Desenvolvimento de Ferramentas de Bioinformática para Busca de Genes em Bases de Dados de Transcriptômica e Metagenômica

Wendel Ribeiro de Almeida[1], Beatriz Simas Magalhães[2], Fabiano Cavalcanti Fernandes[1]

[1]Instituto Federal de Brasília (IFB) – Taguatinga, DF, 72146-000, Brasil

[2]Centro de Análises Proteômicas e Bioquímicas, Pós-Graduação em Ciências Genômicas e Biotecnologia, Universidade Católica de Brasília (UCB), Brasília, DF, 70790-160, Brasil

wendelribi@gmail.com, beatrizs@ucb.br, fabiano.fernandes@ifb.edu.br

Abstract. *The study of the biotechnological potential of a biological system now benefits from the latest development of large-scale sequencing platforms for DNA and RNA that can interrogate and give an insight into the complexity of a biological system. In addition, gene expression analysis is no longer limited to a known set of RNA transcripts, but extends to the sequencing of the total RNA. Therefore, this project has developed a software module that remove the occurrence of manual errors and improves the search for genes in transcriptomic and metagenomic databases.*

Resumo. *O estudo do potencial biotecnológico de um sistema biológico, hoje se beneficia do recente desenvolvimento de plataformas de sequenciamento em larga escala de DNA e RNA em que se pode interrogar e se obter uma visão ampla da complexidade de um sistema biológico. Além disso, a análise da expressão gênica não se limita mais a um conjunto conhecido de RNA transcritos, mas estende-se ao sequenciamento do RNA total. Assim sendo, o presente projeto desenvolveu um módulo de software que possibilita a eliminação da ocorrência de erros manuais e a automatização da busca de genes em bases de dados de metagenômica e transcriptômica.*

1. Introdução

Na busca por novos genes em bases de dados de metagenômica e transcriptômica, o pesquisador parte muitas vezes de padrões de genes semelhantes existentes em organismos próximos. Quando não se conhece genes semelhantes, o pesquisador parte de padrões conhecidos de sequências de aminoácidos, tais como peptídeos de sinal, para identificar e mapear vias metabólicas comuns no novo organismo estudado. A tradução reversa (Figura 1), conversão de aminoácidos em respectivos possíveis nucleotídeos em cada posição de códon, quando feita de forma manual pode incorrer em diversos erros causando, portanto uma busca por genes equivocados. Um módulo de software que realize a tradução reversa de forma automática pode auxiliar o pesquisador na eliminação de tarefas manuais e consequentemente na redução de erros na busca por genes.

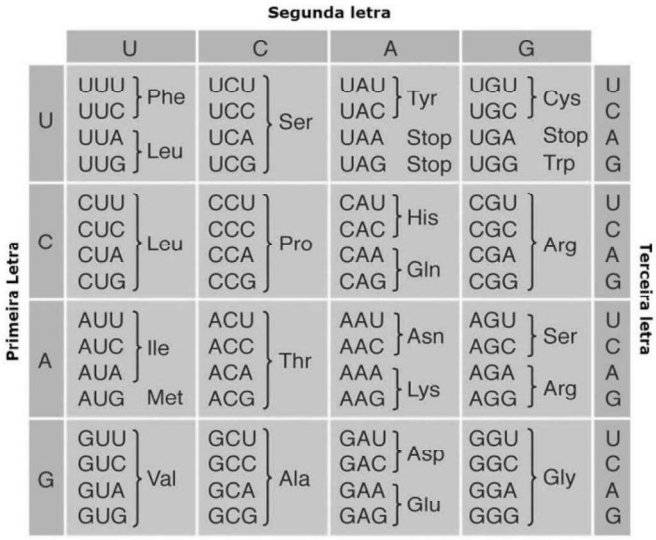

Figura 1. Relação de códons de RNA para tradução e tradução reversa

2. Referencial Teórico

Segundo Prosdocimi (2007, p.39):

> A molécula de DNA (*Deoxyribonucleic Acid* - Ácido Desoxirribonucleico) é estática e está presente, com a mesma constituição, em todas as células do organismo. A decifração desse conteúdo estático de DNA é tarefa da genômica. Já o conteúdo de RNA (*Ribonucleic Acid* - Ácido Ribonucleico) de uma determinada célula depende do tempo e das condições à qual ela está sendo submetida. O transcriptoma mede a parte do genoma que está sendo utilizada em um determinado momento. E essa parte do genoma expresso é diferente para cada tipo celular. Existem genes que são expressos apenas na pele, outros no cérebro e alguns nos ossos. Alguns genes são ainda mais expressos quando a célula está submetida a um choque térmico, à restrição calórica ou à falta de oxigênio. Enquanto o genoma é apenas um, existem vários transcriptomas possíveis para uma mesma espécie.

Que pondera ainda a complementaridade entre as análises de genoma e de transcriptoma na solução de questões específicas. Ainda de acordo com Malkaram, Hassan e Zempleni (2012, p. 658):

> Historicamente, a quantificação de mRNA (RNA mensageiro) tem estado na vanguarda da transcriptômica. Mais recentemente RNA pequenos, como micro RNA e RNA não-codificante, têm atraído considerável atenção. *Microarrays* e sequenciadores de última geração são as tecnologias analíticas primárias na investigação de transcriptômica; os usuários podem utilizar-se de várias plataformas de análise. A tecnologia de *microarray* é madura e bem estabelecida. No entanto, existem algumas dificuldades com a análise de dados e a reprodutibilidade dos resultados, especialmente no contexto de pesquisa nutricional, devido à natureza complexa das relações entre os nutrientes e os genes-alvo. O perfil de transcrição utilizando em *microarrays* tem sido utilizado para identificar alvos celulares para muitos macronutrientes e micronutrientes, e também para caracterizar diferenças de expressão de genes em diferentes condições nutricionais.

3. Material e Métodos

As sequências de RNA ou *reads* foram obtidas pela plataforma 454 Roche GS FLX – *Titanium* do Centro de Genômica de Alto Desempenho do Distrito Federal, consórcio formado pela UCB/UnB/Embrapa/PCDF/GDF. No total foram obtidas 940.000 leituras de sequência pelo método NGS (*Next Generation Sequencing* – Sequenciamento de Nova Geração) para o transcriptoma de *Phyllomedusa distincta*.

3.1. Computador e Sistema Operacional

Foi utilizado um computador do tipo PC, processador Intel® Pentium com 500GB de disco rígido e 4GB de memória RAM. O sistema operacional utilizado foi o Ubuntu versão 12.10.

3.2. Linguagens de Programação

Para processar os dados de sequenciamento e análises de RNA foram desenvolvidos módulos de *software* para Bioinformática. As linguagens utilizadas foram Perl e Bash. Perl ("*Practical Extraction And Report Language*") é uma linguagem de programação estável e multiplataforma usada em diversos setores da indústria de *software*. Perl é uma das linguagens especialmente versáteis no processamento de cadeias (*strings*), manipulação de texto e na busca da ocorrência de padrões, implementado através de expressões regulares [Till 1996].

4. Resultados e Discussão

No processo de busca de genes em novos organismos, parte-se usualmente de padrões pré-existentes conhecidos, tais como genes que geram inibidores de protease, dermaseptinas, dermacisteínas, bradicininas, etc. Para em seguida partir para uma identificação de formas semelhantes ou novas. Para tanto, o pesquisador utiliza sequências de aminoácidos que compõem os peptídeos e em seguida realiza uma tradução reversa manual dos mesmos para seus equivalentes nucleotídeos. A tradução reversa, quando feita de forma manual, está sujeita a ocorrência de erros humanos na conversão dos aminoácidos para os códons (Tabela 1). Foi desenvolvido um script em PERL que automatiza essa conversão (Tabela 2).

Tabela 1 – Exemplos de genes conhecidos, traduções reversas com e sem erros e proteínas excretadas

Nome	Tradução reversa automatizada	Tradução reversa com erros manuais	Aminoácidos
Inibidor de Protease	ACXTACCCXAACGAGTGCCT	ACXTACACXAACGAGTGX	TYPNECLL
Dermatoxina	GGXCTXCTXTGXGGXATACT	GGXCTXCTXTGXGTXATACTXAAC	GLLSGILN

Tabela 2 – Trecho de código fonte de *script* em Perl para tradução reversa

```
1 sub COMBINA {
2   if ($exec!=0){if ($numero==0) {$array[$numero]++;
3                 if ($array[$numero]>$codons[$numero]) { while ($array[$numero]>$codons[$numero]){
4                       $array[$numero]=0;
5                       $numero++;
6                        if ($numero>$salva){$numero=0;}
7                       $array[$numero]++;}
8                     while ($numero>$salva){
9                       $numero=0;
10                      $array[$numero]++;
11                      if ($array[$numero]>$codons[$numero]) {
12                         $array[$numero]=0;} $numero++;}
13                }else {$numero++;
14                   while ($numero>$salva){$numero=0;
15                      $array[$numero]++;
16                      if ($array[$numero]>$codons[$numero]){$array[$numero]=0;}
17                      $numero++;}}$numero=1;}}
```

5. Considerações Finais e Trabalhos Futuros

Pesquisadores frequentemente executam tarefas manuais de tradução reversa de padrões conhecidos de aminoácidos para mapear vias metabólicas conhecidas em novos organismos. Erros podem ocorrer com frequência na tradução reversa, com isso, foi identificado o processo de tradução reversa e construído um software que automatiza essa tarefa, facilitando o estudo de novos organismos e a busca por novos genes. Concluímos que a utilização de uma ferramenta para execução desta tarefa acaba se tornando essencial para um pesquisador evitar trabalhos exaustivos e ocorrência de erros manuais.

Como continuação do trabalho pretende-se criar uma banco de dados NoSQL (*Not Only SQL* – Não apenas SQL) baseado em grafos com o mapeamento dos genes descobertos da *Phyllomedusa distincta* e dos padrões pré-existentes conhecidos *a priori*, tais como genes que geram inibidores de protease, dermaseptinas, dermacisteínas, bradicininas, etc.

6. Referências

Malkaram, S. A., Hassan, Y. I. e Zempleni, J. (2012). Online tools for bioinformatics analyses in nutrition sciences. *Adv Nutr*, v. 3, n. 5, p. 654-65.

Prosdocimi, F. Introdução à Bioinformática. UFRJ, Rio de Janeiro, mar. 2014. Disponível em: <http://www2.bioqmed.ufrj.br/prosdocimi/FProsdocimi07_CursoBioinfo.pdf>. Acesso em: 12 mar. 2014.

Till, D. (1996) “Teach Yourself Perl 5 in 21 days”, Sams Publishing, Indianapolis, IN, 2nd. edition.

7. Financiamento

O projeto contou com financiamento do CNPq/IFB na forma de bolsa de iniciação científica de ensino médio PIBIC-EM.