UNIVERSIDADE FEEVALE

SAMUEL ANTONIO KLEIN

**FERRAMENTA EDUCATIVA PARA ALGORITMOS DE ALINHAMENTO EM BIOLOGIA COMPUTACIONAL**

##### (Título Provisório)

###### Anteprojeto de Trabalho de Conclusão

Novo Hamburgo

2014

SAMUEL ANTONIO KLEIN

**FERRAMENTA EDUCATIVA PARA ALGORITMOS DE ALINHAMENTO EM BIOLOGIA COMPUTACIONAL**

(Título Provisório)

Anteprojeto de Trabalho de Conclusão de Curso, apresentado como requisito parcial

à obtenção do grau de Bacharel em

Ciência da Computação pela

Universidade Feevale

Orientador: Ricardo Ferreira de Oliveira

Novo Hamburgo

2014

# RESUMO

A bioinformática é um campo da biologia relacionada à computação que está experimentando um rápido crescimento e desenvolvimento devido à necessidade de manipular-se com grande quantidade de dados genéticos e bioquímicos. A bioinformática abrange um número amplo de tópicos, sendo um dos principais o alinhamento de sequências. Alunos da área de biologia, com pouca experiência em informática, têm razoável dificuldade no aprendizado de algoritmos de alinhamento. Para aperfeiçoar a aprendizagem de alunos nas disciplinas que envolvem ferramentas de bioinformática pretende-se criar um ambiente educativo voltado para ilustrar as etapas e processos do alinhamento de cromossomos e proteínas. Com este ambiente visa-se proporcionar aos estudantes de Bioinformática uma ferramenta. Este trabalho apresentará os métodos de alinhamento utilizados em disciplinas de bioinformática culminando com o desenvolvimento de uma ferramenta educativa que proporcione uma clara compreensão de como estes algoritmos trabalham para o aluno com pouca formação na área de computação.

Palavras-chave: Bioinformática. Biologia Computacional. Apoio ao ensino de Bioinformática. Biologia Molecular. Alinhamento de Sequências.

SUMÁRIO

MOTIVAÇÃO 5

OBJETIVOS 7

METODOLOGIA 8

CRONOGRAMA 10

BIBLIOGRAFIA 12

# MOTIVAÇÃO

A bioinformática é um campo da biologia relacionada à computação que está experimentando um rápido crescimento e desenvolvimento devido à necessidade de manipular-se com grandes quantidades de dados genéticos e bioquímicos. Estes dados são provenientes do trabalho de vários pesquisadores e são armazenados em grandes bancos de dados de acesso público. Segundo Larson (2006, p.116), a bioinformática possui mais de 140 programas que compõem os pacotes de análise e comparação de sequência de proteínas e nucleotídeos que são encontrados em banco de dados como *EMBL, GenEMBL, PIR-Protein, Restriction Enzyme Database* e *GenBank*.

Nos bancos de dados há também uma grande variedade de informações sobre estruturas moleculares, expressão gênica diferencial, diversidade genética, evolução, etc. que podem ser extraídas pela bioinformática. Um dos grandes desafios é o desenvolvimento de procedimentos pelos quais esses dados podem ser “inseridos” e "extraídos" em bancos de dados secundários, pelos pesquisadores. (Santos; Ortega 2014, p.7).

O número de cursos que contemplam a disciplina de bioinformática tem crescido rapidamente, uma vez que um número crescente de ferramentas de biologia computacional tem sido usado por biólogos moleculares, bioquímicos, cientistas da computação e médicos, entre outros. Para Lesk (2005, p.14, tradução nossa), por exemplo,

A ciência da computação é um campo novo e próspero, com o objetivo de fazer uso mais eficaz de *hardware* de tecnologia da informação. Certas áreas de ciência da computação teórica colidir mais diretamente em bioinformática. Deixe-nos considera-los com referência a um problema específico biológica, recuperar a partir de uma base de dados todas as sequências semelhantes a uma sequência de sonda.

Segundo Pevzner e Shamir (2011, p.18), a bioinformática tornou-se parte essencial da biologia moderna, principalmente quando se realiza pesquisas biológicas com as ferramentas de bioinformática, como BLAST ou *Basic Local Alignment Search Tool*. De acordo com Santos e Ortega (2014, p.4),

Atualmente a bioinformática é imprescindível para a manipulação dos dados biológicos. Ela pode ser definida como uma modalidade que abrange todos os aspectos de aquisição, processamento, armazenamento, distribuição, análise e interpretação da informação biológica.

O constante avanço da ciência e das tecnologias analisa-se que o ensino de Biologia e Ciências ainda permanecem, em muitos casos, limitados às aulas expositivas com mínima participação dos alunos. A utilização de outras modalidades didáticas tais como: audiovisuais, ferramentas computacionais, etc. (LEPIENSKI; PINHO, 2014, p. 2), gerando uma necessidade em material, conteúdo digital e ferramenta de apoio.

O uso de ferramentas didáticas no estudo de bioinformática torna-se uma necessidade importante para compreensão dos comportamentos gerados pelos alinhamentos, de forma auxiliar o melhor entendimento do funcionamento de ferramenta já consolidada no mercado, o uso de amostras menores e exemplificado o funcionamento do processo alinhamento.

O processo didático para elaboração do conceito de alinhamento de nucleotídeos e de proteínas trata-se da exploração de algoritmos de manipulação de forma detalhada, trazendo todo o passo de cada elemento e componente envolvido.

A comparação de pares ou mais sequencias é uma necessidade para identificar alguma serie caracteres ou padrões de caracteres que possui uma mesma ordem. Na figura 1 segue o exemplo de um alinhamento das palavras “ALINHAMENTO” e “ALINHARSEQUENCIA”, demonstrando um resultado da ferramenta utilizado.

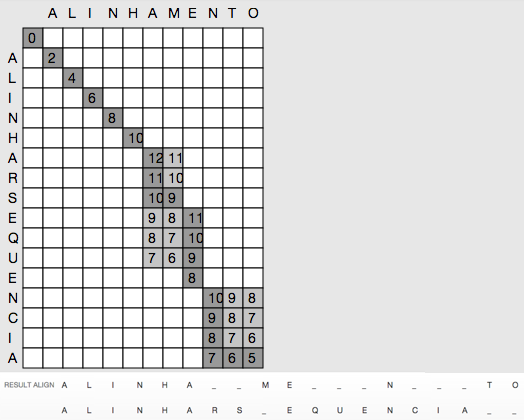


Figura 1 – Exemplo de alinhamento global

O processo de alinhamento de pares de sequências se define em matriz de pontos (dot matrix), programação dinâmica e dicionário de palavras ou k-tuplas (BLAST), no alinhamento utilizado matriz de pontos (dot matrix), estão divididos em dois mais importantes que são o global e o local.

Segundo Baxevanis e Ouellette (2001, p. 202), A ideia básica por trás de programação dinâmica vem da observação de outros caminhos provisórios que termina em um ponto ao longo do caminho verdadeiro ideal deve ser ele mesmo o caminho ideal que levaram a esse ponto. Assim, o caminho ótimo pode ser encontrado, por extensão periódica de outros caminhos ideias.

O algoritmo de alinhamento global (Needleman-Wunsch), resultado no cálculo da matriz de valores em que os comportamentos resultam em uma amostra dos pares alinhados de forma central. O que conhecemos como o algoritmo de Needleman-Wunsch é uma aplicação de uma estratégia de melhor caminho chamado de programação dinâmica para o problema de encontrar o melhor alinhamento (NEEDLEMAN; WUNSCH, 1970).

Outro alinhamento derivado do global é o alinhamento local (Smith-Waterman), esta forma de alinhamento igualmente envolve a geração de uma matriz, mas localiza os pares de forma deslocada do centro (SETUBAL, J.; MEIDANIS J., 1997).

O estudo para entender os comportamentos do resultado de um alinhamento muitas vezes torna-se complexo para um aluno de biologia, pois não compreende a essência entre os dados disponibilizados e os resultados apresentados.

De acordo com Faleiro, Andrade e Junior (2011, 47), o grande avanço obtido até os dias de hoje, a rapidez com que surgem novas técnicas da biologia molecular e o gigantesco volume de dados e informações produzidos pelos projetos nessa área exigem que a bioinformática esteja em constante evolução, para um aperfeiçoamento de ferramentas de bioinformática para o desenvolvimento.

# OBJETIVOS

Objetivo geral:

Analisar os principais algoritmos de alinhamento contemplados em cursos de bioinformática e criar uma ferramenta educativa que ilustra as etapas do processo do alinhamento selecionado. Com base na ferramenta implementada analisar a mesma com auxílio de uso em disciplinas de bioinformática.

Objetivos específicos:

* Realizar revisão bibliográfica sobre o alinhamento de sequências;
* Analisar os algoritmos de alinhamento de sequências: Global, Local e Outros;
* Identificar as principais necessidades dos cursos de bioinformática;
* Determinar softwares de apoio ao ensino de bioinformática já existente;
* Identificar as etapas dos algoritmos onde há maior dificuldade de aprendizado;
* Realizar a modelagem da ferramenta educativa;
* Implementar um protótipo da ferramenta;
* Realizar experimentação da ferramenta em cursos de bioinformática.

# METODOLOGIA

Este trabalho será desenvolvido com base em pesquisas científicas, no intuito de exemplificar e auxiliar no sequenciamento de pares de nucleotídeos ou proteínas, com o uso da bioinformática. No decorrer das atividades serão realizadas avaliações do software com alunos e professores, com o conceito de aprimorar a ferramenta de acordo com as necessidades. A disponibilização da ferramenta será *OpenSource*, assim podendo haver contribuições de outros grupos interessados.

Quanto à natureza, o presente trabalho caracteriza-se como uma pesquisa aplicada uma vez que se pretende usar algoritmos já existentes no projeto de uma ferramenta educativa. Não se almeja criar novos algoritmos ou técnicas de alinhamento, senão, desenvolver meios didáticos para aprimorar o ensino de bioinformática para interessados com pouco conhecimento computacional. A pesquisa terá uma abordagem de caráter qualitativo pois pretende-se homologar a ferramenta junto a profissionais que lecionam a disciplina de bioinformática. No que tange aos objetivos da investigação pode-se afirmar que essa é uma pesquisa descritiva uma vez que não há ferramentas educativas desta natureza em português.

Quanto aos procedimentos, neste trabalho serão realizadas as formas:

a) pesquisa bibliográfica: inicialmente será executada uma pesquisa na literatura, com o objetivo de levantar os algoritmos mais frequentemente utilizados no ensino de bioinformática;

b) levantamento: Será aplicada para a coleta de informações e necessidades de modo a validar, ou não, a ferramenta proposta;

c) estudo de caso: disponibilizar a ferramenta para docentes da disciplina e verificar a satisfação e adequação da mesma.

Inicialmente nos meses entre fevereiro a abril do ano de 2014 será desenvolvido o anteprojeto. Logo após dará início ao desenvolvimento o Trabalho de Conclusão I, em que será levantado o referencial teórico, buscando conhecimento através de livros e artigos sobre técnicas de desenvolvimento para ambiente web e dispositivos móveis, técnicas de desenvolvimento orientado a testes e práticas de uso de algoritmos de bioinformática para alinhamento de sequência.

No início de agosto de anos de 2014 terá início o Trabalho de conclusão II. Com a base do referencial teórico adquirido anteriormente, será desenvolvido uma ferramenta de interface prática para o trabalho no ensino de alinhamento de sequência, a interface está disponível no modelo de página web, este modelo terá ferramenta para manipular e entender os comportamentos das sequencias. Em seguida será testado ambiente no ensino de alinhamento demonstrando suas melhorias para auxiliar o aprendizado.

# CRONOGRAMA

Trabalho de Conclusão I

1. Realizar pesquisa bibliográfica.

2. Elaborar e entregar o anteprojeto.

3. Identificar as principais necessidades da ferramenta.

4. Analisar softwares educacionais de bioinformática.

5. Elaborar os casos de uso da ferramenta.

6. Elaborar a modelagem de classes, eventos e interações.

7. Identificar os principais requisitos da ferramenta em termos de algoritmos e etapas a serem consideradas.

8. Elaborar e entregar o TC I.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Etapa | Meses | | | |
| Mar | Abr | Mai | Jun |
| 1. |  |  |  |  |
| 2. |  |  |  |  |
| 3. |  |  |  |  |
| 4. |  |  |  |  |
| 5. |  |  |  |  |
| 6. |  |  |  |  |
| 7. |  |  |  |  |
| 8. |  |  |  |  |

Trabalho de Conclusão II

9. Iniciar o desenvolvimento do TC II.

10. Dar prosseguimento ao desenvolvimento da ferramenta, com base nas conclusões do TC I. 11. Dar continuidade a modelagem de classes, eventos e interações. Revisar a modelagem como um todo.

12. Desenvolver a ferramenta.

13. Elaborar protótipo e testes da ferramenta.

14. Apresentar a ferramenta a professores de disciplinas de bioinformática.

15. Estabelecer conclusões.

16. Entregar e apresentar o TC II.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Etapa | Meses | | | |
| Ago | Set | Out | Nov |
| 9. |  |  |  |  |
| 10. |  |  |  |  |
| 11. |  |  |  |  |
| 12. |  |  |  |  |
| 13. |  |  |  |  |
| 14. |  |  |  |  |
| 15. |  |  |  |  |
| 16. |  |  |  |  |

# BIBLIOGRAFIA

BAXEVANIS, Andreas D.; OUELLETTE, B.F. Francis. **Bioinformatics:** **A Practical Guide to the Analysis of Genes and Proteins.** 2ª ed. New York US:  WILEY, 2001.

FALEIRO, Fabio Gelape; Andrade, Solange R. M.; JUNIOR, Fabio B. dos R.. **BIOTECNOLOGIA Estado da arte e aplicações na agropecuária.** Embrapa Cerrados, 2011.

KEEDWELL, Edward; NARAYANAN, Ajit. **Intelligent Bioinformatics: The Application of Artificial Intelligence Techniques to Bioinformatics Problems.** New Delhi India: WILEY, 2005.

LARSON, Richard S.. **Bioinformatics and Drug Discovery.** New Jersey US: HUMANA PRESS, 2006.

LEPIENSKI, Luis Marcos; PINHO, Kátia E. Prus. **Recursos didáticos no ensino de biologia e ciências**. Disponível em <http://www.diadiaeducacao.pr.gov.br/portals/pde/arquivos/400-2.pdf>. Acesso em 19 abr. 2014.

LESK, Arthur M.. **Introduction to Bioinformatics.** 2ª ed. New York US: OXFORD, 2005.

NEEDLEMAN, Saul B.; WUNSCH, Christian. **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** J. Mol. Biol. 48, p. 443–453, 1970.

PEVZNER, Pavel; SHAMIR, Ron. **Bioinformatics for Biologists**. Cambridge UK: UNIVERSITY PRESS, 2011.

PRODANOV, Cleber C.; FREITAS, Ernani C. de. **METODOLOGIA DO TRABALHO CIENTÍFICO: Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico.** 2ª ed. Novo Hamburgo: FEEVALE, 2013.

SANTOS, Fabrício; ORTEGA, José Miguel. **Bioinformática aplicada a genômica**. Disponível em <<http://www.icb.ufmg.br/lbem/aulas/grad/tge/bioinfo/bioinfogenomica.pdf>>. Acesso em 23 mar. 2014.

SETUBAL, Carlos; MEIDANIS, João. **Introduction to Computational Molecular Biology.** Boston: PWS, 1997.

YE, Shui Qing**. Bioinformatics: A Practical Approach**. London UK: TAYLOY & FRANCIS, 2008.