Predição de Sítios de Início de Tradução (SIT) usando Redes Neurais Artificiais, K vizinhos mais próximos e Bagging.

Cristiane Neri Nobre

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Bioinformática

Predição de Sítios de Início de Tradução (SIT) usando Redes Neurais Artificiais, K vizinhos mais próximos e Bagging.

Cristiane Neri Nobre

Orientador: *Dr. Antônio Pádua Braga*

Co-orientador: *Dr. José Miguel Ortega*

Exame de Qualificação submetido à Banca Examinadora designada pelo Programa de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais.

Belo Horizonte, Maio de 2005

Sumário

Lista de Figuras

Lista de Abreviaturas

CDS	Seqüências Codificadoras - CoDing Sequence
DNA	Ácido Desoxiribonucleico
cDNA	Ácido Desoxiribonucleico complementar
ESTs	Seqüências de Expressão - Expressed Sequence Tags
IA	Inteligência Artificial
KNN	K-Nearest-Neighbor
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
RNA	Ácido Ribonucléico
RN	Rede Neural Artificial
mRNA	Ácido Ribonucléico Mensageiro
RefSeq	Reference Sequences
RNAs	Redes Neurais Artificiais
SITs	Sítios de Início de Tradução
SVM	Support Vector Machines
Trepan	TREes PArroting Networks
UTRs	UnTranslated Regions

CAPÍTULO 1

Introdução

1.1 Identificação do Sítio de Início da Tradução - Uma breve introdução

Sistemas vivos são conhecidos pelas proteínas que produzem de acordo com sua informação genética. No entanto, somente parte das seqüências dos nucleotídeos carregando esta informação codifica proteínas (CDS -CoDing Sequence - para seqüência codificadora), enquanto outras partes não (UTR - UnTranslated Regions - para regiões não traduzidas) (Zien et al., 2000). Desta forma, dado um fragmento de DNA ou mRNA[1], um problema central da biologia computacional é determinar se ele contém CDS; e, a partir dai, descobrir qual proteína ele codifica.

De acordo com Pedersen (Pedersen e Nielsen, 1997), o reconhecimento de Sítios de Início de Tradução (SITs) em eucariotos nem sempre inicia-se na primeira metionina (AUG) em seqüências de mRNA ou cDNA[2]. Isto faz com que a predição de SITs não seja uma tarefa trivial, especialmente quando analisando ESTs[3] e da-

[1] RNA mensageiro, o produto da transcrição do DNA genômico. O mRNA pode ser editado pela célula para remover íntrons (em eucariotos) ou em outras formas que resultem em diferenças em relação ao DNA genômico transcrito (Gibas e Jambeck, 2001).

[2] Uma seqüência de DNA gerada artificialmente por transcrição reversa do mRNA. O cDNA representa aproximadamente os componentes codificantes da região do DNA genômico que produziu o mRNA (Gibas e Jambeck, 2001).

[3] ESTs são seqüências curtas de cDNA preparadas a partir de mRNA extraído de uma célula em condições específicas ou em fases de desenvolvimento específicas. As ESTs são utilizadas para identificação rápida de genes, e não abrangem a seqüência codificante completa de um gene (Gibas e

dos genômicos, onde o mRNA completo não é conhecido ou, muitas vezes, quando é conhecido não é livre de erros.

O primeiro trabalho para identificação do SIT em seqüências de cDNA data de 1984, quando Kozak (Kozak, 1984) desenvolveu a primeira amostra de consenso para uma grande coleção de dados. Este consenso é GCCACC**atg**G, onde a Guanina (G) aparece na posição +4 (posição que segue o códon ATG) e uma purina, preferencialmente a Adenina (A), aparece na posição -3. Ou seja, estas bases, nestas posições, são extremamente importantes para a identificação correta do SIT.

Em eucariotos, o modelo de escaneamento supõe que os ribosomos se ligam primeiro à região 5' do mRNA e percorre em direção à região 3' até encontrar o primeiro AUG da seqüência (Kozak, 1984; Kozak, 1999; Kozak, 1989). Desta forma, começa-se a tradução dos códons para os aminoácidos. Esta é a teoria mais utilizada para identificação do SIT. No entanto, existem exceções: devido a um contexto pobre (com ruídos, por exemplo), este primeiro AUG pode ser ignorado. A tradução pode ainda acontecer com um códon diferente do AUG, mas isto é raro em eucariotos (Pain, 1996; Kozak, 1999; Hatzigeorgiou, 2002).

Pedersen e Nielsen (Pedersen e Nielsen, 1997) comunicaram um trabalho onde uma Rede Neural Artificial (RN) foi treinada com uma janela de 203 nucleotídeos para uma base de vertebrados e o desempenho da rede foi de 85%.

Zien e colaboradores (Zien et al., 2000) usando Support Vector Machines (SVM), obtiveram resultados melhores para a mesma base de dados usada por Pedersen e Nielsen. A taxa de precisão obtida com o uso de SVM foi de 88.1%.

De acordo com Zeng (Zeng et al., 2002), técnicas de aprendizado de máquina têm sido usadas com bastante sucesso na identificação do SIT. Os autores conseguiram um desempenho de 90% usando a mesma base de dados de Pedersen e Nielsen.

Mais recentemente, (Hatzigeorgiou, 2002) apresenta um programa, DIANA-TIS, usando RNs treinada com seqüências humanas. A base de dados utilizada contêm seqüências de cDNA completa e um desempenho de 94% é obtido.

Em relação aos estudos realizados até o momento, observa-se que:

Jambeck, 2001).

- Esses autores têm trabalhado apenas com metazoários superiores, desconsiderando eucariotos mais primitivos;
- O desempenho dos classificadores utilizados fica próximo de 86% para a grande maioria dos experimentos conduzidos;
- Além disso, foram realizados poucos testes sobre a qualidade das seqüências.

Desta forma, o objetivo deste trabalho é identificar o SIT usando sistemas de aprendizado existentes na literatura: *Perceptron* simples (Zurada, 1992; Braga et al., 2000; Rosenblatt, 1958; Minsky e Papert, 1969), *Multilayer perceptron* (Zurada, 1992; Braga et al., 2000; Sethi e Yoo, 1994), KNN (K-Nearest-Neighbor) (Khan, 2002; Cover e Hart, 1967; Dasarathy, 1991; Jiawei e Micheline, 2001; Morin e Raeside, 1981), *Bagging* com *perceptron* simples e *Bagging* com *Multilayer perceptron* (Zurada, 1992; Braga et al., 2000; Breiman, 1997; Chawla et al., 2003). Para isto, foram extraídas seqüências de cinco organismos (*Mus musculus*, *Rattus norvegicus*, *Homo sapiens*, *Danio rerio* e *Drosophila melanogaster*) sob três diferentes graus de inspeção (revisado, *provisional* e *predicted*), a partir da base de dados RefSeq (Reference Sequences) do NCBI (National Center for Biotechnology Information).[4]

Pretende-se ainda representar o conhecimento adquirido por estes sistemas de aprendizado em forma de árvore de decisão ou regras *if ... then*, paralelamente ao mascaramento específico de algumas das posições do consenso, para identificar quais os sinais mais relevantes e, conseqüentemente, sugestão do mecanismo de reconhecimento dos padrões (Sethi e Yoo, 1994; Quinlan, 1986). Esta é uma das principais razões pelas quais têm sido realizadas pesquisas para a extração de conhecimento de seqüências biológicas (Browne et al., 2003) e (Maddouri e Elloumi, 2002).

Já existem na literatura vários algoritmos de aprendizado, tanto simbólicos quanto conexionistas, que representam o conhecimento por meio de regras e/ou por árvores de decisão. O C4.5 (Quinlan, 1993), um dos algoritmos mais conhecidos da abordagem simbólica, e o Trepan (Mark e Craven, 1996; Andrews e Diederich, 1995;

[4] http://www.ncbi.nlm.nih.gov.

Braga et al., 2000), um dos mais conhecidos da abordagem conexionista, extraem o conhecimento por meio de árvores de decisão e de regras. O CN2 (Clark e Niblett, 1988; Clark e Boswell, 1991), muito conhecido da abordagem simbólica, representa o conhecimento por meio de regras, apenas.

Estes algoritmos já têm sido aplicados a alguns problemas biológicos (incluindo a predição do sítio de *splicing*) (Browne et al., 2003; Maddouri e Elloumi, 2002). No entanto, a sua aplicação ainda é muito recente; na predição de SIT, por exemplo, não existe nenhum trabalho publicado até o momento.

Desta forma, este trabalho também pretende avaliar várias técnicas de classificação complexa utilizando RNAs de forma a classificar, de forma rápida e analítica, aquelas posições que iniciam a codificação ou não de proteínas.

1.1.1 Sistemas de aprendizado

A idéia principal dos sistemas de aprendizado é aprender a partir de exemplos. Um exemplo é um par $(x, f(x))$, onde x é a entrada e $f(x)$ é a saída da função aplicada a x.

Existem basicamente três tipos de sistemas de aprendizado: aprendizado supervisionado, não-supervisionado e aprendizado por reforço. No aprendizado supervisionado, é fornecido um conjunto de exemplos de treinamento da forma $\langle \vec{x}, y \rangle$, onde y representa a variável a ser identificada e $\vec{x}$ é o vetor de entrada representando as características relevantes para se determinar y. O objetivo do aprendizado supervisionado é induzir y a partir do vetor $\vec{x}$. Isto é, deve ser construído um modelo $\hat{y} = f(\vec{x})$, de função desconhecida f, que permita predizer valores de y para exemplos desconhecidos.

No aprendizado não supervisionado, é fornecido um conjunto de exemplos de treinamento, mas cada exemplo consiste somente do vetor $\vec{x}$, sendo que o valor de y não é fornecido. O objetivo do aprendizado não supervisionado é construir um modelo que represente as regularidades do conjunto de treinamento. A natureza dos modelos construídos pelos algoritmos não supervisionados varia muito de método para método. Por exemplo, existem métodos não supervisionados que exploram seus

dados de treinamento estimando as funções de distribuição de probabilidade (Silverman, 1986), construindo categorizações hierárquicas e reduzindo os dados para um espaço de dimensão menor representando a maioria de suas variáveis.

O aprendizado por reforço é descrito como o processo iterativo de mapeamento de um sinal de entrada em um sinal de saída através de um processo de tentativa e erro, onde busca-se maximizar a medida de desempenho do sistema através de sinais de reforço (ou enfraquecimento) das respostas fornecidas (Zurada, 1992) e (Braga et al., 2000).

Este trabalho está baseado em aprendizado supervisionado. Mais especificamente, focalizar-se-á em aprendizado de classificação supervisionado. Nesse caso, o objetivo é predizer um valor discreto $\hat{y}$ a partir de um vetor $\vec{x}$. Em outras palavras, cada vetor $\vec{x}$ é destinado para predizer um conjunto específico de classes.

1.1.2 Critérios de avaliação de modelos de aprendizado

A consideração mais importante em sistemas de aprendizado é obter um modelo que tenha um alto grau de generalização. Isto significa que, para o aprendizado supervisionado, deseja-se um modelo que seja capaz de predizer um valor y a partir de um vetor $\vec{x}$ que não esteja preferencialmente nos dados de treinamento. De fato, não é muito difícil predizer um valor correto para y a partir de exemplos que estejam neste conjunto.

Uma outra consideração importante na escolha de um método de aprendizado é a flexibilidade da linguagem usada pelo algoritmo para representá-lo em hipóteses (Mark e Craven, 1996)[5]. É desejável que os modelos de aprendizado usem representações declarativas como forma de representação. O modelo não pode ser restrito a um procedimento particular, nem a tarefas muito específicas.

Acrescidos à generalização, compreensibilidade e flexibilidade representacional, existem outros critérios que são frequentemente utilizados na avaliação de um algoritmo de aprendizado ou das hipóteses que ele produz. Estes critérios incluem a eficiência dos algoritmos na indução de hipóteses, a eficiência do algoritmo em

[5]Uma hipótese corresponde aos resultados obtidos do modelo de aprendizado.

classificar novos exemplos e a facilidade com que o algoritmo pode se adaptar às hipóteses atuais para o mais novo exemplo aprendido.

Existe um número muito grande de razões pelas quais a compreensibilidade é uma consideração importante quando se fala em sistemas de aprendizado. A seguir, serão apresentadas algumas destas razões:

- **Validação:** Frequentemente os usuários necessitam saber como o sistema de aprendizado tomou determinada decisão. A habilidade em analisar as hipóteses aprendidas é muito importante em alguns domínios. Isso é verificado, por exemplo, em problemas biológicos (Maddouri e Elloumi, 2002);
- **Integração entre Sistemas Conexionistas e Simbólicos:** As regras, além de possibilitarem uma descrição simbólica concisa e apurada do classificador, facilitam também a comunicação com outros sistemas baseados em conhecimento, como os sistemas especialistas;
- **Exploração de dados e indução de teorias:** Se não for possível o entendimento dos resultados obtidos a partir de um classificador não será possível haver indução de teorias;
- **Explanação:** O objetivo da explanação é fornecer ao usuário a capacidade de explorar o conhecimento do sistema. Idealmente, é preciso que a explanação responda a questões sobre aspectos relevantes a respeito do conhecimento do sistema e as ações tomadas. Do ponto de vista do usuário, estas respostas precisam ser completas e de fácil compreensão.

1.2 Justificativa

Como já foi mencionado anteriormente, apesar das referências apresentadas, ainda existem poucos trabalhos relacionados ao tema "Predição do SIT". E talvez, devido a isto, algumas necessidades, que são de suma importância neste contexto, não foram trabalhadas (ou o foram de forma parcial). Dentre as principais necessidades, podem-se destacar: i) comparação, a partir do SIT, entre diversos organismos, ii)

estudos sobre eucariotos mais primitivos, iii) estudo formal sobre a qualidade de anotação das seqüências utilizadas e, iv) extração de conhecimento na forma de regras do tipo *if . . . then* para este problema.

Outro ponto observado é que muitas vezes têm-se poucas seqüências de um determinado organismo. Com base nisso, pretende-se analisar se o treinamento com um determinado organismo pode ser utilizado para validar seqüências de outro. Com isso, poderemos comparar organismos com base no estudo do SIT e, dependendo dos resultados, não ficarmos limitados ao tamanho das bases.

Desta forma, este trabalho vem contribuir de forma significativa para a área de bioinformática, em relação a predição do SIT.

1.3 Objetivos

Neste sentido, este trabalho tem como principais objetivos:

- Extrair seqüências de vertebrados (*H. sapiens*, *M. musculus*, *R. norvegicus* e *D. rerio*, e também de eucariotos primitivos (*Saccharomyces cerevisiae*, *Schizosaccharomyce ponbe*, *Caenorhabditis elegans*) e planta (*Arabidopsis thaliana*) do banco de dados público RefSeq do NCBI. Este estudo objetiva também contribuir para a Rede Genoma do Estado de Minas Gerais, utilizando-se o transcriptoma do *Shistosoma mansoni*; em uma fase posterior do trabalho, estenderemos a análise ao transcriptoma de procariotos (*Escherichia coli* e *Xylella fastidiosa*);

- Implementar métodos de aprendizado para predição do SIT;

- Fornecer uma interpretação biológica para as seqüências extraídas;

- Verificar se é possível validar eficientemente o reconhecimento de seqüências de um dado organismo, sendo que um outro organismo foi utilizado na fase de treinamento;

- Com base no item anterior, fazer uma comparação entre os organismos analisados.

- Extrair conhecimento a partir dos classificadores implementados, com o objetivo de sugerir o mecanismo de reconhecimento, tornando a representação dos resultados mais intuitiva;

1.4 Visão geral deste trabalho

Os capítulos restantes deste trabalho serão organizados da seguinte maneira:

O capítulo 2 apresenta uma descrição dos sistemas de aprendizado utilizados.

O capítulo 3 apresenta os materiais e métodos utilizados.

O capítulo 4 apresenta os primeiros resultados obtidos. Uma análise detalhada destes resultados também será apresentada.

E, finalmente, no capítulo 5 serão apresentadas as conclusões e propostas de continuidade deste trabalho.

CAPÍTULO 2

Descrição dos Classificadores Utilizados

2.1 Redes Neurais Artificiais

Existe uma grande variedade de arquiteturas de RNs e métodos de aprendizado supervisionado e não-supervisionado. A arquitetura utilizada neste trabalho foi uma rede *feedforward* aplicada a um problema de classificação (identificação do SIT). Desta forma, a seção seguinte trata deste tipo específico de Rede Neural.

2.1.1 Classificação por Redes Neurais Artificiais

A figura 2.1 apresenta uma rede *feedforward* composta por várias camadas de unidades de processamento simples. O estado de uma unidade é determinado pela ativação representada por um número real, variando entre [0,1] ou [-1,1].

A estrutura *feedforward* pode ser descrita como um conjunto de unidades de processamento disposto em camadas que se interconectam, uma após a outra, chegando até uma camada de saída. Na forma mais simples, conhecida como uma rede de uma única camada, tem-se uma camada de entrada por onde são apresentados os padrões ou os estímulos para a rede vindos do ambiente e uma camada de saída que de fato realiza as computações na rede. O termo única camada é associado justamente a esta característica pois as unidades na camada de entrada não realizam qualquer computação, servindo apenas para a passagem dos sinais de

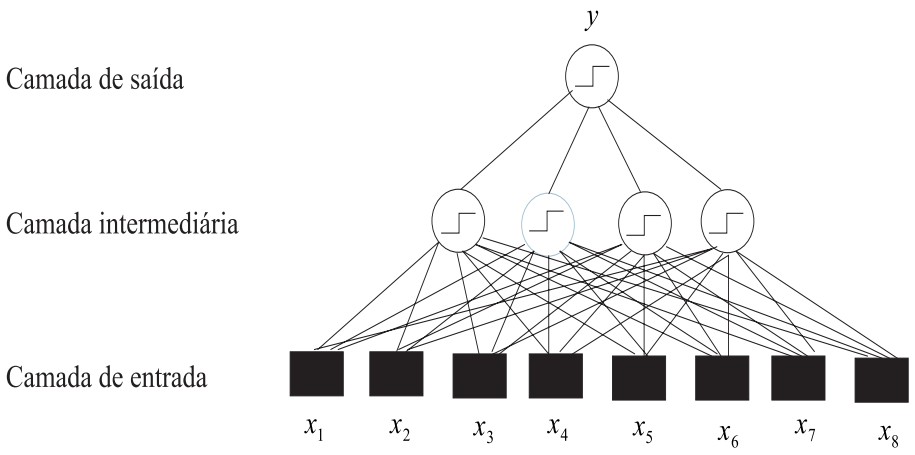


Figura 2.1: Uma rede neural *feedforward* com uma única saída. As unidades na camada de entrada correspondem aos atributos do problema. A unidade de saída corresponde às predições realizadas pela rede. As unidades intermediárias realizam o mapeamento entre as unidades de entrada e saída.

entrada à camada de saída (Zurada, 1992; Braga et al., 2000).

Uma rede para classificação que tenha somente unidades de entrada e de saída é capaz de representar somente regiões de decisão lineares limitadas por seu espaço de entrada. Para representar limites mais complexos é necessário adicionar unidades intermediárias à rede. Estas unidades transformam o espaço de entrada em outro fazendo com que as unidades de saída façam discriminações não lineares.

Desta forma, a entrada para as unidades da camada intermediária e de saída de uma rede *feedforward* é dada pela equação 2.1:

$$y = \sum_{j} w_{ij} x_i + \theta_i. \tag{2.1}$$

onde w_{ij} é o peso entre a unidade j a unidade i, x_j é o valor das unidades de entrada j e θ_i é o limiar da unidade i. O limiar de uma unidade é um parâmetro que determina a predisposição da unidade em ter uma ativação alta (ou baixa) antes de receber sinais de ativação de outras unidades. A ativação de uma unidade intermediária ou de saída é resultante de uma função de ativação, também conhecida por função de transferência. As funções logística e também hiperbólica são as mais utilizadas.

A primeira limita os valores de ativação numa faixa de [0,1]. A segunda limita os valores em [-1,1].

2.1.2 Aprendizado com Redes Neurais Artificiais

As Redes Neurais Artificiais possuem a capacidade de aprender a partir de exemplos e de fazer interpolações e extrapolações do que aprenderam (Braga et al., 2000). O objetivo do aprendizado conexionista é determinar a intensidade das conexões entre neurônios. Um conjunto de procedimentos bem definido para adaptar os parâmetros de uma RNA para que a mesma possa aprender uma determinada função é chamado de algoritmo de aprendizado. Existem vários algoritmos de aprendizado e eles diferem entre si, basicamente, pela maneira através da qual o ajuste dos pesos é feito.

A utilização de uma RNA na solução de uma tarefa passa inicialmente por uma fase de aprendizagem, onde a rede extrai informações relevantes de padrões de informações apresentados para a mesma, criando assim uma representação própria para o problema. A etapa de aprendizagem consiste em um processo iterativo de ajuste de parâmetros da rede, os pesos das conexões entre as unidades de processamento, que guardam, ao final do processo, o conhecimento que a rede adquiriu do ambiente em que está operando. Uma definição geral do que vem a ser aprendizagem pode ser expressa da seguinte forma (Mark e Craven, 1996; Andrews e Diederich, 1995): Aprendizagem é o processo pelo qual os parâmetros de uma rede neural são ajustados através de uma forma continuada de estímulo pelo ambiente no qual a rede está operando, sendo o tipo específico de aprendizagem realizada definido pela maneira particular como ocorrem os ajustes realizados nos parâmetros. Diversos métodos para treinamento de redes foram desenvolvidos, podendo estes serem agrupados nos seguintes paradigmas: aprendizado supervisionado e aprendizado não supervisionado, aprendizado por reforço e aprendizado por competição. Os três primeiros paradigmas falados anteriormente já foram descritos na seção 1.1.1 e o aprendizado por competição é um caso particular do aprendizado não supervisionado.

Com foi falado no início desta seção este trabalho está baseado em aprendizado

supervisionado e mais especificamente, focalizar-se-á em aprendizado de classificação supervisionado. O objetivo então será predizer um valor discreto $\hat{y}$ a partir de um vetor $\vec{x}$.

2.2 Bagging

A idéia principal do classificador *Bagging* é selecionar (com reposição) n pontos do conjunto de treinamento e usar m classificadores (que neste caso foi o perceptron e MLP) para classificar os padrões de teste. O classificador ótimo será definido pelo voto majoritário. Ou seja, apresenta-se um padrão teste para os m classificadores. A partir das respostas, escolhe-se qual a classe mais votada para este padrão. Isto é repetido para todos os padrões do conjunto de teste (Breiman, 1997; Chawla et al., 2003).

A vantagem do *Bagging* é reduzir a variância ou instabilidade do classificador utilizado. Desta forma, *Bagging* tende a cancelar parte do ruído enquanto varia entre os membros das amostras.

2.3 KNN

Esta metodologia de classificação consiste na identificação de grupos de indivíduos com características similares e seu posterior agrupamento (*clustering*) (Cover e Hart, 1967; Dasarathy, 1991; Jiawei e Micheline, 2001; Morin e Raeside, 1981).

O método aqui implementado, KNN, ou *K Nearest Neighbor* (K vizinhos mais próximos) utiliza o conceito de distância entre amostras, considerando-se uma amostra um vetor (linha) contendo várias colunas (variáveis). Neste sentido, uma medida de similaridade entre pontos poderia ser baseado nas várias distâncias entre um ponto a ser classificado em uma dada classe e pontos de classes conhecidas.

Isto quer dizer que, se duas classes **A** e **B** possuem vários pontos em seus domínios, dado um ponto desconhecido x, este ponto será classificado em função da quantidade de pontos cujas distâncias forem as menores possíveis em relação às classes **A** e **B**.

Assumindo i amostras conhecidas e duas classes 0 e 1, o algoritmo de classificação pelo método KNN pode ser descrito como indicado na tabela 2.1 (Cover e Hart, 1967; Dasarathy, 1991):

```
Início
    Ler n amostras de classes conhecidas (ac)
    Ler amostra de classe desconhecida (ad)
    iniciar i = 1
    Repita
        Computar a distância de ad para aci
        Atribuir a distância e a classe de aci ao vetor de distâncias
        Incrementar i = i + 1
    Até (computar todas as distâncias, isto é, i > n)
    Ordenar o vetor de distâncias por ordem de distância
    Selecionar as K primeiras posições do vetor de distâncias
    iniciar i = 1
    Iniciar classe0 = 0
    Iniciar classe1 = 0
    Repita
        Comparar a classe
        Se (classe igual a 0)
            Incrementar classe0 = classe0 + 1
        Senão
            Incrementar classe1 = classe1 + 1
        Fim Se
        Incrementar i = i + 1
    Até (computar todas as K amostras, isto é, i > K)
Fim do Algoritmo
```

Tabela 2.1: Uma visão geral do algoritmo do método classificação KNN.

Desta forma, dado um padrão de teste (desconhecido) x, sua classificação é realizada da seguinte maneira:

- Inicialmente, calcula-se a distância entre x e todos os padrões de treinamento;
- Verifica-se a quais classes pertencem os K padrões mais próximos;
- A classificação é feita associando-se o padrão de teste à classe que for mais freqüente entre os K padrões mais próximos de x.

Há vários tipos de distâncias que normalmente são adotadas para implementar esse classificador. As distãncias de Manhattan, Minkowsk (ou norma Lk) e Euclidiana são as mais utilizadas (Khan, 2002). Neste trabalho utilizou-se a distância Euclidiana.

No próximo capítulo, serão descritos os materiais e métodos utilizados neste trabalho.

CAPÍTULO 3

Materiais e Métodos

3.1 Conjunto de Dados

Para a identificação do SIT, foram obtidas seqüências positivas (que iniciam a tradução dos mRNAs em proteínas) e negativas (que não iniciam tradução; por segurança, contém o ATG fora de fase de leitura considerada correta (de acordo com a figura 3.2) de bases de dados curadas *RefSeq* do NCBI para os seguintes organismos: *Mus musculus* (camundongo), *Rattus norvegicus* (rato), *Homo sapiens* (homem), *Danio rerio* (peixe) e *Drosophila melanogaster* (mosca). Estas seqüências foram utilizadas para alimentar os classificadores mencionados nos capítulos anteriores.

As seqüências de referência (RefSeq) fornecem uma coleção não redundante de seqüências de DNA, RNA e proteínas. No RefSeq, a redundância do GenBank[1] é retirada.

As principais características do RefSeq são:

- Não redundância;
- Seqüências de nucleotídeos e proteínas explicitamente linkados;

[1] GenBank é o banco de dados de seqüências genéticas do NIH (National Institutes of Health), uma coleção anotada de todas as seqüências de DNA disponíveis publicamente. GenBank e PIR (Protein Information Resource) são bases de dados norte-americanas, EMBL (European Molecular Biology Laboratory) é uma base de dados européia, DDBJ (DNA Data Bank of Japan) é uma base japonesa e a Swiss-Prot é suíça.

- Validação dos dados e consistência no formato dos arquivos (normalmente dois caracteres, seguido pelo undescore e seis dígitos);
- Curado pelo staff do NCBI e colaboradores, com *status* de revisão, indicados em cada registro.

As seqüências de referência existem sob três níveis de confiança: *reviewed*, *provisional* e *predicted*, que corresponde à diminuição do grau de inspeção. As seqüências *reviewed*, por serem revisadas por membros do *staff* do NCBI e seus colaboradores, são, de maneira geral, as melhores seqüências disponíveis de um determinado organismo.

A figura 3.1 apresenta um fragmento de um arquivo original extraído do NCBI.

```
LOCUS       IL2                1047 bp    mRNA    linear   PRI 31-JAN-2003
DEFINITION  Homo sapiens interleukin 2 (IL2), mRNA.
ACCESSION   NM_000586
VERSION     NM_000586.2  GI:28178860KEYWORDS
SOURCE      Homo sapiens (human)
ORGANISM  Homo sapiens          Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
...
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff.
...
CDS         295..756

1   cgaattcccc tatcacctaa gtgtgggcta atgtaacaaa gagggatttc acctacatcc
61  attcagtcag tctttggggg tttaaagaaa ttccaaagag tcatcagaag aggaaaaatg
121 aaggtaatgt tttttcagac aggtaaagtc tttgaaaata tgtgtaatat gtaaaacatt
181 ttgacacccc cataatattt ttccagaatt aacagtataa attgcatctc ttgttcaaga
241 gttccctatc actctcttta atcactactc acagtaacct caactagcga ccgaatgaat
301 atgaatggac tcctgtcttg cattgcacta agtcttgcac ttgtcacaaa cagtgcacct
...
```

Figura 3.1: Fragmento de um arquivo RefSeq no formato original. Este arquivo contém informações tais como organismo, número de acesso, nível de confidência, posição de início do CDS, dentre outras.

Através deste arquivo, podemos verificar que o organismo em questão é o *Homo sapiens* (seqüência revisada), com número de acesso igual NM_000586 e cujo CDS é iniciado no nucleotídeo 295; ou seja, o início da tradução desta seqüência começa na posição 295 do mRNA.

A partir de arquivos neste formato, seqüências positivas e negativas foram extraídas através de um *parsing*, que selecionava seqüências positivas de 13 bases próximo ao códon ATG anotado (9 bases antes e 1 depois do códon inicial) e seqüên-

cias negativas usando códons fora da fase de leitura, conforme ilustrado na figura 3.2.

Em (Pedersen e Nielsen, 1997) existe uma indicação de que o tamanho da janela varia entre 13 a 203 nucleotídeos e que o tamanho ideal (que fornece os melhores resultados) é de 203 (100 bases antes e 100 depois do códon inicial). Este artigo mostra que o melhor desempenho obtido foi de 85% para os vertebrados analisados (*Bos taurus* (vaca), *Gallus gallus* (galinha), *Homo sapiens* (homem), *Mus musculus* (camundongo), *Oryctolagus cuniculus* (rato), *Sus scrofa* (porco) e *Xenopus laevis* (rã africana)).

Todavia, neste trabalho, optamos por usar uma janela com apenas 13 bases, favorecendo o estudo do padrão apresentado originalmente por Kokaz (Kozak, 1984; Kozak, 1986). Como os resultados estavam bastante promissores testou-se com este tamanho, apenas. No entanto, serão testados neste trabalho janelas com tamanhos ainda menores com o objetivo principal de verificar a possibilidade de redução do tempo de treinamento e, principalmente, o efeito do mascaramento de algumas posições que podem revelar-se muito importantes.

A figura 3.2 mostra como estas seqüências foram extraídas a partir dos arquivos originais.

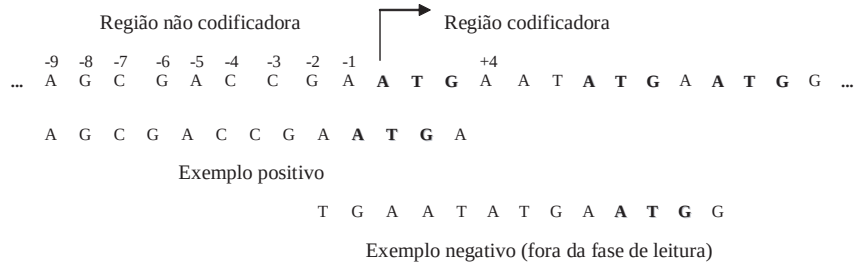


Figura 3.2: Construção das seqüências positivas e negativas usando-se uma janela de 13 nucleotídeos com códon ATG começando-se na décima posição. As sequências negativas foram obtidas fora da fase de leitura.

Testes anteriores foram realizados obtendo-se seqüências negativas na mesma fase de leitura e percebeu-se que o desempenho diminuiu, aproximadamente, em 10%.

Estas seqüências foram extraídas apenas de arquivos contendo a posição de início do CDS maior ou igual a 10 (para se conseguir as 9 bases antes do ATG). Desta forma, todas as seqüências contendo CDS abaixo deste número foram desconsideradas. A tabela 3.1 apresenta a porcentagem de dados que foi desconsiderada durante o processo de extração. Como pode ser observado, de maneira geral, a maior porcentagem foi para o *Rattus norvegicus* revisado, onde 98% das seqüências foram desconsideradas.

Tabela 3.1: Porcentagem de seqüências que foi desconsiderada durante o processo de extração das bases de dados.

Organismos	Inspeção	Porcentagem
Homo sapiens	Revisado	3%
	Provisional	10%
	Predicted	4%
Mus musculus	Revisado	6%
	Provisional	17%
	Predicted	2%
Rattus norvegicus	Revisado	98%
	Provisional	18%
	Predicted	15%
Danio rerio	Provisional	11%
	Predicted	2%
Drosophila melanogaster	Provisional	4%

Analisando-se as seqüências extraídas, percebeu-se ainda que havia um grande número de repetições. Neste sentido, eliminaram-se as repetições e os testes foram refeitos, com o objetivo de analisar o impacto desta eliminação sobre o desempenho dos classificadores.

A tabela 3.2 apresenta o número de seqüências positivas e negativas obtidas, com e sem repetição, para todos os organismos estudados.

Tabela 3.2: Número de seqüências (Revisadas, *provisional* e *predicted*), com e sem repetições.

Organismos	Inspeção	Com repetição Pos - Neg	Sem repetição Pos - Neg
Homo sapiens	Revisado	73 - 17	55 - 16
	Provisional	2879 - 2593	2740 - 2509
	Predicted	130 - 117	128 - 116
Mus musculus	Revisado	120 - 120	98 - 92
	Provisional	1718 - 1718	1659 - 1669
	Predicted	3775 - 3775	3666 - 3680
Rattus norvegicus	Revisado	19 - 17	18 - 16
	Provisional	3928 - 3928	3770 - 3871
	Predicted	132 - 132	131 - 131
Danio rerio	Revisado	1 - 1	1 - 1
	Provisional	2510 - 2510	2460 - 2468
	Predicted	2375 - 2375	2336 - 2339
Drosophila melanogaster	Provisional	329 - 256	224 - 175

Em agosto de 2004, época em que estes dados foram obtidos, havia apenas um seqüência revisada para o organismo *Danio rerio* e apenas seqüências *provisional* para a *Drosophila melanogaster*.

Para todos os classificadores utilizados (cujos desempenhos serão mostrados no capítulo seguinte), 60% das bases de dados foram utilizadas no treinamento e o restante (40%) foram utilizadas para teste.

3.2 Codificação utilizada

As entradas foram apresentadas aos classificadores *Perceptron* simples, *Multilayer perceptron*, KNN, *Bagging* com *perceptron* simples e *Bagging* com *multilayer perceptron* (descritos no capítulo 2) codificando-se as seqüências (positivas e negativas) em uma *string* binária, usando um esquema de codificação onde cada nucleotídeo é representado por 4 dígitos binários: A=0001, C=0010, G=0100 e T=1000 (codificação com espaço). Desta forma, todos os classificadores utilizados neste trabalho possuem 52 entradas (13 bases * 4 *bits*) e uma saída (1 ou 0). Ou seja, a saída é 1 se a seqüência contêm o códon AUG inicializador da tradução ou 0, no caso de a seqüência não conter este inicializador.

Para os cinco organismos estudados (*Mus musculus*, *Rattus norvegicus*, *Homo sapiens*, *Danio rerio* e *Drosophila melanogaster*) e os três tipos de seqüências (reviewed, *provisional* e *predicted*), diferentes redes *feedforward* foram testadas durante o treinamento:

- Rede *feedforward* sem unidades ocultas (*perceptron*);
- Redes *feedforward* com unidades ocultas, variando-se de 1 a 15 neurônios na camada oculta.

O objetivo do primeiro teste foi verificar se o problema em questão era linearmente separável. Com isso, além de analisarmos a complexidade do problema, o tempo de treinamento, de maneira, geral seria muito menor. Os primeiros resultados mostram que o desempenho da rede ficava em torno de 68% para o organismo *Rattus norvegicus*, por exemplo. Isto indica que o problema não é linearmente separável, demandando uma rede com uma ou mais camadas ocultas. Estes resultados estão apresentados no capítulo seguinte.

O capítulo, a seguir, apresentará os principais resultados obtidos até o momento.

CAPÍTULO 4

Resultados e Discussões

Os resultados, a seguir, serão divididos em três pontos principais: o primeiro será quanto ao desempenho dos classificadores; o segundo, quanto à análise das seqüências positivas e negativas e o terceiro será quanto a semelhanças entre os organismos, baseados no SIT.

4.1 Desempenho dos classificadores

Para se calcular o desempenho dos classificadores, três medidas foram avaliadas: a acurácia total (Ac), a sensibilidade (Sens) e a especificidade (Espec) (Zeng et al., 2002; Haifeng e Tao, 2004).

A acurácia é definida pela equação 4.1:

$$Ac = 100 * \frac{VP + VN}{VP + VN + FN + FP} \tag{4.1}$$

A sensibilidade e a especificidade são definidas por:

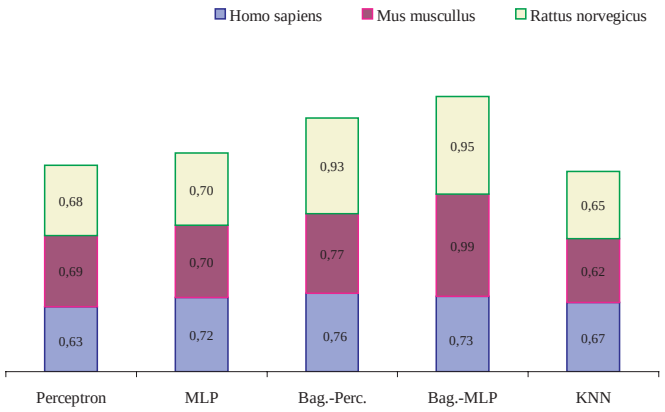
$$Sens = 100 * \frac{VP}{VP + FN} \tag{4.2}$$

$$Espec = 100 * \frac{VN}{VN + FP} \tag{4.3}$$

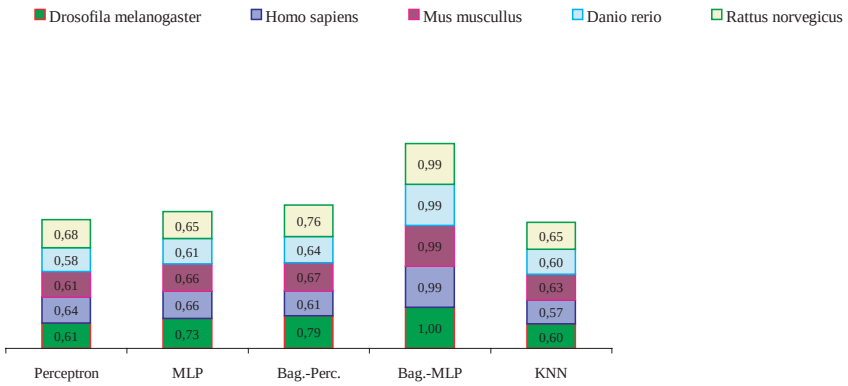
onde, VP, VN, FP e FN denotam o número de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos, respectivamente.

A acurácia refere-se à porcentagem do número de acertos da classe positiva e negativa juntas. A taxa de sensibilidade, conhecida como taxa de verdadeiro-positivo, refere-se à porcentagem de acertos dentro da classe positiva. A taxa de especificidade, também conhecida como taxa de verdadeiro-negativo, refere-se à porcentagem do número de acertos dentro da classe negativa.

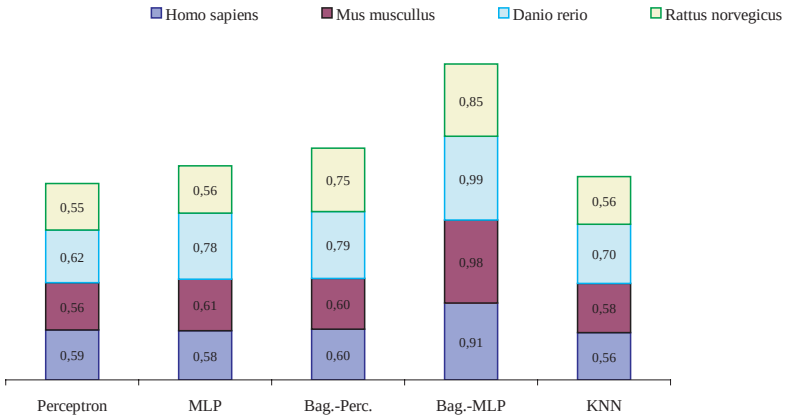
A figura 4.1 apresenta a taxa de acurácia de todos os classificadores implementados neste trabalho. Percebe-se claramente uma variação no valor da acurácia entre os diferentes classificadores. Vale a pena ressaltar que os melhores resultados foram obtidos utilizando-se o *Bagging* com *multilayer perceptron*. Como já citado, *Bagging* tende a suavizar a resposta dos membros do conjunto de classificadores, melhorando-se assim a generalização.



(a) Seqüências revisadas.



(b) Seqüências *provisional*.



(c) Seqüências *predicted*.

Figura 4.1: Desempenho dos classificadores para as seqüências revisadas, *provisional* e *predicted*. Estes resultados foram obtidos utilizando-se 60% dos dados no treinamento e o restante, 40%, nos testes.

Estes resultados foram obtidos com 2 neurônios na camada oculta e treinada com o algoritmo trainBFG (do Matlab 7.0), com 30 classificadores. A acurácia total foi calculada sobre a média entre 10 diferentes execuções. Vale destacar que, por questão de tempo, obtou-se por utilizar o *Bagging* com apenas 30 classificadores. De maneira geral, a acurácia aumenta até um determinado número de classificadores utilizados; depois disso, tem-se uma pequena diferença no desempenho (dados não mostrados). No entanto, como os resultados estavam bastante satisfatórios, obtou-se por utilizar 30 classificadores apenas.

Um ponto observado é quanto à acurácia encontrada para o *Homo sapiens* com seqüências revisadas. Observando-se o tamanho destas seqüências percebe-se um ligeiro desbalanceamento entre as classes. Vê-se que a taxa de sensibilidade, neste caso, é maior (87,3%) do que a taxa de especificidade (64,6%), sugerindo que os classificadores podem estar aprendendo mais a classe positiva do que a negativa. Testes para fazer um balancemento das classes será uma próxima etapa a ser realizada.

A acurácia obtida para o *Rattus norvegicus* com seqüências revisadas e *predicted* também é ligeiramente inferior às obtidas para os outros organismos. Novos testes serão realizados para tentar descobrir a razão deste desempenho e o que poderá ser feito para se obter melhores resultados.

Visto que os melhores resultados foram obtidos a partir do *Bagging* com *multilayer perceptron*, estas três taxas foram calculadas apenas para este classificador. Desta forma, a tabela 4.1 apresenta, para este classificador, as porcentagens da acurácia, sensibilidade e especificidade para os cinco organismos analisados, levando-se em consideração o conjunto de dados com seqüências repetidas.

Tabela 4.1: Taxas de acurácia(Ac), sensibilidade(Sens) e especificidade(Espec) para as seqüências com repetição.

Organismos	Inspeção	Tamanho Pos - Neg	Ac	Sens	Espec
Homo sapiens	Revisado	73 - 17	75,9%	87,3%	64,6%
	Provisional	2879 - 2593	99,4%	98,9%	99,7%
	Predicted	130 - 117	93,1%	88,8%	97,4%
Mus musculus	Revisado	120 - 120	98,5%	99,2%	95,7%
	Provisional	1718 - 1718	99,8%	99,8%	99,8%
	Predicted	3775 - 3775	98,5%	98,1%5	98,9%
Rattus norvegicus	Revisado	19 - 17	89,2%	95,2%	83,2%
	Provisional	3928 - 3928	99,9%	99,9%	99,9%
	Predicted	132 - 132	81,9%	79,7%	84,0%
Danio rerio	Provisional	2510 - 2510	99,5%	99,9%	99,1%
	Predicted	2375 - 2375	99,7%	99,9%	99,5%
Drosophila melanogaster	Provisional	329 - 256	99,1%	98,8%	98,5%

Como descrito no capítulo anterior, desejava-se verificar se o fato de o conjunto de dados possuir muitas seqüências repetidas poderia afetar o desempenho dos classificadores. Para isto, foram eliminadas as repetições dos conjunto de dados e os resultados obtidos estão apresentados na tabela 4.2.

Tabela 4.2: Taxas de acurácia(Ac), sensibilidade(Sens), especificidade(Espec) para as seqüências sem repetição.

Organismos	Inspeção	Tamanho Pos - Neg	Ac	Sens	Espec
Homo sapiens	Revisado	55 - 16	73,5%	74,3%	72,6%
	Provisional	2740 - 2509	99,3%	98,7%	99,9%
	Predicted	128 - 116	91,8%	91,4%	92,2%
Mus musculus	Revisado	98 - 92	99,9%	100%	99,9%
	Provisional	1659 - 1669	99,6%	99,8%	99,3%
	Predicted	3666 - 3680	98,6%	98,4%	98,9%
Rattus norvegicus	Revisado	18 - 16	95,2%	90,4%	100%
	Provisional	3770 - 3871	99,9%	99,9%	99,9%
	Predicted	131 - 131	85%	83,7%	86,5%
Danio rerio	Provisional	2460 - 2468	99,7%	99,9%	99,6%
	Predicted	2336 - 2339	99,8%	99,9%	99,7%
Drosophila melanogaster	Provisional	224 - 175	100%	100%	100%

Pelas tabelas 4.1 e 4.2, percebe-se que estas taxas tem pouca ou quase nenhuma alteração. Com isso, podemos usar apenas os dados sem as repetições, visto

que o tempo de processamento da rede é ligeiramente menor neste caso. É importante notar que os resultados por nós obtidos são sensivelmente melhores que os já relatados na literatura e que a acurácia e especificidade em torno de 99% constitui uma aplicação muito relevante para o uso em análises de seqüências, pois confere certeza quase absoluta a um operador que esteja lidando com uma seqüência de cada vez.

4.2 Análise das seqüências positivas e negativas

Um dos propósitos deste trabalho é estudar as seqüências positivas e negativas extraídas. É plausível supor que o sinal utilizado para reconhecimento do SIT possa depender de (i) um padrão reconhecido para início da tradução e um padrão randômico em torno do negativo ou (ii) um padrão positivo somado a algum sinal específico em torno do ATG negativo. O consenso sugerido por Kozak (Kozak, 1984; Kozak, 1999; Kozak, 1989) desde o início das investigações sobre esse tema sugere fortemente que o mecanismo de reconhecimento não se deve a sinalização exclusiva em torno do padrão negativo e, portanto, essa possibilidade é descartada. Neste sentido, dois tipos de análise foram realizadas:

- Primeiramente, foi feita uma análise, por posição, da frequência de nucleotídeos de ambas as seqüências;
- Foi realizada, também, uma análise da freqüência dos trinucleotídeos.

As seções seguintes apresentam os resultados destas análises.

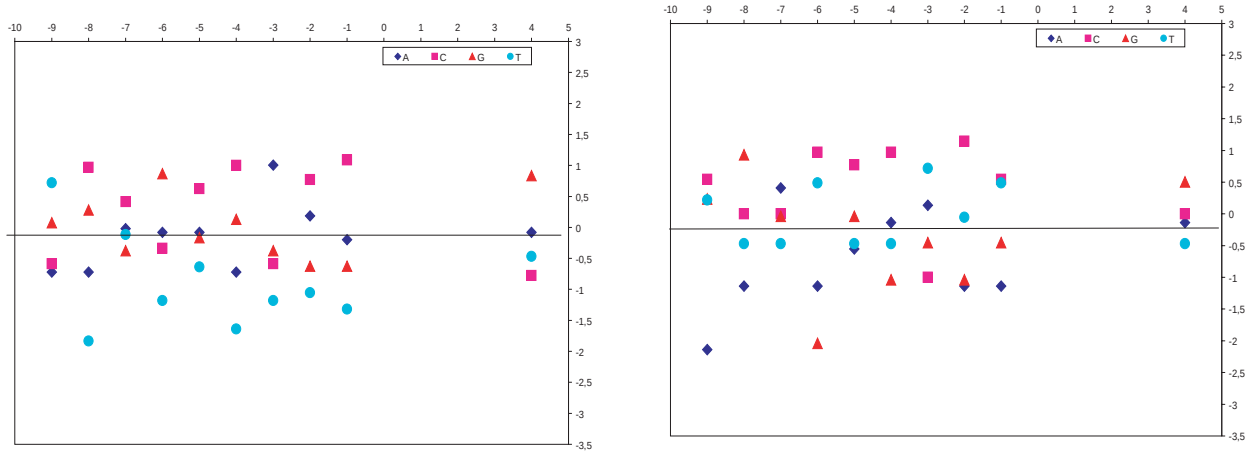
4.2.1 Frequência de bases das seqüências

Com o objetivo de verificar a freqüência de bases existentes nas seqüências positivas e negativas, foram geradas as figuras 4.2, 4.3, 4.4, 4.5 e 4.6. Estas figuras foram obtidas a partir das sequências sem repetições e estão de acordo com a equação 4.4:

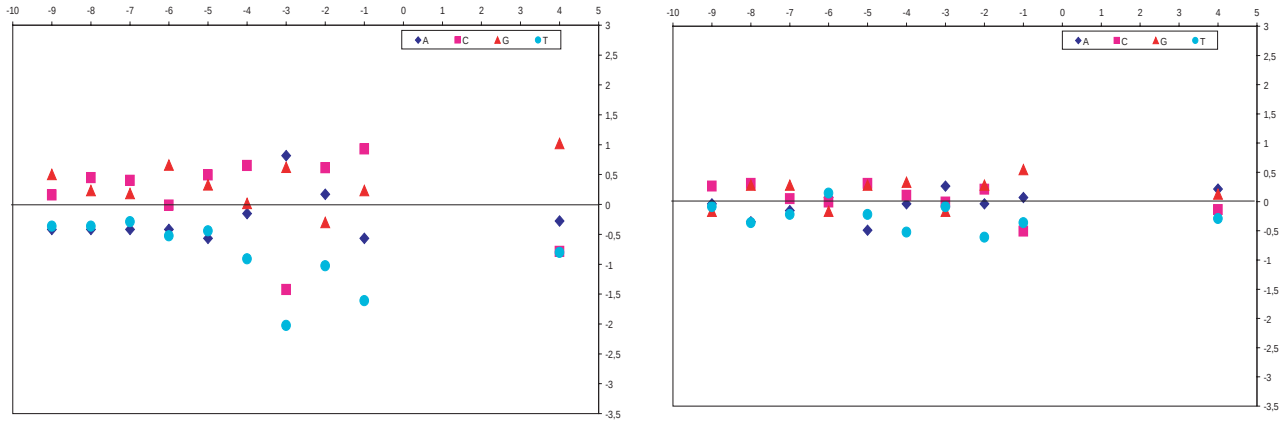
$$Rel = \log_2 \frac{FreqdeBaseEmCadaPos}{FreqDeBasenocDNA} \tag{4.4}$$

onde, FreqdeBaseEmCadaPos é a freqüência das bases A, C, G, e T das posições de de -9 a -1 e +4 das seqüências de 13 bases extraídas (as posições do ATG foram desconsideradas) e FreqDeBasenocDNA é a freqüência destas mesmas bases nos cDNAs em questão. Nesse tipo de experimento, é importante possuir um conjunto de dados que seja suficiente para demonstrar um padrão randômico quando ele é a realidade. Assim, marcamos com um "*"nos *captions* das figuras casos onde o número de seqüências utilizadas foi inferior a 1000. Com isto, não estamos dizendo que 1000 seja o número mínimo para comprovar esta aleatoriedade das seqüências negativas e muito menos que esse seja o número ideal para se obter melhores desempenhos dos classificadores. Mesmo porque, neste último caso, quando se trata de Redes Neurais, nem sempre o mais importante é o tamanho da amostra e sim a qualidade desta. Ou seja, é possível termos poucas sequências e mesmo assim obtermos desempenhos satisfatórios, desde que esta amostra (mesmo que pequena) consiga representar a totalidade dos dados. Veja, por exemplo, o caso da *Drosophila melanogaster* que tem apenas 329 sequências positivas e 256 negativas; neste caso, o classificador obteve-se uma acurácia de 99,1%. Isto sugere que estas sequências são tão boas quanto às do *Danio rerio*, por exemplo, com tamanhos muito maiores (todas acima de 2000).

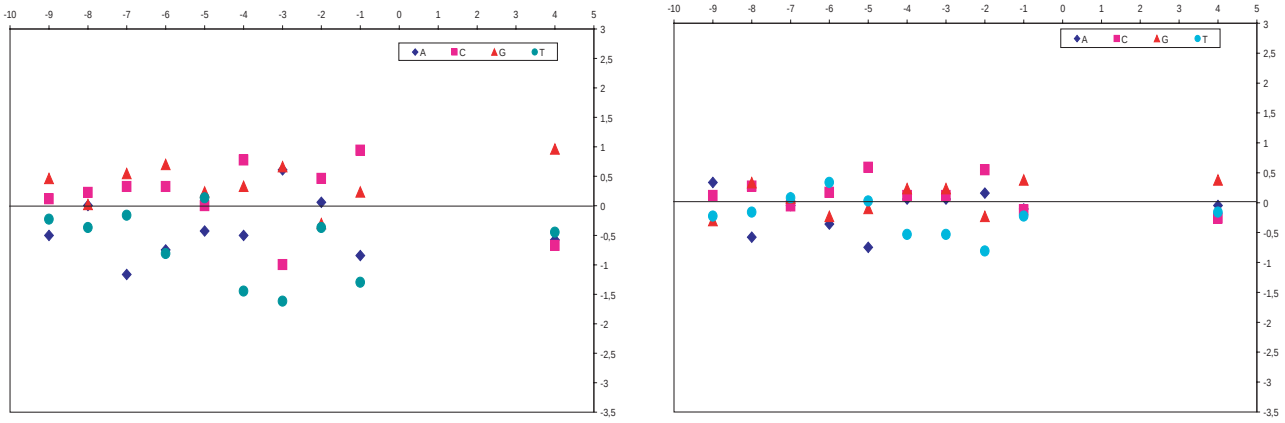
No entanto, preliminarmente, preferimos basear nossas análises nas figuras que não estão marcadas com "*". Estudos futuros tentarão definir qual a margem de separação entre as duas classes, tentando descobrir desta forma, um número suficiente de amostras que validem estes dados. Uma possibilidade inicial para isso seria o uso de Support Vector Machines (SVM) (Vapnik, 1995; Vapnik, 1998; Zien et al., 2000).



Homo sapiens revisado *.

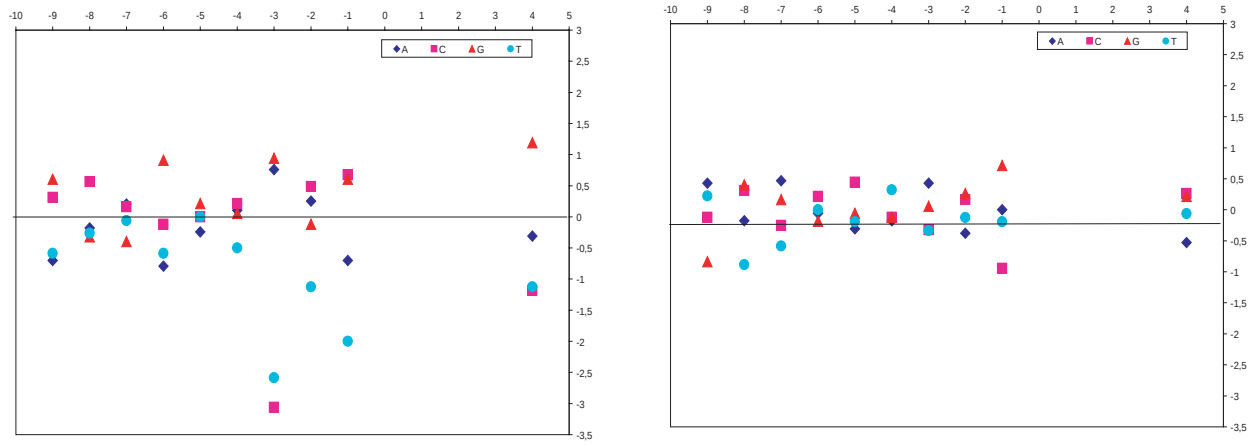


Homo sapiens provisional.

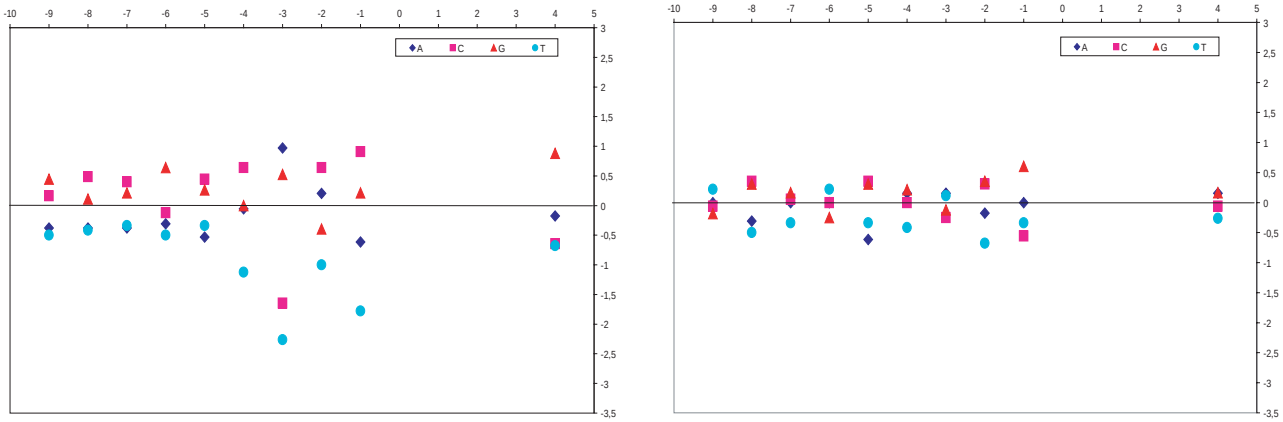


Homo sapiens predicted *.

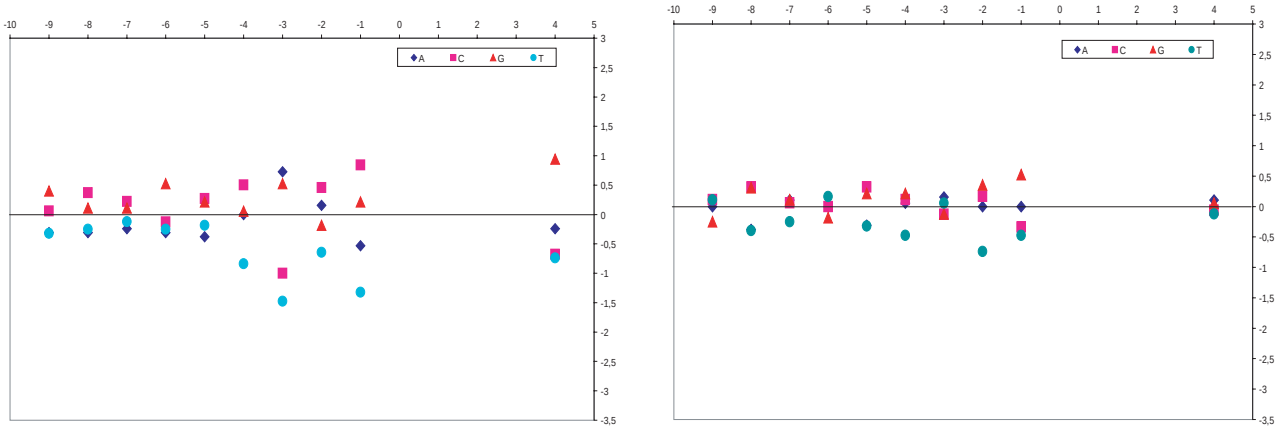
Figura 4.2: Freqüência de bases das seqüências positivas e negativas (sem repetições), respectivamente, do *Homo sapiens*.



Mus musculus revisado *.

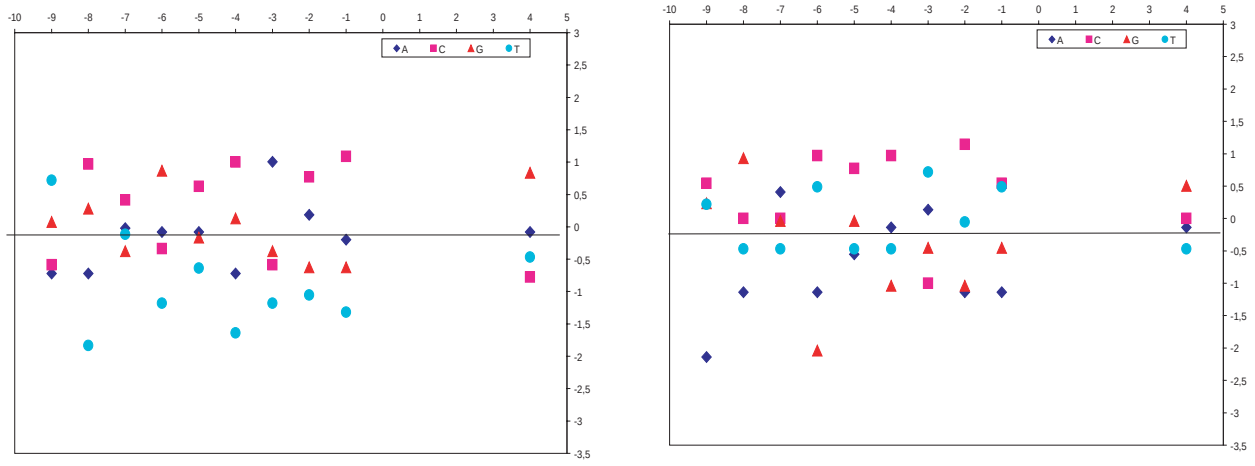


Mus musculus provisional.

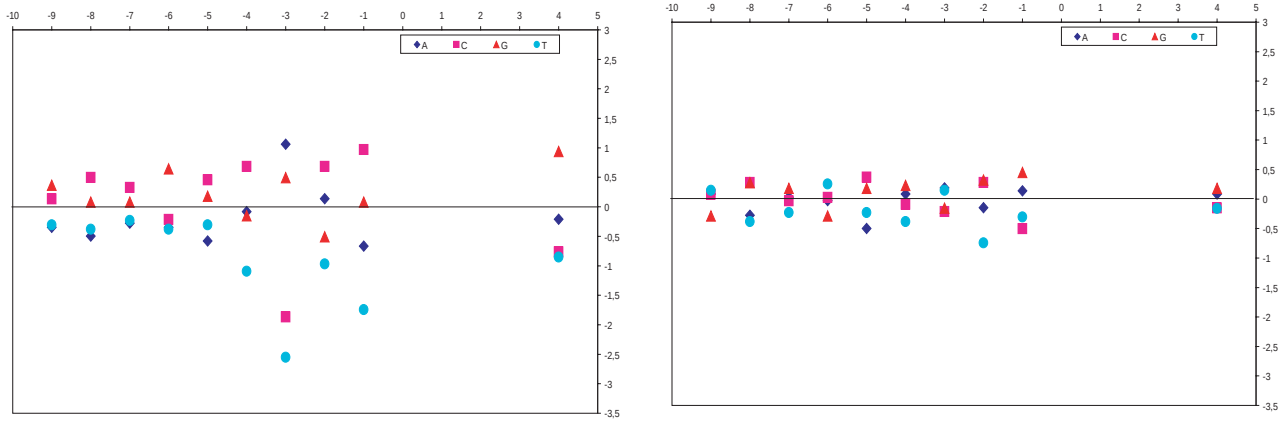


Mus musculus predicted.

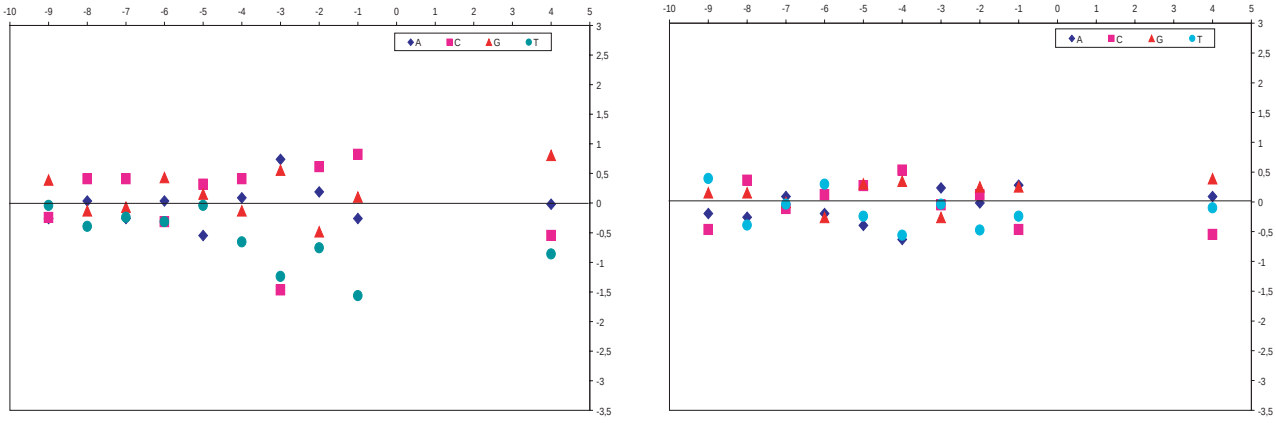
Figura 4.3: Freqüência de bases das seqüências positivas e negativas (sem repetições), respectivamente, do *Mus musculus*.



Rattus norvegicus revisado *.

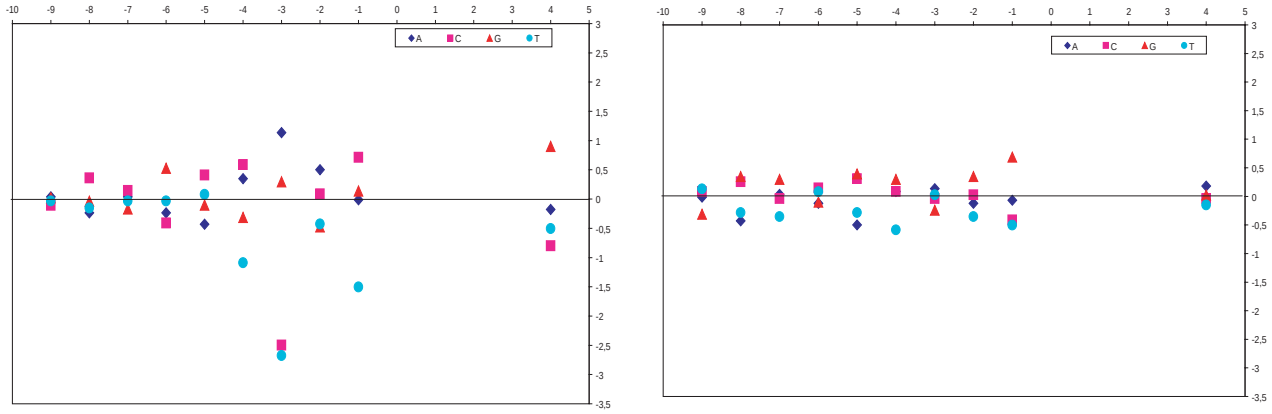


Rattus norvegicus provisional.

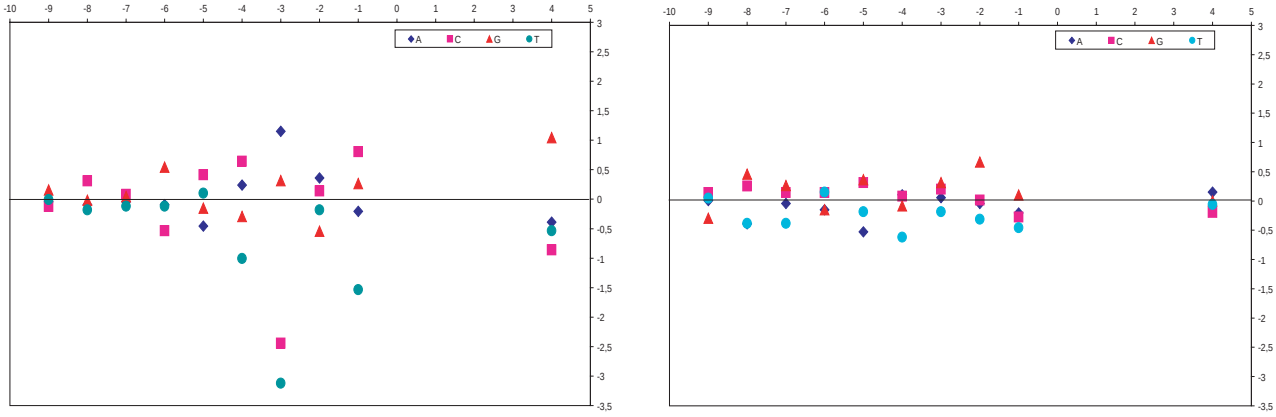


Rattus norvegicus predicted *.

Figura 4.4: Freqüência de bases das seqüências positivas e negativas (sem repetições), respectivamente, do *Rattus norvegicus*.

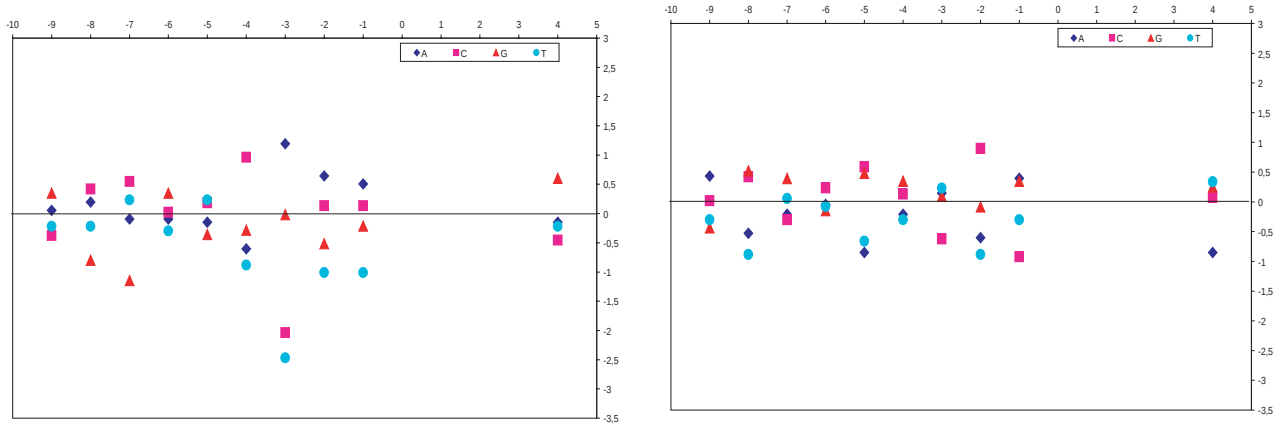


Danio rerio provisional.



Danio rerio predicted.

Figura 4.5: Freqüência de bases das seqüências positivas e negativas (sem repetições), respectivamente, do *Danio rerio*.



Drosophila melanogaster provisional *.

Figura 4.6: Freqüência de bases das seqüências positivas e negativas (sem repetições), respectivamente, do *Drosophila melanogaster*. O "*"sinaliza bases menores de 1000 seqüências.

De acordo com estas figuras, podemos perceber que as seqüências positivas estão de acordo com o padrão do consenso de Kozak, mostrando que a presença da purina (Adenina ou Guanina) na posição -3 é muito importante para a identificação correta do SIT. Uma outra posição importante é a +4, onde normalmente aparece uma Guanina.

Quanto às seqüências negativas, percebe-se pelas figuras que não existe um padrão, à exceção de uma leve tendência à presença da base G imediatamente anterior ao ATG (posição -1) ser um pouco superior à freqüência de G no mRNA, em quase todos os casos. Ou seja, em geral a freqüência de cada base fica perto da ocorrência dela no mRNA, que é sempre em torno de 25% (dados não mostrados). Isso explica porque os dados plotados neste caso, ficam próximo de zero. As únicas bases que apresentam comportamento diferente são as revisadas do *Homo sapiens* e *Rattus norvegicus*. Nestes dois casos, percebemos que o comportamento não se mostra aleatório. Todavia, como argumentado, essas bases possuem um tamanho provavelmente insuficiente para essa análise. Outra observação interessante é que, tanto nas seqüências negativas como positivas, a presença da Timina é aparentemente menor que no mRNA como um todo, sugerindo que sua ausência nos padrões

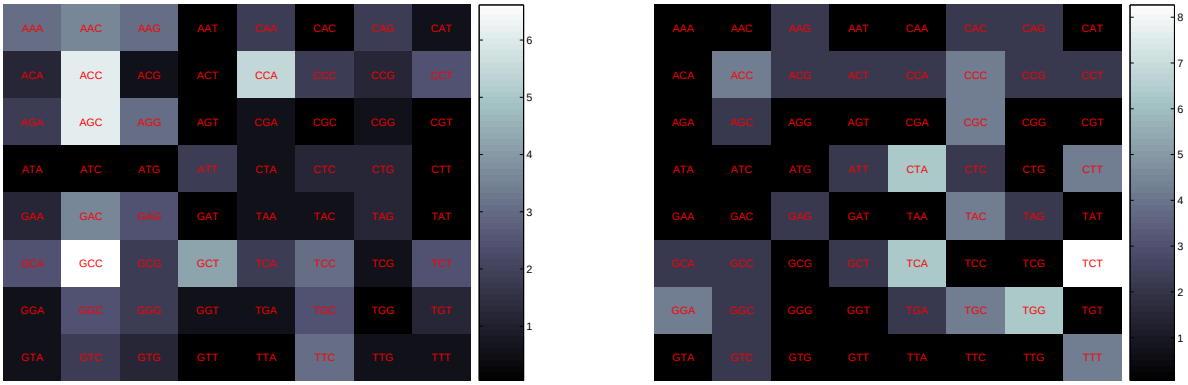
positivos não é determinante. No entanto, se compararmos a quantidade de Timina nestas seqüências (positivas e negativas), podemos perceber que as negativas possuem uma quantidade ligeiramente maior desta base.

Um outro fato interessante é que a Adenina (A) não aparece na mesma proporção que a Timina (T); o mesmo acontece que a Citosina e Guanina. Isso se deve ao fato de estarmos trabalhando com fitas simples de cDNA.

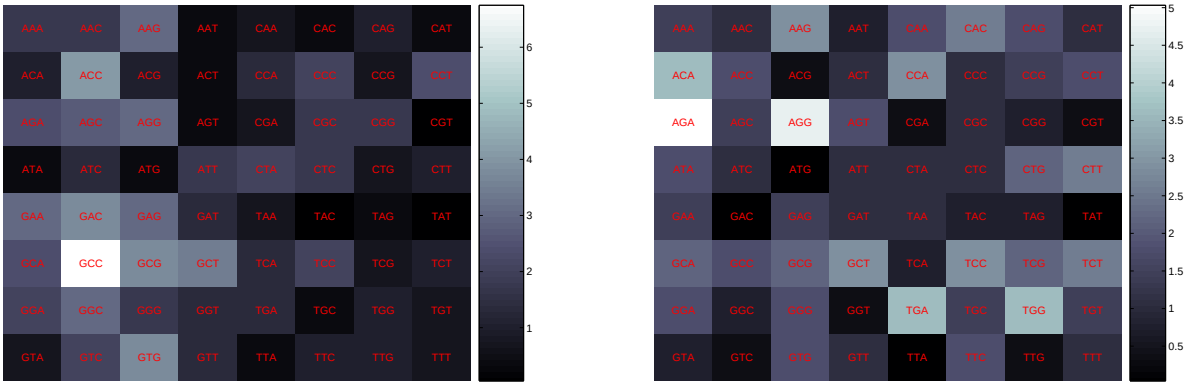
4.2.2 Freqüência de trinucleotídeos

Com o objetivo de analisar a região antes do códon inicializador da tradução (as 9 bases antes do AUG), foi realizada, também, uma análise da freqüência de trinucleotídeos neste contexto. Para estes testes, foram consideradas as seqüências sem repetições.

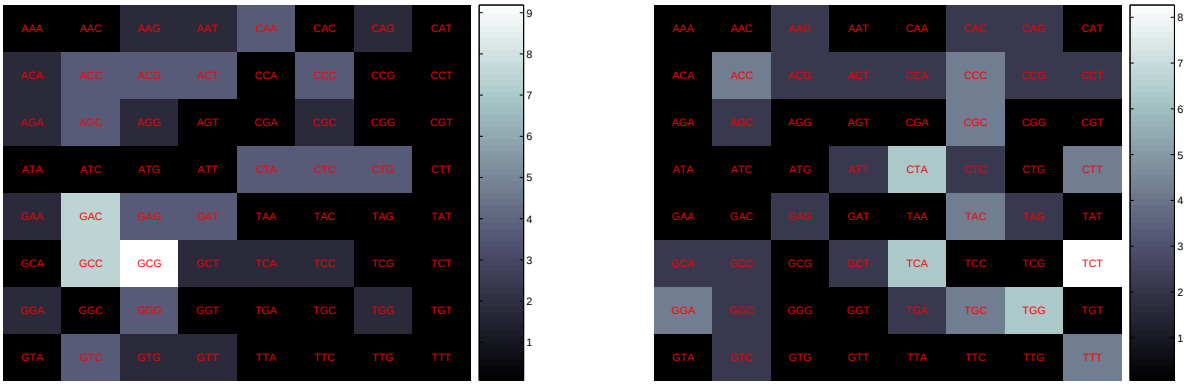
O programa percorre as seqüências de três em três bases; ou seja, o deslizamento na janela é feito de três em três posições. Para exemplificar, considere a seqüência: "**CGGACGCAT**"; neste caso, teríamos a seguinte composição: CGG=1, ACG=1 e CAT=1 e não teríamos o trinucleotídeo GGA, por exemplo. Desta forma, composições de todas as seqüências foram calculadas e as freqüências obtidas para cada organismo (com seqüências revisadas, *provisional* e *predicted*, respectivamente) são apresentadas nas figuras 4.7, 4.8 e 4.9. Novamente, como o número de seqüências utilizado pode afetar o experimento, adicionamos a marça "*"àquelas determinações feitas com amostras inferiores a 1000 seqüências. Como já foi dito na seção 4.2.1, testes serão realizados para se tentar descobrir um conjunto de dados que seja suficiente.



Homo sapiens *.

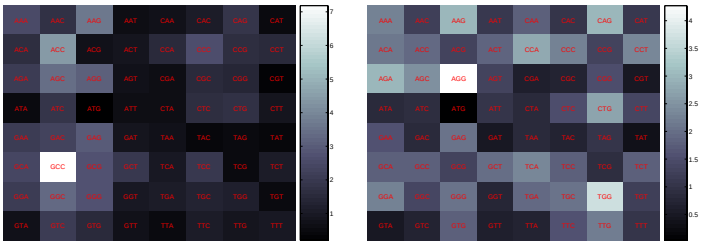


Mus musculus *.

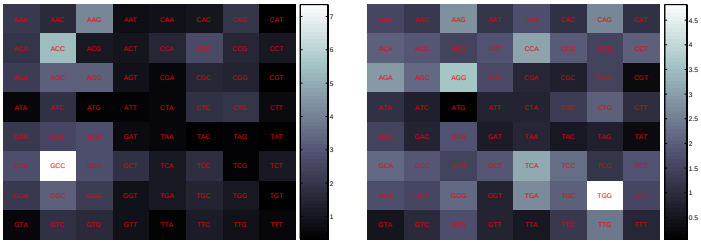


Rattus norvegicus *.

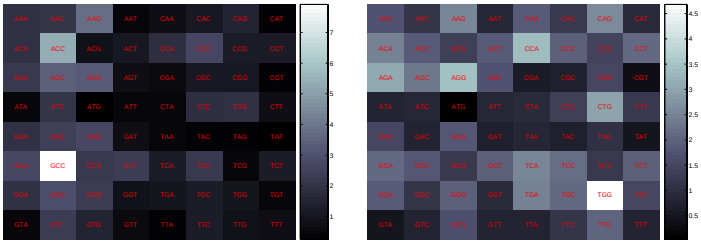
Figura 4.7: Freqüência de trinucleotídeos das seqüências (revisadas) positivas e negativas (sem repetições), respectivamente, dos organismos estudados. O "*"sinaliza bases com menos de 1000 seqüências.



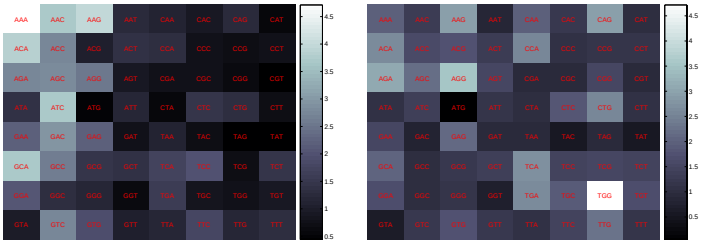
Homo sapiens.



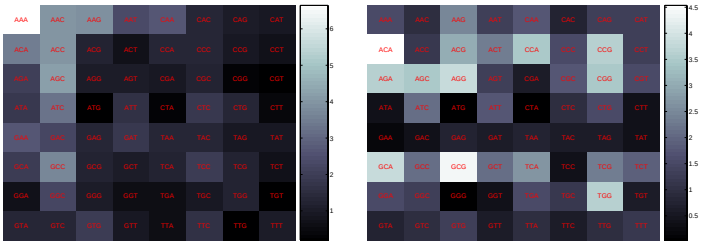
Mus musculus.



Rattus norvegicus.

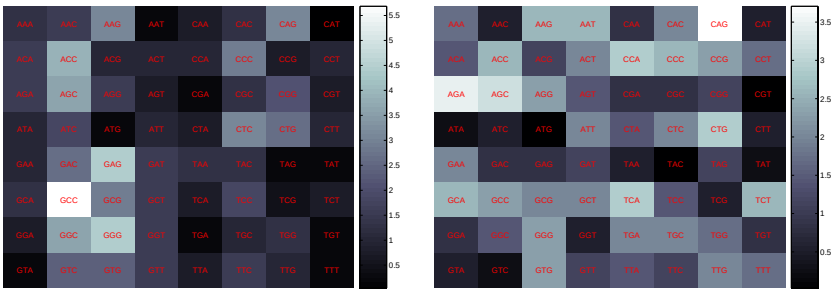


Danio rerio.

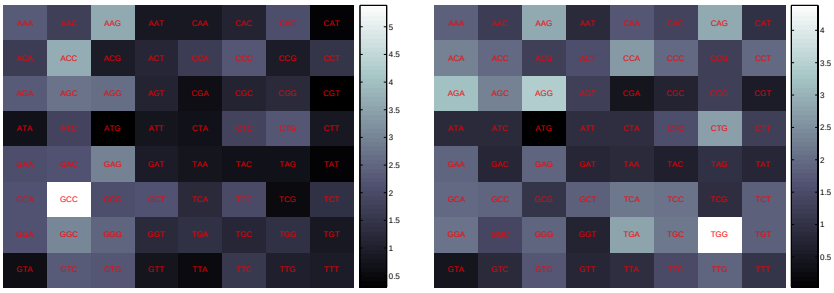


Drosophila melanogaster *.

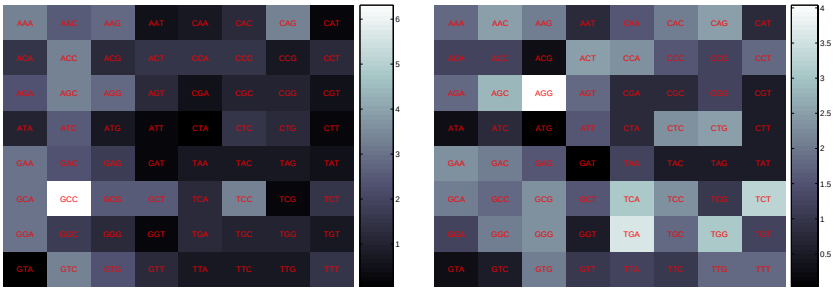
Figura 4.8: Freqüência de trinucleotídeos das seqüências (*provisional*) positivas e negativas, respectivamente, dos organismos estudados. O "*"sinaliza bases com menos de 1000 seqüências.



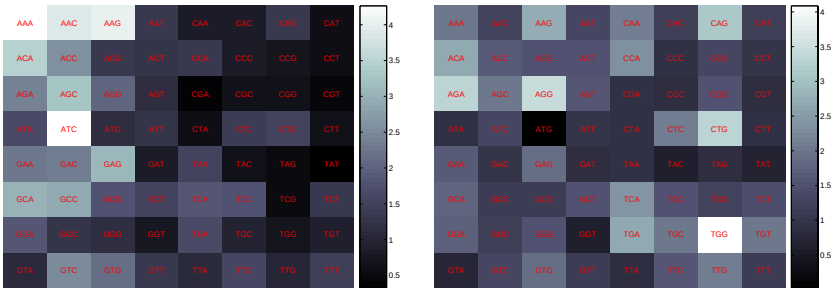
Homo sapiens *.



Mus musculus.



Rattus norvegicus *.



Danio rerio.

Figura 4.9: Freqüência de trinucleotídeos das seqüências (*predicted*) positivas e negativas (sem repetições), respectivamente, dos organismos estudados.

Pelas figuras, é possível perceber que a freqüência de trinucleotídeos das seqüências positivas normalmente é diferente das seqüências negativas para todos os organismos estudados. De maneira geral, as maiores freqüências nos positivos se concentram no lado esquerdo e para os negativos, no lado direito (onde tem-se mais Timina). Estes resultados estão de acordo com os encontrados no estudo sobre a frequência de bases das seqüências, onde se percebia uma presença maior da Timina nas seqüências negativas, em relação às seqüências positivas.

Outro ponto importante é que os testes apontam, nos positivos, uma maior freqüência dos trinucleotídeos GCC e GCG (próximo de 9%) para os organismos *Mus musculus*, *Rattus norvegicus* e *Homo sapiens*. Para o *Danio rerio* e a *Drosophila melanogaster*, entretanto, o trinucleotídeo AAA aparece com maior freqüência. Observando-se estas frequências percebe-se que estes dois organismos têm uma freqüência bem diferente dos outros três. Nos negativos, entretanto, os trinucleotídeos TGG e TCT são os mais freqüentes.

Ainda neste sentido, foram analisadas também as posições onde estes trinucleotídeos são mais freqüentes. Ou seja, uma vez sabendo, por exemplo, que o GCC é o mais freqüente no Homo sapiens queremos descobrir também qual é a posição onde se tem mais GCC. Será se este trinucleotídeo fica no início da seqüência de 13 ou mais próximo da metionina (AUG)?

Para o caso do Homo sapiens, considerando-se as 55 seqüências revisadas positivas, encontramos 11 ocorrências do trinucleotídeo GCC. Destas 11 ocorrências, 1 está localizada nas posições de -9 à -7, 5 nas posições de -6 à -4 e 5 nas últimas três posições antes do AUG. A tabela 4.3 apresenta a freqüência dos trinucleotídeos mais freqüentes nestas três posições para todos os organismos analisados. Apenas as freqüências referentes às seqüências positivas serão apresentadas, visto que a freqüência obtida, neste caso, é bastante significativa. Além disso, os mesmos testes foram realizados com as seqüências negativas e observou-se uma freqüência aleatória, em relação à posição, entre os trinucleotídeos mais freqüentes (dados não mostrados).

Tabela 4.3: Freqüência, nas posições de -9 à -7 (Pos 1), -6 à -4 (Pos 2) e de -3 à -1 (Pos3), de trinucleotídeos mais freqüentes (Tri+) para os cinco organismos analisados.

Organismos	Inspeção	Tri+	Pos 1	Pos 2	Pos 3
Homo sapiens	Revisado	GCC	9%	45,5%	45,5%
	Provisional	GCC	20,5%	26,2%	53,2%
	Predicted	GCC	18%	22,7%	59%
Mus musculus	Revisado	GCC	40%	15%	45%
	Provisional	GCC	18,9%	30,6%	50,4%
	Predicted	GCC	20,9%	24%	54,9%
Rattus norvegicus	Revisado	GCG	0%	60%	40%
	Provisional	GCC	19%	27%	52,7%
	Predicted	GCC	24%	20%	56%
Danio rerio	Provisional	AAA	23,4%	14,8%	61,7%
	Predicted	AAA	18,2%	23,2%	58,8%
Drosophila melanogaster	Provisional	AAA	15%	8%	75%

Como pode ser observado, para todos os organismos, os trinucleotídeos são mais freqüentes na posição mais próxima do AUG. Esses dados, somados à análise independente de cada posição (figuras 4.2 à 4.6) ilustram a importância da utilização da RNA que pondera as 10 posições simultaneamente.

4.3 Comparação entre organismos

Como foi comentado no primeiro capítulo deste trabalho, é comum, para alguns organismos, termos poucas seqüências disponíveis. Neste trabalho inclusive, tivemos apenas uma seqüência revisada para o *Danio rerio* e poucas para o *Rattus norvegicus*.

Neste sentido, testes foram realizados com o objetivo de verificar se seria possível validar seqüências de um determinado organismo cujo treinamento foi realizado com outro. Assim queremos saber se podemos treinar com seqüências do *Mus musculus*, por exemplo, e validar com o *Homo sapiens*.

Para isto, primeiramente, uma rede foi treinada com seqüências revisadas do *Rattus norvegicus* e testada, também com seqüências revisadas, do *Mus musculus*, *Homo sapiens* e do próprio *Rattus norvegicus*. Estes resultados estão apresentados na figura 4.10.

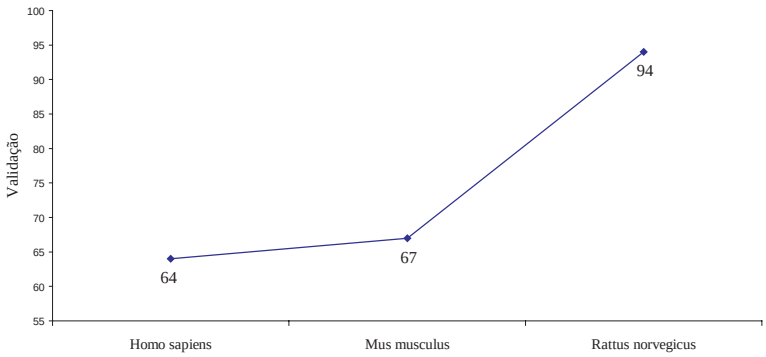


Figura 4.10: Validação das seqüências revisadas dos organismos *Homo sapiens*, *Mus musculus* e *Rattus norvegicus* em relação às seqüências também revisadas do *Rattus norvegicus*.

Esta figura sugere que o Rattus norvegicus se aproxima mais do *Mus musculus* do que do *Homo sapiens*. Embora o número de seqüências e a acurácia da determinação utilizando-se seqüências de *R. norvegicus* não tenha sido a melhor possível, é importante notar que a acurácia era de 95% para a validação de *R. norvegicus*, repetindo-se o resultado, e que acurácia para *M. musculus* quando o treinamento era feito com seqüências do próprio organismo era muito superior ao obtido neste experimento (98,5% e 99.9%, tabelas 4.1 e 4.2, respectivamente, contra 67%, figura 4.10). Estes experimentos foram realizados com condições de menor amostragem, para evidenciar procedimentos mais críticos, onde se tem poucas seqüencias disponíveis. Experimentos em andamento realizam a mesma comparação utilizando-se as seqüências provisórias, em maior quantidade.

Um próximo teste foi realizado treinando-se com seqüências revisadas do *Mus musculus* e testando-se com seqüências *provisional* dos organismos *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus*, *Danio rerio* e *Rattus norvegicus*. Estes resultados estão apresentados na figura 4.11.

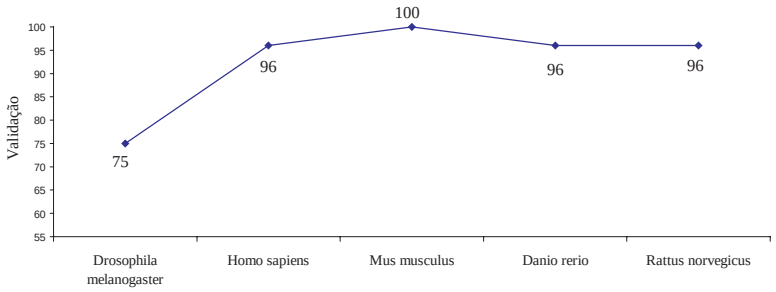


Figura 4.11: Validação das seqüências *provisional* da *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus*, *Danio rerio* e *Rattus norvegicus* em relação às seqüências revisadas do *Mus musculus*.

É interessante perceber que a *Drosophila melanogaster* distoa do restante dos organismos e que o restante é bem próximo do *Mus musculus*.

Finalmente, um último teste foi realizado treinando-se com seqüências *provisional* da *Drosophila melanogaster* e testando-se com seqüências também *provisional* da *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus*, *Danio rerio* e *Rattus norvegicus*. Estes resultados estão apresentados na figura 4.12.

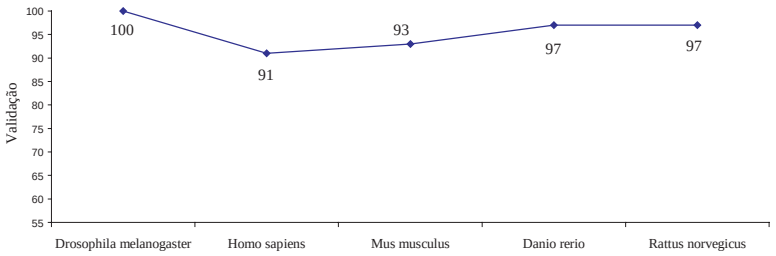


Figura 4.12: Validação das seqüências *provisional* dos organismos *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus*, *Danio rerio* e *Rattus norvegicus* em relação às seqüências também *provisional* da *Drosophila melanogaster*.

No entanto, será necessário uma maior quantidade de testes para que estas comparações possam ser melhor exploradas.

O capítulo seguinte apresenta as conclusões e propostas para trabalhos futuros.

CAPÍTULO 5

Conclusões e propostas de continuidade

5.1 Conclusões

A identificação do SIT é hoje um problema que necessita de muitas pesquisas. Trabalhar com diferentes organismos (metazoários superiores e eucariotos mais primitivos), fazer uma análise detalhada das seqüências, melhorar o desempenho dos classificadores obtidos até então, fazer uma comparação entre diversos organismos, com base no SIT e representar o conhecimento adquirido por estes classificadores, de forma que seja fácil de ser interpretado são algumas das necessidades para quem trabalha com a biologia molecular computacional.

Alguns métodos já foram desenvolvidos até o momento, mas o desempenho obtido varia entre 85% e 90%.

Neste trabalho, foram implementados vários classificadores existentes na literatura com o objetivo de aumentar este desempenho. O *Bagging* com *Multilayer perceptron* mostrou-se mais adequado a este problema, fornecendo uma validação de 99% para a maioria dos organismos analisados (exceto para o *Homo sapiens* com seqüências revisadas e para o *Rattus norvegicus* com seqüências revisadas e *predicted*.

Estes resultados foram obtidos utilizando-se a seguinte metodologia:

- Primeiramente, destaca-se a qualidade das bases analisadas. Todas elas foram extraídas do RefSeq que são sequências de referência;
- O critério utilizado para a extração das sequências negativas. Com isso, eliminaram-se bastante falsos negativos.
- Escolha dos classificadores utilizados.

Para as seqüências revisadas do *Homo sapiens*, uma das possíveis razões para o desempenho obtido (75,9%) é o desbalanceamento das classes. Testes relativos a este desbalanceamento e estudos mais aprofundados sobre as seqüências (qualidade e quantidade) do Rattus norvegicus serão realizados.

Ainda, quanto à análise das sequências, mostramos que as positivas estão de acordo com o consenso de Kozak (viés da Guanina (G) na posição +4 uma purina, preferencialmente a Adenina (A), na posição -3). Quanto às negativas, mostramos que elas possuem um comportamento aleatório.

Quanto ao treinamento com um organismo e validação com outro, os primeiros testes mostram que isso é possível entre determinados organismos; possivelmente os mais semelhantes entre si. Mostram também (juntamente com a análise das seqüências), que a *Drosophila melanogaster* se diferencia dos outros organismos. No entanto, testes mais rigorosos deverão ser realizados.

5.2 Sugestões de continuidade de trabalho

Alguns pontos neste trabalho ainda precisam ser aperfeiçoados; enquanto outros, precisam ser desenvolvidos. Os itens, a seguir, são sugestões para a continuidade deste trabalho que tem previsão de término no início de 2007, quando terá sido completado o prazo de quatro anos.

- Investir em técnicas para melhorar o desempenho dos classificadores que estão abaixo de 90%;
- Realizar estudos com o objetivo de descobrir qual o número suficiente de seqüências necessárias em alguns dos experimentos realizados;

- Expandir os testes para eucariotos mais primitivos, a partir dai, fazer uma comparação entre estes e metazoários superiores;

- Fazer uma análise mais criteriosa de similaridades entre organismos com base no SIT. Hoje um grande problema da genômica é a falta de dados para alguns organismos. Neste trabalho inclusive, tivemos apenas uma sequência revisada do *Danio rerio*. Se comprovarmos a semelhança entre organismos, talvez possamos validar organismos que possuem poucas seqüências com base em organismos que possuem um número de sequências mais significativo;

- Extrair regras do classificador *Bagging* com *Multilayer perceptron* para que o conhecimento adquirido seja de fácil entendimento. Uma das grandes contribuições desse trabalho é obter uma base de conhecimento a partir dos resultados obtidos. Este conhecimento, que pode ser representado por meio de regras *if-then*, por exemplo, poderá ser utilizado para classificar seqüências positivas e negativas arbitrárias. Elas também poderão ser utilizadas para melhorar o conhecimento dos especialistas, uma vez que as regras geradas podem criar novas relações a partir dos dados. Com isso, estaremos dando uma grande contribuição para a área de Predição de SIT, que ainda é pouco estudada.

Referências

Andrews, R. e Diederich, J. (1995). A survey and critique of techniques for extracting rules from trained artificial neural networks. *Neurocomputing Research Centre*.

Braga, A. P., Carvalho, A., e Ludermir, T. B. (2000). *Redes neurais artificiais: teoria e aplicações*. Livros Técnicos e Científicos.

Breiman, L. (1997). Bagging predictions. *Machine Learning,* 24:123–140.

Browne, A., Hudson, B. D., Whitley, D., Ford, M., e Picton, P. (2003). Biological data mining with neural networks: Implementation and application of a flexible decision tree extraction algorithm to genomic problem domains. *Neurocomputing: Special Issue on Bioinformatics*, 57:275–293.

Chawla, N. V., Moore, T. E., Hall, L. O., Bowyer, K. W., Kegelmeyer, W. P., e Springer, C. (2003). Distributed Learning With Bagging-like Performance. *Pattern Recognition Letters*, 24:455–471.

Clark, P. e Boswell, R. (1991). Rule induction with cn2: Some recent improvements. *Machine Learning - Proceedings of the Fifth European Conference*, 3:151–163.

Clark, P. e Niblett, T. (1988). The cn2 induction algorithm. *Machine Learning*, 3:261–283.

Cover, T. e Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Information Theory*, 13:21–27.

Dasarathy, B. (1991). Nearest-neighbor classification techniques. *IEEE Computer Society Press*.

Gibas, C. e Jambeck, P. (2001). *Desenvolvendo bioinformática: ferramentas de software para aplicações em biologia*.

Haifeng, L. e Tao, J. (2004). A class of edit kernels for svms to predict translation initiation sites in eukaryotic mrnas. *ACM Press*, pages 262–271. New York, NY, USA.

Hatzigeorgiou, A. (2002). Translation initiation start prediction in human cdnas with high accuracy. *Bioinformatics*, 18:343–350.

Jiawei, H. e Micheline, K. (2001). *Data Mining: Concepts and Techniques*. Morgan Kaufmann,.

Khan, Maleq, D. Q. P. W. (2002). K-nearest neighbor classification on spatial data streams using p-trees. *Advances in knowledge discovery and data mining: 6th Pacific Asia conference*, pages 6–8.

Kozak, M. (1984). Compilation and analysis of sequences upstream from the translational start site in eukaryotic mrnas,. *Nucl. Acids Res*, 12:857–872.

Kozak, M. (1986). Point mutations define a sequence flanking the aug initiator codon that modulates translation by eukaryotic ribosomes. *Cell*, 44:283–292.

Kozak, M. (1989). The scanning model for translation: an update. *J. Cell. Biol.*, 108:229–241.

Kozak, M. (1999). Initiation of translation in prokaryotes and eukaryotes. *Gene*, 234:187–208.

Maddouri, M. e Elloumi, M. (2002). A data mining approach based on machine learning techniques to classify biological sequences. *Knowledge Based Systems*, 15(4):217–223.

Mark, W. e Craven, M. (1996). Extracting comprehensible models from trained neural networks. *Available as CS Technical Report 1326*.

Minsky, M. e Papert, S. (1969). Perceptron: An introduction to computational geometry. *MIT Press*. Cambridge.

Morin, R. e Raeside, D. (1981). A reappraisal of distance-weighted k-nearest neighbor classification for pattern recognition with missing data. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-11(3):241–243.

Pain, V. (1996). Initiation of proteins synthesis in eukaryotes cells. *Eur. J. Biochem*, 236:747–771.

Pedersen, A. e Nielsen, H. (1997). Neural network prediction of translation initiation sites in eukaryotes: perspectives for est and genome analysis. *Proc. 5th International Conference on Intelligent Systems for Molecular Biology*, pages 226–233.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Quinlan, R. (1986). Induction of decision trees. *Machine Learning*, 1.

Rosenblatt, R. (1958). The perceptron. a probabilistic model for information storage and organization in the brain. *Psychological Review*, (65):386–408.

Sethi, I. e Yoo, J. (1994). Symbolic approximation of feedforward neural networks. *Pattern Recognition in Pratice*, 4. North-Holland, New, York, NY.

Silverman, B. W. (1986). Density estimation for statistics and data analysis. *Clapman and Hall*. New York.

Vapnik, V. (1995). *The Nature of Statistical learning theory*. Springer-Verlag. New York.

Vapnik, V. (1998). *Statistical learning theory*. John Wiley & Sons. New York.

Zeng, F., Yap, R., e Wong, L. (2002). Using feature generation and feature selection for accurate prediction of translation initiation sites. *Genome Inform Ser Workshop Genome Informatics*, 13:192–200.

Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lemmen, C., Smola, A., Lengauer, T., e Muller, K. R. (2000). Engineering support vector machine kernels that

recognize translation initiation sites. *Bioinformatics*, 16:799–807.

Zurada, J. M. (1992). *Introduction to Artificial Neural Systems*. Prentice Hall.