# Sumário

# 1 Resumo

Atualmente muitas pesquisas estão sendo conduzidas no intuito de entender a expressão gênica. Dentro destas pesquisas, estão as que se empenham na identificação de elementos regulatórios. Uma vez identificado e entendido o papel de um elemento regulatório, estudos podem ser feitos utilizando esse elemento (ou conjunto de elementos), como por exemplo o melhoramento genético de um organismo. Para facilitar estas pesquisas muitos algoritmos foram desenvolvidos. Entretanto nenhuns dos algoritmos apresentados até o momento são livres de erros. Neste projeto serão conduzidas pesquisas para a identificação dos algoritmos mais expressivos e a implementação dos mesmos, integrando-os em uma ferramenta onde o usuário poderá fazer comparações entre as diversas abordagens, identificando os resultados mais satisfatórios. Os estudos serão conduzidos no genoma da soja, uma cultura amplamente utilizada e importante para a economia nacional.

# 2 Objetivos

Objetivo geral:

- Desenvolver uma ferramenta de identificação de elementos regulatórios, integrando as principais abordagens de predição.

Objetivos específicos:

- Implementação dos algoritmos de predição que farão parte da ferramenta.
- A ferramenta proposta poderá auxiliar pesquisadores antecipando os estudos experimentais de regulação da expressão gênica.

# 3 Justificativas

A identificação dos elementos regulatórios através de uma ferramenta computacional poderá auxiliar diversas pesquisas na área biológica antecipando os estudos experimentais de regulação da expressão gênica. Isto pode ser uma grande aliada em estudos de grande importância atualmente, como por exemplo o melhoramento genético, que vem sendo objeto de estudos a alguns anos, mostrando-se eficiente na modificação genética de grãos como soja, milho e arroz utilizados na alimentação. Desta forma, é possível obter-se culturas resistentes a condições adversas como baixa temperatura, seca e alta salinidade. A implementação da ferramenta de integração e dos principais algoritmos de predição, permitira fixar e utilizar os conhecimentos obtidos nas diversas áreas da ciência da computação estudadas até o momento. Como técnicas de programação, engenharia de software, algoritmos e estruturas de dados.

# 4 Fundamentação Teórico-metodológica

Em cada célula dos organismos vivos, para que ocorra a transcrição de um gene é necessário a ação conjunta dos fatores de transcrição e dos elementos regulatórios. Os fatores de transcrição são proteínas que se ligam nos elementos regulatórios. Já os elementos regulatórios são pequenos segmentos de DNA localizados em uma região antes do gene que será transcrito. Essa região é chamada de região promotora (ou reguladora), na figura 4.1, podemos observar a região promotora e os elementos regulatórios, onde os fatores de transcrição se conectarão.

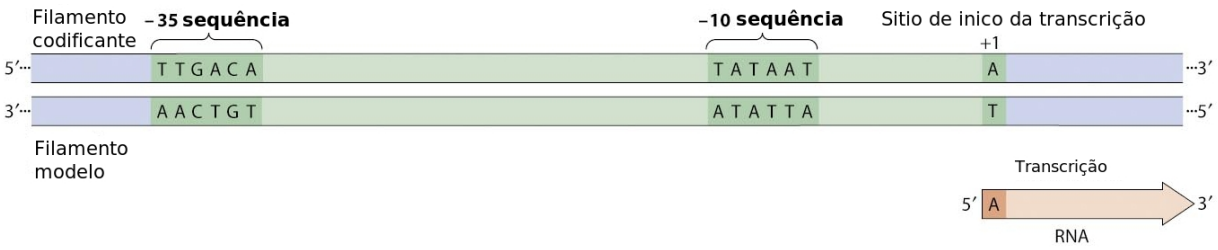


Figura 4.1: Região promotora

É de grande importância a identificação dos elementos regulatórios. Uma vez que eles estão ligados com a expressão de um gene. O conhecimento agregado com a identificação desses elementos pode levar a melhoramentos genéticos de organismos importantes para o consumo e a economia mundial, como por exemplo de grãos como o arroz, soja e milho.

No empenho de encontrar elementos regulatórios diversos algoritmos foram propostos. Das e Dai (DAS; DAI, 2007) os classificaram em três grupos:

- Os baseados em sequências promotoras de genes que são regulados pelos mesmos fatores de transcrição (genes co-regulados); estes métodos se concentram em apenas um único genoma.

  Este ainda é subdividido em : predição probabilistica e predição baseada em palavras.

- Os que utilizam sequências promotoras de genes ortólogos, que são sequências de DNA similares a várias espécies, indicando que estas espécies derivaram de um ancestral comum, também chamados de métodos de rastros filogenéticos.
- Os métodos que combinam rastros filogenéticos e sequências promotoras de genes coregulados.

Esses algoritmos utilizam de sequências de DNA como entrada. A sequência de DNA é formada pelo alfabeto (A, C, G e T). Cada letra representa uma base nitrogenada no DNA, que formará extensas palavras sem nenhuma restrição de posições. Para que possa ser manipulada no computador essas sequências são geralmente representadas por arquivos texto com a extensão *.fasta* (figura 4.2).

```
>HUMAN
CAGGTTATCAGCAACAACACAGTCATATCCATTCTCAATTAGCTCTACCACAGTGTGTGAACCAATGTATCCAGCACCAC
CTGTAACCAAAACAATTTTAGAAGTACTTTCACTTTGTAACTGAGCTGTCATTTATATTGAATTTTCAAAAATTCTTACTTT
TTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATACATATCCATATCTAATCTTACTT
ATATGTTGTGGAAATGTAAAGAGCCCCATTATCTTAGCCTAAAAAAAACCTTCTCTTTGGAACTTTCAGTAATACGCTTAAC
GCTCATTGCTATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTCCGTGC
TCCTCGTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAAT
CTAGCTTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGGCCCCACAAACCTTCAAATTAACGAATCAAATTAACA
CCATAGGATGATAATGCGATTAGTTTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATCTATTA
>RAT
CGGTTTAGCATCATAAGCGCTTATAAATTTCTTAATTATGCTCGGGCACTTTTCGGCCAATGGTCTTGGTAATTCCTTTGCGC
TAGAATTGAACTCAGGTACAATCACTTCTTCTGAATGAGATTTAGTCATTATAGTTTTTTCTCCTTGACGTTAAAGTATAGAG
TATATTAACAATTTTTTGTTGATACTTTTATGACATTTGAATAAGAAGTAATACAAACTGAAAATGTTGAAAGTATTAGTTAAA
>MOUSE
CATTAATTTTGCTTCCAAGACGACAGTAATATGTCTCCTACAATACCAGTTTCGCTGCAGAAGGCACATCTATTACATTTACTG
AGCATAACGGGCTGTACTAATCCAAGGAGGTTTACGGACCAGGGGAACTTTCCAGATTCAGATCACAGCAATATAGGACTAG
```

Figura 4.2: Exemplo de um arquivo de sequência de DNA

Esses arquivos de sequência de DNA podem ser encontrados embancos de dados públicos como SoyDB, Phytozome e o TRANSFAC.

Dentro dessas sequências, estão os elementos regulatórios (figura 4.3). Mas diversos fatores como, a remoção e a mutação de uma base e talvez o principal: a não padronização dos elementos regulatórios, tornam difícil a sua identificação.

## 4.1 Algoritmos de predição de elementos regulatórios

Apesar dos avanços para encontrar elementos regulatórios, a busca *in silico* não é tão precisa quanto, por exemplo, a classificação de genes, gerando muitos resultados falsos. Isto deixa em aberto um vasto campo para ser explorado (ROMBAUTS et al., 2003). Nos parágrafos a seguir serão discutidos alguns algoritmos.

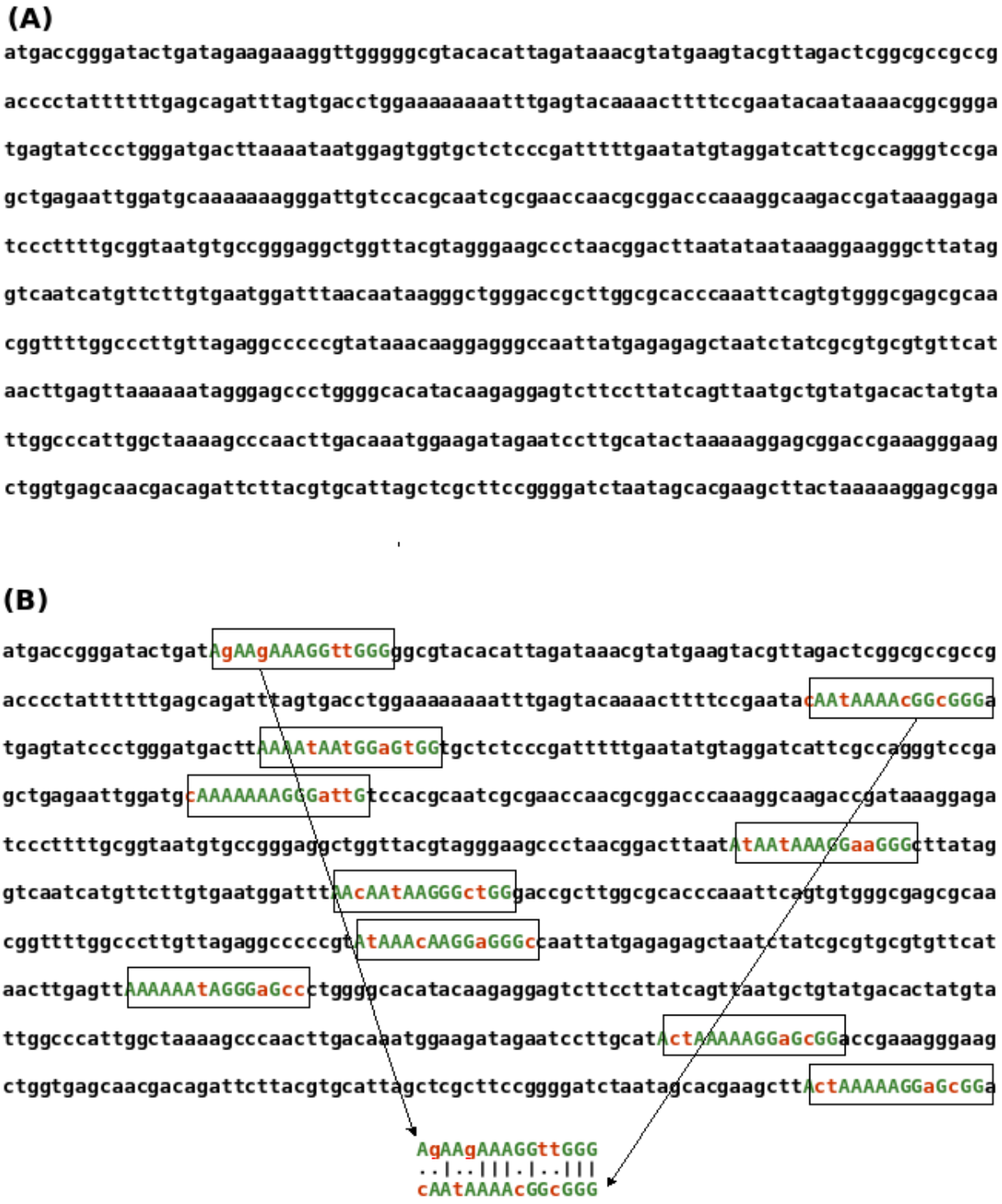


Figura 4.3: (A)Sequência regulatória (B)Elementos regulatórios identificados

Para identificar os elementos regulatórios em leveduras (*Saccharomyces cerevisiae*) (HELDEN; ANDRé; COLLADO-VIDES, 1998), propuseram o Oligo-Analysis, um algoritmo baseado na predição em palavras. Neste algoritmo foram usadas sequências promotoras de genes co-regulados, e para a identificação dos elementos foi utilizado o método de enumeração exaustiva, que computa todas as possíveis subsequências possíveis através das sequências

promotoras.

Blanchette *et al.* (BLANCHETTE; TOMPA, 2002), utilizaram da abordagem de predição baseada em rastros filogenéticos, para desenvolver o FootPrinter. Assim como todos os outros algoritmos que se baseiam nessa abordagem, eles assumem que os elementos regulatórios são regiões conservadas que não sofreram muitas mutações ao longo da evolução. Este algoritmo tem como entrada as sequências promotoras de várias espécies e a árvore filogenética das espécies relacionadas. As sequências de cada espécie são inseridas nas folhas da árvore, então são feitas varias comparações das sequências, a começar das folhas. As sequências de tamanho $k$, mais conservadas são promovidas para o nível acima da árvore, são feitas novas comparações até ser atingido a raiz da árvore, chegando a sequências ótimas, que passarão por mais comparações, desta vez da raiz até as folhas, finalizando o algoritmo com os elementos preditos.

Sinha *et al.* (SINHA; BLANCHETTE; TOMPA, 2004), propuseram um algoritmo combinando as técnicas probabilísticas e de rastros filogenéticos. O algoritmo permite a entrada de sequências promotoras de genes ortólogos, relacionadas com a árvore filogenética definida pelo usuário. As sequências dos elementos regulatórios podem ser conservadas ou não conservadas, o algoritmo as trata de maneira diferente. O algoritmo permite uma flexibilidade, na escolha da árvore filogenética, podendo escolher espécies distantemente relacionadas, que além da identificação dos elementos conservados entre as espécies possibilita a identificação de elementos que não estão relacionados com as espécies ortólogas.

# 5 Contribuições e ou resultados Esperados

Com o desenvolvimento da ferramenta espera-se possibilitar ao usuário, identificar os elementos regulatórios através das diversas abordagens. Será possível comparar os resultados entre as abordagens e identificar os mais expressivos. Utilizando de algoritmos com uma baixa taxa de falsos positivos, espera-se que a ferramenta seja confiável para o uso na predição de elementos regulatórios.

# 6 Procedimentos Metodológicos/Métodos e Técnicas

Para o desenvolvimento deste trabalho serão realizadas as seguintes atividades:

- Estudos dos algoritmos para a predição dos elementos regulatórios.

  Nesta etapa serão estudadas as técnicas de predição de elementos regulatórios.

- Implementação de métodos para a predição de elementos regulatórios.

  Nesta fase, serão implementados os métodos mais significativos para a predição dos elementos regulatórios. Se algum algoritmo apresentar-se muito complexo exigindo um tempo de implementação que extrapolará o cronograma e havendo uma implementação do mesmo, de código livre, está será integrada na ferramenta, juntamente com os algoritmos implementados.

- Integração dos algoritmos implementados, construindo uma ferramenta única para a predição dos elementos regulatórios.

  Neste ponto, todos os algoritmos implementados serão agrupados em uma única ferramenta, para facilitar a utilização do usuário. A arquitetura da ferramenta proposta é descrita na figura 6.1, onde podemos visualizar o esquema de funcionamento.

- Validação dos elementos encontrados no genoma da soja.

  Neste passo os elementos encontrados serão comparados com os elementos já existentes, avaliando a precisão dos métodos implementados.

Para implementação dos algoritmos será utilizada a linguagem de programação Python, devido ao grande desempenho desta linguagem e por oferecer um conjunto de métodos que facilita a manipulação de *strings*.

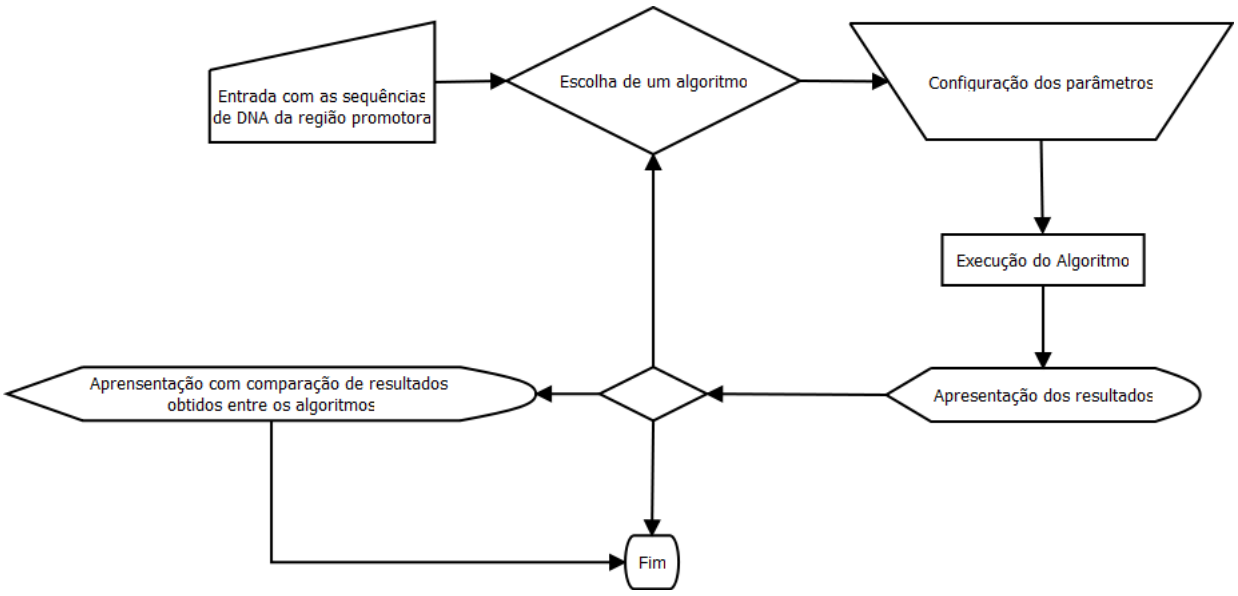


Figura 6.1: Esquema da funcionamento da ferramenta

# 7 Cronograma de Desenvolvimento

O cronograma de desenvolvimento dos trabalhos será:

| Mês | Descrição | Indicador físico |
|---|---|---|
| Abril | Revisão bibliográfica | Texto com 5 páginas contendo a revisão bibliográfica |
| Maio | Escolha dos algoritmos para serem implementados | Texto explicativo dos algoritmos |
| Junho | Implementação dos algoritmos escolhidos<br>Implementação dos algoritmos escolhidos | Relatório das atividades desenvolvidas |
| Agosto | Implementação dos algoritmos escolhidos | Relatório das atividades desenvolvidas |
| Setembro | Termino da implementação dos algoritmos, início da integração e testes dos algoritmos | Relatório das atividades desenvolvidas |
| Outubro | Termino e testes da ferramenta de integração | Relatório das atividades desenvolvidas |

# Referências Bibliográficas

BLANCHETTE, M.; TOMPA, M. Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting. *Genome Research*, v. 12, n. 5, p. 739–748, maio 2002. ISSN 1088-9051. Disponível em: <http://dx.doi.org/10.1101/gr.6902>.

DAS, M.; DAI, H. K. A survey of DNA motif finding algorithms. *BMC Bioinformatics*, v. 8, n. Suppl 7, p. S21+, 2007. ISSN 1471-2105. Disponível em: <http://dx.doi.org/10.1186-/1471-2105-8-S7-S21>.

HELDEN, J. van; ANDRé, B.; COLLADO-VIDES, J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies1. *Journal of Molecular Biology*, Centro de Investigación sobre Fijación de Nitrógeno, Universidad Nacional Autónoma de México, AP565A Cuernavaca, Morelos, 62100, México. jvanheld@ebi.ac.uk, v. 281, n. 5, p. 827–842, set. 1998. ISSN 00222836. Disponível em: <http://dx.doi.org/10.1006/jmbi.1998.1947>.

ROMBAUTS, S. et al. Computational Approaches to Identify Promoters and cis-Regulatory Elements in Plant Genomes. *Plant Physiol.*, Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, B-9000 Gent, Belgium., v. 132, n. 3, p. 1162–1176, jul. 2003. ISSN 0032-0889. Disponível em: <http://dx.doi.org/10.1104/pp.102.017715>.

SINHA, S.; BLANCHETTE, M.; TOMPA, M. PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, Center for Studies in Physics and Biology, The Rockefeller University, New York, NY 10021, USA. saurabh@lonnrot.rockefeller.edu, v. 5, n. 1, p. 170+, out. 2004. ISSN 1471-2105. Disponível em: <http://dx.doi.org/10.1186/1471-2105-5-170>.