

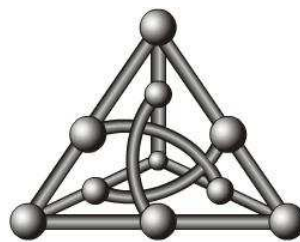
---

# Classificação de Seqüências de RNAs não-codificantes através da Distância de Compressão Normalizada

Habib Asseiss Neto  
Sérgio Ronaldo Alves de Sousa Júnior

---

Departamento de Computação e Estatística  
Universidade Federal de Mato Grosso do Sul



**Orientador: Prof. Dr. Said Sadique Adi**

Campo Grande, Dezembro de 2008

# Sumário

<b>1</b>	<b>Introdução</b>	<b>1</b>
<b>2</b>	<b>Conceitos Biológicos</b>	<b>3</b>
2.1	DNA e RNA . . . . .	3
2.2	cRNA e ncRNA . . . . .	4
<b>3</b>	<b>Conceitos Computacionais</b>	<b>7</b>
3.1	Cadeias . . . . .	7
3.2	Compressores . . . . .	8
3.2.1	Gzip . . . . .	9
3.2.2	Bzip2 . . . . .	9
3.2.3	LZMA . . . . .	9
3.3	Comparação de Sequências . . . . .	10
3.3.1	Similaridade . . . . .	10
3.3.2	Distância de Compressão Normalizada . . . . .	11
<b>4</b>	<b>Objetivos</b>	<b>13</b>
<b>5</b>	<b>Metodologia</b>	<b>14</b>
<b>6</b>	<b>Resultados</b>	<b>16</b>
<b>7</b>	<b>Conclusão</b>	<b>18</b>

## **Resumo**

Sabe-se hoje que, além dos RNAs mensageiros, as células de qualquer organismo incluem outros tipos de RNAs, denominados genericamente de RNAs não-codificantes. Os ncRNAs são classificados em tipos de acordo com a função que desempenham dentro da célula. Propomos, neste trabalho, uma metodologia de classificação de ncRNAs baseada na comparação das seqüências desses elementos biológicos. Essa comparação é realizada por meio da Distância de Compressão Normalizada, uma medida derivada dos conceitos de Teoria da Informação, mais especificamente do conceito da Complexidade de Kolmogorov. Uma outra medida de comparação, denominada similaridade, é também utilizada na tarefa de classificação, e seus resultados comparados com aqueles obtidos através da Distância de Compressão Normalizada.

# Capítulo 1

## Introdução

A Biologia Computacional, ou Bioinformática, pode ser definida como o desenvolvimento e uso de técnicas computacionais e matemáticas para auxiliar na solução de problemas da biologia [14]. Um dos vários temas abordados por essa área é o da análise de seqüências genômicas. Nesse contexto, um problema de grande importância é o da classificação. Nele, dado um conjunto de seqüências de interesse, estamos interessados em classificá-las em grupos de acordo com suas características e funcionalidades. Classificações como essas permitem compreender vários aspectos biológicos relacionados às seqüências genômicas e, conseqüentemente, ao papel que elas desempenham dentro do organismo.

Neste trabalho, estamos interessados especificamente no problema da classificação de RNAs não-codificantes (ncRNAs). A motivação para classificar seqüências desse tipo está na importância que essas moléculas vêm ganhando no cenário atual. Essa importância deve-se a descobertas recentes sobre as funções dos ncRNAs dentro das células. Apesar disso, estudos que tratam de moléculas desse tipo ainda são pouco quando comparados ao de outras seqüências genômicas.

Para a classificação de seqüências de ncRNAs, desenvolvemos uma metodologia baseada na comparação de seqüências. Essa metodologia possui como medida de comparação a Distância de Compressão Normalizada (DCN)[5], medida essa fundamentada nos conceitos da Teoria da Informação. Mais especificamente, a DCN foi desenvolvida com base na Complexidade de Kolmogorov que, apesar de ser uma noção não-computável, pode ser aproximada via compressão.

Nossa metodologia apresentou bons resultados na classificação de um total de 1342 seqüências de ncRNAs em cinco grupos distintos: *miRNA*, *snoRNA*, *snRNA*, *snmRNA* e *RNase P RNA*. No intuito de avaliar de forma mais detalhada a qualidade dos resultados obtidos via DCN, eles foram comparados àqueles gerados por meio de uma outra medida de comparação, denominada similaridade. Os resultados dessa última foram piores que aqueles mostrados pela DCN.

Este texto está dividido em capítulos onde se incluem noções biológicas e computacionais, além desta introdução e conclusão sobre os estudos e resultados obtidos. No próximo capítulo, são abordados os conceitos relacionados à biologia necessários para uma

melhor compreensão do problema estudado. Logo em seguida, os entendimentos básicos necessários para a compreensão da solução computacional do problema são fornecidos. Os conceitos de Distância de Compressão Normalizada e similaridade são apresentados na seção 3.3. O objetivo deste estudo é tratado no capítulo 4 e então a metodologia proposta é explanada no capítulo 5. Por fim, os resultados obtidos por nossos testes são apresentados e uma conclusão sobre eles é dada.

# Capítulo 2

## Conceitos Biológicos

Neste capítulo são abordados os principais conceitos biológicos usados no decorrer deste documento. A maioria das definições utilizadas a seguir foram obtidas em [14].

### 2.1 DNA e RNA

Dentre os vários elementos que compõem as células de qualquer organismo vivo estão dois tipos de ácidos nucleicos: o **ácido ribonucléico** (RNA) e o **ácido desoxirribonucléico** (DNA). Ambos são formados por moléculas denominadas nucleotídeos, que por sua vez constituem-se de um radical fosfato, um açúcar e uma dentre cinco bases nitrogenadas, denominadas **Adenina**, **Citosina**, **Guanina**, **Timina** e **Uracila**. Apesar de algumas similaridades composicionais, os RNAs e os DNAs diferenciam-se tanto em termos estruturais quanto em termos funcionais.

Enquanto o RNA é uma cadeia simples de nucleotídeos, o DNA é uma cadeia dupla, cujos filamentos encontram-se dispostos em uma estrutura helicoidal, e ligados um ao outro por meio de pontes de hidrogênio que se formam entre pares de bases ditas **complementares**: a base A para-se com a base T e a base C com a base G. Vale salientar também que os nucleotídeos componentes do DNA incluem somente as bases Adenina, Citosina, Guanina e Timina, enquanto que os nucleotídeos componentes do RNA incluem as bases Adenina, Citosina, Guanina e Uracila. As diferenças estruturais entre as moléculas de DNA e RNA podem ser visualizadas na Figura 2.1.

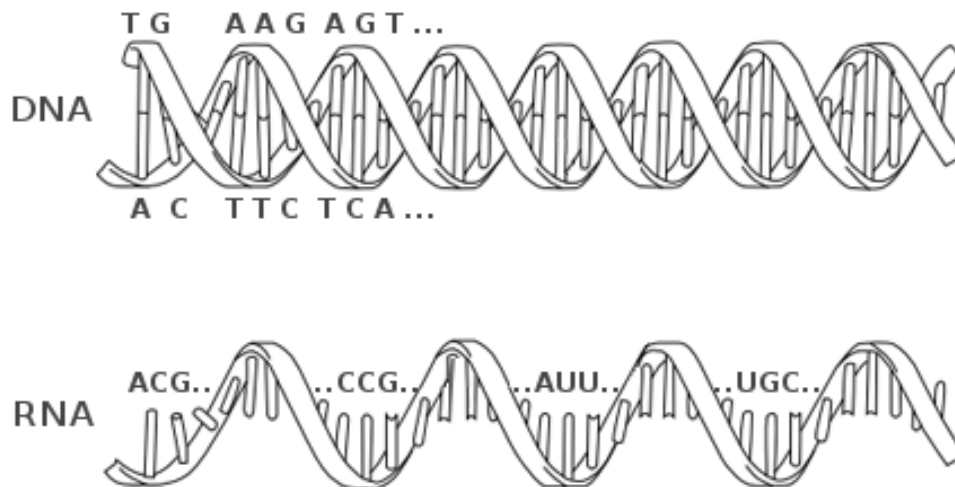


Figura 2.1: Estruturas de uma molécula de DNA e uma molécula de RNA.

Com relação às diferenças funcionais entre o RNA e o DNA, sabe-se hoje que o último possui, essencialmente, uma única função, a de codificar informação, enquanto que o RNA desempenha funções variadas dentro da célula. Essa diversidade de papéis reflete-se na imensa variedade de RNAs existentes. De acordo com a função que desempenha, os RNAs podem ser dividido em duas classes principais: os **RNAs codificantes** (cRNAs) e os **RNAs não-codificantes** (ncRNAs). Na realidade, existe um único RNA codificante, denominado RNA mensageiro (mRNA), que contém a informação para codificação de proteínas. Sobre os RNAs não-codificantes, os mais conhecidos são: o RNA transportador (tRNA), responsável pelo transporte de aminoácidos e o RNA ribossomal (rRNA), que possui papel estrutural. Além desses, vários outros tipos de ncRNAs existem, e mais detalhes sobre eles serão dados na seção seguinte.

Uma outra molécula importante para o desenvolvimento de qualquer ser vivo é a proteína. Ela é formada a partir da união de aminoácidos. Esses, por sua vez, são moléculas orgânicas formadas por átomos de carbono, hidrogênio, oxigênio e nitrogênio unidos entre si. As proteínas possuem as mais diversas funções nos mais diversos organismos. As milhares de enzimas que um organismo possui, por exemplo, são todas proteínas com funções importantes.

Tanto as seqüências de RNAs quanto as de proteínas são sintetizadas a partir do DNA, e conhecidas genericamente como **transcritos**.

## 2.2 cRNA e ncRNA

Certas partes ao longo da molécula da DNA possuem informações utilizadas na síntese de moléculas de RNA. Essas partes são conhecidas como **genes**. Como já foi mencionado, os RNAs podem ser classificados em dois tipos distintos: RNAs codificantes e RNAs não-codificantes. Da mesma forma, os genes podem ser classificados em genes codificantes e não-codificantes de acordo com o RNA sintetizado a partir deles.

Os RNAs codificantes incluem, basicamente, um RNA denominado Mensageiro (mRNA). Após sintetizada, essa molécula é processada e dá origem a uma proteína. Esse processamento inclui uma fase de *splicing* e uma fase de tradução. Tais fases serão explicadas a seguir e podem ser melhor compreendidas pela visualização da Figura 2.2.

Na fase de *splicing*, o RNA sintetizado sofre um processo de maturação, onde algumas porções dessa molécula são eliminadas (**íntrons**). As partes restantes, denominadas **éxons** ligam-se formando o RNA mensageiro maduro. No final do processo, o mRNA é constituído apenas por seqüências que codificam os aminoácidos de uma proteína.

A fase de tradução é responsável por converter a molécula de mRNA em proteína. Isso é realizado por meio de um outro RNA denominado transportador (tRNA). Ele lê as informações do mRNA de três em três bases. Cada tripla de base é chamada de **códon**. A tradução é realizada em uma estrutura celular denominada ribossomo. Os RNAs transportadores são posicionados corretamente junto aos RNAs mensageiros e as ligações peptídicas são catalisadas entre os aminoácidos para a síntese de proteínas.

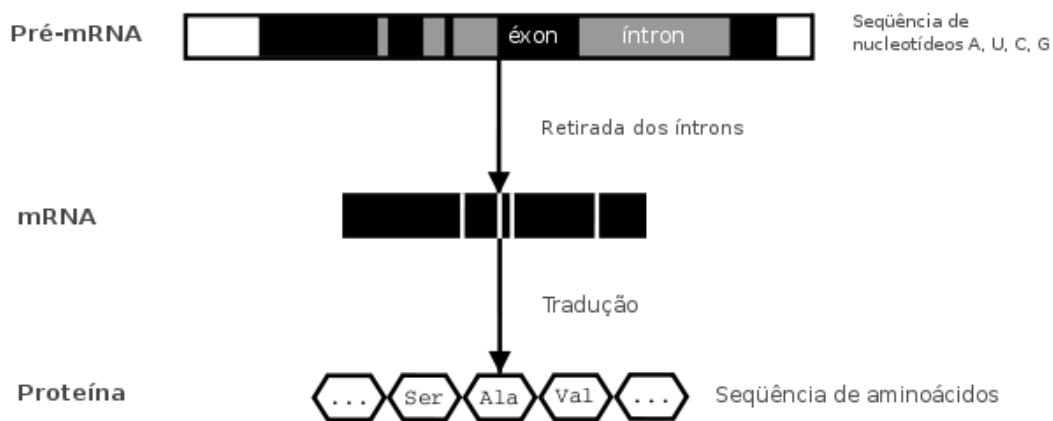


Figura 2.2: O processo de síntese de proteína a partir do DNA

Sobre os RNAs não-codificantes, eles podem ser de vários tipos e desempenhar funções variadas dentro da célula. Dentre os principais ncRNAs estão: tRNA, rRNA, snoRNA, miRNA, siRNA, piRNA. A função de cada uma dessas moléculas é descrita a seguir:

- **RNA transportador (tRNA):** são utilizados como moléculas tradutoras da informação de cada códon componente do mRNA em um aminoácido específico a ser adicionado à proteína sendo formada. O tRNA desempenha essa função através de duas regiões: o anticódon, que é responsável pelo reconhecimento de códons específicos do mRNA, e o final 3', ao qual o aminoácido correspondente ao mRNA é anexado;

- **RNA ribossomal (rRNA)**: esse é o componente central do ribossomo (“maquinário” fabricante de proteína de todas as células vivas). A função do rRNA é prover um mecanismo para decodificar o mRNA em aminoácidos e interagir com os tRNAs durante a tradução. As moléculas de rRNA têm diversos papéis na síntese de proteína: papel catalítico, papel de reconhecimento e papel estrutural;
- **small nucleolar RNA (snoRNA)**: é uma classe de pequenas moléculas que realizam modificações químicas em rRNAs, além de outros ncRNAs, tal como o tRNA. Essas modificações possuem como principal objetivo promover a maturação desses ncRNAs, transformando-os em moléculas ativas. A origem desses genes ainda não está clara, mas acredita-se que eles originam-se dos íntrons do mRNA;
- **microRNA (miRNA)**: aparenta estar relacionado com a regulação gênica. As moléculas de miRNA são parcialmente complementares a uma ou mais moléculas de mRNA e sua principal função é reduzir a expressão de genes codificantes, inibindo a tradução de mRNAs;
- **small interfering RNA (siRNA)**: possui o mesmo papel do miRNA, porém reduz a expressão de genes codificantes degradando o mRNA ao invés de inibir sua tradução;
- **piwi-interacting RNA (piRNA)**: é uma classe de pequenas moléculas de RNA existentes basicamente em células dos mamíferos. Assim como os miRNAs e os snoRNAs, os piRNAs também estão relacionados com a regulação gênica. Mais especificamente, eles atuam no silenciamento de genes capazes de se auto-duplicar no interior do genoma.

# Capítulo 3

## Conceitos Computacionais

Neste capítulo serão introduzidos alguns dos principais conceitos computacionais utilizados durante este trabalho. Alguns dos conceitos utilizados a seguir foram obtidos em [14].

### 3.1 Cadeias

Um alfabeto  $\Sigma$  é um conjunto finito de símbolos ou caracteres. Uma cadeia  $X_1X_2\dots X_n$  construída sob  $\Sigma$  é uma seqüência finita e ordenada de caracteres de  $\Sigma$ . Os seguintes conceitos estão associados à noção de cadeia:

- **Comprimento:** o comprimento de uma cadeia é definido como a quantidade de caracteres contidos nessa cadeia. A cadeia `abcd`, por exemplo, possui comprimento igual a 4;
- **Concatenação:** para quaisquer duas cadeias  $X$  e  $Y$ , a concatenação de  $X$  e  $Y$ , representada por  $XY$ , é definida como a justaposição de  $Y$  em  $X$ , ou seja, a seqüência de caracteres em  $X$  seguida pela seqüência de caracteres em  $Y$ . Por exemplo, se  $X = \text{abcd}$  e  $Y = \text{efgh}$ , então a concatenação de  $X$  e  $Y$  é  $XY = \text{abcdefgh}$ ;
- **Prefixo:** o prefixo de uma cadeia  $X = X_1\dots X_n$  é qualquer cadeia  $X' = X_1\dots X_m$ , onde  $m \leq n$ . Um prefixo próprio de uma cadeia  $X = X_1\dots X_n$  é qualquer cadeia  $X' = X_1\dots X_m$ , onde  $m < n$ ;
- **Sufixo:** o sufixo de uma cadeia  $X = X_1\dots X_n$  é qualquer cadeia  $X' = X_k\dots X_n$ , onde  $k \geq 1$ . Um sufixo próprio de uma cadeia  $X = X_1\dots X_n$  é qualquer cadeia  $X' = X_k\dots X_n$ , onde  $k > 1$ ;
- **Subcadeia:** uma cadeia  $Y$  é dita subcadeia de uma cadeia  $X$  se existem duas cadeias  $U$  e  $V$  tais tais que  $X = UYV$ . Em outras palavras, podemos definir  $Y$  como sendo um prefixo de um sufixo de  $X$  ou, equivalentemente, um sufixo de um prefixo de  $X$ .

No contexto deste trabalho, estamos interessados nos conceitos de cadeia para a representação de seqüências de DNA e RNA. O alfabeto utilizado então é constituído de 5 símbolos diferentes: A, C, T, G e U. Um exemplo de uma cadeia DNA  $D$  é  $D = \text{ACGTACTGATGAGACGGAGCAG}$ , enquanto que um exemplo de uma cadeia de RNA  $R$  é  $R = \text{UACGUAUGCAUGAUCGUA}$ .

## 3.2 Compressores

A operação de compressão consiste na redução do tamanho da representação computacional de um dado [3]. Dentre as principais motivações para essa operação estão a economia de espaço em dispositivos de armazenamento e o ganho de desempenho em transmissão de dados. A operação de compressão é realizada através de algoritmos designados para esse fim. Cada um desses algoritmos implementa um ou mais métodos de compressão específicos. Sobre esses métodos, eles podem ser genericamente classificados em duas categorias: com perdas e sem perdas de dados.

Um método de compressão é classificado como sem perdas se a informação obtida após a descompressão é idêntica à informação que se tinha antes da compressão. Esses métodos podem ser aplicados, por exemplo, a textos e programas de computador, onde uma perda mínima de dados acarreta o seu não funcionamento ou torna os dados incompreensíveis. Por outro lado, algumas situações permitem que perdas de dados pouco significativas ocorram. Em sons e filmes, por exemplo, algumas perdas não são significativas ao olho e ouvido humano. Nesses casos, os dados obtidos após a descompressão não são necessariamente idênticos aos originais.

Além da classificação decorrente da perda ou não de informações, os métodos de compressão podem também ser classificados de acordo com outras propriedades. Dentre as principais categorias para classificação de métodos de compressão estão:

- **Simétricos e assimétricos:** quando as tarefas de compressão e descompressão são feitas executando-se algoritmos com complexidades semelhantes, diz-se que o método de compressão é simétrico. O LZW (*Lempel-Ziv-Welch*) [21] é um bom exemplo de algoritmo simétrico. Por outro lado, quando o algoritmo de compressão tem maior complexidade que o de descompressão, ou vice-versa, dizemos que o método de compressão é assimétrico. Isso ocorre, por exemplo, com o LZ77 [22];
- **Adaptativos e não-adaptativos:** quando as regras de compressão variam de acordo com os dados e à medida que eles vão sendo processados, dizemos que os métodos são adaptativos. Caso contrário, os métodos são não-adaptativos;
- **Métodos estatísticos:** os métodos estatísticos utilizam como informação principal as freqüências dos símbolos no fluxo de dados, e alteram a representação de cada símbolo ou grupo de símbolos de acordo com essas freqüências ou através de alguma outra característica probabilística. Dessa forma, visam reduzir o número de bits usados para representar cada símbolo ou grupo de símbolos. Dois dos exemplos

mais conhecidos de algoritmos baseados em métodos estatísticos são o algoritmo de Huffman [9] e o da Codificação Aritmética [18];

- **Métodos baseados em dicionários:** os métodos baseados em dicionários utilizam-se de dicionários ou estruturas similares de forma a eliminar repetições de símbolos ou frases redundantes. Como exemplos de algoritmos baseados nesse método, temos os algoritmos da família Lempel-Ziv, como LZW, LZ77 e LZ78 [22]. Apesar da diferença entre os métodos estatísticos e baseados em dicionários, os programas mais usados de compressão associam uma técnica baseada em dicionário com uma técnica estatística.

Neste trabalho, focalizamos nossas atenções em três programas específicos de compressão, classificados na categoria sem perdas. São eles: Gzip [6], Bzip2 [19] e LZMA [17]. A descrição desses programas encontra-se a seguir.

### 3.2.1 Gzip

O Gzip (abreviação para GNU-Zip) é um programa clássico, criado em 1993, para a compressão de arquivo. É um software livre criado para substituição do programa “*compress*” utilizados em sistemas Unix antigos. É baseado no algoritmo *DEFLATE*, que é uma combinação de dois outros algoritmos, *Huffman* e *Lempel-Ziv* (LZ77). A quantidade de compressão obtida depende do tamanho da entrada e da distribuição de subcadeias comuns. Tipicamente, o tamanho de um código-fonte ou de um texto escrito em português ou inglês torna-se de 60 a 70% menor após a compressão pelo Gzip.

### 3.2.2 Bzip2

O Bzip2 utiliza diversas técnicas de compressão, que são executadas em uma certa ordem durante a compressão e na ordem inversa durante a descompressão. O programa utiliza uma técnica que converte seqüências de caracteres considerados freqüentes em cadeias de símbolos idênticos. Assim como no Gzip, a codificação de *Huffman* também é utilizada. Este compressor é conhecido por ser lento na compressão, porém a descompressão é relativamente rápida.

### 3.2.3 LZMA

O LZMA (*Lempel-Ziv-Markov chain-Algorithm*) é um algoritmo de compressão em desenvolvimento desde 1998. Ele usa um método de compressão baseado em dicionário, muito similar ao LZ77, e tem como característica uma alta taxa de compressão. O LZMA é essencialmente o algoritmo *DEFLATE* (o mesmo utilizado no Gzip), mas conta com um tamanho de dicionário maior: 32 MB ao invés de 32 KB.

### 3.3 Comparação de Seqüências

Na biologia computacional, a comparação de seqüências é uma das operações de maior importância, pois serve como base para outros processamentos mais complexos. De maneira informal, essa operação consiste em determinar quão semelhantes são duas seqüências. Por trás dessa tarefa aparentemente simples, existe o conceito de medida de distância. Essa medida corresponde a um número real ou inteiro que determina o grau de semelhança entre duas seqüências. Neste trabalho, estamos interessados em duas medidas específicas: a similaridade e a Distância de Compressão Normalizada.

#### 3.3.1 Similaridade

A definição de similaridade está intimamente relacionada à noção de alinhamento. Informalmente, alinhar duas seqüências consiste em inserir espaços entre os seus caracteres de modo que elas fiquem do mesmo tamanho. Feito isso, as seqüências podem ser colocadas uma em cima da outra, criando-se assim uma correspondência entre os seus caracteres. Essa disposição das seqüências modificadas (com os espaços) é o que chamamos de **alinhamento**. Cada par de caracteres alinhados é chamado de coluna do alinhamento, e pode incluir dois caracteres iguais, dois caracteres diferentes e um caractere e um espaço. Um exemplo de alinhamento entre as seqüências  $s = \text{GACTAGCTACATTTTCGAGC}$  e  $t = \text{AGCCCGTAATATGC}$  é mostrado abaixo. Nele, os espaços estão sendo representados por ‘\_’.

```
GACTAGCTACATTTTCGAGC
-AGCC-CGTAATAT-G--C
```

Dado um alinhamento  $\mathcal{A}$ , é possível associar uma pontuação a ele por meio de uma função que atribui um valor inteiro para cada par de caracteres possíveis de compor as suas colunas. Essa pontuação corresponde à soma dos valores de cada coluna. Considerando uma função de pontuação  $w$  tal que  $w(x, x) = 1$ ,  $w(x, y) = -1$  e  $w(x, -) = -2$ , a pontuação do alinhamento dado acima é  $-10$ .

Dentre todos os alinhamentos possíveis entre duas seqüências, aquele de maior pontuação total é chamado de **alinhamento ótimo**. A pontuação desse alinhamento é denominada **similaridade**. Considerando novamente as seqüências  $s$  e  $t$  anteriores, um dos possíveis alinhamentos ótimos entre elas tem pontuação  $-6$  e é mostrado a seguir.

```
GACTAGCTACATTTTCGAGC
AGCCCG-TA-ATAT---GC
```

O melhor alinhamento entre duas seqüências  $s$  e  $t$ , de tamanhos  $m$  e  $n$ , respectivamente, pode ser calculado por meio de um algoritmo de programação dinâmica baseado na seguinte recorrência:

$$M[i, j] = \max \begin{cases} M[i, j - 1] + w(-, t[j]), \\ M[i - 1, j - 1] + w(s[i], t[j]), \\ M[i - 1, j] + w(s[i], -). \end{cases} \quad (3.1)$$

Na recorrência acima,  $M$  é uma matriz de dimensão  $m \times n$ . O valor de similaridade buscado encontra-se na última posição dessa matriz, e o alinhamento propriamente dito pode ser construído “voltando-se” em  $M$  até a sua primeira posição.

### 3.3.2 Distância de Compressão Normalizada

A Distância de Compressão Normalizada possui como base a Complexidade de Kolmogorov. A Complexidade de Kolmogorov de uma seqüência  $X$ , denotada por  $K(X)$ , é definida como o tamanho do menor programa binário que gera  $X$ . Intuitivamente,  $K(X)$  representa a quantidade mínima de informação necessária para gerar  $X$ . A Complexidade de Kolmogorov condicional  $K(X|Y)$  é definida como o tamanho do menor programa binário que, a partir de uma cadeia dada  $Y$ , gera  $X$ . Observe que a Complexidade de Kolmogorov de uma seqüência  $X$ ,  $K(X)$ , corresponde à Complexidade de Kolmogorov condicional  $K(X|\lambda)$ , onde  $\lambda$  é a palavra vazia.

A função de distância a seguir, proposta por Vitányi *et al.* em [2], utiliza-se da noção de Complexidade de Kolmogorov condicional para calcular quão próximas são duas seqüências  $X$  e  $Y$ . Essa função é chamada de **Distância de Informação Normalizada** ( $DIN$ ).

$$DIN(X, Y) = \frac{\max\{K(X|Y), K(Y|X)\}}{\max\{K(X), K(Y)\}}$$

Infelizmente, a Complexidade de Kolmogorov é uma noção não-computável, e em aplicações práticas ela é aproximada pelo tamanho da seqüência comprimida, o que pode ser calculado por um algoritmo de compressão. Para uma seqüência  $X$ , então, é possível aproximar  $K(X)$  por meio de  $C(X)$ , onde  $C$  é um compressor real e  $C(X)$  é o tamanho da seqüência  $X$  comprimida por  $C$ . Podemos, então, gerar uma aproximação da função  $DIN$ .

Note que o denominador de  $DIN$  é facilmente aproximado para  $\max\{C(X), C(Y)\}$ . O numerador, no entanto, envolve complexidade condicional, contendo os termos  $K(X|Y)$  e  $K(Y|X)$ . Em [5], uma aproximação para esta parte da fração é sugerida por Steven de Rooij da seguinte maneira:  $\max\{K(X|Y), K(Y|X)\} \approx \min\{C(XY), C(YX)\} - \min\{C(X), C(Y)\}$ . Podemos substituir  $\min\{C(XY), C(YX)\}$  por apenas  $C(XY)$ , pois, a partir de testes em [4],  $C(XY)$  e  $C(YX)$  apresentaram pouca diferença.

A função que aproxima a função  $DIN$  utilizando um compressor real  $C$  é chamada de **Distância de Compressão Normalizada** ( $DCN$ ).

$$DCN(X, Y) = \frac{C(XY) - \min\{C(X), C(Y)\}}{\max\{C(X), C(Y)\}}$$

A DCN nos retorna um valor não-negativo  $n$ , onde  $0 \leq n \leq 1 + \epsilon$ , representando quão diferentes as seqüências  $X$  e  $Y$  são. Valores menores de DCN representam seqüências mais similares. O erro  $\epsilon$  no limite superior é devido a imperfeições nas técnicas de compressão utilizadas.

Note que na fórmula da DCN, o tamanho da seqüência  $XY$  comprimida é utilizado. Essa compressão possibilita abstrair como a seqüência  $X$  pode ser comprimida através da seqüência  $Y$ , onde  $Y$  serve como base de informação para ajudar, através de partes em comum, a comprimir  $X$ . Logo, quanto mais partes em comum  $X$  e  $Y$  possuem, sendo assim consideradas mais semelhantes, o DCN nos retorna um valor mais próximo de zero. Quanto mais próximo de zero o valor de DCN for, significa que houve menos trabalho, ou seja, mais informações de  $Y$  foram aproveitadas para comprimir  $X$ .

# Capítulo 4

## Objetivos

Considerando a importância dos ncRNAs e o fato de a funcionalidade de muitos deles ainda ser desconhecida, torna-se importante o desenvolvimento de ferramentas computacionais que auxiliem no estudo desses transcritos. O foco principal de nosso trabalho é desenvolver e implementar um método que auxilie na caracterização de ncRNAs. Mais especificamente, estamos interessados na classificação de ncRNAs desconhecidos com base na sua comparação com seqüências já caracterizadas.

Ainda como parte dos objetivos, pretendemos comparar as seqüências de ncRNAs por meio de duas medidas de comparação distintas e então avaliar a eficiência de cada uma delas na tarefa de classificação.

# Capítulo 5

## Metodologia

Para atingirmos os objetivos mencionados, os seguintes passos foram executados:

1. Identificação de duas bases de dados distintas de seqüências de ncRNAs a partir das quais foram criados dois conjuntos distintos: o conjunto de testes, contendo as seqüências a serem classificadas, e o conjunto de possíveis representantes, contendo as seqüências que irão compor os conjuntos de classificação. As bases identificadas foram *NONCODE* [13] e *functional RNAdb* [10];
2. Uma vez identificadas as bases de dados, verificou-se quais tipos de ncRNAs aparecem em ambas e a quantidade de cada um deles nessas bases. De acordo com essa verificação, os seguintes tipos de ncRNAs foram selecionados: miRNA, snoRNA, snRNA, snmRNA, RNase P RNA. Os ncRNAs desses tipos foram então extraídos da base *RNAdb* no intuito de compor os conjuntos de classificação. Da base *NONCODE*, foram extraídos os ncRNAs, dos tipos mencionados, a compor o conjunto de testes;
3. Considerando a existência de várias seqüências semelhantes dentro de cada conjunto de classificação, construímos, para cada um deles, um subconjunto composto de seqüências ditas representantes. Ou seja, de seqüências que representam o conjunto de maneira mais abrangente possível e sem redundância. A escolha dos representantes para cada um dos tipos de ncRNAs de interesse foi feita comparando-se, por meio da DCN, todas as seqüências de um certo tipo entre si. Para cada par de seqüências  $x$  e  $y$ , verifica-se então se o valor  $DCN(x,y)$  é menor que 0.4. Nesse caso, assume-se que as duas seqüências são semelhantes e não há necessidade de ambas pertencerem ao conjunto de representantes. Portanto uma delas é eliminada desse conjunto. Caso contrário assume-se que as duas seqüências representam ncRNAs distintos (mas do mesmo tipo), e por isso ambas devem pertencer ao conjunto de representantes;
4. Com os conjuntos de classificação, as seqüências do conjunto de testes foram classificadas como segue: cada seqüência é comparada, via o valor da DCN, com todos os representantes de cada tipo. A cada comparação, verifica-se o valor do DCN e

armazena-se o menor valor encontrado até então. No momento em que não houver mais representantes a comparar, verifica-se a seqüência representante mais próxima (menor valor de DCN) da seqüência a ser classificada. Classifica-se então essa última como sendo do mesmo tipo da primeira.

Para a execução dos passos acima descritos foi desenvolvido um programa em *Bash script*, a ser executado em ambientes Unix, que se utiliza de comandos padrão do próprio sistema operacional. A execução do programa se dá através de uma chamada com os seguintes argumentos, nesta ordem: compressor, conjuntos de classificação, conjunto de testes e limite de DCN.

O compressor informado como parâmetro é uma cadeia que pode assumir os seguintes valores: “lzma”, “bzip2” ou “gzip”. Os conjuntos de classificação e de testes são representados como diretórios no sistema de arquivos. As seqüências de ncRNA são arquivos de texto contendo apenas a cadeia de caracteres que as representam. Os diretórios de classificação e de testes são formados por subdiretórios que representam os tipos de ncRNAs, e estes, por sua vez, guardam as seqüências existentes em seus respectivos tipos.

Sempre que duas seqüências precisam ser comparadas, o compressor escolhido é utilizado para comprimir as seqüências em questão, além da concatenação delas. A partir do tamanho dos arquivos comprimidos, a similaridade é calculada através da fórmula de DCN, já descrita anteriormente.

# Capítulo 6

## Resultados

No intuito de avaliar a exatidão da DCN, assim como a nossa metodologia, um total de 1342 seqüências de ncRNAs pertencentes ao homem (*Homo sapiens*) e ao camundongo (*Mus musculus*) foram classificadas como miRNA, snoRNA, snRNA, snmRNA ou RNase P RNA de acordo com o último passo da nossa metodologia. A exatidão dos resultados obtidos foi mensurado calculando-se os seguintes valores: verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos para cada um dos cinco tipos de ncRNAs. Além disso, uma medida que agrega os valores mencionados anteriormente, denominada **Acerto**, é utilizada para medir a qualidade dos resultados da classificação. Essa medida pode ser expressa como:

$$Acerto = \frac{VP}{VP + FN}$$

Tipo	VP	VN	FP	FN	Acerto (%)
miRNA	286	866	25	165	63
snoRNA	695	452	160	35	95
snRNA	97	1199	46	0	100
snmRNA	5	1275	14	48	9
RNase P RNA	11	1328	3	0	100
<b>Média</b>	<b>218.8</b>	<b>1024</b>	<b>49.6</b>	<b>49.6</b>	<b>73.4</b>

Tabela 6.1: Resultados apresentados através de um limitante 0.4 para a DCN. As siglas VP, VN, FP e FN significam, respectivamente, Verdadeiro Positivo, Verdadeiro Negativo, Falso Positivo e Falso Negativo

A Tabela 6.1 mostra que nossa técnica de classificação, baseada na Distância de Compressão Normalizada, apresenta bons resultados. Nesta tabela, apresentamos os resultados

obtidos através de uma escolha de limitante de DCN igual a 0.4, utilizando o compressor LZMA. A determinação desse valor e desse compressor foi feita empiricamente.

Apesar do conjunto que contém snmRNAs ter um baixo número de verdadeiros positivos, isso não acontece com os outros tipos. O restante dos conjuntos apresenta bons resultados, próximo ou igual à pontuação máxima, exceto no caso do conjunto que contém seqüências miRNA, que apresentou um comportamento mediano, com 63% de acerto.

Um dos possíveis motivos para a baixa qualidade do resultado para o snmRNA é que suas seqüências apresentam uma grande variação de comprimento. Como a DCN utiliza a compressão das seqüências para obter a similaridade, o comprimento é um fator fundamental no tamanho do arquivo resultante após a compressão. Logo, duas seqüências do mesmo tipo deveriam obter arquivos comprimidos de tamanhos semelhantes, o que não acontece com a maioria das seqüências de snmRNA.

A exatidão da Distância de Compressão Normalizada na tarefa de classificação de ncRNAs fica mais evidente quando comparada à classificação feita utilizando-se a medida de similaridade. Realizamos os mesmos experimentos para essa medida e obtivemos os resultados que podem ser conferidos na Tabela 6.2. É importante lembrar que os representantes de cada tipo de ncRNA utilizado neste experimento são os mesmos utilizados nos testes via DCN. Pode-se notar que os resultados apresentam valores muito inferiores aos do teste anterior, pois o acerto máximo conseguido por tipo é de 55%, que é o caso das seqüências do tipo snoRNA. Além disso, nenhuma das seqüências de RNase P RNA foram classificadas corretamente, pois seu valor verdadeiro positivo é 0. Isso implica em uma taxa de acerto de 0% para este tipo. Devido a esses fatores, a média de acerto total é 27.4%.

Tipo	VP	VN	FP	FN	Acerto (%)
miRNA	84	857	34	367	18
snoRNA	402	494	118	328	55
snRNA	23	1096	149	74	23
snmRNA	22	884	405	31	41
RNase P RNA	0	1226	105	11	0
<b>Média</b>	<b>106.2</b>	<b>911.4</b>	<b>162.2</b>	<b>162.2</b>	<b>27.4</b>

Tabela 6.2: Resultados apresentados através da comparação de seqüências via similaridade. As siglas VP, VN, FP e FN significam, respectivamente, Verdadeiro Positivo, Verdadeiro Negativo, Falso Positivo e Falso Negativo

# Capítulo 7

## Conclusão

Neste trabalho, propomos uma metodologia para classificação de seqüências de ncRNAs. Essa metodologia basea-se na comparação de seqüências. A medida de comparação denomina-se Distância de Compressão Normalizada e fundamenta-se na Complexidade de Kolmogorov. Apesar de ser uma medida apenas teórica, a Complexidade de Kolmogorov de uma seqüência pode ser aproximada via o tamanho da seqüência comprimida.

A metodologia apresentada para classificação de ncRNAs mostrou-se satisfatória, gerando bons resultados para os testes realizados. Os resultados apresentados mostram que, de um total de 1342 seqüências, 1094 delas foram classificadas corretamente através da nossa metodologia. O bom desempenho da DCN na classificação de seqüências de ncRNAs torna-se mais evidente quando a comparamos com a similaridade. Utilizando essa medida, apenas 531 foram corretamente classificadas.

Dada a importância biológica da classificação de ncRNAs, pretendemos continuar o estudo desse problema. Isso envolve a avaliação de outras medidas de comparação assim como a análise de possíveis variações de nossa metodologia.

# Referências Bibliográficas

- [1] Adi, S. S. *Identificação de Genes por Comparação de Sequências*. PhD thesis, Universidade de São Paulo (USP), 2005.
- [2] Bennett, C. H.; Gács, P.; Li, M.; Vitányi, P. M. B. and Zurek, W. Information distance. *IEEE TIT: IEEE Transactions on Information Theory*, 44, 1998. [3.3.2](#)
- [3] Blelloch, G. E. *Introduction to Data Compression*, 2001. Canegie Mellon University. [3.2](#)
- [4] Cilibrasi, R. The CompLearn Toolkit, 2003. *IEEE Transactions on Information Theory*. [3.3.2](#)
- [5] Cilibrasi, R. and Vitányi, P. M. B. Clustering by compression. *Information Theory, IEEE Transactions on*, 51(4):1523–1545, 2005. [1](#), [3.3.2](#)
- [6] Deutsch, P. L. Gzip file format specification. RFC 1952, May 1992. [3.2](#)
- [7] Dias Correia, J. H. R. Funcionalidades dos RNA não codificantes (ncRNA) e pequenos RNA reguladores, nos mamíferos. *REDVET Revista Electrónica de Veterinária*, 8(10), 2007.
- [8] Eddy, S. R. Non-coding rna genes and the modern rna world. *Nat Rev Genet*, 2(12):919–929, December 2001.
- [9] Huffman, D. A. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952. [3.2](#)
- [10] Kin, T.; Yamada, K.; Terai, G.; Okida, H.; Yoshinari, Y.; Ono, Y.; Kojima, A.; Kimura, Y.; Komori, T. and Asai, K. . frnadb: a platform for mining/annotating functional rna candidates from non-coding rna sequences. *Nucleic Acids Res*, 35(Database issue), January 2007. [1](#)
- [11] Kocsor, A.; Kertész-Farkas, A.; Kaján, L. and Pongor, S. Application of compression-based distance measures to protein sequence classification: a methodological study. *Bioinformatics*, 22(4):407–412, February 2006.
- [12] Li, M.; Chen, X.; Li, X.; Ma, B. and Vitányi, P. M. B. The similarity metric. *Information Theory, IEEE Transactions on*, 50(12):3250–3264, 2004.

- [13] Liu, C.; Bai, B.; Skogerbo, G.; Cai, L.; Deng, W.; Zhang, Y.; Bu, D.; Zhao, Y. and Chen, R. Noncode: an integrated knowledge database of non-coding rnas. *Nucleic Acids Res*, 33(Database issue), January 2005. [1](#)
- [14] Meidanis, J. and Setubal, J. C. *Introduction to Computational Molecular Biology*. PWS Publishing Co., 1997. Instituto de Computação - Universidade de Campinas (UNICAMP). [1](#), [2](#), [3](#)
- [15] Mena-Chalco, J. P. Identificação de regiões codificantes de proteína através da transformada modificada de Morlet. Master's thesis, Instituto de Matemática e Estatística da Universidade de São Paulo, 2005.
- [16] Pang, K. C.; Stephen, S.; Engström, P. G.; Tajul-Arifin, K.; Chen, W.; Wahlestedt, C.; Lenhard B.; Hayashizaki, Y. and Mattick, J. S. Rnadb-a comprehensive mammalian noncoding rna database. *Nucleic Acids Research*, 33(Supplement 1):D125+, January 2005.
- [17] Pavlov, I. LZMA SDK (Software Development Kit), 1998. <http://www.7-zip.org/sdk.html>, 02-12-2008. [3.2](#)
- [18] Rissanen, J. J. and Langdon, G. G., Jr. Arithmetic coding. *IBM Journal of Research and Development*, 23(2):149–162, 1979. [3.2](#)
- [19] Seward, J. bzip2 and libbzip2, A program and library for data compression, 1996. [3.2](#)
- [20] Storz, G. et al. An expanding universe of noncoding rnas. *Science (New York, N.Y.)*, 296(5571):1260–1263, May 2002.
- [21] Welch, T. A. A technique for high-performance data compression. *Computer*, 17(6):8–19, 1984. [3.2](#)
- [22] Ziv, J. and Lempel, A. A universal algorithm for sequential data compression. *Information Theory, IEEE Transactions on*, 23(3):337–343, 1977. [3.2](#)