2 4
W	4 3 3
Y	4 1 2
V	3 4 2

Two-way conversion table of amino acid residues to quaternary numeral system triplets.

Table 2. Performance comparison of similarity search tools on the same query dataset (1323 ORFs) against different protein similarity search databases

Database	Total sequences	Total aa	Running time		Percentage of sequences over relative score threshold found by RAFTS^3 compared with BLAST						
			BLAST	RAFTS^3	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Pfam	15,929,002	5,169,768,107	26,323 s	66 s	89%	93%	95%	96%	97%	97%	98%
NR	19,689,576	6,752,058,980	35,668 s	68 s	87%	90%	92%	94%	96%	98%	98%

The total number of sequences and amino acids included in each database are shown in the “Total sequences” and “Total aa” columns, respectively. The relation RAFTS^3/BLAST stands by the percentage of the amount of sequences found by RAFTS^3 over those found with BLAST by each relative score threshold.

Table 3 - Performance comparison of similarity search tools on the same query dataset (1000 sequences) against different protein similarity search databases

Database	Total sequences	Total aa	Running time		Percentage of sequences over relative score threshold found by RAFTS^3 compared with BLAST						
			BLAST	RAFTS^3	0.3	0.4	0.5	0.6	0.7	0.8	0.9
UniRef50	6,784,251	2,189,361,886	10,128 s	46 s	81%	80%	77%	80%	87%	92%	95%
NR	19,689,576	6,752,058,980	28,351 s	52 s	88%	92%	93%	94%	95%	96%	95%

The total number of sequences and amino acids included in each database are shown in the “Total sequences” and “Total aa” columns, respectively. The relation RAFTS^3/BLAST stands by the percentage of the amount of sequences found by RAFTS^3 over those found with BLAST by each relative score threshold.

Table 4. Comparison of top 4 results of BLAST and RAFTS3 for genes of Hebaspirillum seropedicae SmR1

RAFTS^3			BLAST		
Relative Score	**E-Value**	**Subject Sequence**	**Relative Score**	**E-value**	**Subject Sequence**
Query - NCBI Reference Sequence: YP_003776253.1					
1	0	dinitrogenase reductase [*Herbaspirillum seropedicae* SmR1]	1	0	dinitrogenase reductase [*Herbaspirillum seropedicae* SmR1]
0.91	0	nitrogenase iron protein [*Burkholderiales bacterium* JOSHI_001]	0.91	0	nitrogenase iron protein [*Burkholderiales bacterium* JOSHI_001]
0.88	0	nitrogenase iron protein [*Dechlorosoma suillum* PS]	0.90	0	nitrogenase reductase [*Leptothrix cholodnii* SP-6]
0.87	0	nitrogenase iron protein [*Azoarcus sp.* KH32C]	0.90	0	nitrogenase reductase [*Methylococcus capsulatus* str. *Bath*]

RAFTS^3			BLAST		
Relative Score	**E-Value**	**Subject Sequence**	**Relative Score**	**E-value**	**Subject Sequence**
Query - NCBI Reference Sequence: YP_003773449.1					
0.99	0	major facilitator superfamily (MFS) transporter protein [*Herbaspirillum seropedicae* SmR1]	0.99	0	major facilitator superfamily (MFS) transporter protein [*Herbaspirillum seropedicae* SmR1]
0.72	0	hypothetical protein PA2G_05719 [*Pseudomonas aeruginosa* 2192]	0.72	0	major facilitator superfamily [*Burkholderia multivorans* CGD2M]
0.72	0	hypothetical protein PACG_05164 [*Pseudomonas aeruginosa* C3719]	0.71	0	major facilitator transporter [*Burkholderia multivorans* ATCC 17616]
0.68	0	major facilitator transporter [*Pseudomonas putida* DOT-T1E]	0.71	0	major facilitator family transporter [*Burkholderia multivorans* CGD1]

RAFTS^3			BLAST		
Relative Score	**E-Value**	**Subject Sequence**	**Relative Score**	**E-value**	**Subject Sequence**
Query - NCBI Reference Sequence: YP_003773452.1					
1	0	hypothetical protein Hsero_0014 [*Herbaspirillum seropedicae* SmR1]	1	0	hypothetical protein Hsero_0014 [*Herbaspirillum seropedicae* SmR1]
0.25	1.00E-113	diguanylate cyclase/phosphodiesterase with extracellular sensor [*Thioalkalimicrobium aerophilum* AL3]	0.36	4.00E-157	PAS/PAC sensor-containing diguanylate cyclase/phosphodiesterase [*Leptothrix cholodnii* SP-6]
0.24	4.00E-101	PAS:GGDEF protein [*Oceanospirillum* sp. MED92]	0.29	4.00E-126	hypothetical protein SCD_02084 [*Sulfuricella denitrificans* skB26]
0.19	1.00E-81	diguanylate cyclase (GGDEF) domain-containing protein [*Bradyrhizobium* sp. WSM1253]	0.27	4.00E-124	diguanylate cyclase/phosphodiesterase with PAS/PAC sensor(s) [*Sideroxydans lithotrophicus* ES-1]

The query sequences are indicated by their accession number on NCBI. The subject sequences are ordered by the criteria of relative score on RAFTS^3 and E-value on BLAST.

Table 5 - Comparison of the number of occurrences of where the best hit of each tool occur in the Top 50 of the other

	BLAST best hits		RAFTS^3 best hits	
Rank position	Number of hits	Mean of scores	Number of hits	Mean of scores
1	429	0.89	429	0.89
2	105	0.91	118	0.86
3	44	0.83	65	0.84
4	34	0.90	35	0.83
5	25	0.86	25	0.81
6	14	0.83	16	0.74
7	10	0.89	9	0.84
8	9	0.88	13	0.78
9	8	0.87	10	0.81
10	8	0.86	7	0.66
11	3	0.77	2	0.72
12	2	0.95	5	0.57
13	3	0.74	6	0.71
14	8	0.75	8	0.87
15	1	0.77	6	0.70
16	5	0.89	3	0.47
17	2	0.91	2	0.63
18	1	0.94	2	0.63
19	3	0.60	3	0.69
20	1	1.00	1	0.59
21	1	0.81	2	0.42
22	1	0.99	2	0.68
23	2	0.79	0	0.00
24	0	0.00	2	0.87
25	1	1.00	3	0.91
26	0	0.00	1	0.97
27	1	0.98	2	0.36
28	2	0.73	1	0.99
29	0	0.00	1	0.99
30	0	0.00	2	0.74
31	0	0.00	3	0.46
32	1	0.80	1	0.96
33	0	0.00	2	0.75
34	0	0.00	1	0.57
35	0	0.00	1	0.96
36	3	0.97	0	0.00
37	0	0.00	2	0.96
38	0	0.00	0	0.00
39	1	0.95	0	0.00
40	1	0.98	0	0.00
41	0	0.00	0	0.00
42	0	0.00	2	0.78

43	0	0.00	0	0.00
44	0	0.00	4	0.68
45	0	0.00	0	0.00
46	0	0.00	1	0.53
47	0	0.00	1	0.92
48	0	0.00	0	0.00
49	0	0.00	1	0.53
50	0	0.00	0	0.00

The BLAST best hits columns presents where and how many of the 1000 BLASTp best hits against the NR database happen in RAFTS^3 Top 50 result report and their mean relative score. In the same way, the columns RAFTS^3 best hits stands for where and how many RAFTS^3 best hits happen in BLASTp Top 50 results.

Table 6 - Comparison of Top 50 results of BLAST and RAFTS^3 for 5 random genes

Query gene	Rank	BLASTp		RAFTS^3	
		Score	GI	Score	GI
hypothetical protein [Nocardiopsis prasina]	1	0.72	54026316	0.68	108799494
hypothetical protein [Nocardiopsis prasina]	2	0.71	300786153	0.68	126435148
hypothetical protein [Nocardiopsis prasina]	3	0.68	126435148	0.72	54026316
hypothetical protein [Nocardiopsis prasina]	4	0.68	379709308	0.71	386772711
hypothetical protein [Nocardiopsis prasina]	5	0.68	108799494	0.70	379736307
hypothetical protein [Nocardiopsis prasina]	6	0.71	386772711	0.70	319949673
hypothetical protein [Nocardiopsis prasina]	7	0.68	363419515	0.71	300786153
hypothetical protein [Nocardiopsis prasina]	8	0.67	312139946	0.67	379748180
hypothetical protein [Nocardiopsis prasina]	9	0.70	325964444	0.66	379755468
hypothetical protein [Nocardiopsis prasina]	10	0.70	319949673	0.70	325964444
hypothetical protein [Nocardiopsis prasina]	11	0.66	148271386	0.67	254821758
hypothetical protein [Nocardiopsis prasina]	12	0.66	342858396	0.66	379763014
hypothetical protein [Nocardiopsis prasina]	13	0.70	379736307	0.68	363419515
hypothetical protein [Nocardiopsis prasina]	14	0.69	302526736	0.68	379709308
hypothetical protein [Nocardiopsis prasina]	15	0.65	239985975	0.69	302526736
hypothetical protein [Nocardiopsis prasina]	16	0.67	378816503	0.67	378816503
hypothetical protein [Nocardiopsis prasina]	17	0.66	379755468	0.67	397679577
hypothetical protein [Nocardiopsis prasina]	18	0.64	120405695	0.67	392136326
hypothetical protein [Nocardiopsis prasina]	19	0.66	379763014	0.66	148271386
hypothetical protein [Nocardiopsis prasina]	20	0.68	382944866	0.61	359774163
hypothetical protein [Nocardiopsis prasina]	21	0.68	382944705	0.60	385651505
hypothetical protein [Nocardiopsis prasina]	22	0.68	392068192	0.68	382944866
hypothetical protein [Nocardiopsis prasina]	23	0.68	392185753	0.68	382944705
hypothetical protein [Nocardiopsis prasina]	24	0.63	182440470	0.67	312139946
hypothetical protein [Nocardiopsis prasina]	25	0.67	363999376	0.69	374610654
hypothetical protein [Nocardiopsis prasina]	26	0.67	392136326	0.64	333989082
hypothetical protein [Nocardiopsis prasina]	27	0.67	397679577	0.43	377569218
hypothetical protein [Nocardiopsis prasina]	28	0.63	311741875	0.63	311741875

hypothetical protein [Nocardiopsis prasina]	29	0.67	379748180	0.68	392068192
hypothetical protein [Nocardiopsis prasina]	30	0.67	254821758	0.67	325674171
hypothetical protein [Nocardiopsis prasina]	31	0.67	325674171	0.54	392847751
hypothetical protein [Nocardiopsis prasina]	32	0.65	296164157	0.62	118472651
hypothetical protein [Nocardiopsis prasina]	33	0.61	359774163	0.55	365870395
hypothetical protein [Nocardiopsis prasina]	34	0.59	333022807	0.55	392086623
hypothetical protein [Nocardiopsis prasina]	35	0.59	328880653	0.55	358003015
hypothetical protein [Nocardiopsis prasina]	36	0.53	331696237	0.55	386691466
hypothetical protein [Nocardiopsis prasina]	37	0.65	145222559	0.53	163857205
hypothetical protein [Nocardiopsis prasina]	38	0.68	375137662	0.50	344998040
hypothetical protein [Nocardiopsis prasina]	39	0.57	326330248	0.65	145222559
hypothetical protein [Nocardiopsis prasina]	40	0.58	354570978	0.05	395776055
hypothetical protein [Nocardiopsis prasina]	41	0.62	302523345	0.04	386354368
hypothetical protein [Nocardiopsis prasina]	42	0.58	392383971	0.05	375143224
hypothetical protein [Nocardiopsis prasina]	43	0.63	262204158	0.05	83716891
hypothetical protein [Nocardiopsis prasina]	44	0.57	337267501	0.05	257068192
hypothetical protein [Nocardiopsis prasina]	45	0.57	392849102	0.05	167577907
hypothetical protein [Nocardiopsis prasina]	46	0.56	387970235	0.05	241206048
hypothetical protein [Nocardiopsis prasina]	47	0.56	392520598	0.06	218893839
hypothetical protein [Nocardiopsis prasina]	48	0.56	393170990	0.06	373478940
hypothetical protein [Nocardiopsis prasina]	49	0.56	397688108	0.05	359149253
hypothetical protein [Nocardiopsis prasina]	50	0.55	358003015	0.05	168009884
transposase [Bacillus cereus]	1	0.93	206973981	0.91	218896864
transposase [Bacillus cereus]	2	0.92	75763559	0.86	229085868
transposase [Bacillus cereus]	3	0.91	218896864	0.92	229090883
transposase [Bacillus cereus]	4	0.92	391290908	0.93	206973981
transposase [Bacillus cereus]	5	0.92	206973765	0.91	196038692
transposase [Bacillus cereus]	6	0.92	229090883	0.91	196038713
transposase [Bacillus cereus]	7	0.91	196038692	0.92	75763559
transposase [Bacillus cereus]	8	0.91	196038713	0.92	206973765
transposase [Bacillus cereus]	9	0.90	196042445	0.90	196042445
transposase [Bacillus cereus]	10	0.90	75760431	0.83	222094044
transposase [Bacillus cereus]	11	0.88	206973407	0.88	206973407
transposase [Bacillus cereus]	12	0.86	229085868	0.92	391290908
transposase [Bacillus cereus]	13	0.84	228905309	0.84	228905309
transposase [Bacillus cereus]	14	0.83	222094044	0.82	229073617
transposase [Bacillus cereus]	15	0.83	229172505	0.82	229095554
transposase [Bacillus cereus]	16	0.82	229095554	0.83	229172505
transposase [Bacillus cereus]	17	0.82	229073617	0.90	75760431
transposase [Bacillus cereus]	18	0.80	229182210	0.80	229182210
transposase [Bacillus cereus]	19	0.66	225871669	0.55	75762371
transposase [Bacillus cereus]	20	0.65	227811612	0.52	228911455
transposase [Bacillus cereus]	21	0.65	254762474	0.65	227811612
transposase [Bacillus cereus]	22	0.57	301068226	0.66	225871669
transposase [Bacillus cereus]	23	0.55	75762371	0.65	254762474
transposase [Bacillus cereus]	24	0.56	228970158	0.47	75764333
transposase [Bacillus cereus]	25	0.52	228911455	0.57	301068226

transposase [Bacillus cereus]	26	0.55	229106961	0.46	75759724
transposase [Bacillus cereus]	27	0.54	229119272	0.44	75763688
transposase [Bacillus cereus]	28	0.54	206973911	0.56	228970158
transposase [Bacillus cereus]	29	0.54	221642251	0.47	10956343
transposase [Bacillus cereus]	30	0.50	228924890	0.47	301068223
transposase [Bacillus cereus]	31	0.48	228906894	0.47	165873444
transposase [Bacillus cereus]	32	0.47	75764333	0.47	254739166
transposase [Bacillus cereus]	33	0.46	75759724	0.04	371777874
transposase [Bacillus cereus]	34	0.44	75763688	0.04	163786962
transposase [Bacillus cereus]	35	0.47	47568876	0.03	229580250
transposase [Bacillus cereus]	36	0.47	301068223	0.03	399003489
transposase [Bacillus cereus]	37	0.47	10956343	0.03	336315012
transposase [Bacillus cereus]	38	0.47	165873444	0.03	397602092
transposase [Bacillus cereus]	39	0.47	254739166	0.03	328859802
transposase [Bacillus cereus]	40	0.42	10956376	0.04	355670767
transposase [Bacillus cereus]	41	0.45	23099091	0.03	255954341
transposase [Bacillus cereus]	42	0.40	254687682	0.04	390944355
transposase [Bacillus cereus]	43	0.44	383438775	0.03	375163697
transposase [Bacillus cereus]	44	0.44	386712685	0.03	284035464
transposase [Bacillus cereus]	45	0.44	52078969	0.03	325912143
transposase [Bacillus cereus]	46	0.44	383439073	0.03	218192290
transposase [Bacillus cereus]	47	0.41	261409030	0.03	301776841
transposase [Bacillus cereus]	48	0.41	315647012	0.04	340380971
transposase [Bacillus cereus]	49	0.40	261409860	0.03	296278265
transposase [Bacillus cereus]	50	0.41	372455285	0.03	12697963
iroE [Klebsiella pneumoniae]	1	1.00	262044290	1.00	262044290
iroE [Klebsiella pneumoniae]	2	0.99	397743876	0.99	238894719
iroE [Klebsiella pneumoniae]	3	0.99	386034811	0.99	397743876
iroE [Klebsiella pneumoniae]	4	0.99	397345874	0.99	386034811
iroE [Klebsiella pneumoniae]	5	0.99	152970230	0.99	152970230
iroE [Klebsiella pneumoniae]	6	0.99	397446209	0.99	365141280
iroE [Klebsiella pneumoniae]	7	0.99	238894719	0.99	397345874
iroE [Klebsiella pneumoniae]	8	0.99	365141280	0.99	397446209
iroE [Klebsiella pneumoniae]	9	0.87	288935507	0.87	290509545
iroE [Klebsiella pneumoniae]	10	0.87	290509545	0.86	206579000
iroE [Klebsiella pneumoniae]	11	0.86	206579000	0.87	288935507
iroE [Klebsiella pneumoniae]	12	0.52	378978774	0.49	375002495
iroE [Klebsiella pneumoniae]	13	0.53	376400045	0.48	353606873
iroE [Klebsiella pneumoniae]	14	0.52	375261636	0.49	366059620
iroE [Klebsiella pneumoniae]	15	0.52	397658746	0.48	16761559
iroE [Klebsiella pneumoniae]	16	0.52	376395752	0.51	261341410
iroE [Klebsiella pneumoniae]	17	0.51	261341410	0.49	204929666
iroE [Klebsiella pneumoniae]	18	0.51	376385392	0.49	353661331
iroE [Klebsiella pneumoniae]	19	0.51	157145930	0.48	363551476
iroE [Klebsiella pneumoniae]	20	0.50	295096492	0.49	238909547
iroE [Klebsiella pneumoniae]	21	0.50	376383330	0.50	295096492
iroE [Klebsiella pneumoniae]	22	0.50	376383956	0.49	168238684

iroE [Klebsiella pneumoniae]	23	0.49	397167683	0.47	353596784
iroE [Klebsiella pneumoniae]	24	0.47	317053692	0.49	366083251
iroE [Klebsiella pneumoniae]	25	0.49	366059620	0.47	392819761
iroE [Klebsiella pneumoniae]	26	0.49	322614400	0.48	168823183
iroE [Klebsiella pneumoniae]	27	0.48	365850221	0.48	205353732
iroE [Klebsiella pneumoniae]	28	0.49	366083251	0.03	264676321
iroE [Klebsiella pneumoniae]	29	0.49	353661331	0.49	197249038
iroE [Klebsiella pneumoniae]	30	0.48	336247161	0.03	237799500
iroE [Klebsiella pneumoniae]	31	0.49	168262124	0.50	376383956
iroE [Klebsiella pneumoniae]	32	0.49	168233729	0.04	330817286
iroE [Klebsiella pneumoniae]	33	0.49	375002495	0.04	294010958
iroE [Klebsiella pneumoniae]	34	0.49	204929666	0.50	157418222
iroE [Klebsiella pneumoniae]	35	0.49	238909547	0.43	213424548
iroE [Klebsiella pneumoniae]	36	0.49	353617579	0.04	383777638
iroE [Klebsiella pneumoniae]	37	0.49	197249038	0.49	386598810
iroE [Klebsiella pneumoniae]	38	0.49	168238684	0.03	325271609
iroE [Klebsiella pneumoniae]	39	0.48	168823183	0.03	386818396
iroE [Klebsiella pneumoniae]	40	0.48	56414734	0.03	188534867
iroE [Klebsiella pneumoniae]	41	0.50	338767125	0.03	385653217
iroE [Klebsiella pneumoniae]	42	0.50	222104800	0.03	167829434
iroE [Klebsiella pneumoniae]	43	0.48	379050444	0.03	303327182
iroE [Klebsiella pneumoniae]	44	0.48	353606873	0.03	328770351
iroE [Klebsiella pneumoniae]	45	0.48	16761559	0.04	87309978
iroE [Klebsiella pneumoniae]	46	0.48	205353732	0.05	344172297
iroE [Klebsiella pneumoniae]	47	0.48	375124587	0.05	323358023
iroE [Klebsiella pneumoniae]	48	0.48	363551476	0.04	254413004
iroE [Klebsiella pneumoniae]	49	0.48	353570067	0.04	227875198
iroE [Klebsiella pneumoniae]	50	0.48	198243714	0.04	152968329
PREDICTED: secernin-3 [Myotis lucifugus]	1	0.93	301769725	0.91	344268354
PREDICTED: secernin-3 [Myotis lucifugus]	2	0.93	281348304	0.92	194222320
PREDICTED: secernin-3 [Myotis lucifugus]	3	0.92	194222320	0.91	296490700
PREDICTED: secernin-3 [Myotis lucifugus]	4	0.93	57110813	0.91	115495695
PREDICTED: secernin-3 [Myotis lucifugus]	5	0.91	296490700	0.93	281348304
PREDICTED: secernin-3 [Myotis lucifugus]	6	0.91	115495695	0.93	301769725
PREDICTED: secernin-3 [Myotis lucifugus]	7	0.91	344268354	0.93	57110813
PREDICTED: secernin-3 [Myotis lucifugus]	8	0.90	109100111	0.90	109100111
PREDICTED: secernin-3 [Myotis lucifugus]	9	0.89	296204486	0.89	380791835
PREDICTED: secernin-3 [Myotis lucifugus]	10	0.89	38504671	0.89	38504671
PREDICTED: secernin-3 [Myotis lucifugus]	11	0.89	380791835	0.89	111601409
PREDICTED: secernin-3 [Myotis lucifugus]	12	0.89	332209358	0.87	395837305
PREDICTED: secernin-3 [Myotis lucifugus]	13	0.89	397507616	0.87	348585759
PREDICTED: secernin-3 [Myotis lucifugus]	14	0.89	291391765	0.89	114581821
PREDICTED: secernin-3 [Myotis lucifugus]	15	0.89	111601409	0.89	332209358
PREDICTED: secernin-3 [Myotis lucifugus]	16	0.89	114581821	0.87	351715131
PREDICTED: secernin-3 [Myotis lucifugus]	17	0.87	395837305	0.89	397507616
PREDICTED: secernin-3 [Myotis lucifugus]	18	0.87	351715131	0.89	296204486
PREDICTED: secernin-3 [Myotis lucifugus]	19	0.87	348585759	0.89	291391765

PREDICTED: secernin-3 [Myotis lucifugus]	20	0.84	297264350	0.84	302058287
PREDICTED: secernin-3 [Myotis lucifugus]	21	0.84	302058287	0.84	332209360
PREDICTED: secernin-3 [Myotis lucifugus]	22	0.84	397507618	0.84	297264350
PREDICTED: secernin-3 [Myotis lucifugus]	23	0.84	332209360	0.83	332814760
PREDICTED: secernin-3 [Myotis lucifugus]	24	0.83	354505419	0.84	397507618
PREDICTED: secernin-3 [Myotis lucifugus]	25	0.83	332814760	0.83	354505419
PREDICTED: secernin-3 [Myotis lucifugus]	26	0.80	61969660	0.80	61969660
PREDICTED: secernin-3 [Myotis lucifugus]	27	0.80	149022245	0.80	15929748
PREDICTED: secernin-3 [Myotis lucifugus]	28	0.80	15929748	0.72	119631548
PREDICTED: secernin-3 [Myotis lucifugus]	29	0.80	74153182	0.80	74153182
PREDICTED: secernin-3 [Myotis lucifugus]	30	0.77	395519795	0.77	395519795
PREDICTED: secernin-3 [Myotis lucifugus]	31	0.74	383087724	0.80	149022245
PREDICTED: secernin-3 [Myotis lucifugus]	32	0.75	126326616	0.75	149639530
PREDICTED: secernin-3 [Myotis lucifugus]	33	0.74	327283502	0.57	62914002
PREDICTED: secernin-3 [Myotis lucifugus]	34	0.75	149639530	0.57	148695178
PREDICTED: secernin-3 [Myotis lucifugus]	35	0.72	119631548	0.74	327283502
PREDICTED: secernin-3 [Myotis lucifugus]	36	0.74	224055113	0.63	301610221
PREDICTED: secernin-3 [Myotis lucifugus]	37	0.65	41054327	0.64	348519679
PREDICTED: secernin-3 [Myotis lucifugus]	38	0.64	163914461	0.64	163914461
PREDICTED: secernin-3 [Myotis lucifugus]	39	0.63	301610221	0.02	358391805
PREDICTED: secernin-3 [Myotis lucifugus]	40	0.63	94482839	0.02	390350007
PREDICTED: secernin-3 [Myotis lucifugus]	41	0.64	348519679	0.02	302916981
PREDICTED: secernin-3 [Myotis lucifugus]	42	0.62	61557143	0.03	340372503
PREDICTED: secernin-3 [Myotis lucifugus]	43	0.57	62914002	0.03	325145043
PREDICTED: secernin-3 [Myotis lucifugus]	44	0.58	47218098	0.02	358366518
PREDICTED: secernin-3 [Myotis lucifugus]	45	0.57	148695178	0.03	326917505
PREDICTED: secernin-3 [Myotis lucifugus]	46	0.56	311272660	0.02	322710652
PREDICTED: secernin-3 [Myotis lucifugus]	47	0.55	395826584	0.03	342874004
PREDICTED: secernin-3 [Myotis lucifugus]	48	0.55	327275780	0.02	147864006
PREDICTED: secernin-3 [Myotis lucifugus]	49	0.56	126308319	0.02	308469783
PREDICTED: secernin-3 [Myotis lucifugus]	50	0.55	326934081	0.02	345487083
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	1	0.86	395824169	0.95	325495571
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	2	0.88	160221327	0.88	160221327
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	3	0.87	47523442	0.88	27806027
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	4	0.87	344271937	0.87	47523442
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	5	0.88	27806027	0.87	344271937
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	6	0.85	300797824	0.86	395824169
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	7	0.95	325495571	0.85	332229985
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	8	0.85	149047896	0.84	20070193
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	9	0.85	20522231	0.85	300797824
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	10	0.85	74142710	0.84	297685326
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	11	0.84	354499096	0.83	10945629
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	12	0.83	10945629	0.84	109110256
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	13	0.84	1805353	0.84	216409744
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	14	0.85	332229985	0.86	351702108
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	15	0.87	126352395	0.85	301769265
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	16	0.84	109110256	0.84	384940122

PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	17	0.85	397473205	0.84	2077920
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	18	0.84	384940122	0.87	126352395
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	19	0.84	20070193	0.85	20522231
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	20	0.84	297685326	0.85	74142710
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	21	0.84	297270159	0.84	1805353
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	22	0.84	2077920	0.76	355567920
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	23	0.84	216409744	0.84	354499096
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	24	0.85	348570102	0.85	348570102
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	25	0.85	301769265	0.85	149047896
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	26	0.86	351702108	0.84	297270159
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	27	0.83	345805854	0.85	397473205
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	28	0.76	355567920	0.83	345805854
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	29	0.71	49036491	0.66	334311611
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	30	0.76	332832881	0.71	49036491
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	31	0.66	334311611	0.76	332832881
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	32	0.60	4586618	0.68	296190799
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	33	0.59	115334528	0.57	395505691
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	34	0.59	45384188	0.59	115334528
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	35	0.58	115529250	0.60	4586618
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	36	0.59	168479587	0.59	45384188
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	37	0.56	4104218	0.59	168479587
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	38	0.54	291565556	0.56	345326142
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	39	0.54	4126870	0.56	4104218
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	40	0.56	345326142	0.58	115529250
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	41	0.54	224809509	0.46	327290547
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	42	0.54	148224522	0.44	24158439
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	43	0.68	296190799	0.45	66356139
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	44	0.53	44355486	0.46	15145791
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	45	0.57	395505691	0.53	44355486
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	46	0.68	149047895	0.47	1947098
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	47	0.68	425578	0.47	281351212
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	48	0.68	148694876	0.43	218683821
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	49	0.67	220401	0.45	14010847
PREDICTED: steroidogenic factor 1 isoform X2 [Camelus ferus]	50	0.48	350537337	0.46	13492975

The table shows the Top 50 results of BLASTp and RAFTS3 for 5 sequences randomly selected from the test dataset compared against the NR. The subject sequences are indicated by their GI number and ordered by the default criteria of each tool, the relative score of each one was calculated.

APÊNDICE III - ORGANISMOS REFERÊNCIA UTILIZADOS NO EASYGENE

ACCESSION	ORGANISMO	ORGANISMO REFERENCIA
NC_000853	Thermotoga maritima MSB8	Thermotoga maritima
NC_000913	Escherichia coli str. K-12 substr. MG1655	Escherichia coli K-12
NC_000917	Archaeoglobus fulgidus DSM 4304	Archaeoglobus fulgidus DSM 4304
NC_000919	Treponema pallidum subsp. pallidum str. Nichols	Treponema pallidum
NC_000964	Bacillus subtilis subsp. subtilis str. 168	Bacillus subtilis
NC_002696	Caulobacter crescentus CB15	Caulobacter crescentus
NC_002932	Chlorobium tepidum TLS	Chlorobium tepidum
NC_002936	Dehalococcoides ethenogenes 195	Escherichia coli K-12
NC_002939	Geobacter sulfurreducens PCA	Geobacter sulfurreducens PCA
NC_002967	Treponema denticola ATCC 35405	Treponema denticola ATCC 35405
NC_002977	Methylococcus capsulatus str. Bath	Escherichia coli K-12
NC_003112	Neisseria meningitidis MC58	Neisseria meningitidis serogroup A Z2491
NC_003295	Ralstonia solanacearum GMI1000	Ralstonia solanacearum
NC_003910	Colwellia psychrerythraea 34H	Escherichia coli K-12
NC_004129	Pseudomonas protegens Pf-5	Escherichia coli K-12
NC_004368	Streptococcus agalactiae NEM316	Streptococcus agalactiae NEM316
NC_007146	Haemophilus influenzae 86-028NP	Haemophilus influenzae Rd
NC_007503	Carboxydothermus hydrogenoformans Z-2901	Escherichia coli K-12
NC_008261	Clostridium perfringens ATCC 13124	Clostridium perfringens
NC_009525	Mycobacterium tuberculosis H37Ra	Mycobacterium tuberculosis H23Rv
NC_009707	Campylobacter jejuni subsp. doylei 269.97	Campylobacter jejuni
NC_010380	Streptococcus pneumoniae Hungary19A-6	Streptococcus pneumoniae
NC_010729	Porphyromonas gingivalis ATCC 33277	Escherichia coli K-12
NC_011498	Helicobacter pylori P12	Helicobacter pylori 26695