DESENVOLVIMENTO DE FERRAMENTA WEB PARA PREDIÇÃO DE SÍTIOS DE INÍCIO DE TRADUÇÃO EM SEQUÊNCIA DE mRNA

Diego Dias de Matos[1], **Lívia Márcia Silva**[1]

[1]Departamento de Ciência da Computação – Universidade Presidente Antônio Carlos (UNIPAC)
Rua Palma Bageto Viol S/N – Barbacena – MG – Brasil

diegodias1990@hotmail.com, livimarcia@yahoo.com.br

Abstract. *This article aims to develop a web tool for the recognition of translation start site of proteins. This will be used and validated methods already proposed by other authors, making evident the need for tool with this functionality.*

Resumo. *Este artigo visa o desenvolvimento de uma ferramenta web para o reconhecimento de Sítio de Início de Tradução de proteínas. Para tal serão utilizados métodos já propostos e validados por outros autores, ficando em evidência a necessidade de ferramenta com essa funcionalidade.*

1. Introdução

Um dos grandes problemas da comunidade científica biológica é a predição de Sítios de Início de Tradução (SITs), que consiste em localizar o início da sequência codificadora de proteínas. Kozak(1984) detectou que na maioria dos casos, a previsão do SIT é determinada pela presença da primeira metionina(ATG) na sequência de mRNA; porém posteriormente Pedersen e Nielsen (1997) detectaram que o reconhecimento do SIT em eucariotos nem sempre começa no primeiro ATG, e a escolha pelo códon inicial da CDS(CoDing Sequence) depende da posição e também do contexto da sequência.

A fim de predizer o início da tradução de uma sequência será analisado a região *upstream*, região que antecede o SIT, ou seja, inicia-se juntamente com a sequência e termina exatamente no nucleotídeo anterior ao SIT; e região *downstream*, região que sucede o SIT, ou seja, inicia-se no nucleotídeo subsequente ao SIT e termina juntamente com a sequência.

Dependendo da posição de início da síntese na fita de mRNA, o trio de nucleotídeos selecionado para a síntese poderá variar, alterando-se também os aminoácidos que serão traduzidos. Com tudo, a predição de SIT é uma tarefa complexa, por essa razão, métodos computacionais de busca de padrões podem ser utilizados a fim de extrair o conhecimento implícito envolvido nesse processo(NOBRE; ORTEGA; BRAGA, 2007).

O trabalho proposto visa o desenvolvimento de uma ferramenta web para localizar o sítio de início de tradução de proteínas em sequências de mRNA, ou seja, dada uma sequência de mRNA localizar onde está o provável SIT.

O controle do início da tradução é um dos mais importantes processos na regulação da expressão genética (NAKAGAWA et al., 2008). Com isso tem-se que a

determinação de SITs não é uma tarefa fácil, e ao mesmo tempo é de grande relevância para a inferência genética. Segundo Liu et al. (2004), uma alta acurácia na localização de SITs pode ser útil para um melhor entendimento da proteína que é gerada das sequências de nucleotídeos. Portanto, o estudo proposto permitirá:

1. Classificação de fragmentos de sequência de mRNA, determinando se o mesmo possui ou não um sítio de início de tradução, a fim de obter melhor acurácia.
2. Disponibilizar ferramenta para predizer o SIT.

Sabe-se que em uma sequência de mRNA o início da tradução quase sempre inicia-se em um códon ATG (start codon) e termina em um dos stop codons (TAA, TAG, TGA). No entanto, a tradução pode não se iniciar no primeiro ATG da sequência (PEDERSEN; NIELSEN, 1997; HATZIGEORGIOU, 2002). Dependendo de onde for localizado o SIT isso pode mudar o aminoácido que será gerado, por isso a precisão em localizar o SIT é tão importante. Para isso métodos computacionais de busca de um conjunto de características devem ser utilizados a fim de extrair o conhecimento implícito envolvido nesse processo.

2. Estado da arte

Essa seção visa apresentar os principais métodos de previsão de SIT. No final será demonstrado como alguns métodos são disponibilizados na web.

2.1. Métodos para previsão de SIT

2.1.1. Reconhecimento através de redes neurais artificiais

Um dos trabalhos mais antigos em previsão de SIT foi realizado por Stormo, Schneider e Gold (1982). Eles utilizaram codificação de 4 bits (A= 1000, C = 0100, G= 0010 e T = 0001) e janelas de 51, 71 e 101 nucleotídeos centradas no ATG. Pedersen e Nielsen (1997) treinaram redes neurais artificiais com base de dados de vertebrados e obtiveram excelentes resultados. O trabalho é de tamanha importância, que até hoje estudiosos utilizam de sua base para realizar comparações e validar suas metodologias. Das sequências, apenas as que possuíam *downstream* maior ou igual a 10 nucleotídeos e *upstream* maior ou igual a 150 nucleotídeos foram selecionadas. Essas sequências foram então filtradas para remover aquelas pertencentes a uma mesma família gênica, genes homólogos de diferentes organismos e sequências repetidas. Essa base de dados, bem como uma descrição detalhada da mesma, encontra-se disponível em http://www.cbs.dtu.dk/databases/NetStart/.

Pedersen e Nielsen (1997) utilizaram uma Rede Neural Artificial (RN), perceptron de múltiplas camadas, treinada com uma janela de 203 nucleotídeos centrada no ATG. E utilizaram 0, 1, 2, 5, 10, 20, 30 e 50 neurônios na camada intermediária. Eles obtiveram sensibilidade (porcentagem de acertos dentro da classe positiva), especificidade (porcentagem de acertos dentro da classe negativa) e acurácia de 78%, 87% e 85%, respectivamente. O sistema desenvolvido por eles, denominado NetStart, também encontra-se disponível em http://www.cbs.dtu.dk/services/NetStart/.

Os autores ainda realizaram testes para descobrir quais características são importantes para distinguir SIT de não-SIT. Eles descobriram que a posição -3 é crucial na identificação do SIT, fato que já havia sido identificado por Kozak (1987).

Hatzigeorgiou (2002) apresenta um programa com alta acurácia na previsão de SIT, ele também utilizou sequências de mRNA humano, alcançando acurácia de 94%. Foram utilizados dois módulos: consensus-ANN (analisa a vizinhança imediata do candidato a SIT) e coding-ANN (avalia as regiões *upstream* e *downstream* do candidato).

O módulo consensus-RN avalia o SIT candidato e sua vizinhança mais imediata por meio de uma janela de 12 nucleotídeos. As sequências foram extraídas a partir das posições -7 a +5 e a codificação de 4 bits (adotada em trabalhos anteriores) foi utilizada.

O módulo coding-RN foi utilizado para avaliar as regiões *upstream* e *downstream* do SIT candidato e trabalha com janelas de 54 nucleotídeos.

Assim, para avaliar as sequências dos 54 nucleotídeos, estas são transformadas em um vetor de 64 unidades correspondendo à frequência de determinado códon na sequência. Os dois módulos propostos são integrados da seguinte forma: dados um ATG candidato, o consensus-RN é aplicado para calcular um consenso S1. O coding-RN é aplicado então para obter um score S2 da região *upstream*. Isso se repete também para a região *downstream* e o score S3 é calculado. O score final para o candidato a SIT é então obtido por S1 X (S3 - S2) Esse cálculo é realizado para todos os ATGs da molécula e o primeiro ATG que oferecer um score acima de 0.2 é considerado o SIT da sequência.

2.1.2. Reconhecimento através de SVM (Support Vector Machines)

Zien et al (2000) trabalharam com a mesma base de dados de Pedersen e Nielsen porém utilizando SVM, com isso alcançaram uma acurácia de 88,1%. Eles mostraram como obter melhorias usando uma nova função de *Kernel*, chamada *locality-improved Kernel* com uma pequena janela em cada posição. O *locality-improved Kernel* enfatiza correlações entre as posições da sequência que são próximas entre si , e um tamanho de 3 nucleotídeos *upstream* e *downstream* foi empiricamente determinado como ótimo. Ou seja, a modificação consistiu em privilegiar correlações locais entre nucleotídeos, enquanto as correlações com nucleotídeos de posições distantes foram consideradas de pouca importância ou inexistentes. Com esta função de Kernel, eles obtiveram sensibilidade, especificidade e acurácia de 69,9%, 94,1% e 88,1%, respectivamente. Eles mostraram, assim, que através de funções simples de Kernel é possível conseguir uma acurácia melhor do que aquela obtida por Pedersen e Nielsen utilizando RN. Mais tarde, Zien et al. (2000) melhoraram estes resultados através de uma função de Kernel mais sofisticada, também chamada de Kernel de Salzberg. Por meio desse Kernel , eles obtiveram uma acurácia de 88,6% para a mesma base de dados. Li e Jiang (2004) utilizaram duas novas propostas para identificação do SIT. Primeiro, eles introduziram uma classe de novos Kernels baseados em string Edit distance, chamados de edit Kernels, para serem utilizados como SVM. Em um segundo momento, eles converteram a região *downstream* de um ATG em uma sequência de aminoácidos antes de aplicar o SVM. Eles mostraram que a abordagem adotada por eles é significativamente melhor (sensibilidade = 99,92%, especificidade = 99,82% e acurácia de 99,9% para a base de dados utilizadas por Pedersen e Nielsen (1997). Nobre, Ortega e Braga (2007) realizaram experimentos para descoberta de SIT utilizando-se 12 nucleotídeos na região *upstream* e *downstream*, além de SVM com funções simples de Kernel. Eles apresentaram uma nova metodologia para codificação. Assim ao invés de codificar individualmente cada nucleotídeo, a codificação

foi feita por trinca, com janela deslizante de tamanho 3. Para balanceamentos dos dados, utilizaram o algoritmo Smote (CHAWLA et al.,2002) para replicação das amostras da classe minoritária. Os autores utilizaram a base de dados de Pedersen e Nielsen , obtendo 95,63% de acurácia.

2.1.3. Reconhecimento através de análise estátistica

O trabalho pioneiro de Kozak (1984), uma análise estatística sobre as sequências de 211 mRNAs de células eucarióticas, revelou que algumas posições das sequências de mRNAs, relativas ao SIT, são muito conservadas. A posição -3, ou seja, três nucleotídeos *upstream* do SIT, apresenta uma purina, nucleotídeo A (Adenina) ou G (Guanina), sendo em 79% das sequências, o nucleotídeo A. Há um predomínio do nucleotídeo C nas posições -1, -2, -4 e -5. Seguindo este raciocínio, determinou-se um consenso, denominado consenso de Kozak, ilustrado pela Figura 1.

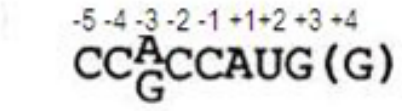


Figura 1. Consenso de Kozak Fonte: Kozak (1984).

Observou-se também que em 95% dos casos, por ela selecionados, o SIT inicia-se no primeiro ATG da sequência. No trabalho desenvolvido por Kozak (1984) foi identificada uma proporção de 79% de Adenina (A) na posição -3 (e 18% de Guanina (G)); já Cavener e Ray (1991), utilizando 2.595 sequências, constataram 58% de probabilidade de ser A na referida posição. Nakagawa et al. (2008) realizaram análises comparativas entre 47 espécies, incluindo animais, fungos, plantas e protistas, revelando a existência de consensos para diferentes espécies. Com base nessa análise, foram identificadas as seguintes regiões de consenso: presença de uma purina (A ou G) na posição -3, presença de A ou C na posição -2, presença de C na posição +5. A posição -3 já havia sido descoberta por Kozak (1984), sendo, portanto confirmada nesse estudo.

2.1.4. Reconhecimento através de avaliação de características

Até o momento, os métodos apresentados extraem informações para previsão do SIT basicamente através do conhecimento contido na própria sequência. No entanto, a partir de 2002, surgem métodos que utilizam características extraídas a priori como entradas para os classificadores. Zeng et al. (2002) empregam a técnica chamada de k-grams e poucos refinamentos para produzir características candidatas. Um k-gram é um padrão de k letras consecutivas, que podem ser aminoácidos ou nucleotídeos. Um k-gram também pode ser restrito àqueles que estão em fase com o ATG codificante. Cada k-gram e sua frequência no fragmento da sequência transformam-se em uma característica candidata. Uma outra técnica para produzir características candidatas é a idéia da posição específica do k-gram. Ou seja, essa técnica identificada em qual posição do fragmento da sequência

o k-gram aparece. Utilizando-se os nucleotídeos, existem 4k possíveis k-grams para cada valor de k. Alguns valores típicos para k são 1, 2, 3, 4 ou 5. Partindo-se do fato de que o processo biológico de traduzir nucleotídeo em aminoácidos a partir de 3 nucleotídeos (também chamado códon) inicia-se no SIT, Zeng, Yap e Wong (2002) utilizam 3-grams, tanto da região *upstream* quanto da região *downstream* da molécula. Utilizam também 1-gram para considerarem cada posição especificamente. Pelo fato de que o número de características geradas ser muito grande, Zeng, Yap e Wong (2002) propõe metodologia para selecionar as características mais importantes, utilizando o método de seleção de características baseado em correlação. Esses autores selecionaram (9) nove características, descritas abaixo:

1. Posição -3;
2. ATG *upstream* em fase;
3. TAA *downstream* em fase;
4. TAG *downstream* em fase;
5. TGA *downstream* em fase;
6. CTG *downstream* em fase;
7. GAC *downstream* em fase;
8. GAG *downstream* em fase;
9. GCC *downstream* em fase.

Utilizando-se dessas características, Zeng, Yap e Wong (2002) obtiveram sensibilidade de 84,3%, especificidade de 86,1% precisão de 66,3% e acurácia de 85,7% utilizando classificador Baysiano. Já por meio de SVM, encontraram sensibilidade 73,9%, especificidade de 93,2%, precisão de 77,9% e acurácia de 88,5%; utilizando redes neurais obtiveram sensibilidade de 77,6%, especifidade de 93,2%, precisão de 78,8% e acurácia de 89,4%; e sensibilidade de 74%, especificidade de 94,4%, precisão de 81,1% e acurácia de 89,4% ao utilizar árvores de decisão. Li et al. (2003) e Liu e Wong (2003) investigaram uma abordagem alternativa de extração de características baseada em aminoácidos. Os seguintes k-grams foram gerados Liu et al. (2004):

1. X-up, número de vezes que o aminoácido X aparece na região *upstream*.
2. X-down, número de vezes que o aminoácido X aparece na região *downstream*.
3. XY-up, número de vezes que dois aminoácidos XY aparecem como uma subtring na região *upstream*.
4. XY-down, número de vezes que dois aminoácidos XY aparecem como uma subtring na região *downstream*.

Liu et al. (2004) também geraram características booleanas a partir dos segmentos de sequências extraídas da base de Pedersen e Nielsen (1997): presença ou ausência de um ATG na região *upstream*; presença ou ausência do nucleotídeo ”A”ou ”G”na posição -3; presença ou ausência de ”G”na posição +4. Foram selecionadas 100 características. As nove principais características estão listadas abaixo e interessantemente sete delas correspondem às características selecionadas por Zeng, Yap e Wong (2002):

- ATG-up corresponde a ATG *upstream* em fase;
- STOP- *down* corresponde a TAA, TAG e TGA *downstream* em fase;
- Pos-3 AouG corresponde à posição -3;
- L-down corresponde à CTG *downstream* em fase;
- D-down corresponde à GAC *downstream* em fase;

- E-down corresponde GAG *downstream* em fase;
- A-down corresponde à GCC *downstream* em fase.

Silva et. al(2010) propôs um método de balanceamento do tipo *undersampling* baseado em Clusterização, M-Clus, além de uma nova metodologia que adiciona características às sequências e que melhora o desempenho do classificador a partir da inclusão do conhecimento obtido pelo modelo. Por meio dessa metodologia, as taxas de desempenho utilizadas, acurácia, sensibilidade, especificidade e acurácia ajustada são superiores a 9% (*Mus musculus*). A precisão aumenta significamente, de 43,05% para 82,05% (*Mus musculus*) e de 13,54% para 35,63% (*Rattus novergicus*), adotando a inclusão do conhecimento obtido pelo modelo. Para resolução do problema, faz-se necessário o investimento em técnicas de balanceamento de classes, além de uma metodologia criteriosa que melhora visivelmente os resultados. Ao utilizar o método de balanceamento M-Clus há um aumento significativo na taxa de sensibilidade, de 51,39% para 91,55% e de 47,45% para 88,09%, para os organismos *Mus musculus* e *Rattus novergicus*, respectivamente. A inclusão de algumas características durante o treinamento, tais como a presença de ATG na região *upstream* do Sítio de Início de Tradução, melhora a taxa de sensibilidade em aproximadamente 9% para o organismo *Mus musculus* e 6% para o *Rattus novergicus*.

2.2. Reconhecimento através de ferramentas web

2.2.1. NETSTART

Essa aplicação web como citado anteriormente foi criada por Pedersen e Nielsen (1997), eles utilizaram redes neurais artificiais para prever qual é o códon de iniciação da sequência de mRNA. As redes neurais foram treinadas com duas de bases de dados: de Vertebrados e da A.Thaliana(planta). A ferramenta encontra-se no endereço http://www.cbs.dtu.dk/services/NetStart/, e foi disponibilizada pelos autores para trabalhar com sequências no formato FASTA. Para utilizar o software às sequências podem ser inseridas de duas maneiras , são estas:

- Colar uma única sequência (apenas os nucleotídeos) ou uma ou mais sequências na formatação FASTA.
- Selecionar um arquivo FASTA em seu disco local.

Ambas as formas poderão ser utilizadas simultaneamente. Toda a sequência especificada será processada. O alfabeto de entrada permitido é A, C, G, T, U e X (desconhecido), todos os outros símbolos que forem informados serão convertidos em X antes do processamento. Os caracteres T e U são tratados como equivalentes. Dependendo da sequência que se queira testar deve-se selecionar, "Vertebrate"ou "A. Thaliana". O formato de saída está representado da seguinte forma. Cada sequência de entrada será mostrada com um início previsto, seguido de uma tabela mostrando as posições e as pontuações de todas as instâncias "ATG"na sequência. Nas linhas abaixo, os códons ATG previstos são indicados com a letra "i"(initiation), outros exemplos de ATG são indicados pela letra "N"(non-start). Os pontos são definidos para todos os outros elementos da sequência. As pontuações variam de 0.0 até 1.0, quando superiores a 0.5 representam um provável início de tradução. A Figura 2 apresenta um exemplo de saída da ferramenta.

```
[...]TCAATCCCATCCGAGCCATTGTGGACAACATGAAGGTGAAGCCGAATCCGAACAAAACCG[...]
[...].............................i............................[...]
Pos     Score      Pred
------------------------
 98     0.711      Yes
123     0.408      -
129     0.317      -
153     0.223      -
222     0.151      -
239     0.727      Yes
284     0.770      Yes
```

Figura 2. Resultado do NetStart retirado do teste com *mus musculus* (camundongo).

2.2.2. TIS MINER

DNAFSMiner foi construído com as tecnologias de mineração e dados estatísticos. A metodologia consiste em três etapas:

- Gerar recursos candidatos a partir das sequências;
- Seleção de características relevantes dos candidatos;
- Integração dos recursos selecionados, com SVM (support vector machines) para construir um sistema de classificação e previsão.

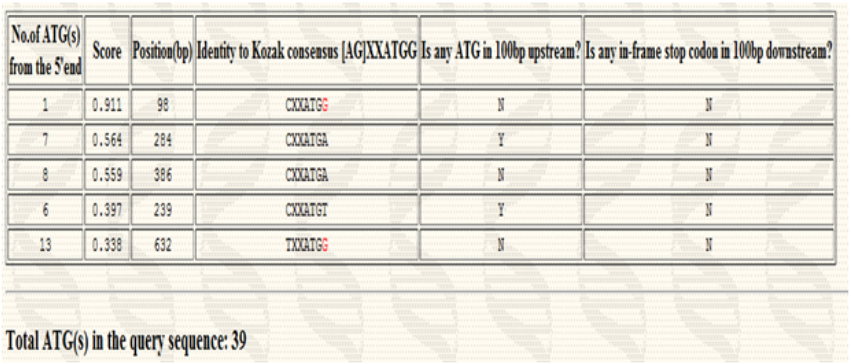
O TIS Miner foi treinado com 3312 sequências de mRNA vertebrados extraídas do GenBank (liberação de 95). Os dados foram analisados por Pedersen et al. A precisão de treinamento do modelo de classificação é de 92,45% à 80,19% de sensibilidade e especificidade de 96,48%. Para a utilização do softwares às sequências podem ser inseridas de duas maneiras, são estas:

- Colar uma única sequência (apenas os nucleotídeos) ou uma sequência no formato FASTA.
- Selecionar um arquivo FASTA em seu disco local.

Porém existe um limite máximo de 50.000 bps por sequência para evitar um longo tempo de espera para os usuários. O *Number of predictions* (número de previsões) é o número que será exibido de ATGs superiores na página de resultado, como configuração, o padrão é 5. A Figura 3 representa a página de saída do TIS Miner que é uma tabela com 6 colunas. Abaixo segue um exemplo:

Interpretação dos resultados:

- No.of ATG(s) from the 5' end - O número i nesta coluna da tabela indica que o candidato correspondente é o i ATG candidato da extremidade 5'.
- Score- Esta coluna mostra a pontuação (variando de (0,1)) da previsão de que "o candidato correspondente é um SIT". Quanto maior a pontuação, maior a probabilidade do candidato correspondente ser um verdadeiro SIT. Se o limite for definido como 0,6 (ou seja , se o escore de predição de um candidato é maior do que 0,6, então será prevista um SIT verdadeiro, caso contrário, será previsto como um não-SIT), a acurácia, sensibilidade, especificidade, e precisão são 72,2%, 54,6%, 89,7% e 84,1%, respectivamente.



No.of ATG(s) from the 5'end	Score	Position(bp)	Identity to Kozak consensus [AG]XXATGG	Is any ATG in 100bp upstream?	Is any in-frame stop codon in 100bp downstream?
1	0.911	98	CXXATGG	N	N
7	0.564	284	CXXATGA	Y	N
8	0.559	386	CXXATGA	N	N
6	0.397	239	CXXATGT	Y	N
13	0.338	632	TXXATGG	N	N

Total ATG(s) in the query sequence: 39

Figura 3. Resultado do TisMiner retirado do teste com Mus musculus (camundongo).

- Position (pb) - Esta coluna é a posição do candidato correspondente na sequência de ácido nucléico apresentados.
- Identity to Kozak consensus [AG]XXATGC - De acordo com a matriz de peso (Kozak) foi desenvolvido para predição de SIT, um resíduo de G tende a seguir um SIT verdadeiro, enquanto se espera encontrar um A ou G três nucleotídeos *upstream* de um SIT verdadeiro. Esta coluna mostra como a ATG candidato se encaixa nesse consenso.
- Is any ATG in 100bp *upstream*? - Essa coluna indica se um existe um ATG dentro de 100 pbs na região *upstream* do candidato.
- Is any in-frame stop codon dentro de 100bp *downstream*? - Essa coluna indica se existe um stop codon dentro de 100 pbs na região *downstream* do candidato.

3. Metodologia

Essa seção trata da descrição de toda a metodologia utilizada para o desenvolvimento do trabalho, a saber: (1) a forma de extração de sequências positivas (que codificam proteínas) e negativas (que não codificam proteínas) do mRNA (2) os classificadores utilizados; (3) a ferramenta criada. As seções seguintes descrevem cada uma destas fases.

3.1. Extração das sequências positivas e negativas

Para se utilizar o classificador SVM, sequências positivas (SIT) e negativas (não-SIT) foram extraídas através de uma ferramenta implementada, LOCALIZE, com janela do seguinte tamanho: -10+30. As sequências foram extraídas apenas de arquivos contendo a quantidade mínima de nucleotídeos da região *upstream* da janela considerada. Dessa forma, todas as sequências que não continham esse número foram desconsideradas. A Figura 4 apresenta exemplos de extração de sequências positivas e negativas, dada uma molécula de mRNA. O SIT é determinado pelo ATG destacado em vermelho, e é representado pelas posições +1, +2 e +3. A Figura 4 (a) apresenta um exemplo de uma sequência positiva. As partes (b) e (c) da Figura 4 apresentam exemplos de sequências negativas.

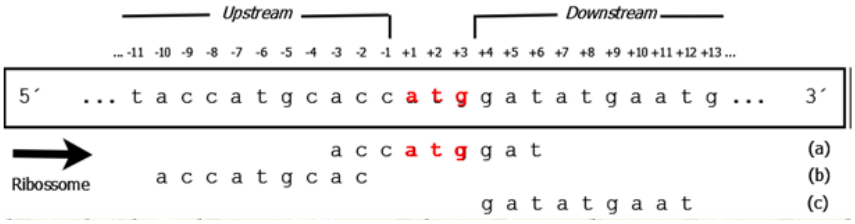


Figura 4. O ribossomo escaneia a sequência de mRNA da região 5' para a região 3' Encontrar um códon de ATG com contexto apropriado. A parte (a) da figura apresenta um exemplo de extração de sequências positivas (TIS), partes (b) e (c) apresentam sequências negativas. Fonte: Silva et. al (2010)

As sequências foram codificadas utilizando-se o esquema de codificação de 4 bits (A= 1000, C = 0100, G= 0010 e T = 0001). Dessa forma, sem considerar a inclusão de características e considerando-se, por exemplo, janela de tamanho -10+30, o classificador SVM possui 160 entradas (40 bases x 4 bits = 160), isto porque o tamanho das janelas é 10 nucleotídeos *upstream* e 30 nucleotídeos *downstream* centrados no ATG; e uma saída (0 ou 1), ou seja, a saída 1 representa que a sequência contém o códon ATG inicializador da tradução, e a saída 0 representa que a sequência não contém esse inicializador.

3.2. Support Vector Machine - SVM

Nesse trabalho, utilizou-se o algoritmo SVMlight, implementado por Joachims(1999) e disponível em http://svmlight.joachims.org/. Originalmente denominado "classificador de margem ótima", ela foi introduzida em (BOSER; GUYON; VAPNIK, 1992) para aplicação em problemas de classificação binária. Em Cortes e Vapnik (1995), sendo chamado de "máquina de vetores de suporte", foi proposta uma maneira de se lidar eficientemente com os exemplos que são notadamente incorretos, isto é, que estão fora da região de sua classe. A denominação "máquina de vetores de suporte", ou Support Vector Machine (SVM), enfatiza a importância que os vetores mais próximos da margem de separação representam, uma vez que eles determinam a complexidade da SVM (BURBIDGE; BUXTON, 2001). O motivo da escolha do algoritmo SVMlight para implementar o treinamento do SVM foi baseado nos seguintes fatores (NOBRE; ORTEGA; BRAGA, 2007):

- É projetado para operar com grande número de dados de treinamento não tendo problema com quantidade de dados armazenados na memória;
- O tempo de processamento para grandes tarefas é muito satisfatório;
- Trabalha com problemas de todos os tipos: classes separáveis, classes não separáveis e ainda problemas com muita interseção (ruído) entre as classes.

3.3. Validação

A fim de avaliar e analisar a ferramenta proposta, será utilizada a seguinte metodologia de validação.

De acordo com Silva et al(2010) a acurácia mede a proporção de predições corretas, conforme a Equação abaixo.

$$Ac = 100 * \frac{TP + TN}{TP + TN + FN + FP}$$

onde TP, TN, FP e FN denotam o número de verdadeiros positivos, verdadeiros negativos, falso positivos e falso negativos, respectivamente.

A precisão mede a proporção dos possíveis SIT que são certamente SIT, conforme a Equação abaixo.

$$Pr = 100 * \frac{TP}{TP + FP}$$

A sensibilidade, também conhecida como taxa de verdadeiro-positivo, refere-se à porcentagem de acertos dentro da classe positiva, ou seja, mede a proporção de SIT que foi corretamente classificada como SIT, conforme a Equação abaixo.

$$Se = 100 * \frac{TP}{TP + FN}$$

A especificidade, também conhecida como taxa de verdadeiro-negativo, refere-se à porcentagem de acertos dentro da classe negativa, ou seja, mede a proporção de não-SIT que foi corretamente reconhecida como não-SIT, conforme a Equação abaixo.

$$Sp = 100 * \frac{TN}{TN + FP}$$

Já a acurácia ajustada é obtida como sendo a média das medidas de sensibilidade e especificidade, conforme a Equação abaixo.

$$Adj = \frac{Sensibilidade + Especificidade}{2}$$

3.4. LOCALIZE

A ferramenta criada LOCALIZE, possui *layout* leve conforme a Figura 5. Possui um campo para a inserção de sequências e possui a opção de selecionar um documento do tipo FASTA para a predição além de possuir um campo para selecionar se exibe ou não as sequências na tela de respostas, caso o usuário selecione ambas as opções será descartado o documento e será feita a predição através do campo de texto. Atualmente o sistema possui um limite de 500 sequências, porém esse valor é alterável mesmo com o software em funcionamento.

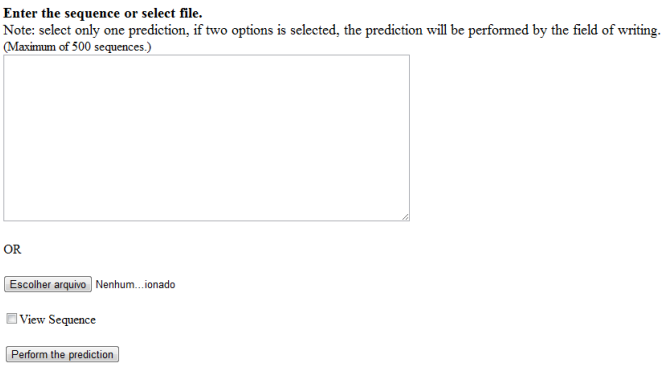


Figura 5. Layout do Sistema.

O seu funcionamento no momento dá-se de maneira bem simples, para a leitura do arquivo utiliza-se um StreamReader, e a leitura dos dados é feita com os métodos ReadLine (lê o arquivo linha a linha) e ReadToEnd (lê o arquivo todo de uma vez). Depois de feita a leitura do arquivo ou a leitura da sequência, chama-se um método para a codificação das sequências, então o sistema gera um arquivo de testes com a sequência codificada. Com o arquivo de testes criado o sistema realiza a chamada ao SVMlight que gerará um arquivo com os resultados. Depois de concluído a predição o sistema redireciona a página principal para a página de resultados e exibe o resultado conforme a Figura 6 e caso o usuário queira salvar os resultados poderá ser feito pelo botão Save.

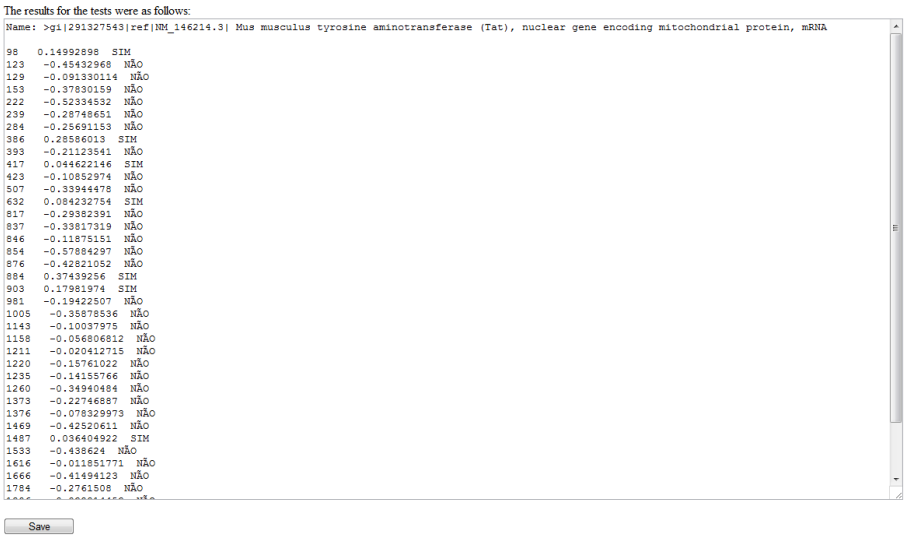


Figura 6. Tela de respostas do sistema.

4. Resultados

A fim de ilustrar o funcionamento da ferramenta, esta seção exibe os resultados dos testes realizados com o sistema de predição. A sequência de mRna utilizada nos testes foi o Mus musculos.

4.1. Validar Metodologia

O primeiro teste realizado foi com o intuito de validar qual ATG seria usado para realizar os testes, o primeiro ATG positivo ou o ATG positivo com o maior score. Foram testadas 40 sequências sendo elas a de numero 1 até a de numero 40. Os resultados obtidos estão na tabela 1.

Tabela 1. Resultado Metodologia

	Quantidade de Acertos	Porcentagem
Score	20	50%
Primeiro	27	67,50%

Como pode ser visto no resultado a escolha pelo primeiro ATG positivo teve porcentagem superior ao de maior Score, tornando-se assim a escolha para os testes.

4.2. Testes

Para a escolha das sequências testadas não houve padrão, foram retirados blocos de 10 sequências aleatoriamente. O numero de ATG‘s positivos e negativos, podem ser vistos na tabela 2.

Tabela 2. Número de ATG‘s

TP	FP	TN	FN
24	10	65	6

Tabela 3. Resultados

Acurácia	84,7619
Precisão	70,5882
Sensibilidade	80,0000
Especificidade	86,6667

Os resultados de acurácia, precisão, sensibilidade e especificidade podem ser conferidos na tabela 3.

Pode-se observar que os níveis estão bem elevados mostrando um bom resultado da ferramenta. A tabela 4 faz um comparativo entre a ferramenta criada e as já existentes.

Tabela 4. Comparativo entre Ferramentas

Method	TP	FP	TN	FN	Ac	Pr	Se	Sp
NetStart	72	21	113	15	83,71	77,42	82,76	84,33
TIS Miner	82	15	134	11	89,26	84,54	88,17	89,93
Localize	24	10	65	6	84,76	70,59	80,00	86,67

Pode-se observar que a ferramenta criada possui acurácia e especificidade maiores que o NetStart e pouco inferiores as do TIS Miner, ferramentas essas largamente conhecidas e avançadas, perdendo apenas na precisão e na sensibilidade, no entanto por valores bem pequenos.

Com esses testes pode-se observar que esta ferramenta esta em um nível muito bom, mostrando assim a eficácia da ferramenta.

5. Conclusão

Como visto nesse trabalho, a tarefa de previsão de SIT não é um problema trivial de ser resolvido. Inúmeros métodos têm sido explorados e avaliados com este fim.

Apesar de já existirem vários softwares de predição, estes possuem algumas limitações, como por exemplo o fato de sempre exibir as sequências testadas, pois em alguns casos, isso polui muito o visual, dificultando a visualização das informações realmente importantes, o número de sequências que é possível de se testar de uma só vez também é muito baixo, o maior é o NetStart com no máximo 50 sequências e outro ponto de defasagem é a impossibilidade de se salvar os resultados diretamente, obrigando o usuário a copiar e colar o que deseja.

O software proposto tem o intuito de adequar-se a essas limitações, possibilitando a escolha de exibir ou não as sequências, possui o número máximo de sequências configurável, ou seja, mesmo com o software em funcionamento é possível alterar o número máximo de sequências de acordo com a capacidade do servidor, além do mais é possível fazer o download dos resultados, salvando onde o usuário necessite.

Além da ferramenta que veio a auxiliar as predições de maneira mais simples, pode-se concluir que a metodologia utilizada contribui de maneira significativa para a previsão de SIT (Silva et. al, 2010).