FERRAMENTAS DE BIOINFORMÁTICA: MANIPULAÇÃO DE SEQÜÊNCIAS E RECUPERAÇÃO DE REGIÕES FLANQUEADORAS DE UM ALVO.

Diógenes C. D. Ribeiro [1]

[1]Univap/IPD, dicuri@univap.br

Resumo- A bioinformática é uma área multidisciplinar, que envolve os profissionais das ciências biológica e computacional. Este trabalho mostra o desenvolvimento de três ferramentas, utilizando técnicas de biologia computacional, teoria das pequenas amostragens e modelagem de dados utilizando UML (*Unified Modeling Language*). Estes programas foram desenvolvidos em plataforma Linux, *open source*, utilizando a linguagem de programação Perl. Os programas aqui desenvolvidos para agilizar e facilitar algumas análises e comparações, alinhando e localizando regiões de *upstream* e *dowstream* de uma seqüência, possibilitando assim que o usuário possa desenhar iniciadores ou buscar subseqüências à montante ou à jusante.

Palavras-chave: Bioinformática, Scripts, Iniciadores DNA, Ferramentas on-line.
Área do Conhecimento: II – CIÊNCIAS BIOLÓGICAS.

Introdução

A bioinformática é um campo emergente da pesquisa que utiliza ferramentas computacionais avançadas para o armazenamento, análise e apresentação de dados biológicos e moleculares. Ela surgiu da necessidade de gerenciamento da quantidade maciça de dados gerados pelos projetos de seqüênciamento, e inclui métodos de análise de seqüências moleculares e pesquisa em banco de dados [1] e [2]. Um segmento que vem se destacando dentro da bioinformática é a biologia computacional que envolve, vários profissionais, de formação biológica e computacional. A biologia computacional é um novo conceito que pode ser entendido de maneira mais ampla como sendo a aplicação de técnicas e ferramentas da informática aos problemas biológicos.

Entre estas diversas áreas, a biologia molecular aliada à genética é a área que mais utiliza essas técnicas computacionais, das quais pode-se destacar a teoria da computação, principalmente através da formulação de algoritmos para solução dos diversos e novos problemas surgidos nos últimos anos, e a teoria e prática de banco de dados, que são necessárias para lidar com a imensa quantidade de dados gerados por projetos genomas e proteomas. O ácido desoxirribonucléico (DNA) é a unidade básica dos genes, que contém as informações necessárias para o desenvolvimento dos organismos vivos. A esse conjunto de genes é dado o nome de genoma. Se fosse possível varrer um genoma. a procura de todos os seus genes, poderiam ser descritas todas as características DNA dependentes que seu possuidor virá a ter, sendo este o objetivo de qualquer projeto genoma [3].

O envolvimento de técnicas computacionais, especialmente o desenvolvimento de algoritmos eficientes torna-se indispensável para uma boa análise dos dados gerados, e assim a finalização com sucesso de cada projeto. A riqueza dessa parceria, informática e biologia molecular, criaram a necessidade de profunda interação entre especialistas de computação e biólogos. Atualmente essa interação ainda tem sido difícil, dada a "lacuna cultural", que existe entre essas duas áreas. Segundo [2], a diminuição dessa "lacuna" é um dos maiores desafios.

Este trabalho aplica técnicas computacionais objetivando levantar o maior número de informações biológicas e estatísticas possíveis, ajudando assim o usuário a comparar, verificar e entender o significado de seus dados. Através de análises computacionais foram desenvolvidas três ferramentas para auxiliar de maneira fácil e rápida a análise de seqüências no formato (FASTA). A ferramenta é disponível *on-line*, com livre acesso para os usuários remotos e no Laboratório de Biologia Molecular e Genomas.

Materiais e Métodos

Sem dúvida, a comparação de seqüências é a operação básica mais importante na área da biologia computacional, servindo de base para muitas outras manipulações mais elaboradas. Ela consiste em encontrar trechos semelhantes às seqüências de entrada *query* em um banco de dados em questão. Contudo, por trás desta aparente simplicidade, esconde-se uma vasta gama de problemas distintos, com formalizações diversas, muitos deles exigindo algoritmos e

estruturas de dados próprios para sua execução eficiente [3]. A busca por identidade é um processo de comparação de seqüências, com função desconhecida, com todas as seqüências constantes em um banco de dados, com funções já conhecidas, na tentativa de inferir uma função para essas "novas" seqüências, por avaliação de equivalências e por suas anotações biológicas como descrito no próprio banco de dados e na literatura.

Os desenvolvimentos de scripts para a solução dos problemas biológicos computacionais são específicos e para isso utilizamos a linguagem de programação *Pratical Extraction and Report Languag*", acrônimo *Perl* [6], sob a plataforma Linux [7] esta linguagem permite a criação de aplicações personalizadas, possibilitando a resoluções dos problemas enfrentados no dia-a-dia durante o decorrer dos projetos de biologia computacional. Devido o sistema operacional Linux ser um sistema operacional confiável, seguro e estar disponível de maneira livre, seu uso foi disseminado no meio acadêmico. Porém, existem algumas regras determinadas pelo consórcio GNU's Not Unix (GNU) (GNU é um acrônimo recursivo).

Dentro do projeto "Ferramentas de Bioinformáticas", foram desenvolvidos três *scripts*, para facilitar e agilizar algumas tarefas, como, por exemplo, encontrar regiões de igualdade e calculo de porcentagens das bases das seqüências, fazer o reverso de uma seqüência e a sua complementar e por último localizar a posição da seqüência dentro de outra sequencia, indicando ao usuário as regiões antes e depois do alinhamento, isso possibilita ao usuário a desenhar iniciadores ou buscar sequencias de, promotores ou atividades, com uma maior segurança, e saber exatamente onde esta seqüência se alinha no seguimento de DNA. O endereço a ferrametas estão disponíveis no endereço:
http://cloneone.univap.br/Tools/index.html.

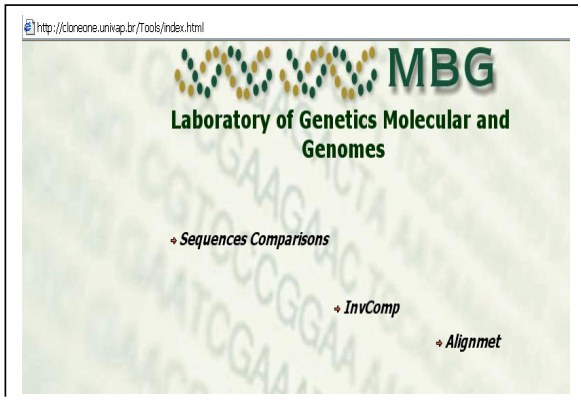

Fig.01 Página principal do projeto, indicando os links de acesso.

Sequences Comparisons, este *script* compara duas seqüências, utilizando vetorização computacional, e para os cálculos das bases foi implementado a teoria das pequenas amostras, para calcular a porcentagem das bases em relação ao tamanho total da seqüência. Ao identificar regiões de igualdade, o *script* armazena as posições do vetor, juntamente com as bases, para que estas sejam calculadas e apresentadas ao usuário posteriormente.

Fig. 02 Interface do script Sequences Comparisons**.**

A figura 01, mostra a interação com o usuário é bem simples, basta entrar com as seqüências em que se deseja efetuar a comparação e os cálculos e *clikar* no botão *submeter.*

Aligment, este *script* foi desenvolvido para que o usuário possa saber exatamente em que posição de um determinado segmento encontra-se a sua seqüência, através do alinhamento o *script* posiciona a *query* mostrando ao usuário a posição *upstream* e *dowstream* do alinhamento*.* Dentro do *script* é possível que o usuário escolha a base de dados que ele deseja comparar, qual opção usar, determinar quantos pares de base antes e depois deseja visualizar, e qual grau de similaridade desejada para um melhor resultado. Com esses parâmetros o *script* irá mostrar as regiões anteriores e posteriores do gene, referente à base de dados escolhida, isso caso haja o alinhamento. Desses resultados o usuário pode determinar com maior precisão onde deve desenhar o seu *primer*, ou simplesmente localizar outros elementos de interesses. Os bancos podem ser atualizados, de acordo com a necessidade do(s) projetos e usuários.

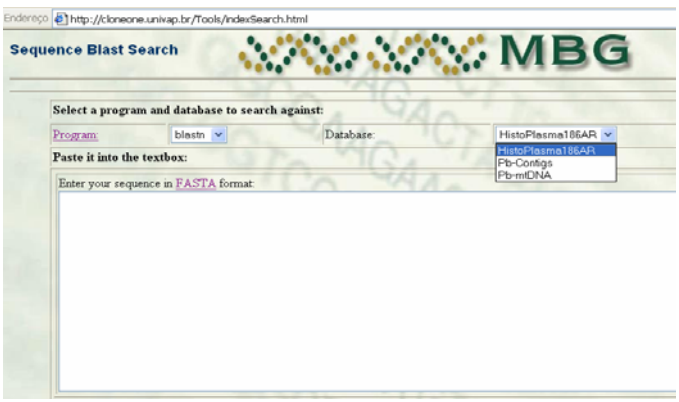

Fig. 03 Interface do script de alinhamento e localização de genes em Databases específicos.

possa ter a flexibilidade de sua análise.

InvComp, a função deste *script* é agilizar e facilitar a simples tarefa de fazer a invenção e a complementaridade de uma seqüência de nucleotídeos, assim o usuário pode visualizar identificando qual a orientação de sua seqüência, facilitando também no momento de desenhar seus iniciadores ou analisar as ORFs.

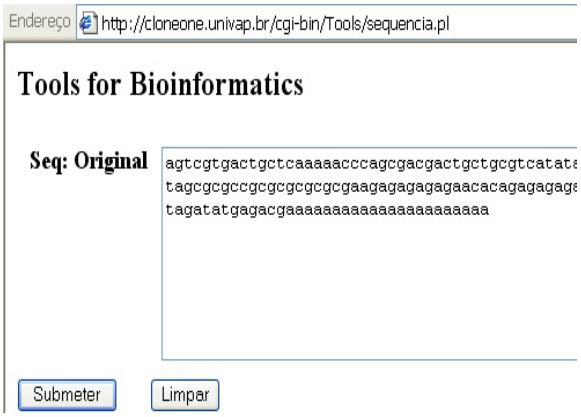

Fig. 04 Interface do script.

Resultados

Os *scripts* desenvolvidos estão em uso, no Laboratório de Biologia Molecular e Genomas da UNIVAP, sendo que o objetivo de cada script foi alcançado.

O resultado do *script* ***Sequences Comparisons*** é mostrado na fig. 05, os cálculos são apresentados, indicando a quantidade de cada base, com sua porcentagem respectiva. Logo abaixo são mostradas as regiões de igualdade, entre as seqüências, se houver.

Cabeçalho =	> gi|23567|	> gi |09876|
Total: (pb) =	149	160
A =	43 - 28.8590604026846	30 - 18.75
C =	35 - 23.489932885906	37 - 23.125
G =	38 - 25.503355704698	54 - 33.75
T =	26 - 17.4496644295302	32 - 20

GATCGCTAGCGCTCTCTCTCTCCACACACACAGGCGGCATATAGAGAGGA (24 - 132)
GATCGCTAGCGCTCTCTCTCTCCACACACACAGGCGGCATATAGAGAGGA

Fig. 05 Resultado da comparação entre as seqüências.

Aligment este *script*, exibirá o alinhamento de sua *query* e as regiões *upstream* e *dowstream*. Cada região é representada por uma cor, sendo: azul a região *upstream* (antes do gene), rosa a região de alinhamento e vermelho a região de *downstream* (depois do gene).

Fig. 06 O resultado de um simples alinhamento, e localização dentro do gene da Database de *Histoplasma capsulatum* mostrando as regiões upstrem e dow strem.

InvComp exibe a sequencia original, juntamente com as modificações estabelecidas, com a inverção, complementariedade e inversa complementar.

Fig. 07 Resultado do script, mostrando a seqüência original, e sua respectiva, inversa, complementar e reversa complementar.

Discussão

O segundo assunto mais importante em pesquisa a banco de dados, comparações entre seqüências é o programa (*script*) computacional usado para estes fins. A qualidade do algoritmo irá ditar a eficiência e a precisão dessas buscas. Isto tem sido desenvolvida ao longo dos anos (Altschul et al, 1994; Shuler et al, 1994 1).

Outro aspecto importante é a facilidade em que os usuários buscam na utilização destas ferrametnas computacionais, e foi penssando nestes aspéctos que os *scripts* desenvolvidos foram modelados e implementados de maneira simples.

Estes *scripts* são apenas o início, hava vista a necessidade crescente de tecnologia que a biologia computacional nescessita.

Conclusão

Este projeto objetivou apenas algumas necessidades, as mais básicas análises de comparações e alinhamentos entre seqüências nucleotídicas, por este motivo seus algoritmos foram desenvolvidos para análises em pequenas escalas, sendo comparações de pares de seqüênciamento.

A interface gráfica simples mostrou-se realmente eficaz, permitindo o uso instintivamente dos comandos.

Alguns ajustes ainda se fazem necessários, por exemplo, na formação das casas decimais do script ***InvComp***, referentes aos cálculos das bases, a não repetição de mais de um alinhamento no ***Sequences Comparisons*** e a inserção automática de bancos no ***Aligment.***

Agradecimentos

Agradecimento muito especial ao meu grande mestre e amigo, Dr. Francisco G. da Nóbrega, pelo apoio nos momentos difíceis, pela liberdade intelectual e por ter me ensinado as fantásticas e apaixonantes estradas do conhecimento e da ciência como um todo.

A todos do Laboratório de Biologia Molecular e Genomas, que me apoiaram e ajudaram neste trabalho.

Também a minha noiva, pela ajuda e paciência nos momentos difíceis e a minha família.

Referências

[1] Andrade, M. A. e Sander, C., Bioinformatics: from genome data to biological knowledge.Curr. Opin. Biotechnol. 8: 675-683, 1997.

[2] Persson, B., Bioinformatics in protein analysis. EXS. 88: 215-231, 2000.

[3] Meidanis, J., Setubal, J. Uma introdução à biologia computacional, UFPE-DI, Recife-PE, 1-2, 1994.

[4] Erickson, D., Haching the genome, Scientific American, 266(4): 128-137, 1992.

[5] Meidanis, J., Setubal, J. Uma introdução à biologia computacional, UFPE-DI, Recife-PE, 1-2, 1994.

[6] Perl, Internet site address: http://www.perl.org/ acessado em 12/03/2005.

[7] Linux, Internet site address: http://www.linux.org/ acessado em 25/12/2004.

[8] Altschul, S. F., *et al*, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, Nuc. Acid. Res., 17(25): 3389-3402, 1997.

[9]Shuler, G. D., et al, Sequence similarity searching using the BLAST family of programs. In *Current Protocols in Human Genetic* (ed. N. Dracopoli), vol. 2, suppl. 4, unit 11.3, John Wiley, New York, 1994.