UNIVERSIDADE FEDERAL DE ALFENAS
DEPARTAMENTO DE CIÊNCIAS EXATAS
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Júnio César Rosa

SISTEMA WEB PARA IDENTIFICAÇÃO E BUSCA DE ILHAS CPG NA BASE HCGP E BUSCAS INDIVIDUAIS

Alfenas, 29 de Agosto de 2013.

UNIVERSIDADE FEDERAL DE ALFENAS
DEPARTAMENTO DE CIÊNCIAS EXATAS
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

SISTEMA WEB PARA IDENTIFICAÇÃO E BUSCA DE ILHAS CPG NA BASE HCGP E BUSCAS INDIVIDUAIS

Júnio César Rosa

Monografia apresentada ao Curso de Bacharelado em Ciência da Computação da Universidade Federal de Alfenas como requisito parcial para obtenção do Título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Nelson José Freitas da Silveira

Alfenas, 29 de Agosto de 2013.

Júnio César Rosa

SISTEMA WEB PARA IDENTIFICAÇÃO E BUSCA DE ILHAS CPG NA BASE HCGP E BUSCAS INDIVIDUAIS

A Banca examinadora abaixo-assinada aprova a monografia apresentada como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação pela Universidade Federal de Alfenas.

Prof. Luiz Eduardo da Silva

Universidade Federal de Alfenas

Profa. Cibele Marli Cação Paiva Gouvêa

Universidade Federal de Alfenas

Prof. Dr. Nelson José Freitas da Silveira (Orientador)

Universidade Federal de Alfenas

Alfenas, 29 de Agosto de 2013.

À minha família

AGRADECIMENTO

Agradeço aos meus pais que sem o apoio e a ajuda constante eu jamais teria conseguido realizar esse trabalho. Agradeço a minha irmã Juliana e meu cunhado Paulo Henrique, que apesar de algumas desavenças sempre estiveram do meu lado e me ajudaram a ser uma pessoa melhor.

Agradeço a todos os meus amigos pelo apoio, as horas divertidas, as horas difíceis, pela paciência e por sua lealdade, vocês me ajudaram a superar desafios e problemas que a vida me impôs no decorrer desses anos de estudo. Em especial agradeço aos meus queridos amigos, Tatiane Fernandes Figueiredo, Mariana Dehon Costa e Lima, Wilson Sasaki Junior, Jéssica Dias, João Antonio, César, Patrick Itamar da Silva Barros, Thiago Ferreira, a minha prima Amanda e ao meu grande amigo e companheiro de todas as horas Felipe Marques de Carvalho e a sua família também, vocês acabaram se tornando a minha segunda família nesses anos, e certamente tudo o que eu conquistei foi graças a minha família e a grande amizade que eu cultivei com vocês.

Agradeço a todos os meus professores que contribuíram imensamente nessa jornada, tanto de forma intelectual, quanto de forma amigável atendendo sempre a todas a minhas dúvidas e questionamentos e possibilitando um ponto de partida quando eu estava sem rumo. Em especial agradeço ao meu professor e orientador Nelson José Freitas da Silveira pela paciência, dedicação e por todo conhecimento confiado a mim no desenvolvimento desse trabalho.

Agradeço aos funcionários da UNIFAL, que certamente fizeram a diferença com seu bom humor e educação, vocês também se tornaram meus amigos, o simples bom dia que eu escutava de vocês fez uma diferença muito grande. Em especial agradeço aos porteiros Luciane, Sandro, Roberto, Anderson, Túlio, aos vigilantes Junior, Henrique, Rodrigo, Elvio, Marcos, Leandro, a faxineira Luciana e as faxineiras que sempre me ofereciam café.

Enfim, agradeço a todos que de alguma forma contribuíram para a realização desse trabalho.

"Não viva para que a sua presença seja notada, mas sim para que a sua falta seja sentida."

Bob Marley

RESUMO

Este trabalho apresenta o desenvolvimento de um sistema web para classificação de Ilhas CpG em diversos tecidos tumorais, usando banco de dados à base de sequências do Genoma Humano do Câncer (HCGP). Utilizando técnicas de alinhamento de sequências entre a sequência fornecida pelo usuário e as sequências presentes no HCGP. O pacote EMBOSS, utilizado no sistema, busca alinhar as sequências do banco e retornar para o usuário as sequências que tenham semelhança significativa entre as do banco e do usuário para com as do banco. Com as sequências retornadas pelo sistema será possível classificar o grupo de tecidos tumorais que a sequência fornecida pelo usuário pertence, dizendo também se a sequência se enquadra nas especificações de Ilhas CpG.

Palavras-Chave: Ilhas CpG, Metilação de DNA, EMBOSS, Banco de dados HCGP, Marcadores Moleculares, Alinhamento de Sequências.

ABSTRACT

This work presents a development of a web-based system for classification of CpG islands in various tumor tissues, using database-based sequences of the Human Cancer Genome Project (HCGP). Using techniques of sequence alignment between the sequence provided by the user and the sequences present in HCGP. The EMBOSS package, used in the system, seeks to align the sequences of the bank and is returned to the user sequences that have significant similarity between the bank and the user towards the bank. With the sequences returned by the system will be possible to classify the group of tumor tissues that the sequence provided by the user belongs, also saying that the sequence fits the specifications of CpG islands.

Keywords: CpG islands, DNA Methylation, EMBOSS, Database HCGP, Molecular Markers, Sequence alignment.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES

DNA	Ácido Desoxirribonucléico
CpG	Citosina ponte Guanina
HCGP	Human Cancer Genome Project
ESTs	Expressed Sequence Tags
EMBOSS	European Molecular Biology Open Software Suite
PHP	Hypertext Preprocessor
pb	Pares de bases
cDNA	DNA complementar
RNA	Ácido Ribonucléico
SNP	Polimorfismo de Dinucleotídeos Únicos
SQL	Structured Query Language
mDNA	DNA Mitocondrial
TFI	Transcript Finishing Initiative
GUI	Graphical User Interface
HTML	Hypertext Markup Language
dbEST	Database Expressed Sequence Tags

SUMÁRIO

1
Introdução

Neste capítulo será apresentada uma visão geral sobre o tema que será tratado neste trabalho. Na Seção 1.1 são apresentadas a justificativa e a motivação do trabalho. Na Seção 1.2 é discutido o problema que envolve o tema proposto. Na Seção 1.3 são mostrados quais são os objetivos que este trabalho se propõe a realizar na monografia.

Projeto Genoma é o nome de um trabalho conjunto realizado por diversos países visando desvendar o código e um organismo (podendo ser animal, vegetal, de fungos, bactérias ou de um vírus) através do seu mapeamento (Pevsner, 2009).

O sequênciamento genômico tem por objetivo determinar a seqüência de bases nitrogenadas (adenina, citosina, guanina e timina) que compõem a molécula de DNA (genoma) de um organismo vivo que, em última análise, determina os genes que caracterizam este organismo.

A técnica de seqüenciamento foi desenvolvida por Sanger, na década de 70, e, desde então, tem sido constantemente aperfeiçoada visando sempre uma maior precisão, automação e a diminuição dos custos atuais que permitiram o seqüenciamento de DNA em larga escala. Apesar do grande desenvolvimento tecnológico nesta área, ainda nos dias de hoje só é possível o seqüenciamento de pequenos fragmentos de DNA, entre 500 a 800 bases.

O sequenciamento genômico, utilizando técnicas de sequenciamento de DNA é auxiliado por métodos computacionais que ajudam a armazenar a informação do genoma em banco de dados proporcionando uma rápida expansão do conhecimento sobre os processos biológicos (Carraro & Kitajima, 2002).

Dentre esses processos encontra-se o estudo da organização da cromatina que desempenha um papel fundamental na determinação de padrões da expressão gênica: regiões na eucromatina menos compactadas são as mais acessíveis para a transcrição, enquanto que as regiões de heterocromatina altamente compactadas são refratários para transcrição. A epigenética é uma área que estuda essas

transcrições (Bender, 2004). A epigenética é definida como o estudo das modificações do DNA e histonas que são herdáveis e não alteram a sequência de bases do DNA. Entre as modificações que as histonas podem sofrer, estão: metilação, fosforilação e acetilação (de Oliveira *et al*, 2010).

Acetilação descreve uma reação que introduz um grupo funcional acetila em um composto orgânico. Deacetilação é a remoção do grupo acetila.

Fosforilação é a adição de um grupo fosfato (PO4) a uma proteína ou outra molécula. A fosforilação é um dos principais participantes nos mecanismos de regulação das proteínas.

Metilação refere-se mais especificamente à substituição de um átomo de hidrogênio pelo grupo metila.

Entretanto, na molécula de DNA, ocorre apenas a metilação. Esta consiste na adição de um grupamento metil na citosina que geralmente precede a uma guanina (dinucleotídeo CpG), e está presente principalmente em regiões promotoras dos genes (de Oliveira *et al*, 2010).

A maior contribuição da ciência brasileira ao genoma humano foi trazida pelo Projeto Genoma Humana do Câncer (Human Genome Cancer Project – HCGP) desenvolvido por 29 laboratórios de sequenciamento e um centro de bioinformática (Kimura & Baía, 2002).

Um dos objetivos do HCGP é a contribuição para a anotação de genes no genoma humano, principalmente para identificar sequências de genes relacionados ao câncer (Brentani *et al*, 2003).

> Foram sequenciados mais de um milhão de fragmentos gênicos expressos (expressed sequences tags, ESTs), provenientes de diferentes tumores humanos. Atualmente, diversos projetos estão em desenvolvimento utilizando informações geradas no HCGP e abrangem observação da expressão diferenciada dos genes em diferentes tumores, caracterização completa de genes específicos, assim como o estudo funcional e estrutural dos produtos proteicos. É promissora a perspectiva de que num futuro próximo, diferentes resultados provenientes destas investigações possam trazer benefícios preventivos, prognósticos e clínicos em câncer e outras doenças (Kimura & Baía, 2002).

1.1 Justificativa e Motivação

A metilação (acréscimo de um grupo metil no DNA) anormal do DNA é fortemente implicada no desenvolvimento do câncer e afeta a expressão de mais de uma centena de genes supressores de tumor ou relacionados com a regulação da proliferação e da apoptose e com o reparo do DNA (Valentini, 2008).

Existem várias técnicas para revelar variabilidade em nível de DNA (Milach, 1998). Marcadores genéticos têm sido usados para estudar quantitativamente traços hereditários por quase 70 anos (Stuber *et al*, 1992). Características de DNA que diferenciam dois ou mais indivíduos e são herdadas geneticamente são conhecidas como marcadores moleculares (Milach, 1998). Recentemente, marcadores moleculares, particularmente marcadores de polimorfismo de DNA têm fornecido aos geneticistas uma fonte essencialmente ilimitada de formas para estudar as características quantitativas de traços hereditários, para manipulação e melhoramento de plantas e animais (Stuber *et al*, 1992).

Alterações em sequências, níveis de expressão e estrutura ou função de proteína tem sido associadas com todo tipo de câncer. Marcadores moleculares podem ser úteis para detectar o câncer, determinando diagnósticos, prognósticos e monitorando o progresso da doença ou resposta terapêutica (Sidransky, 2002).

A progressão do câncer é acompanhada pelo acúmulo de alterações genéticas. Estes levam a padrões de expressão alterada e modificações na estrutura e função da proteína. Mudanças que ocorrem exclusivamente ou, mais comumente em células cancerosas, em comparação com seu tecido normal de origem, podem ser detectadas por biópsia ou nos fluídos corporais, e são utilizados como marcadores moleculares de câncer. Estes marcadores são úteis para detectar o câncer em estágios iniciais, a avaliação da carga tumoral, a progressão da doença e a determinação da resposta à terapia (Sidransky, 2002).

A precisa quantificação do status de metilação de ilhas CpG (conseguir contar precisamente a quantidade de ilhas CpG metiladas) pode, sem dúvida resultar no desenvolvimento de um marcador molecular poderoso para diagnóstico e prognóstico de neoplasias e ainda levar à identificação de novos alvos terapêuticos (Valentini, 2008).

1.2 Problematização

Este trabalho propõem uma ferramenta web para responder as seguintes questões:

É possível desenvolver um sistema web que quantifique precisamente as ilhas CpG de diversos tecidos tumorais, utilização manipulações em bancos de dados ?

Com base em marcadores moleculares identificados, é possível a partir de um alinhamento local contra tais marcadores, identificar a entrada do usuário como um possível marcador ?

1.3 Objetivos

1.3.1 Gerais

O objetivo principal desse trabalho é a criação de um sistema web capaz de analisar uma sequência de bases de DNA fornecida pelo usuário, identificando se a mesma é uma ilha CpG, podendo também realizar um alinhamento dessa sequência com ilhas CpG da base de dados HCGP. O alinhamento de sequências realizado pelo programa EMBOSS retornará o grau de similaridade, identidade e os gaps retornados entre a sequência digitada pelo usuário e a sequência presente no banco de dados HCGP. Através de alinhamentos de sequências é possível selecionar a sequência no banco que apresenta o maior grau de similaridade determinado pelo resultado do programa EMBOSS e exibi-la na tela para o usuário onde ele poderá fazer download do arquivo texto com os dados de execução do programa EMBOSS. Com esses dados será possível classificar o grupo de cancêr que a sequência pertence.

O sistema web também possibilitará a exibição das sequências de DNA que são ilhas CpG da base de dados HCGP para o usuário, assim como a possibilidade do alinhamento local de duas sequências de DNA, ambas informadas pelo usuário.

1.3.2 Específicos

- Estudar a linguagem PHP para desenvolvimento do sistema web.
- Conseguir identificar Ilhas CpG em diversos tecidos tumorais, através da manipulação de caracteres e comparar os resultados para testes e melhorias.
- Realizar alinhamento de sequências digitadas pelo usuário e compará-las com o banco de dados HCGP para realizar a seleção de marcadores moleculares para a sequência informada pelo usuário.
- Disponibilizar a ferramenta para uso geral de especialistas e cientistas que trabalham nessa área.

1.4 Organização da Monografia

O Capítulo 2 é dedicado a mostrar todos os conceitos utilizados para o desenvolvimento desse trabalho. O Capítulo 3 fala sobre banco de dados, mostra os tipos de banco de dados utilizados na bioinformática, e fala sobre trabalhos relacionados ao banco de dados HCGP. O Capítulo 4 fala sobre as linguagens de programação e sobre o pacote EMBOSS utilizado no desenvolvimento desse trabalho. O Capítulo 5 faz um estudo de caso da ferramenta desenvolvida nesse trabalho. O Capítulo 6 são apontadas as conclusões desse trabalho. O Capítulo 7 apresenta as referências bibliográficas usadas no desenvolvimento desse trabalho. O Capítulo 8 é um anexo explicando o algoritmo de alinhamento local de Smith-Waterman usado no desenvolvimento desse trabalho. O Capítulo 9 é um anexo que apresenta um manual de uso para a ferramenta desenvolvida.

2
Fundamentação Teórica

Este capítulo tem por finalidade descrever os seguintes assuntos. Na seção 2.1 é feita uma pequena abordagem ao Projeto Genoma. A seção 2.2 fala sobre as ilhas Cpg's e os critérios para uma sequência ser uma ilha cpg. A seção 2.3 fala sobre alinhamento de sequências. A seção 2.3.1 explica o alinhamento local e seus esquema de pontuações. A seção 2.4 comenta sobre o algoritmo de Smith-Waterman. A seção 2.5 fala sobre EST's.

2.1 Projeto Genoma Humano

> Com o início do Projeto Genoma Humano em 1990 e subsequente disponibilização de sequenciadores automáticos de DNA capazes de gerar dados genômicos em grande escala, os bancos de dados e ferramentas de análise tiveram de se adaptar a este volume crescente de informações. Sequências de nucleotídios são adicionadas aos bancos de dados (como o GenBank) na ordem de milhares de pares de bases (pb) por segundo todos os dias (Santos & Ortega, 2003).

Na área de bioinformática, essas inúmeras sequências de DNA chegam a possuir, 400 a 1000 pb, podendo fundir com sequências cada vez maiores chamadas contigs, através de ferramentas que avaliam a qualidade dessas sequências assim como a sua superposição para que finalmente sejam disponibilizados segmentos cromossômicos inteiros de alta qualidade (Santos & Ortega, 2003).

Segundo Santos & Ortega, 2003 "Para a cobertura total de um genoma com boa qualidade estima-se que este deva ser sequenciado ao equivalente a dez vezes seu tamanho em pares de bases".

> Dados biológicos provenientes do conhecimento genômico são demasiadamente complexos em comparação aos dados oriundos de outras áreas científicas. A partir do conhecimento fundamental do genoma objetiva-se compreender o conjunto de peças que atuam no funcionamento complexo de todo o organismo. Porém, no momento, isso somente é possível por partes. Busca-se entender as estruturas moleculares das proteínas, as interações entre várias proteínas, bem como destas com as demais moléculas biológicas (DNA, carboidratos,

lipídios, etc), as diversas vias metabólicas celulares e o papel da variabilidade genética representada pelas várias formas de cada proteína. Toda essa informação disponibilizada pela ciência genômica (Figura 1) só é possível de ser organizada, analisada e interpretada com o apoio da informática (Santos & Ortega, 2002).

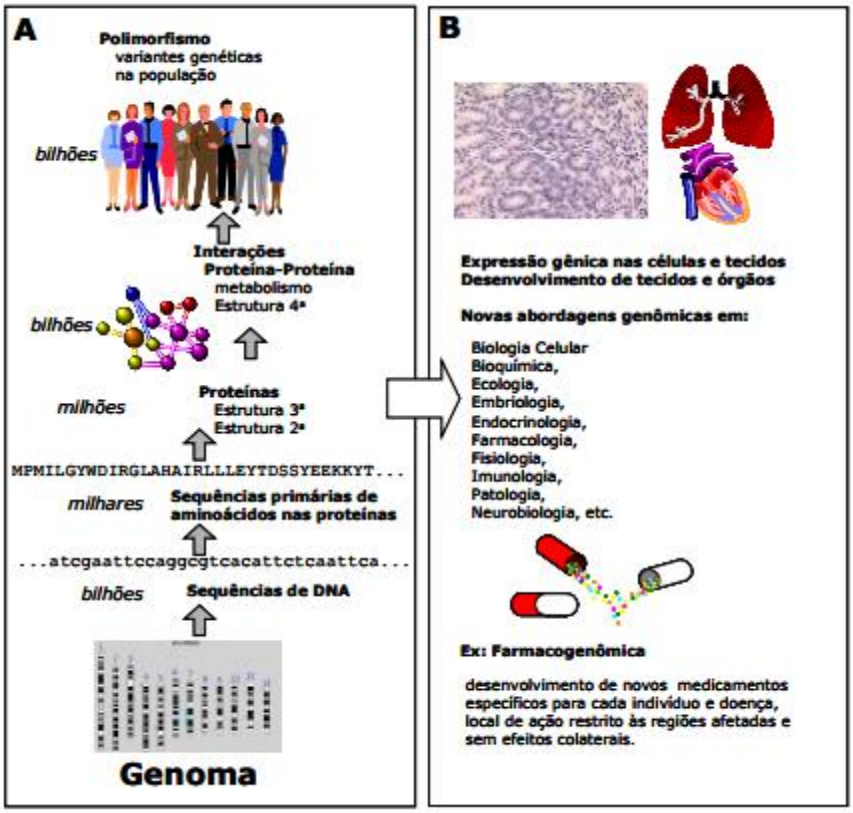


Figura 1 - Acúmulo de dados biológicos (A) e aplicações do conhecimento genômico (B).

Fonte: Santos & Ortega, 2003

Na Figura 1 podemos notar que os dados biológicos da população mostram proteínas na ordem de milhões de pares de bases, que são separadas e codificadas em sequências primárias de aminoácidos na ordem de milhares de pares de bases, o mesmo acontece com o DNA que possui a ordem de bilhões de pares de bases, e é codificado na ordem de milhares de pares de base.

Essa codificação ajuda na hora de fazer novos estudos genômicos e no desenvolvimento de novos medicamentos.

2.2 Ilhas CpG

A Metilação do DNA é um importante evento epigenético que controla várias funções do genoma, dentre essas funções podem ser citadas: regulação, estabilização e manutenção da expressão genética (Bird, 2002 apud Fang *et al*, 2006), recombinação durante a meiose, controle da replicação, regulação da diferenciação celular e inativação do cromossomo X. Entretanto, a aberração no padrão de metilação no promotor de um gene pode levar à perda de função desse gene e ser muito mais frequente do que a mutação genética. (de Oliveira *et al*, 2010) Ilhas CpG são regiões cromossômicas, frequentemente localizadas na posição 5' da cadeia genética, que têm uma alta densidade de DNA não metilado (Fang *et al*, 2006). A Figura 2 apresenta um exemplo de metilação do DNA.

O ácido desoxirribonucléico (ADN, em português: ácido dexirribonucleico; ou DNA, em inglês: deoxyribonucleic acid) é um composto orgânico cujas moléculas contêm as intruções genéticas que coordenam o desenvolvimento e funcionamento de todos os seres vivos e alguns vírus, e que transmitem as características hereditárias de cada ser vivo.

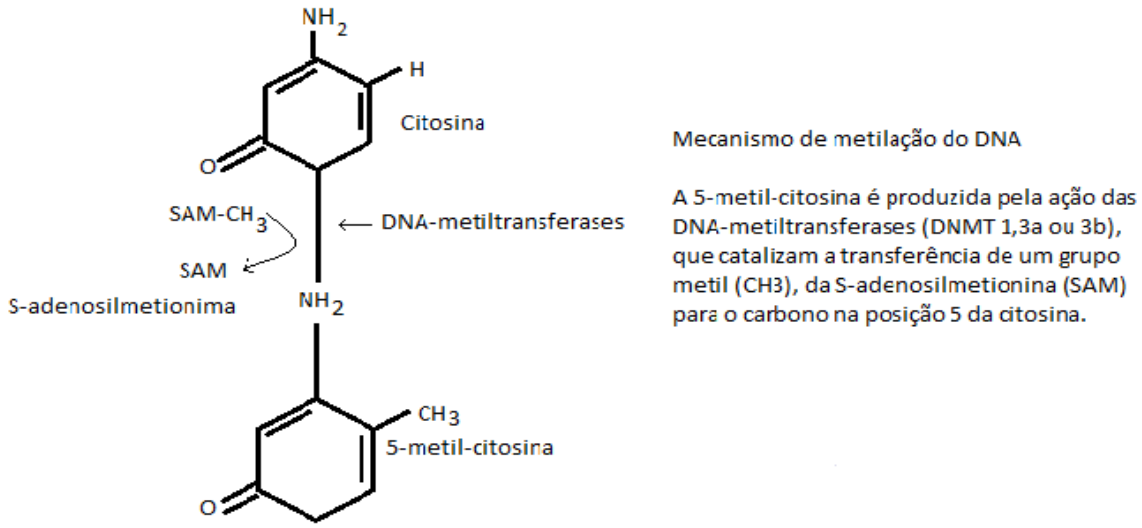


Figura 2 – Metilação do DNA

Fonte: http://dc150.4shared.com/doc/rMY3odZX/preview.html

Metilação de DNA é o mecanismo que permite que determinados genes (e não outros) se manifestem dentro de células normais especializadas, visto que todas as células do corpo trazem a mesmo genoma. O que as diferencia é a manifestação de um ou de outro gene durante o desenvolvimento e em toda a sua vida para desativar genes desnecessários. Na metilação de bactérias o grupo metil

(CH3) são adicionados a determinados locais do DNA para impedir que ele seja destruído por enzimas de restrição.

Em contraste, pequenas porções de DNA, chamadas ilhas CpG, são comparativamente ricas em nucleotídeos CpG e quase sempre estão livres de metilação. Estas ilhas CpG estão frequentemente localizadas na região promotora dos genes humanos e a metilação dentro das ilhas está associada à inativação da transcrição do gene correspondente. Em células cancerígenas, a metilação anormal de DNA desativa genes que normalmente evitariam divisões celulares impróprias. A Figura 3 mostra um esquema exemplificando a metilação nas ilhas CpG.

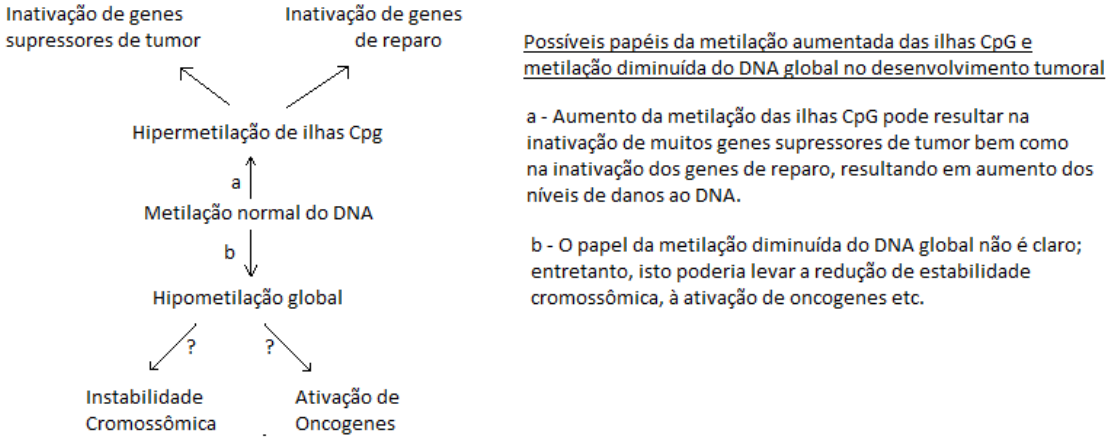


Figura 3 – Papéis da Metilação no Desenvolvimento Tumoral

Fonte: http://dc150.4shared.com/doc/rMY3odZX/preview.html

Os dinucleotídeos CpG aparecem esparsos pelos genomas eucariotos ou agrupados em regiões definidas como ilhas CpG. A literatura nos mostra que a maioria dos dinucleotídeos CpG esparsos está metilada, ao contrário das ilhas CpG que estão desmetiladas. Essas ilhas são frequentes em regiões promotoras de certos genes, incluindo genes housekeeping (os genes constitutivos que são necessários para a manutenção da função celular básica).

Definindo: "Ilhas CpG" são regiões do DNA maior que 300 pb (pares de bases), contendo aproximadamente 50% de bases C e G e com uma presença esperada de aproximadamente 60% de dinucleotídeos CpG. (Morgan et al, 2004 apud de Oliveira et al, 2010).

Os critérios originais de uma ilha CpG foram uma sequência de DNA maior que 300 pb com um conteúdo G + C ≥ 50%, e uma razão CpG observação/expressão (o/e) de ≥ 0,6 ver fórmula abaixo (Gardiner-Garden & Frommer, 1987). Se a razão for maior que 0,6 significa que a sequência de DNA analisada possui metilação aberrante das ilhas CpG podendo ocasionar câncer.

O dinucleotídeo CpG geralmente é muito deficiente em genomas de mamíferos, mas as ilhas CpG muitas vezes podem ser encontradas dispersas no genoma, particularmente perto ou dentro dos genes. Estas ilhas CpG são críticas na regulação da expressão gênica e diferenciação celular (Bird, 2002 apud Wang & Leung, 2004).

$$Obs/Exp = (CG^{*}N)/(C^{*}G)$$

Observando a fórmula podemos concluir que o valor do produto de número de genes CG's vezes o tamanho da sequência dividido pelo produto do número de C's vezes o número de G's não pode ser maior que 0,6.

2.3 Alinhamento de Sequências

Utilizando método de alinhamento de sequências é possível fazer uma análise comparativa entre resíduos de duas ou mais sequências para determinar se as sequências apresentam semelhança suficiente para uma verificação de homologia (Baxevanis & Ouellette, 2004) homologia entre sequências implicam que as entidades envolvidas tenham uma origem com um ancestral comum (Russo & Matioli, 2001). O termo Similaridade é uma quantidade observável que pode ser expressa como, por exemplo, a identidade por cento ou alguma outra medida adequada (Baxevanis & Ouellette, 2004).

Ao observar um grau surpreendentemente alto de similaridade de sequência entre dois genes ou proteínas, podemos inferir que eles compartilham uma história evolutiva comum, e a partir disto é possível prever que elas também devem ter funções biológicas semelhantes (Baxevanis & Ouellette, 2004).

No campo da Bioinformática o alinhamento de sequências possui uma diversidade de aplicações, sendo considerada uma das operações mais importantes desta área. Esse processo consiste em achar o grau de similaridade entre duas ou mais sequências (Prosdocimi *et al*, 2002).

Similaridade e homologia são conceitos diferentes. O alinhamento de sequências busca indicar o grau de similaridade entre sequências, já homologia parte de um princípio de cunho evolutivo, e não possui mensuração ou gradação. Duas sequências são homólogas se possuem um ancestral ou história evolutiva comum, se a hipótese evolutiva não se comprova, simplesmente não são homólogas (Prosdocimi *et al*, 2002).

Existem vários programas de computador que realizam o alinhamento de sequências e a grande maioria deles é utilizado on-line, sem a necessidade de instalação. Por exemplo: ClustalW, Multialin, FASTA, BLAST 2 sequences, etc (Prosdocimi *et al*, 2002).

> Alinhamento de sequências consiste em introduzir espaços ou lacunas (gaps) (ver Figura 4) entre os monômeros de uma ou mais sequências a fim de obter o melhor alinhamento possível. A qualidade de um alinhamento é determinada pela soma dos pontos obtidos por cada unidade pareada (match) menos as penalidades pela introdução de gaps e posições não pareadas (mismatch) (Prosdocimi *et al, 2002).*

Figura 4 – Alinhamento de duas sequências de proteínas

Fonte: Prosdocimi et al, 2002

Podemos observar na Figura 4 que as pontuações se baseiam por acertos (quando um gene é igual ao outro) chamado de match, erros (quando um gene é diferente de outro) chamado de mismatch, e gaps (quando existe espaços vazios) na Figura 4 é denotado por -.

2.3.1 Alinhamento Local

Muitas vezes uma sequência de proteína ou família de proteínas fornecem pistas sobre a função de um gene recém-sequenciado. Como tiras de DNA e banco de dados de aminoácidos continuam a crescer em tamanho, tornam-se cada vez comum a análise de novos genes sequenciados e proteínas, devido à maior chance de encontrar tais homologias (Altschul *et al,* 1990).

Existem várias ferramentas para pesquisar bancos de dados de sequência, mas todos buscam alguma medida de similaridade entre as sequências para distinguir relações biologicamente significativas. Esses métodos buscam atribuir pontuações para inserções, exclusões e substituições e calculam o alinhamento de duas sequências para o conjunto menos dispendioso de tal mutações. Tal alinhamento pode ser pensado como minimizando a distância evolutiva ou maximizando a semelhança entre duas sequências comparadas. Em qualquer dos casos, o custo deste alinhamento é uma medida de similaridade (Altschul *et al,*

1990).

Em métodos heurísticos a medida de similaridade não é explicitamente definida como um conjunto de custo mínimo de mutações, mas está implícito no algoritmo em si. Por exemplo o programa FASTP (Lipman & Pearson, 1985; Pearson & Lipman, 1988 apud Altschul *et al,* 1990) primeiro encontra regiões localmente similares entre duas sequências baseadas na suas identidades, mas sem gaps (lacunas) e em seguida, pontua essas regiões usando uma medida de similaridade entre seus resíduos. Apesar de sua aproximação indireta de medidas de evolução mínimas, ferramentas heurísticas como o FASTP tem se tornado muito populares e tem identificado muitos relacionamentos distantes entre sequências, mas significantes biologicamente (Altschul *et al*, 1990).

Medidas de similaridade podem ser geralmente classificadas como global ou local. Algoritmos de similaridade global buscam otimizar a melhor pontuação no alinhamento de duas sequências, que podem incluir grandes extensões de baixa similaridade (Needleman & Wunsch, 1970 apud Altschul *et al,* 1990). Algoritmos de similaridade local buscam apenas as maiores cadeias, onde a semelhanças são maiores, conservando as subsequências (Altschul *et al*, 1990). Por exemplo num alinhamento local entre uma sequência s e t, o alinhamento só ocorre na subcadeia de s e na subcadeia de t, a pontuação é calculada apenas onde as subcadeias de s e t estão alinhadas (ver Figura 5) (Ticona & Carlos, 2003).

Medidas de similaridade locais são normalmente utilizadas para pesquisas em banco de dados, onde cDNAs podem ser comparados com sequências parciais de genes e onde regiões de proteínas com relacionamentos distantes podem compartilhar a mesma região isolada de semelhança (Altschul *et al*, 1990).

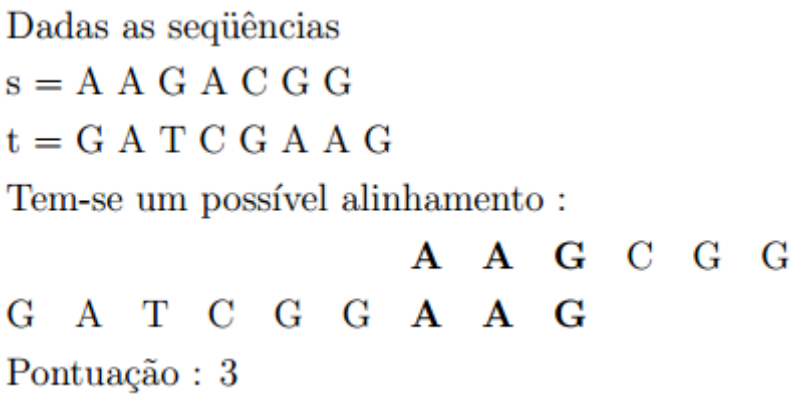


Figura 5 – Exemplo de alinhamento local entre duas sequências

Fonte: Ticona & Carlos, 2003

2.4 Algoritmo de Smith-Waterman

O algoritmo de Smith-Waterman é desenvolvido para achar o alinhamento local ótimo entre duas sequências. Foi proposto por Smith e Waterman em 1981 e implementado por Gotoh em 1982 (Manavski & Valle, 2008).

O algoritmo de Smith-Waterman é um algoritmo bem conhecido em bioinformática, que encontra o alinhamento ótimo entre duas sequências de DNA ou de proteína (Smith & Waterman, 1981 apud Li *et al*, 2007). Determinar o quão bem duas sequências podem se alinhar é importante na descoberta de genes homólogos e estudar a história evolutiva das moléculas e espécies (Page, 1998 apud Li *et al*, 2007). No entanto, o algoritmo de Smith-Waterman não é utilizado para procurar bancos de dados de sequência, pois é muito lento quando executado para muitas sequências (Li *et al*, 2007).

Esse algoritmo usa técnicas de programação dinâmica para encontrar as pontuações de similaridade local, porém o alinhamento pode iniciar e terminar em qualquer lugar das duas sequências, desde que possuam a melhor pontuação de similaridade (Pearson, 1991). Programação dinâmica tipicamente resolve problemas de otimização, que são definidos por avaliação de regras em cima de um espaço de busca recursivo. Os elementos desse espaço podem ser alinhamentos, estruturas secundárias de RNA, estrutura de genes, triangulação de polígonos, entre outros (Giegerich, 2000).

O alinhamento das duas sequências é baseada no cálculo de uma matriz de alinhamento. O número das suas colunas e linhas é dada pela quantidade de resíduos na consulta e as sequências do banco de dados, respectivamente. O cálculo baseia-se numa matriz de substituição e uma função gap-penalidade (Manavski & Valle, 2008).

2.5 Expressed Sequences Tags (ESTs)

Para entender a organização do complexo comportamento biológico devemos entender seus processos em termos de seus constituintes moleculares (Kirschner, 2005 apud Nagaraj *et al*, 2007) devemos não só identificar, catalogar e atribuir a função de todos os seus genes e produtos de genes, mas também compreender as interconexões entre DNA, RNA e proteínas. Acompanhando o significativo avanço nas tecnologias highthroughput (microarrays, sequências automatizadas e

espectrometria de massas), transcriptômica (tem por objetivo estudar o transcriptoma, ou seja, o conjunto dos transcritos de mRNA de uma célula), o estudo glabal da transcrição em conjundo com genômica e proteômica (tem por objetivo estudar todas as proteínas produzidas por uma espécie, incluindo a estrutura dos genes, a sua seqüência e outras características dos cromossomos), sem dúvida, contribuiram para uma abordagem de sistema biológicos. Essas tecnologias geraram um dilúvio de dados. Felizmente, ferramentas computacionais eficientes (redes de dados inteligentes, consultas (query) em banco de dados, recuperação de dados, ferramentas de análise e visualização) otimização a mineração dos dados, aceleraram o processo de descoberta (Nagaraj *et al*, 2007).

Expressed Sequence Tags (ESTs) e DNA complementar (cDNA) são sequências que fornecem evidência direta para todas as amostras de transcrição e são atualmente os recursos mais importantes para a exploração do transcriptoma. ESTs são curtas (200-800 bases de nucleotídeos), não editáveis, aleatoriamente selecionados com leituras derivadas de sequências singulares de bibliotecas de cDNA (Nagaraj *et al*, 2007).

ESTs de alto rendimento podem ser geradas com um custo razoavelmente baixo a partir de qualquer posição 5' ou 3' de final de um clone de cDNA. Em 1991 ESTs foram usadas como fonte primária para a descoberta de gene humano (Adams *et al*, 1991 apud Nagaraj *et al*, 2007). Depois disso, houve um crescimento exponencial na geração e acumulação de dados EST em bancos de dados públicos para organismos inumeráveis. No momento, ESTs permitem descoberta de genes, anotação complementar de genoma, ajuda na identificação de estrutura genética, estabelece a viabilidade de transcrições alternativas, guia o polimorfismo de nucleotídeo único (SNP) caracteriza e facilita a análise proteômica [3-5] (Nagaraj *et al*, 2007).

Normalmente são geradas para a identificar novos genes e definir o transcriptoma de diferentes tecidos. Atualmente, cerca de 12 milhões de ESTs foram depositadas no dbEst e aproximadamente 4 milhões são de humanos, contendo ESTs de tecidos normais e cancerígenos. Sua análise criteriosa pode revelar informações importantes, sobre mecanismos envolvidos na evolução do câncer (Pinheiro *et al*, 2002).

3
Banco de Dados Biológicos

Este capítulo apresenta uma descrição sobre o que são banco de dados, fala sobre os tipos de banco de dados usados na bioinformática e apresentas alguns trabalhos relacionados ao banco de dados HCGP que é usado pela ferramenta web;

3.1 Banco de Dados

Na década de 60 a computação consistia no desenvolvimento de aplicações individuais baseadas em arquivos. Com o tempo começaram a aparecer os problemas, pois aumentava a complexidade na manutenção dos programas, tanto no seu desenvolvimento quanto na sincronia dos dados a serem atualizados. Por este motivo, sugiram esforços e investimentos para pesquisar um meio mais barato para obter uma solução mecânica mais eficiente. Em 1970 um pesquisador da IBM, Ted Codd, propôs o banco relacional, onde ele idealizou que o usuário era capaz de acessar informações através de comandos. Estes comandos se tornaram uma linguagem padrão na indústria de banco de dados, linguagem chamada de SQL (Rohden, 2009).

Através do banco de dados o computador poderia atuar como um coordenador central das atividades de toda a empresa, funcionando como um recurso corporativo básico (Rohden, 2009).

Banco de dados são sistemas de coleções de informações que se relacionam de forma que crie um sentido. Normalmente incluem um arquivo de informações, uma organização lógica ou "estruturada" dessas informações e ferramentas para se ter acesso a elas. Bancos de dados da biologia molecular contêm sequências de ácidos nucleicos e de proteínas, estruturas e funções de macromoléculas, padrões de expressão, redes de vias metabólicas e cascatas de regulação (Lesk, 2008).

O mecanismo de acesso a um banco de dados é o conjunto de ferramentas utilizadas para organizar as informações selecionadas do banco de maneira útil, possibilitando através de índices que o usuário consiga encontrar um informação específica dentro do banco de dados (Lesk, 2008).

Através dos índices é possível surgir uma variedade de consultas nos bancos de dados utilizados na bioinformática. Estas incluem:

- A partir de uma sequência, ou fragmento de uma sequência, encontrar sequências similares no banco de dados.
- A partir de uma proteína, ou parte de uma estrutura protéica, encontrar estruturas de proteínas e sequências que sejam similares à estrutura protéica no banco de dados.
- A partir de uma sequência de uma proteína desconhecida, encontrar estruturas no banco de dados que adotem estruturas tridimensionais similares (Lesk, 2008).

3.2 Banco de Dados Relacional

O banco de dados relacional é baseado no princípio de que todos os dados estão dispostos em tabelas e toda sua definição é fundamentada na lógica de predicados e na teoria dos conjuntos (Figura 6). Esse tipo de banco define maneiras de armazenar, manipular e recuperar dados na forma de tabelas, construindo um banco de dados (Rohden, 2009). Os bancos relacionais são constituídos de tabelas, chaves primárias e chaves estrangeiras.

Tabelas é um conjunto não ordenado de linhas (tuplas). Onde cada linha é composta por uma série de campos. E cada campo é identificado por um nome de campo. Campo homônimo de todas as linhas de uma tabela forma uma coluna (Rohden, 2009).

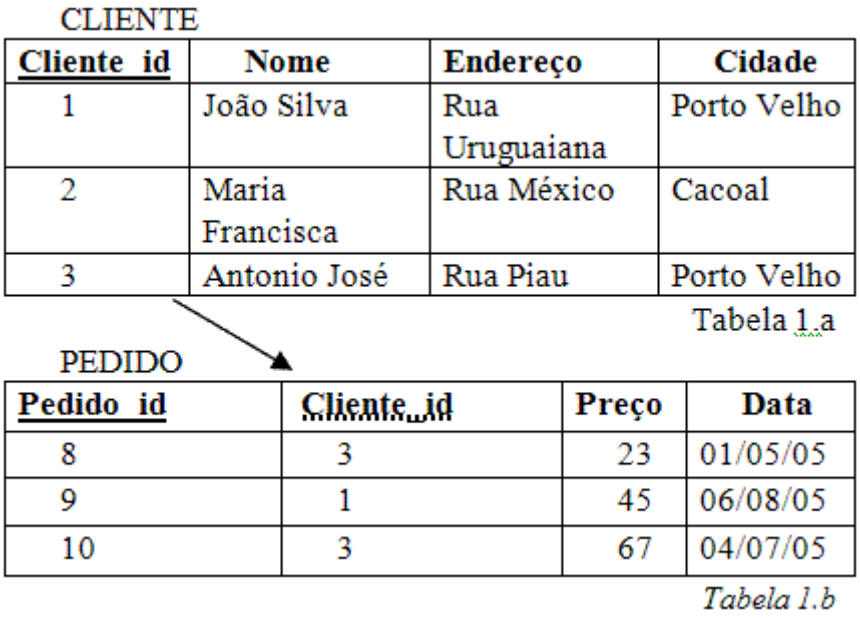
Uma chave primária é uma coluna ou uma combinação de colunas onde os valores dessa coluna distinguem uma linha das demais dentro de uma tabela (Rohden, 2009).

A chave estrangeira permite a implementação do relacionamento entre tabelas em um banco de dados relacional. Os valores de uma chave estrangeira permitem unir diferentes tabelas ajudando na consulta de dados do banco relacional, normalmente chaves estrangeiras são chaves primárias.

Esse tipo de banco foi desenvolvido com o intuito de prover o acesso às informações de forma ágil possibilitando uma maior variedade de abordagens no tratamento das informações (Rohden, 2009).

O criador do modelo relacional Edgar Frank Codd definiu em seu artigo algumas características desse modelo dentre elas temos:

- Toda informação deve ser representada de uma única forma, como dados em uma tabela;
- Todo dado (valor atômico) pode ser acessado logicamente (e unicamente) usando o nome da tabela, o valor da chave primária da linha e o nome da coluna;
- A capacidade de manipular a relação base ou relações derivadas como um operador único não se aplica apenas a recuperação de dados, mas também a inserção, alteração e eliminação de dados;
- Independência lógica de dados (Rohden, 2009).



CLIENTE

Cliente_id	Nome	Endereço	Cidade
1	João Silva	Rua Uruguaiana	Porto Velho
2	Maria Francisca	Rua México	Cacoal
3	Antonio José	Rua Piau	Porto Velho

Tabela 1.a

PEDIDO

Pedido_id	Cliente_id	Preço	Data
8	3	23	01/05/05
9	1	45	06/08/05
10	3	67	04/07/05

Tabela 1.b

Figura 6 – Exemplo de Tabela em um banco de dados relacional

Fonte: http://my.opera.com/maicokrause/blog/2009/04/27/fundamentos-dos-bancos-de-dados-relacionais

Na Figura 6 acima podemos observar a relação entre as tabelas Cliente e Pedido, nela podemos observar que o Pedido_id 8 da tabela Pedido está relacionado ao Cliente_id 3 da tabela Cliente assim é possível relacionar as duas tabelas para obter o endereço correto do cliente, possibilitando a expansão da tabela para detalhamento de informações.

3.3 Banco de Dados Públicos

Devido a magnitude do conjunto de dados produzidos torna-se fundamental a organização desses dados em bancos que permitam acesso on-line. O investimento contínuo na construção de bancos de dados públicos é um dos grandes motivos do sucesso dos projetos genoma e, em especial, do Projeto genoma Humano (Prosdocimi *et al*, 2002).

Bancos de dados que manipulam sequências de nucleotídeos, de aminoácidos ou estruturas de proteínas podem ser classificados em bancos de sequência primários e secundários (Prosdocimi *et al*, 2002).

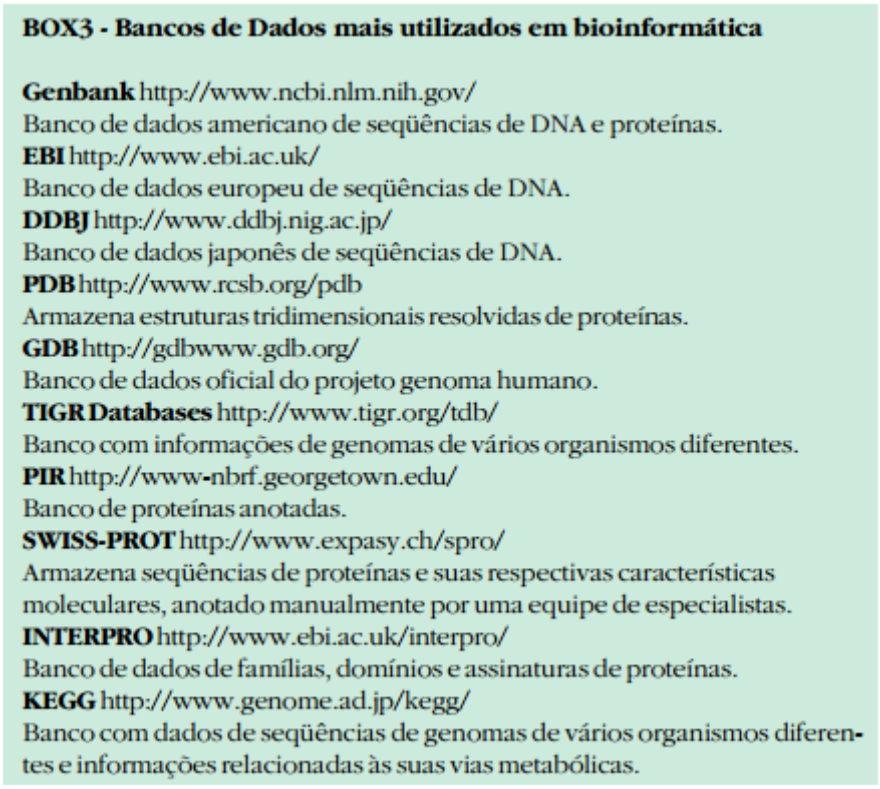
Banco de dados primários são formados por sequências de nucleotídeos, aminoácidos ou estruturas protéicas, sem qualquer processamento. Os principais bancos são o GenBank, o EBI (European Bioinformatics Institute), o DDBJ (DNA Data Bank of Japan) e o PDB (Protein Data Bank) (Prosdocimi *et al*, 2002).

Os bancos de dados secundários, são aqueles que derivam dos primários, ou seja, foram formados usando as informações depositados nos bancos primários. Por exemplo o PIR (Protein Information Resource) ou o SWISS-PROT. O SWISS-PROT é um banco de dados onde as informações sobre sequências de proteínas foram anotadas e associadas a informações sobre seus domínios funcionais, proteínas homologas e outros (Prosdocimi *et al*, 2002).

Embora a sequência de nucleotídeos, a sequência de aminoácidos e a estrutura de proteína sejam formas diferentes de representar o produto de um dado gene, esses aspectos apresentam informações sobre áreas diferentes e são tratados por projetos diferentes, que resultam em bancos específicos. Esses bancos de sequências são classificados como bancos estruturais ou funcionais. Os bancos estruturais mantém dados relativos das estrutura de proteínas (Prosdocimi *et al*, 2002).

Existem também os bancos funcionais, como o KEGG (Kyoto Encyclopedia of Genes and Genomes) que é um dos bancos mais utilizados. Ele disponibiliza links para mapas metabólicos de organismos com genoma completamente ou parcialmente sequenciados a partir de sequências e de busca através palavras-chave (Prosdocimi *et al*, 2002).

Com o grande número de dados biológicos que vem sendo gerados, vários bancos de dados têm surgido e anualmente a revista Nucleic Acids Research publica uma lista atualizada contendo todos os bancos de dados biológicos disponíveis e sua classificação (Prosdocimi *et al*, 2002).



BOX3 - Bancos de Dados mais utilizados em bioinformática

Genbank http://www.ncbi.nlm.nih.gov/
Banco de dados americano de seqüências de DNA e proteínas.
EBI http://www.ebi.ac.uk/
Banco de dados europeu de seqüências de DNA.
DDBJ http://www.ddbj.nig.ac.jp/
Banco de dados japonês de seqüências de DNA.
PDB http://www.rcsb.org/pdb
Armazena estruturas tridimensionais resolvidas de proteínas.
GDB http://gdbwww.gdb.org/
Banco de dados oficial do projeto genoma humano.
TIGR Databases http://www.tigr.org/tdb/
Banco com informações de genomas de vários organismos diferentes.
PIR http://www-nbrf.georgetown.edu/
Banco de proteínas anotadas.
SWISS-PROT http://www.expasy.ch/spro/
Armazena seqüências de proteínas e suas respectivas características moleculares, anotado manualmente por uma equipe de especialistas.
INTERPRO http://www.ebi.ac.uk/interpro/
Banco de dados de famílias, domínios e assinaturas de proteínas.
KEGG http://www.genome.ad.jp/kegg/
Banco com dados de seqüências de genomas de vários organismos diferentes e informações relacionadas às suas vias metabólicas.

Figura 7 – Banco de dados mais usados na bioinformática

Fonte: Prosdocimi et al, 2002

- O banco de dados de seqüência GenBank é de acesso aberto, possui uma coleção comentada de todas as seqüências de nucleotídeos publicamente disponíveis e traduções de proteína. Esta base de dados é produzido e mantido pelo Centro Nacional de Informações sobre Biotecnologia (NCBI), como parte da Colaboração de Bancos de Sequências de Nucleotídeos (INSDC). GenBank, e seus colaboradores recebem sequências produzidas em laboratórios de todo o mundo, possui mais de 100.000 organismos distintos.
- EMBL-EBI, que foi criado em 1980 nos laboratórios da EMBL, em Heidelberg, na Alemanha foi o primeiro banco de dados de seqüência de nucleotídeos do mundo. O objetivo inicial era criar um banco de dados centralizado de sequências de DNA.
- O DNA Data Bank of Japan (DDBJ) é um banco de dados biológicos que recolhe sequências de DNA. Ele está localizado no Instituto Nacional de Genética (NIG) na prefeitura de Shizuoka Japão. Ele também é membro da Colaboração de Bancos de Sequências de Nucleotídeos (INSDC). O banco de dados DDBJ iniciou suas atividades em 1986 no NIG e continua

sendo o único banco de dados de sequências de nucleotídeos na Ásia.

- O Protein Data Bank (PDB) é um repositório para os dados estruturais 3-D de grandes moléculas biológicas, tais como proteínas e ácidos nucleicos. Os dados, são normalmente obtidos por cristalografia de raios-X ou espectroscopia de RMN e são submetidos por biólogos e bioquímicos de todo o mundo, os dados são de livre acesso na Internet através dos sites das suas organizações e membros (PDBe, PDBj, e RCSB).
- O GDB Banco de Dados do Genoma Humano foi é uma comunidade voltada para coleção de dados genéticos humanos. O conjunto GDB além de outros bancos de dados biológicos teve a participação de líderes de classe mundial em genética humana para atuar como colaboradores para os dados. A fim de garantir um elevado grau de qualidade, os registos dentro de GDB foram submetidos a um processo de revisão por pares, não muito diferente de uma publicação tradicional. Devido à colaboração internacional que fez o projeto do genoma humano, GDB recebeu financiamento de várias fontes, tanto na Europa quanto na Ásia.
- O TIGR é um banco de dados de sequências de genomas de várias especíes diferentes.
- O Protein Information Resource (PIR), está localizado no Centro Médico da Universidade de Georgetown (GUMC), é um recurso bioinformático público integrado de apoio à investigação genômica e proteômica, e estudos científicos. PIR foi criada em 1984 pela Fundação Nacional de Pesquisa Biomédica (NBRF) como um recurso para auxiliar os pesquisadores na identificação e interpretação das informações de sequência de proteína.
- UniProtKB/Swiss-Prot é um banco de dados, manualmente anotados, com elevada qualidade, de sequências de proteína não redundante. Ele combina informações extraídas da literatura científica e biocurator-avaliada de análise computacional. O objetivo UniProtKB/Swiss-Prot é fornecer todas as informações conhecidas e relevantes sobre uma proteína particular.
- InterPro é uma base de dados de famílias de proteínas, domínios funcionais e locais em que as características identificáveis encontradas nas proteínas conhecidas podem ser aplicadas a novas sequências de proteínas de forma a caracterizá-las funcionalmente. O conteúdo do InterPro são baseados em torno assinaturas de diagnóstico e as proteínas que eles correspondem significativamente.

- KEGG (Kyoto Encyclopedia of Genes and Genomes) é uma coleção de bancos de dados on-line que lidam com genomas, vias enzimáticas e produtos químicos biológicos. A base de dados regista o caminho de redes de interações moleculares nas células, e as variantes delas específicas para os organismos específicos. Desde julho de 2011, KEGG mudou para um modelo de assinatura e acesso via FTP não é mais livre.

3.4 Tipos de Banco de Dados Biológicos

O Arquivamento e a distribuição de dados na bioinformática são realizados por organizações que utilizam e mantém os dados em bancos de dados específicos. Com a crescente demanda por equipamentos e pessoal, e a mudança de natureza das habilidades necessárias para incluir uma ênfase maior em computação, esse arquivamento passou a ser de responsabilidade de projetos específicos nacionais e até internacionais.

3.4.1 Bancos de Dados de Sequências de Ácidos Nucleicos

O arquivamento mundial de sequências de ácidos nucléicos é uma parceria tríplice entre o National Center for Biotechnology Information, o EMBL Data Library e o DNA Data Bank of Japan. Esses bancos de dados organizam, arquivam e distribuem sequências de DNA e RNA. Os bancos de dados de ácidos nucleicos, são coleções de registros ou entradas. Cada entrada tem a forma de um arquivo texto contendo dados e anotações para uma sequência contígua única. As entradas têm um ciclo de vida no banco de dados. O Banco também possui uma tabela de características que descreve propriedades de regiões específicas de uma anotação (Lesk, 2008).

3.4.2 Banco de Dados de Genoma

Embora as sequências de genomas constituam entradas padrões nos arquivos de sequência de ácidos nucleicos, muitas espécies têm banco de dados especiais que juntam sequências do genoma e suas anotações com outros dados relacionados às espécies (Lesk, 2008).

3.4.3 Bancos de Dados de Sequências de Proteínas

Em 2002, três bancos de dados de proteínas - o Protein Information Resource o SWISS-PROT e o TrEMBL, coordenaram seus esforços para formar o consórcio UnitProt. Hoje, quase todas as informações de sequências de aminoácidos são provenientes da tradução de sequências de ácidos nucleicos. Informações sobre ligantes, pontes dissulfeto, associações entre subunidades, modificações pós-traducionais, glicosilação, efeitos da edição do mRNA, etc. não podem ser obtidas das sequências de genes. Por exemplo, apenas a partir da informação genética não se saberia que a insulina humana é um dímero estabilizado por pontes dissulfeto. Banco de dados de sequências de proteínas coletam essas informações adicionais da literatura e as disponibilizam como anotações relevantes (Lesk, 2008).

3.5 Banco de dados HCGP e Trabalhos Relacionados

O Projeto Genoma Humano do Câncer (HCGP) envolve analisar sistematicamente alterações genômicas em grande número de tumores humanos para determinar ambas as alterações genéticas e epigenéticas comuns e identificar as mudanças que caracterizam subtipos diferentes de tumor.

Um dos principais objetivos do HCGP é a identificação de genes novos de câncer por meio sequenciamento de genoma, especificamente para encontrar mutações que ocorrem com 5% ou maior frequência em uma ampla variedade de tumores humanos. A implicação é que tais mutações levaria a novos alvos terapêuticos.

3.5.1 Polimorfismo de Dinucleotídeos Únicos (SNPs)

Os polimorfismos de nucleotídeos únicos (SNPs) constituem o tipo de variação mais comum do genoma humano. Conceitualmente, SNPs correspondem a posições onde ocorrem duas bases alternativas com frequência considerável na população humana. Embora muitas vezes não haja uma relação direta entre SNPs e o aparecimento de doenças, um número crescente destes tem sido identificado com envolvimento nas bases moleculares de doenças genéticas e do câncer. Além disso, SNPs podem ser usados como marcadores genéticos, úteis em estudos de

desequilíbrio de ligação (Ferreira *et al*, 2011).

Este estudo teve como objetivo principal a construção de um protocolo automatizado de identificação e caracterização de polimorfismos na região codificadora de genes expressos em tumores utilizando o ORESTES gerado pelo HCGP (Ferreira *et al*, 2011).

Um alinhamento pareado de todas as ORESTES contra os genes foi realizado usando o Cross_match, utilizando o resultado como a entrada para o alinhamento ancorado (Ferreira *et al*, 2011).

Com o programa Polybayes foram executados três procedimentos:

1) Alinhamento múltiplo ancorado das sequências ORESTES contra os genes (região codificadora);
2) Detecção e eliminação de parálogos;
3) Identificação de polimorfismos (Ferreira *et al*, 2011).

3.5.2 Transcript Finishing Initiative (TFI)

O projeto TFI, tal como o HCGP foi feito entre uma cooperação entre o instituto Ludwig e a FAPESP. Está sendo desenvolvido por uma rede de 29 laboratórios paulistas, e tem como objetivo validar a estrutura gênica completa de 4.000 genes humanos.

Os resultados de uma iniciativa de terminar a transcrição, realizado com o objetivo de identificar e caracterizar novos transcritos humanos, em que RT-PCR foi utilizada para preencher as lacunas (gaps) entre pares de grupos de ESTs, mapeadas contra a sequência genômica. Cada par de clusters EST foi selecionado para a validação experimental e foi designada uma unidade de acabamento transcrição (TFU). Um total de 489 TFUs foram selecionados para a validação e um rendimento global de 43,1% foi alcançado. Foram gerados um total de 59975 pb de sequências transcritas, organizados em 432 exons, que contribui para a definição da estrutura de 211 transcritos humanos. A estrutura de vários transcritos aqui relatado foi confirmado durante o decurso deste projeto, através da geração de suas correspondentes sequências de cDNA de comprimento total. No entanto, para 21% dos TFUs validadas, uma sequência de cDNA de comprimento total ainda não está disponível em bases de dados públicas, e a estrutura de 69,2% destas TFUs não foi corretamente previsto por programas de computador (Sogayar *et al*, 2004).

4
Metodologia

Este capítulo fala sobre todas as linguagens e pacotes de software utilizados para desenvolver o projeto;

4.1 PHP

Hypertext Preprocessor (PHP) é uma linguagem de script do lado do servidor, geralmente usada para criar páginas web dinâmicas. Originalmente desenvolvida em 1994 para ajudar a simplificar as tarefas básicas do site, o PHP desde então se tornou uma das linguagens de programação mais usadas em sites com quase 20 milhões de instalações em todo o mundo (ver Figura 8) (Pleva, 2013).

PHP pode ser usado como uma linguagem de programação de propósito geral para criar linha de comando e aplicativos de desktop GUI juntamente com as páginas web dinâmicas. (Pleva, 2013).

A primeira versão do PHP foi criada por Rasmus Lerdorf em 1995 e foi chamada: Personal Home Page/Form Interpreter: PHP/FI. De acordo com o artigo de Lerdorf intulado: Do You PHP? o motivo pelo qual ele criou a linguagem foi "puramente pelo caso de necessidade de uma ferramenta para resolver os problemas reais relacionados a internet". PHP é uma linguagem central do HTML. Isso significa que o código PHP está dentro do código HTML (Wright, 2013)

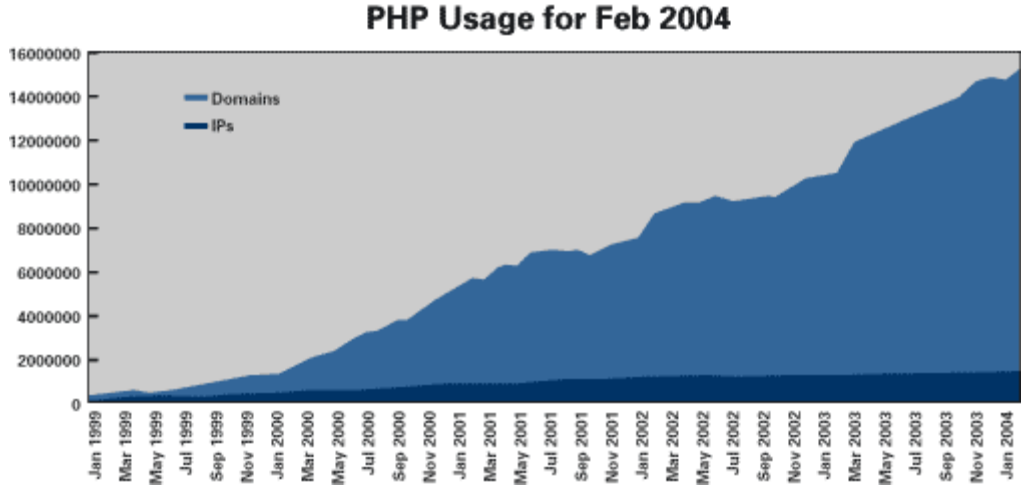


Figura 8 – Uso da linguagem PHP

Fonte: Wright, 2013

4.1.1 Segurança

PHP é conhecida por possuir poucas falhas de segurança dentro do núcleo da linguagem de programação. Muitas das falhas de segurança que foram encontrados ou associadas a PHP vêm de maus hábitos de programação. Esses maus hábitos acontecem quando um programador na criação de uma nova aplicação web esquece uma linha de código, ou pode programar as coisas de uma maneira tal que poderia permitir que um invasor para quebre o script e digite o código-fonte (Pleva, 2013).

4.1.2 O PHP no Projeto

Como a plataforma escolhida para rodar a aplicação desenvolvida no projeto foi a plataforma web, precisava-se de uma linguagem de programação que atendesse este requisito. O PHP foi o escolhido, vendo que além da facilidade de aprendizado e uso da linguagem, essa ferramenta é uma ferramenta muito potente, que atende a todos as necessidades que o projeto venha ter. Além disso, o PHP é uma ferramentas que se familiariza muito bem com o MySQL. Assim, como os dados do HCGP estão armazenados em um banco de dados MySQL, não sofreríamos com incompatibilidade de tecnologias.

4.2 EMBOSS

Na área de análise da sequência, os biólogos acham que os softwares existentes desenvolvidos ficam abaixo de suas necessidades. O grande aumento na área do sequenciamento em larga escala e os desafios da era pós-genômica necessitam de um rápido desenvolvimento de novas aplicações, ou o reforço do software existente (Rice *et al*, 2000).

A pesquisa realizada por Pitt em 1998 descobriu que a maioria dos desenvolvedores de bioinformática usam C e que, 18% estão usando bibliotecas de software livre, e outros 54% gostariam de usar essas bibliotecas no futuro (Rice *et al,* 2000).

Existe uma necessidade de integrar pacotes e bases de dados existentes mais de forma mais eficiente do que os métodos atuais são capazes de fazer. Estes problemas têm sido longamente debatidos pelos membros da EMBnet, com a

conclusão da reunião de 1996 em Helsinque que veio o esforço conjunto para fornecer um pacote de software de análise de sequências integradas (Rice *et al*, 2000).

O pacote EMBOSS (European Molecular Biology Open Software Suite) é o resultado desse esforço. Com mais de 100 aplicações (Tabela 1) e a capacidade para ser executado a partir do comando de linha, através de um navegador web, ou sob interfaces de usuário gráficas mais avançadas. (Rice *et al*, 2000).

A Versão 2.01 do EMBOSS oferece mais de 100 programas de análise e um conjunto de bibliotecas básicas. Além disso, vários pacotes de software disponíveis publicamente têm sido integrados ao ambiente EMBOSS, que são coletivamente denominados EMBASSY. O pacote funciona com a maioria das principais plataformas UNIX (IRIX, Solaris, Tru64 UNIX, Linux, FreeBSD e Macintosh OS X). O software pode ser obtido por FTP anônimo a partir da UK EMBnet. Documentação e suporte por e-mail estão disponíveis através da página web do EMBOSS (Olson, 2002).

Nesse trabalho foi utilizado o programa water do pacote EMBOSS para realizar o alinhamento de duas sequências e para realizar o alinhamento múltiplo de sequências.

Tabela 1 – Algumas das aplicações mais populares do EMBOSS (Rice *et al*., 2000)

Grupo do Programa	Aplicação Selecionada
Alinhamento	Local (matcher, water), e global (stretcher, needle) alinhamento.
Regiões Codificadoras	Uso códon sinônimo (syco), códon estatístico (chips)
Comparação	Comparações de grandes sequências de palavras (dottup, polydot, wordmatch) e alinhamento (supermatcher)
Ilhas Cpg	Relatório (cpgreport) e plotar (cpgplot)
Características de DNA	Repetições (einverted, etandem), Metilação de DNA (dan)
Edição	Utilidades gerais de edição (cutseq, splitter), características (maskfeat), etc
Indexação	Indexação de banco de dados (dbiflat, dbigcg, dbiblast)
Motifs	Procurando prosite (patmatdb, motifsearch), prints (pscan), transfac (tfscan), padrões gerais (fuzznuc, fuzzpro)

Alinhamento múltiplo	Interface com clustalw (emma), display (prettyplot), edição (mse)
Características da proteína	Motifs funcionais (antigenic, sigcleave), motifs estruturais (pepcoil, helixturnhelix), regiões anfipáticas (pepnet, pepwheel), predição transmembranar (TMAP) e display (topo)
Propriedades da proteína	Hydropathy (pepwindow, pepwindowall, octanol), síntese de proteína (digest), geral (pepstats)
Formatos de sequência	Leitura/escrita/conversão de formato (seqret, seqretall) e características da conversão de formato (seqretfeat)
Tradução	Uso de códons (cusp), leitura de quadros (getorf, showorf, backtranseq)
Utilidades	Motif indexação de banco de dados (rebaseextract, tfextract, prosextract, printsextract), listagem de banco de dados (showdb), busca de aplicações (wossname)

4.3 MySql

O MySQL é um sistema de gerenciamento de banco de dados (SGBD), que utiliza a linguagem SQL Structured Query Language (Linguagem de Consulta Estruturada), como interface. É atualmente um dos bancos de dados mais populares , com mais de 10 milhões de instalações pelo mundo.

Um banco de dados permite armazenar, pesquisar, classificar e recuperar dados de forma eficiente. O servidor MySQL controla o acesso aos dados para assegurar que vários usuários possam trabalhar com os dados ao mesmo tempo, fornecer acesso rápido aos dados e assegurar que somente usuários autorizados obtenham acesso. Portanto, o MySQL é um servidor multiusuário e multiencadeado (Welling & Thomson, 2005).

O MySQL está disponível sob um esquema de licença dupla. Você pode usá-lo sob a licença Open Source (GPL – General Public Licence) gratuitamente, contanto que cumpra os termos da licença. No entanto, se quiser distribuir uma aplicação não-GLP que inclua o MySQL, você pode comprar uma licença comercial.

4.3.1 Algumas características do MySql

- Portabilidade: O Mysql pode ser utilizado em muitos sistemas diferentes, como por exemplo sistemas UNIX e Microsoft Windows;
- Suporte total a multi-threads usando threads diretamente no kernel. Isto significa que se pode facilmente usar múltiplas CPUs, se disponível.
- Fornece mecanismos de armazenamento transacional e não transacional.
- É relativamente fácil se adicionar outro mecanismo de armazenamento. Isto é útil se você quiser adicionar uma interface SQL a um banco de dados caseiro.
- Um sistema de alocação de memória muito rápido e baseado em processo(thread).
- Joins muito rápidas usando uma multi-join de leitura única otimizada.
- Tabelas hash em memória que são usadas como tabelas temporárias.
- Funções SQL são implementadas por meio de uma biblioteca de classes altamente otimizada e com o máximo de performance. Geralmente não há nenhuma alocação de memória depois da inicialização da pesquisa.
- Aceita diversos tipos de campos: tipos inteiros de 1, 2, 3, 4 e 8 bytes com e sem sinal, FLOAT, DOUBLE, CHAR, VARCHAR, TEXT, BLOB, DATE, TIME, DATETIME, TIMESTAMP, YEAR, SET e ENUM.
- Um sistema de privilégios e senhas que é muito flexível, seguro e que permite verificação baseada em estações/máquinas. Senhas são seguras porque todo o tráfico de senhas é criptografado quando você se conecta ao servidor.
- Lida com bancos de dados enormes. Foi usado o Servidor MySQL com bancos de dados que contém 50.000.000 registros e sabemos de usuários que usam o Servidor MySQL com 60.000 tabelas e aproximadamente 5.000.000.000 de linhas.
- São permitidos até 32 índices por tabela. Cada índice pode ser composto de 1 a 16 colunas ou partes de colunas. O tamanho máximo do índice é de 500 bytes (isto pode ser alterado na compilação do MySQL). Um índice pode usar o prefixo de campo com um tipo CHAR ou VARCHAR.
- Os clientes podem se conectar ao servidor MySQL usando sockets TCP/IP, em qualquer plataforma. No sistema Windows na família NT

(NT, 2000 ou XP), os clientes podem se conectar usando named pipes. No sistema Unix, os clientes podem se conectar usando arquivos sockets.

4.3.2 O MySql no Projeto

O MySQL possui ligação estritamente estreita com o projeto desenvolvido. Já que a base de dados HCGP que é utilizada como base do projeto, foi criada e preenchida utilizando-se do SGBD (Sistema Gerenciador de Banco de Dados) MySQL. Assim, para que se possa acessar os dados disponibilizados, deve-se ter instalado o MySQL no host que serão utilizados os dados do HCGP.

As alterações de tipos de dados, inserção, remoção e consultas de dados, para testes na base de dados HCGP, podem ser feitas através do servidor MySQL acessando diretamente o terminal no sistema operacional ou utilizando um software que é disponibilizado em um pacote juntamente com o SGBD Mysql, que é o MySQL Workbanck ou MySQL Administrator e MySQL Query Browser.

O projeto desenvolvido realiza várias consultas no banco de dados, dentre essas consultas realizadas temos a consulta que separa possíveis ilhas CpG seguindo um critério pré-estabelecido, como por exemplo o tamanho da sequência tem que ser maior que 300 e o nome da sequência de uma tabela chamada main_sequence tem que ser igual ao nome da tabela sequence. Essa consulta é usada também para realizar o alinhamento de sequências.

5
Estudo de caso

Este capítulo apresenta uma especificação sobre o domínio do problema do projeto, exemplificando como foi o desenvolvimento no cenário do escopo do projeto

5.1 Domínio do Problema

O sistema desenvolvido com a ferramenta web de acesso ao banco de dados HCGP (Human Cancer Genome Project) exerce atividades de modo a ajudar o pesquisador desse banco de dados para um fácil acesso sem que seja necessário o domínio da linguagem SQL pelo usuário do sistema, através desse acesso podemos selecionar as ilhas CpG especificadas pelo usuário exibindo-as e realizar o alinhamento local entre a sequência informada pelo usuário sobre essas ilhas CpG.

Atualmente não é utilizada nenhuma interface que auxilie nas buscas por dados nessa base de dados. A implementação da ferramenta web faz diversas buscas em uma quantidade enorme de dados de modo que se obtenha um resultado rapidamente podendo acessar esses dados de qualquer lugar em que se possua acesso a internet, pois a plataforma adotada web se oferece essa mobilidade.

Tal software contém operações básicas de geração de arquivo texto para o usuário do sistema poder acessar esses dados através de links, podendo fazer download das informações da consulta realizada. De forma geral, uma pessoa entra no site desenvolvido, acessa a ferramenta web de busca, escolhe a consulta desejada, especificando itens das ilhas CpG e do alinhamento desejado, e obtém os resultados, podendo fazer download da consulta pelo site.

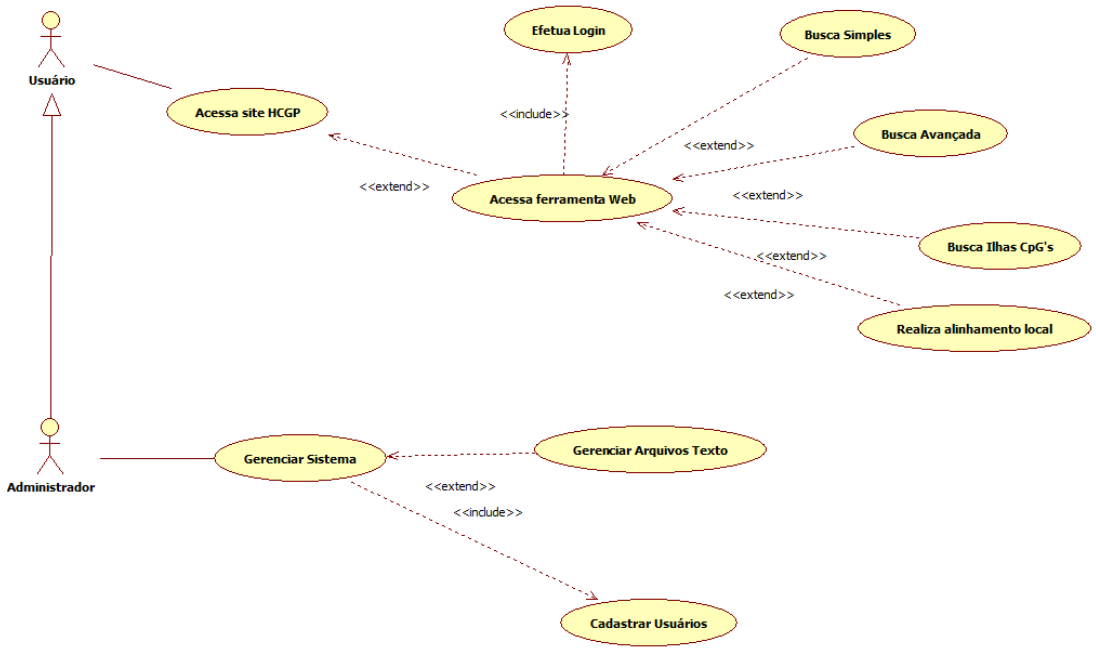


Figura 9 – Estudo do caso do sistema HCGP

A Figura 9 mostra o estudo do caso de todo o sistema web desenvolvido, para o projeto foram usados apenas os casos de uso Busca Ilhas CpG's e Realiza alinhamento local, esses casos consistem no usuário entrar no sistema, ou seja, acessar o site HCGP e assim poder usar as ferramentas de busca de ilhas CpG ou de alinhamento local.

Requisitos do sistema HCGP

Tabela 2 – Lista dos requisitos funcionais do projeto

RF1	Permitir aos usuários acesso para usar a ferramenta de busca de ilhas CpG
RF2	Permitir aos usuários realizarem busca por ilhas CpG
RF3	Permitir aos usuários realizarem alinhamento local entre sequências informadas e sequências de ilhas CpG

Tabela 3 – Lista dos requisitos não funcionais do projeto

RNF1	Funcionamento Web.
RNF2	Linguagem de Desenvolvimento PHP. Utilização do paradigma Orientado a Objetos para a modelagem e implementação do sistema.
RNF3	Utilização do SGBD MySql.
RNF4	A consulta realizada no sistema não deve demorar mais de 3 minutos.

5.2 O Cenário

O desenvolvimento do site que inclui a Ferramenta de Busca do referente trabalho, tem como objetivo introduzir uma facilidade de uso e divulgação do tema em questão. A página inicial contém a seguinte aparência visual mostrada na Figura 10.

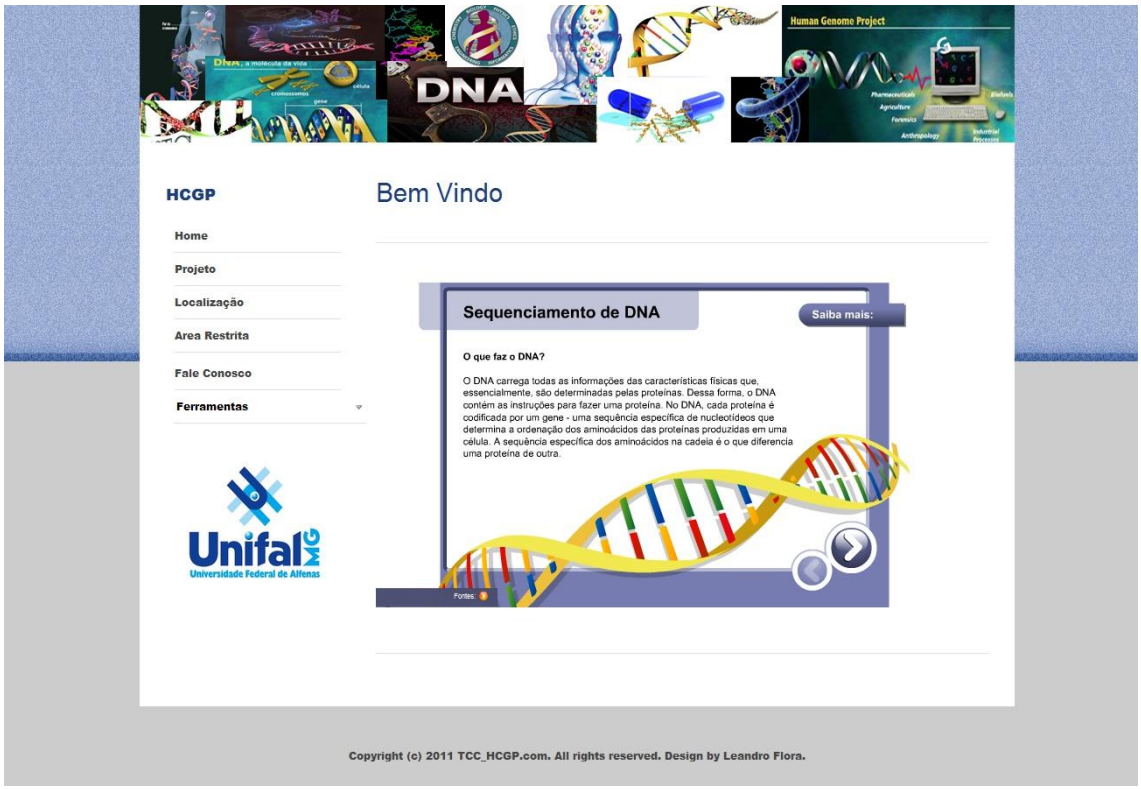


Figura 10 – Homepage da Ferramenta

A Figura 10 acima mostra a página inicial da ferramenta web contendo dentre outras funcionalidades, uma área restrita, uma opção de fale conosco e um menu de ferramentas, esse trabalho tem o foco no menu de ferramentas.

O site é uma forma de o usuário ter uma visualização amigável para poder acessar ao que realmente o interessa. O menu da página contém uma descrição do projeto, a localização de onde foi desenvolvido, uma área restrita que dá acesso aos usuários interessados em acessar a ferramenta de busca ao banco de dados HCGP, um tópico de contato em que o usuário pode solicitar um login no sistema para acessar a ferramenta web ou enviar sugestões e uma página de ferramentas que dá acesso a ferramenta de busca de ilhas CpG e alinhamento local.

5.3 Extração das Consultas

A extração dos dados relacionados no banco de dados foi feita fazendo consultas relacionadas com as características específicas contidas na base de dados do projeto. A integração das tecnologias utilizadas para o desenvolvimento desse trabalho possibilitou uma interface intuitiva para o usuário e uma manipulação rápida dos milhares de dados contidos no banco de dados.

A arquitetura mostrada a seguir descreve o funcionamento desde a requisição da consulta desejada do usuário na página web da ferramenta de busca até o resultado obtido após o acesso no banco de dados.

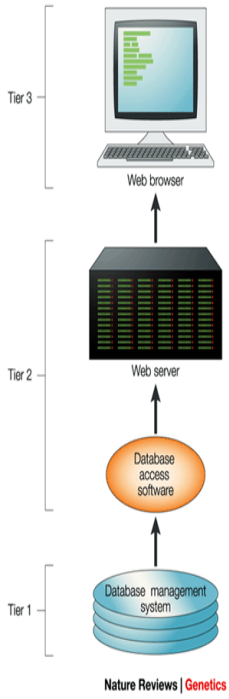


Figura 11 – Ambiente do estudo de caso

Como foi mostrado na Figura 11 acima, o usuário acessa a camada de aplicação, ou seja, acessa o site que contém a ferramenta web de busca no banco de dados HCGP. Após o usuário acessar a ferramenta web, na aba ferramentas, existe um menu expansível que muda quando o usuário clica nela exibindo mais duas opções conforme mostrado na Figura 12 a seguir.

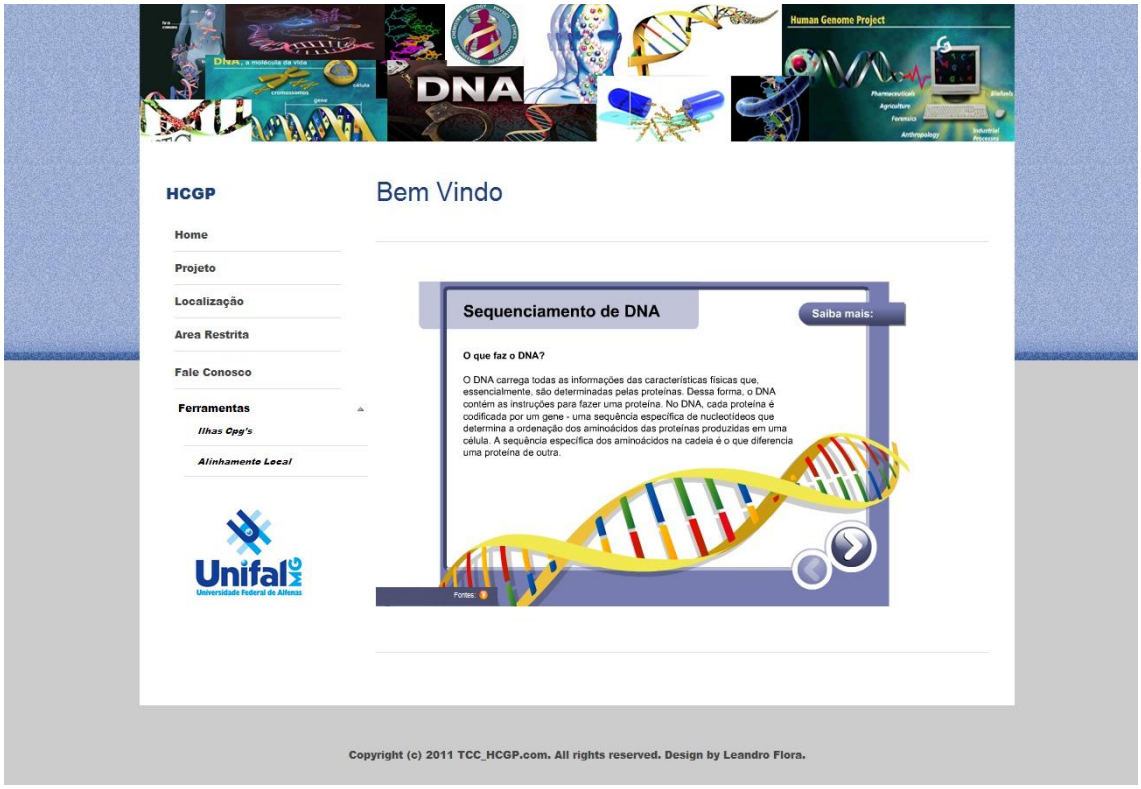


Figura 12 – Ferramenta com os menus de ilhas CpG habilitados

É nessas duas opções que se concentra o foco desse trabalho, com elas é possível fazer uma busca de ilhas CpG na base HCGP, opção Ilhas Cpg's e realizar um alinhamento local entre duas ou mais sequências, ou entre múltiplas sequências, opção Alinhamento Local.

Utilizando essas duas opções, o usuário escolhe o modo de busca para pesquisar efetivamente. A linguagem de programação utilizada faz uma chamada ao sistema gerenciador de banco de dados utilizado que faz a manipulação dos dados requisitados e retorna o resultado em sql obtido, então a linguagem de programação utilizada pelo programador se encarrega de mostrar os dados na tela para o usuário.

5.4 Base de Dados HCGP

A Base de dados Human Câncer Genome Project contém um banco com milhares de dados de sequenciamento genético. Durante a elaboração dessa base de dados o projeto seguiu a uma estrutura que é demostrada na Figura 13.

O Banco de dados do HCGP é composto de 10 tabelas (annotation, estscan, lib_statistics, main_seq, orf_maior, paths, quality, report, sequences,

trans_sequences) que se relacionam de forma a guardar as informações da melhor forma a obter as ESTs. O Modelo de Entidade Relacionamento a seguir demonstra as entidades do banco de dados juntamente com a maneira com que cada tabela se relaciona com a tabela principal e os atributos pertencentes a cada uma.

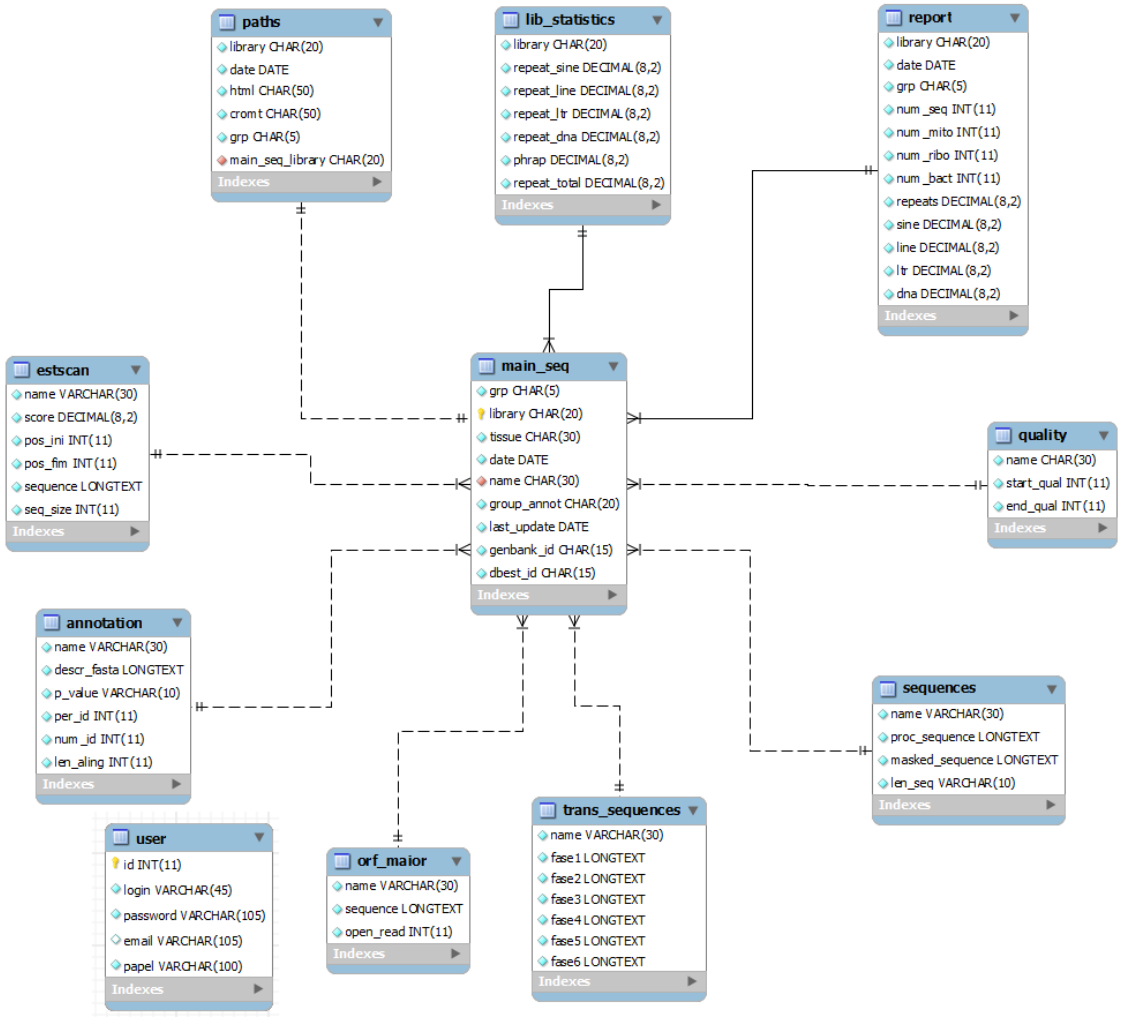


Figura 13 – Modelo Entidade Relacionamento do banco de dados HCGP

5.5 A Ferramenta Web

A ferramenta web contém dois tipos de funcionalidades. A primeira é a busca por ilhas CpG, essa busca procura na base de dados HCGP as ilhas CpG conforme um critério informado pelo usuário, como tecido e group annotation (ver Figura 15), após realizada a consulta é feito um cálculo para verificar se as sequências retornadas pelo banco HCGP realmente são ilhas CpG exibindo numa tabela as sequências que forem ilhas CpG. A tabela contém os seguintes campos: nome,

sequência de DNA, genbank id, group annotation, tecido, tamanho da sequência de DNA e a razão CpG.

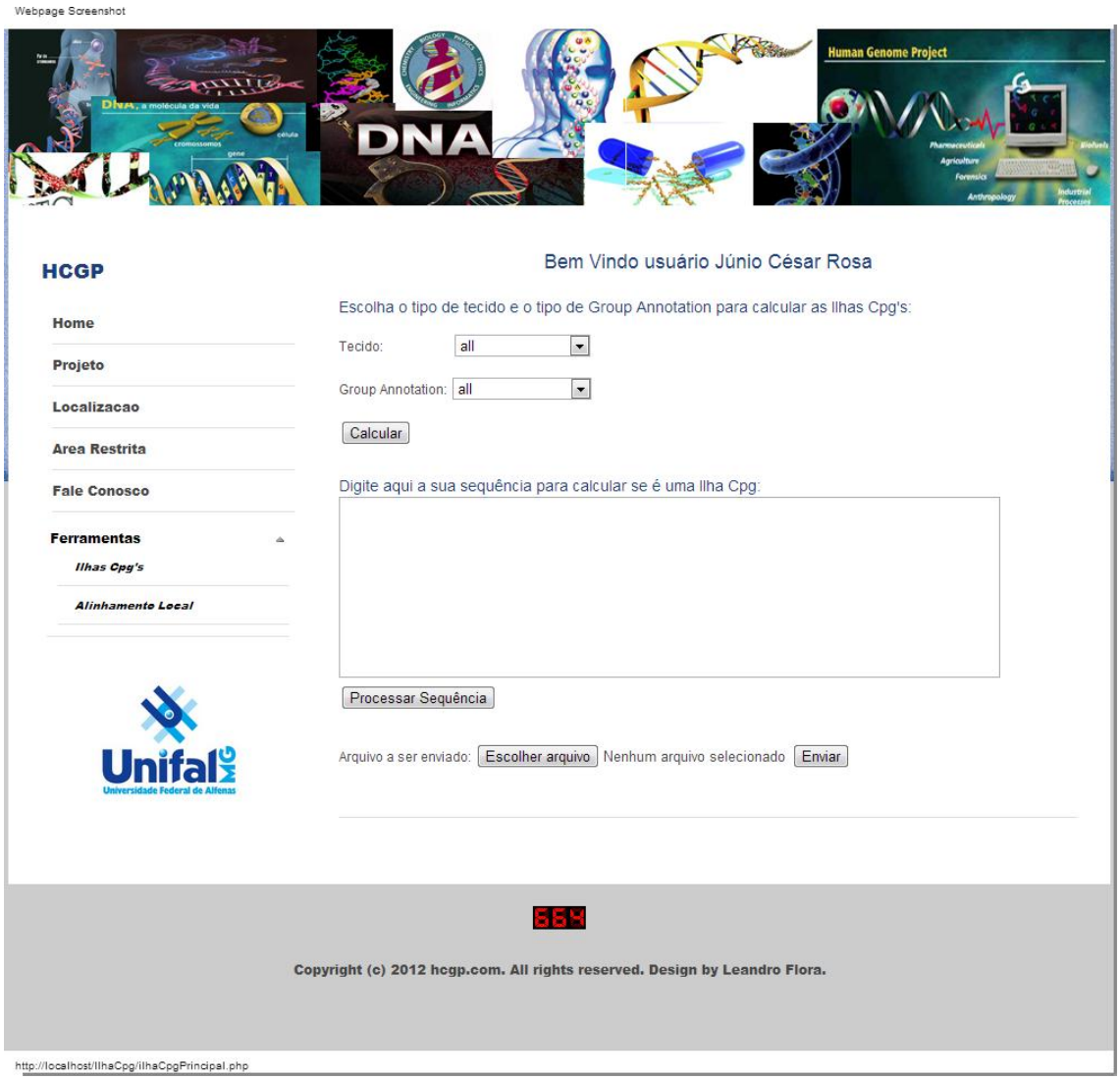


Figura 14 – Ferramenta web de busca por ilhas CpG

Conforme mostrado na Figura 14 acima existe a possibilidade de o usuário também digitar uma sequência e a ferramenta informar se é uma ilha CpG ou não.

Outra possibilidade da ferramenta de ilhas CpG é o processamento de arquivo, o usuário faz upload de um arquivo texto contendo sequências de DNA, esse arquivo é processado e mostra como resultado as sequências contidas no arquivo, e se as sequências do arquivo são ilhas CpG ou não.

A segunda ferramenta é uma ferramenta de alinhamento local de sequências. Essa ferramenta possui duas opções: o alinhamento entre uma sequência informada pelo usuário e as ilhas CpG da base HCGP conforme o critério de tecido e group annotation, e o alinhamento entre duas sequências ambas

informadas pelo usuário (ver Figura 15).

Figura 15 – Ferramenta web de alinhamento de sequências

A ferramenta de alinhamento possibilita a edição de dois campos para regulagem do alinhamento que são os campos Gap-Opening e Gap-Extension, por padrão esses campos valem 10 e 0.5 respectivamente.

O alinhamento é realizado pelo programa water do pacote EMBOSS, e o resultado do alinhamento é a leitura do arquivo de saída gerado por esse programa.

5.6 Trabalhos Futuros

A ferramenta Web HCGP fornece resultados que podem ser usados para diversas outras pesquisas. Pois após ser feita a busca, se o usuário do sistema desejar gerar o arquivo texto com as sequências obtidas, esses dados podem ser usados como entrada em várias outras ferramentas de montagem de sequências podendo gerar ferramentas para selecionar marcadores moleculares.

Marcadores moleculares são amplamente utilizados para estudos de genética populacional, mapeamento e análises de similaridade. O teste exato de FISHER para seleção de marcadores em câncer faz-se usando os p-valores atribuídos aos genes.

6
Conclusões

Este capítulo apresenta as conclusões desta monografia relacionados aos resultados obtidos dos dados do HCGP que foi usado pela ferramenta de busca web;

Os resultados obtidos através do banco de dados HCGP, e da interface desenvolvida para o sistema com acesso através do site feito para a ferramenta (http://www.bcc.unifal-mg.edu.br/hcgp). Pode-se perceber que tanto os resultados quanto ao trabalho foi desenvolvido cumprindo os requisitos e objetivos traçados facilitando a iteração dos dados procurados por pesquisadores que tem interesse nesse banco de dados, juntamente com estudantes da área.

Durante o desenvolvimento do projeto observou-se que os resultados da ferramenta podem variar conforme os critérios pesquisados, por exemplo para uma pesquisa de ilhas CpG que retorna cerca de 8 mil ilhas o tempo para execução para exibir essas informações na página web é muito grande (extende os 20 min). Agora para uma pesquisa de ilhas CpG que busca por todos os tecidos e todos os group annotation, retorna cerca de 45 mil ilhas, os resultados foram inviáveis para serem exibidos para usuário na página web, logo as opções de pesquisa que buscam por todos os tecidos ou todos os group annotation ou as duas opções ao mesmo tempo foram eliminadas do projeto.

Para um alinhamento local que busca por todos os tecidos e todos os group annotation ocorreu um imprevisto, o programa water não suportou um arquivo texto de 45 mil sequências, logo esse tipo de alinhamento foi cancelado

Para um alinhamento que busca por todos os tecidos ou todos os group annotation o tempo de execução foi mundo grande (extendeu os 20 min), logo essa opção de pesquisa foi cancelada do projeto.

Durante a programação do site conclui-se que php não é a melhor linguagem para processamento de sequências, visto que um laço de repetição (for) teve que ser duplicado para pegar corretamente a quantidade de C's, G's e CG's para o cálculo de ilhas CpG.

7 Referências Bibliográficas

Adams, M. D.; Kelley, J. M.; Gocayne, J. D.; Dubnick, M.; Polymeropoulos, M. H.; Xiao, H.; Merril, C. R.; Wu, A.; Olde, B.; Moreno, R. F. & others (1991), 'Complementary DNA sequencing: expressed sequence tags and human genome project', *Science* **252**(5013), 1651--1656.

Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W. & Lipman, D. J. (1990), 'Basic local alignment search tool', *Journal of molecular biology* **215**(3), 403--410.

Baxevanis, A. & Ouellette, B. (2004), *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, Wiley.

Bender, J. (2004), 'DNA methylation and epigenetics', *Annu. Rev. Plant Biol.* **55**, 41--68.

Bird, A. (2002), 'DNA methylation patterns and epigenetic memory', *Genes & development* **16**(1), 6--21.

Brentani, H.; Caballero, O. L.; Camargo, A. A.; da Silva, A. M.; da Silva Jr, W. A.; Neto, E. D.; Grivet, M.; Gruber, A.; Guimaraes, P. E. M.; Hide, W. & others (2003), 'The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags', *Proceedings of the National Academy of Sciences* **100**(23), 13418--13423.

Carraro, D. M. & Kitajima, J. P. (2002), 'Seqüenciamento e bioinformática de genomas bacterianos', *Biotecnologia, Cıkncia & Desenvolvimento* **5**(28), 16--20.

Fang, F.; Fan, S.; Zhang, X. & Zhang, M. Q. (2006), 'Predicting methylation status of CpG islands in the human brain', *Bioinformatics* **22**(18), 2204--2209.

Ferreira, R. A.; Simpson, A. J.; Guimarães, P. E.; Dias, E. & Neto, M. A., 'Identificação de Polimorfismos de Nucleotideos Únicos (SNPs) em Regiões Codificadoras de Genes Humanos', Disponível em: <http://telemedicina.unifesp.br/pub/SBIS/CBIS2002/dados/arquivos/234.pdf> Acesso em: 05 de nov. 2011.

Gardiner-Garden, M. & Frommer, M. (1987), 'CpG islands in vertebrate genomes', *Journal of molecular biology* **196**(2), 261--282.

Giegerich, R. (2000), 'A systematic approach to dynamic programming in bioinformatics', *Bioinformatics* **16**(8), 665--677.

Guizelini, D. (2012), 'Banco de dados biológico no modelo relacional para mineração de dados em genomas completos de procariotos disponibilizados pelo NCBI GenBank', .

Kimura, E. T. & Baia, G. S. (2002), 'ONSA Network and The Human Cancer Genome Project: Contribution to The Human Genome', *Arquivos Brasileiros de Endocrinologia & Metabologia* **46**(4), 325--329.

Kirschner, M. W. (2005), 'Department of Systems Biology Harvard Medical School Boston, Massachusetts 02115', *Cell* **121**, 503--504.

Lesk, A. (2008), *Introdução a bioinformática*, Artmed.

Li, I. T.; Shum, W. & Truong, K. (2007), '160-fold acceleration of the Smith-Waterman algorithm using a field programmable gate array (FPGA)', *BMC bioinformatics* **8**(1), 185.

Lipman, D. J. & Pearson, W. R. (1985), 'Rapid and sensitive protein similarity searches', *Science* **227**(4693), 1435--1441.

Manavski, S. A. & Valle, G. (2008), 'CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment', *BMC bioinformatics* **9**(Suppl 2), S10.

Milach, S. (1998), 'Marcadores de DNA', *BIOTECNOLOGIA Ciencia e Desenvolvimento* **5**, 14--17.

Morgan, H. D.; Dean, W.; Coker, H. A.; Reik, W. & Petersen-Mahrt, S. K. (2004), 'Activation-induced Cytidine Deaminase Deaminates 5-Methylcytosine in DNA and Is Expressed in Pluripotent Tissues IMPLICATIONS FOR EPIGENETIC REPROGRAMMING', *Journal of Biological Chemistry* **279**(50), 52353--52360.

Nagaraj, S. H.; Gasser, R. B. & Ranganathan, S. (2007), 'A hitchhiker's guide to expressed sequence tag (EST) analysis', *Briefings in bioinformatics* **8**(1), 6--21.

Needleman, S. B. & Wunsch, C. D. (1970), 'A general method applicable to the search for similarities in the amino acid sequence of two proteins', *Journal of molecular biology* **48**(3), 443--453.

de Oliveira, Naila Francis Paulo; Planello, Aline Cristiane; Andia, Denise Carleto; Pardo, Ana Paula de Souza (2010), 'Metilação de DNA e câncer.', *Revista brasileira de cancerologia, Rio de Janeiro: Instituto Nacional de Câncer* **56**, 493-499.

Olson, S. A. (2002), 'Emboss opens up sequence analysis', *Briefings in bioinformatics* **3**(1), 87—91.

Page, R. (1998), 'GeneTree: comparing gene and species phylogenies using reconciled trees.', *Bioinformatics* **14**(9), 819--820.

Pearson, W. R. (1991), 'Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms', *Genomics* **11**(3), 635--650.

Pearson, W. R. & Lipman, D. J. (1988), 'Improved tools for biological sequence comparison', *Proceedings of the National Academy of Sciences* **85**(8), 2444--2448.

Pevsner, J. (2009), *Bioinformatics and Functional Genomics*, Wiley.

Pinheiro, D. G.; Cunha, M. A.; Silva, I. T.; Zago, M. A. & Silva Jr, W. A. (2002), Sistema Genérico de Anotação de ESTs, *in* 'VIII Congresso Brasileiro de Informática em Saúde'.

Pleva, J. T., 'PHP: Hypertext Preprocessor', Disponível em: <http://www.uwplatt.edu/csse/Courses/prev/csse411-materials/s12/Justin%20Pleva%20-%20PHP(Hypertext%20Preprocessor).docx> Acesso em: 03 de jul. 2013.

Prosdocimi, F.; Coutinho, G.; Ninnecw, E.; Silva, A. F.; dos Reis, A. N.; Martins, A. C.; dos Santos, A. C. F.; Júnior, A. N. & Camargo Filho, F. (2002), 'Bioinformática: manual do usuário', *Biotecnologia Ciência & Desenvolvimento* **29**, 12--25.

Rohden, Rafael B. 'Banco de Dados: Relacional X Multidimensional'. Departamento de Tecnologia Universidade Regional do Noroeste do Estado do Rio Grande do Sul (UNIJUI), 2009.

Rice, P.; Longden, I.; Bleasby, A. & others (2000), 'EMBOSS: the European molecular biology open software suite', *Trends in genetics* **16**(6), 276--277.

Russo, C. d. & Matioli, S. (2001), 'Como escolher genes para problemas filogenéticos especificos', *Biologia molecular e evolução. Ribeirão Preto: Holos Editora*, 130--136.

Santos, F. R. & Ortega, J. M. (2003), 'Bioinformática aplicada a Genomica', *Melhoramento Genômico, Minas Gerais: UFV*.

Sidransky, D. (2002), 'Emerging molecular markers of cancer', *Nature Reviews Cancer* **2**(3), 210--219.

Smith, T. & Waterman, M. (1981), 'Identification of Common Molecular Subsequences, J', *Molecular Biology* **147**, 195--197.

Sogayar, M. C.; Camargo, A. A. & others (2004), 'A transcript finishing initiative for closing gaps in the human transcriptome', *Genome research* **14**(7), 1413--1423.

Stuber, C. W.; Lincoln, S. E.; Wolff, D.; Helentjaris, T. & Lander, E. (1992), 'Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers.', *Genetics* **132**(3), 823--839.

Ticona, W. G. C. & Carlos, U.-S. (2003), 'Aplicação de Algoritmos Genéticos Multi-Objetivo para Alinhamento de Seqüкncias Biológicas', PhD thesis, Dissertação apresentada ao Instituto de Cikncias Matemáticas e de Computação-ICMC-USP.

Valentini, S. R. (2008), 'Identification and functional characterization of molecular markers involved in head and neck tumor formation by the analysis of the pattern of differential methylation.', *Thematic Projects Opportunities for Health Research in Brazil*, 24-25.

Wang, Y. & Leung, F. C. (2004), 'An evaluation of new criteria for CpG islands in the human genome as gene markers', *Bioinformatics* **20**(7), 1170--1177.

Welling, L. & Thomson, L. (2005), *PHP e MySQL desenvolvimento Web*, Elsevier.

Wright, B., 'PHP: Hypertext Preprocessor', Disponível em: < http://www.cs.ship.edu/~cdgira/courses/CSC434/Fall2004/docs/student_papers/PHP.doc> Acesso em: 03 de jul. 2013.

8 Apêndice A

8.1 O Algoritmo de Smith-Waterman

Para exemplificar o algoritmo de Smith-Waterman suponhamos duas sequências moleculares dadas por A = $a_1 a_2 \cdots a_n$ e B = $b_1 b_2 \cdots b_m$. Uma similaridade (a,b) é dada entre os elementos da sequência a e b. Remoções de tamanho k são dadas por W_k. para achar pares de segmentos com alto grau de similaridade, nós montamos uma matriz H (Smith & Waterman, 1981). Primeiro façamos:

$$H_{k0} = H_{0l} = 0 \ \ para \ \ 0 \leq k \leq n \ \ e \ \ 0 \leq l \leq m$$

Valores preliminares de H têm a interpretação de que H_{ij}, representa a máxima semelhança de dois segmentos que terminam em a_i e b_j respectivamente. Estes valores são obtidos a partir do relacionamento:

$$H_{ij} = max(H_{i-1,j-1} + s(a_i, b_j),\ max_{k \geq 1}(H_{i-k,j} - W_k),\ max_{l \geq 1}(H_{1,j-l} - W_l), 0)$$

Sendo que $l \leq i \leq n \ \ e \ \ l \leq j \leq m$

A fórmula para H_{ij} que se segue considera a possibilidade para o final dos segmentos ocorrerem em qualquer a_i e b_j;

(1) Se a_i e b_j estão associados, a similaridade é:

$$H_{i-1,j-1} + s(a_i, b_j)$$

(2) Se a_i está no final da exclusão de tamanho k, a similaridade é:

$$H_{i-k,j} - W_k$$

(3) Se b_j está no final da exclusão de tamanho l, a similaridade é:

$$H_{i-k,j} - W_l$$

(4) Finalmente, um zero é incluído para evitar cálculos de similaridade negativos, indicando que não existe similaridade entre a_i e b_j (Smith & Waterman, 1981).

O par de segmentos com a máxima similaridade é encontrado pela localização do primeiro elemento máximo de H. Os outros elementos da matriz que conduzem a esse valor máximo são determinados com um procedimento de

rastreamento que termina quando um elemento de H é igual a zero. O procedimento identifica os segmentos bem como produz o alinhamento correspondente. O par de segmentos com o a segunda melhor similaridade é encontrado aplicando o procedimento de rastreamento para o segundo maior elemento de H não relacionado com o primeiro elemento (Smith & Waterman, 1981).

Um exemplo simples é mostrado na Figura 16. Neste exemplo os parametros $s(a_i,b_j)$ e W_k são necessários para escolher a priori das bases estatísticas. Um match, $a_i = a_j$, produzem um $s(a_i,b_j)$ com valores de unidade enquanto que um mismatch produz um minus one-third (menos um terço). Esses valores tem um pontuação por muito tempo, sequências randômicas prováveis de quatro letras por zero. O peso de exclusão deve ser escolhido para ser pelo menos igual à diferença entre um match e um mismatch (Smith & Waterman, 1981).

	Δ	C	A	G	C	C	U	C	G	C	U	U	A	G
Δ	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0
A	0·0	0·0	1·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	1·0	0·0
A	0·0	0·0	1·0	0·7	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	1·0	0·7
U	0·0	0·0	0·0	0·7	0·3	0·0	1·0	0·0	0·0	0·0	1·0	1·0	0·0	0·7
G	0·0	0·0	0·0	1·0	0·3	0·0	0·0	0·7	1·0	0·0	0·0	0·7	0·7	1·0
C	0·0	1·0	0·0	0·0	2·0	1·3	0·3	1·0	0·3	2·0	0·7	0·3	0·3	0·3
C	0·0	1·0	0·7	0·0	1·0	3·0	1·7	1·3	1·0	1·3	1·7	0·3	0·0	0·0
A	0·0	0·0	2·0	0·7	0·3	1·7	2·7	1·3	1·0	0·7	1·0	1·3	1·3	0·0
U	0·0	0·0	0·7	1·7	0·3	1·3	2·7	2·3	1·0	0·7	1·7	2·0	1·0	1·0
U	0·0	0·0	0·3	0·3	1·3	1·0	2·3	2·3	2·0	0·7	1·7	2·7	1·7	1·0
G	0·0	0·0	0·0	1·3	0·0	1·0	1·0	2·0	3·3	2·0	1·7	1·3	2·3	2·7
A	0·0	0·0	1·0	0·0	1·0	0·3	0·7	0·7	2·0	3·0	1·7	1·3	2·3	2·0
C	0·0	1·0	0·0	0·7	1·0	2·0	0·7	1·7	1·7	3·0	2·7	1·3	1·0	2·0
G	0·0	0·0	0·7	1·0	0·3	0·7	1·7	0·3	2·7	1·7	2·7	2·3	1·0	2·0
G	0·0	0·0	0·0	1·7	0·7	0·3	0·3	1·3	1·3	2·3	1·3	2·3	2·0	2·0

Figura 16 – Exemplo de execução do algoritmo de Smith-Waterman

Fonte: Smith & Waterman, 1981

9 Apêndice B

9.1 Manual de funcionamento da ferramenta

Para uma melhor explicação a respeito do sistema desenvolvido, segue a seguir uma sequência de passos a partir da entrada no site até a escolha de uma busca por ilhas CpG ou a realização de um alinhamento local.

Acesso ao site http://www.bcc.unifal-mg.edu.br/hcgp

Acessar o sistema e clicar no link Ferramentas aí vai exibir a seguinte tela.

Figura 17 – Ferramenta web tela inicial

Dai é só clicar na ferramenta web desejada, links Ilhas CpG e Alinhamento Local.

9.2 Manual de funcionamento da ferramenta – Ilhas CpG

Essa seção consiste em mostrar o funcionamento da aba Ilhas CpG mostrada na figura a seguir.

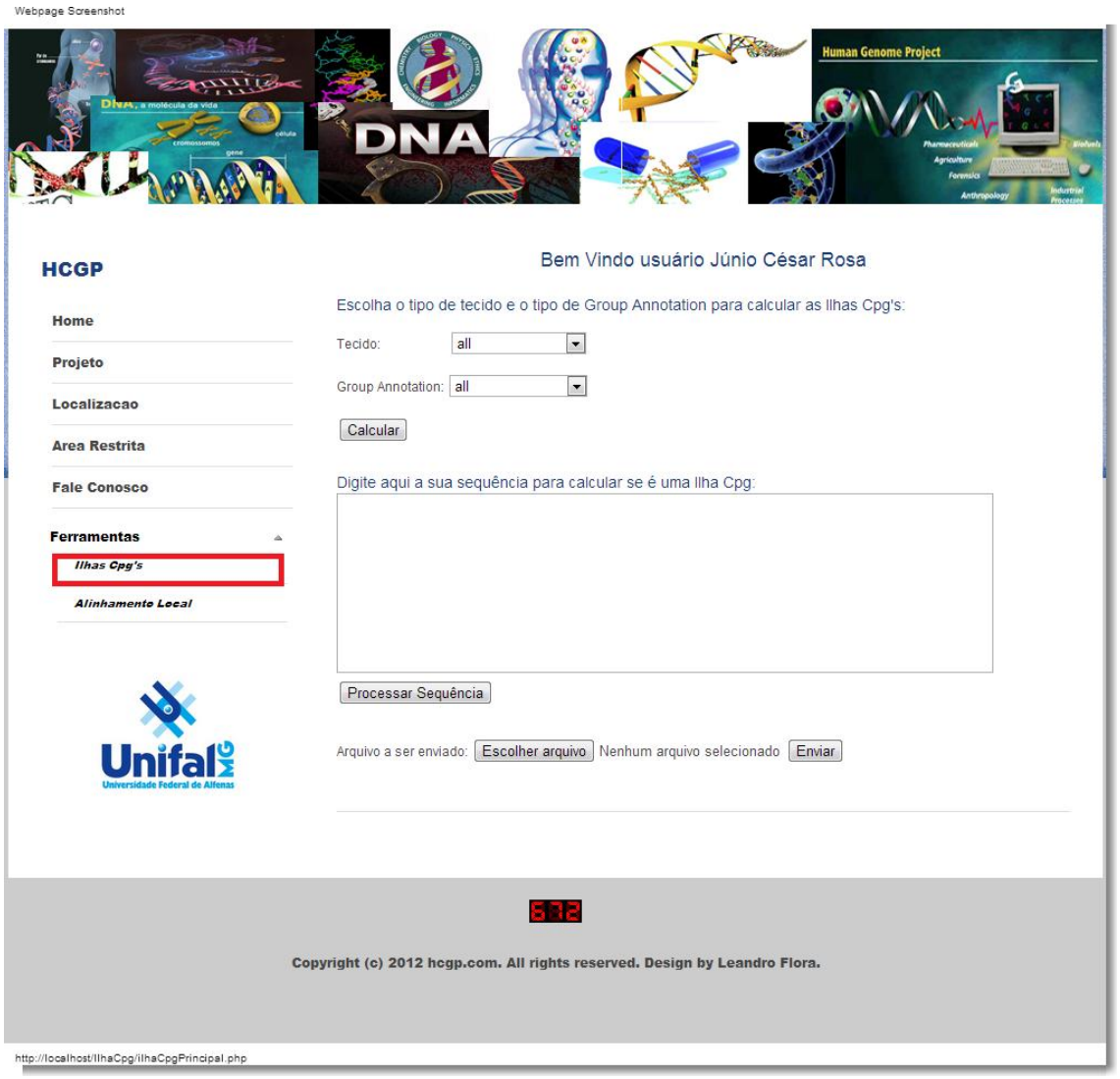


Figura 18 – Página web para busca de ilhas CpG

Essa tela consiste de três funcionalidades: A busca de ilhas CpG na base HCGP, o processamento de uma sequência digitada pelo usuário e o processamento de arquivo texto.

Na primeira funcionalidade que é a de busca por ilhas CpG o usuário escolhe um tecido e um group annotation e clica em Calcular aí vai aparecer a seguinte tela mostrando os resultados da busca para o usuário.

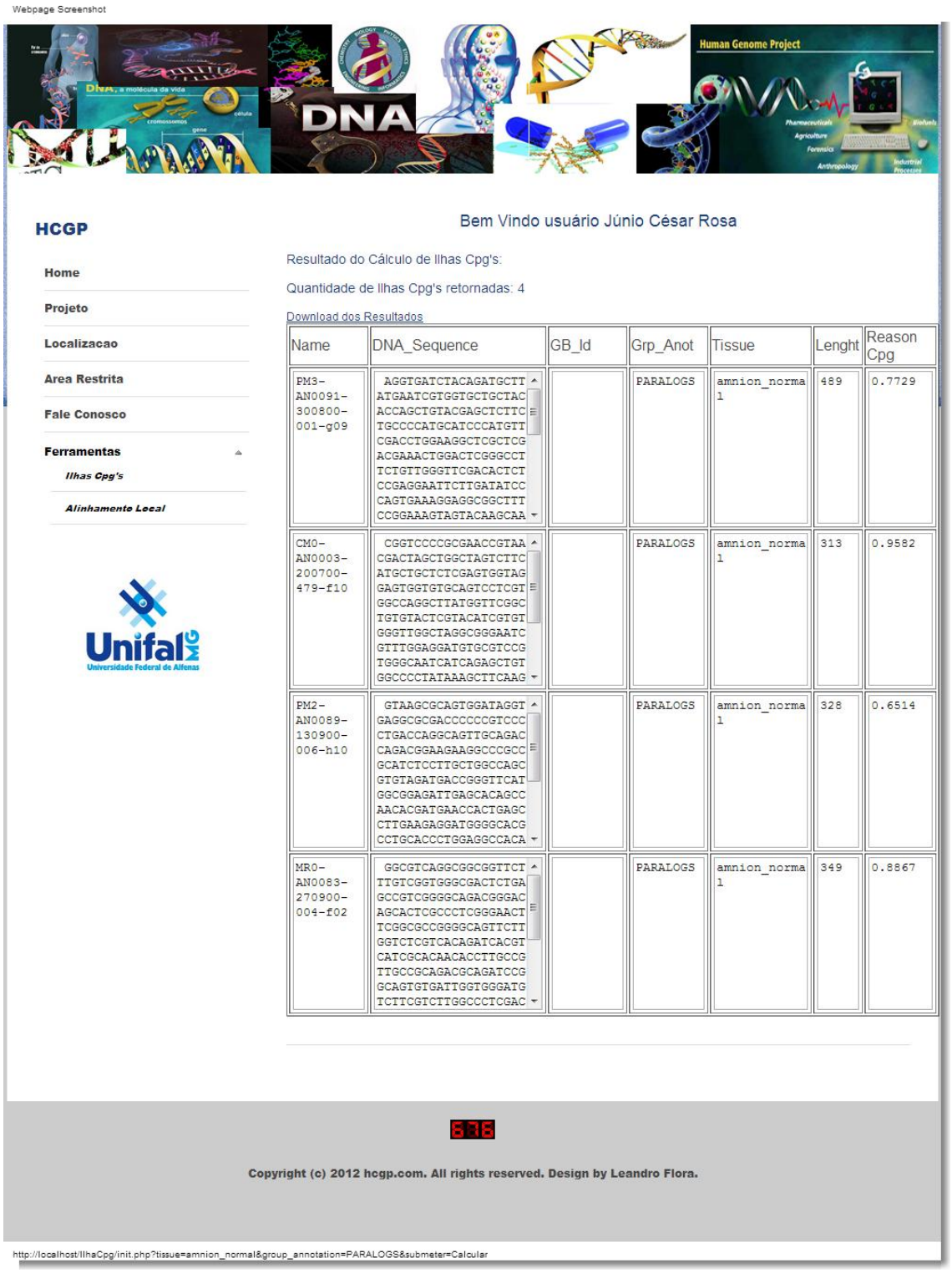


Figura 19 – Exemplo de busca por ilhas CpG na base HCGP

Na segunda funcionalidade o usuário digita uma sequência de DNA na área de texto, conforme mostra a Figura 20 abaixo, e clica em Processar Sequência e vai aparecer uma mensagem indicando se a sequência é uma ilha CpG.

Figura 20 – Segunda funcionalidade na ferramenta de busca de ilhas CpG

A terceira funcionalidade consiste em processar um arquivo texto informado pelo usuário, conforme mostra a Figura 21 abaixo.

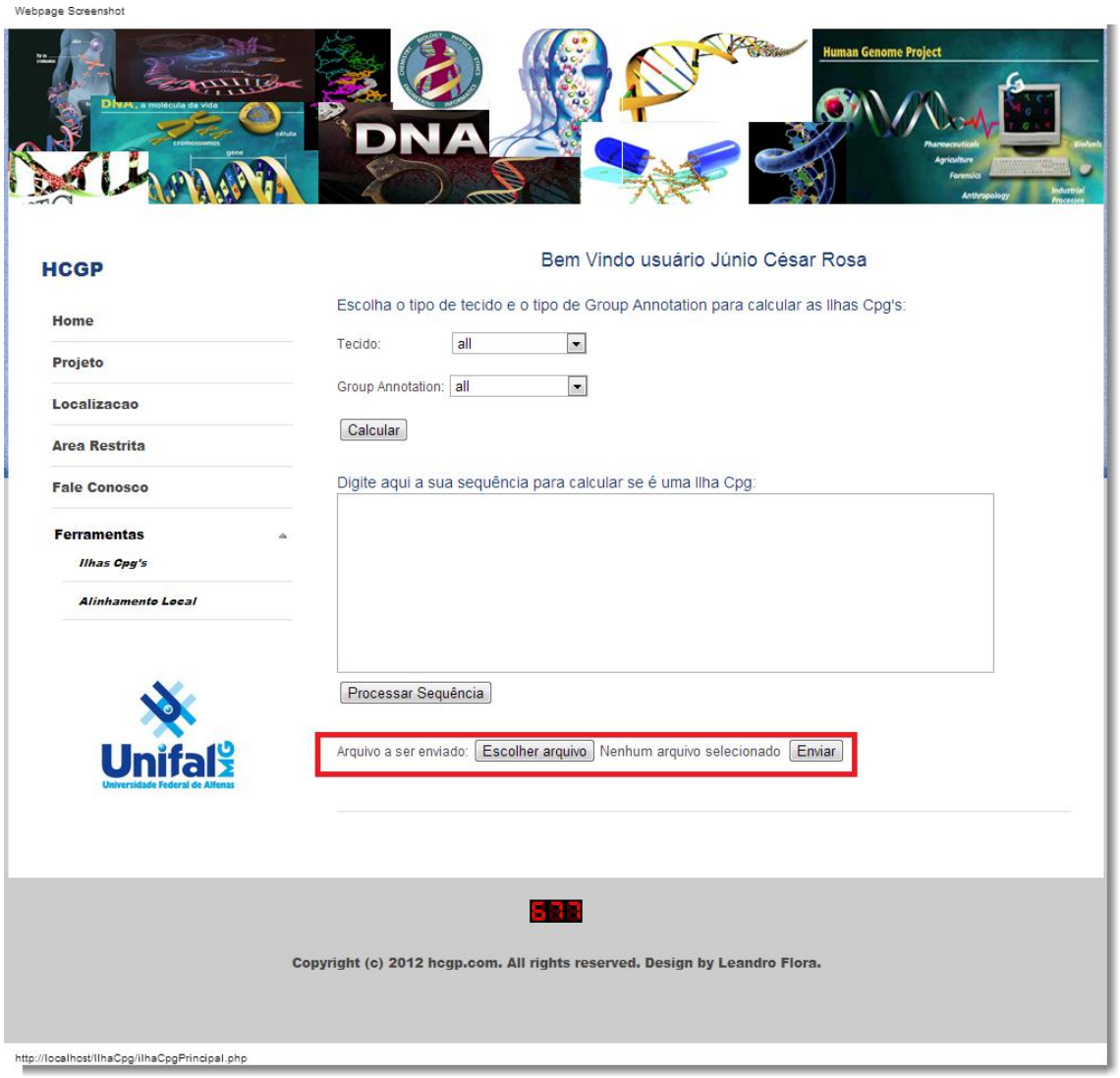


Figura 21 – Terceira funcionalidade na ferramenta de busca de ilhas Cpg's

Esse arquivo texto deve ser um arquivo no formato .txt e deve conter sequências separadas por um enter, conforme mostra a Figura 22 abaixo.

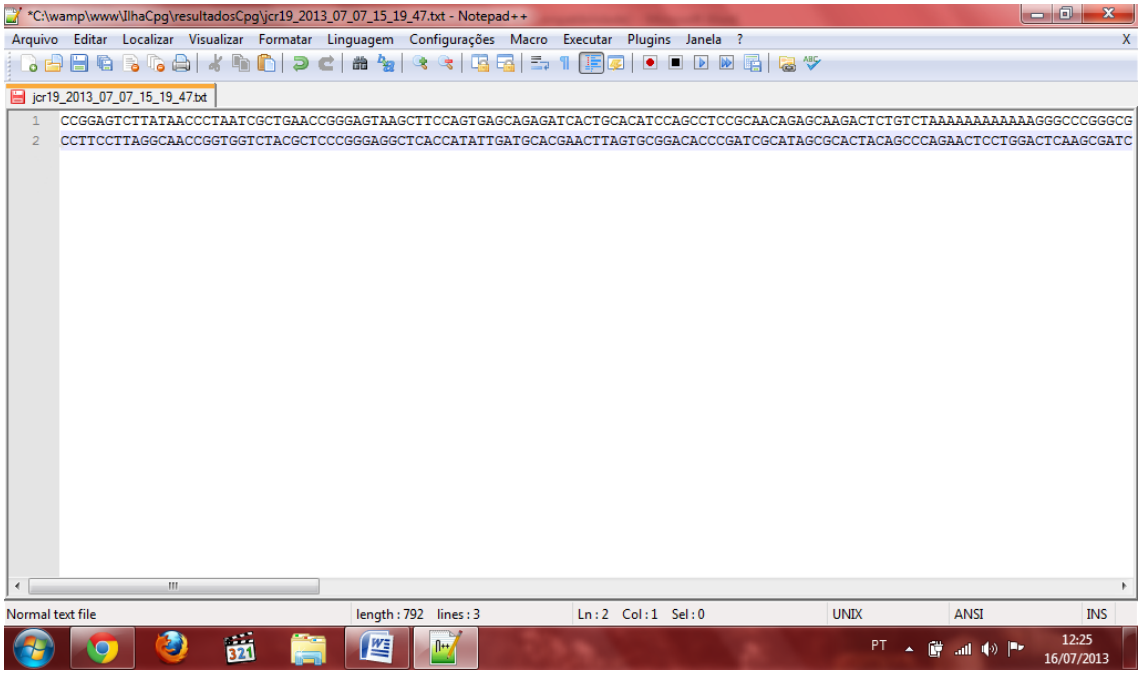


Figura 22 – Exemplo de arquivo texto para execução da terceira funcionalidade da ferramenta de busca de ilhas CpG

Após o usuário fazer o upload do arquivo, clicando em Escolher Arquivo, e clicar em Enviar vai aparecer a seguinte tela para o usuário, contendo o resultado do processamento das sequências.

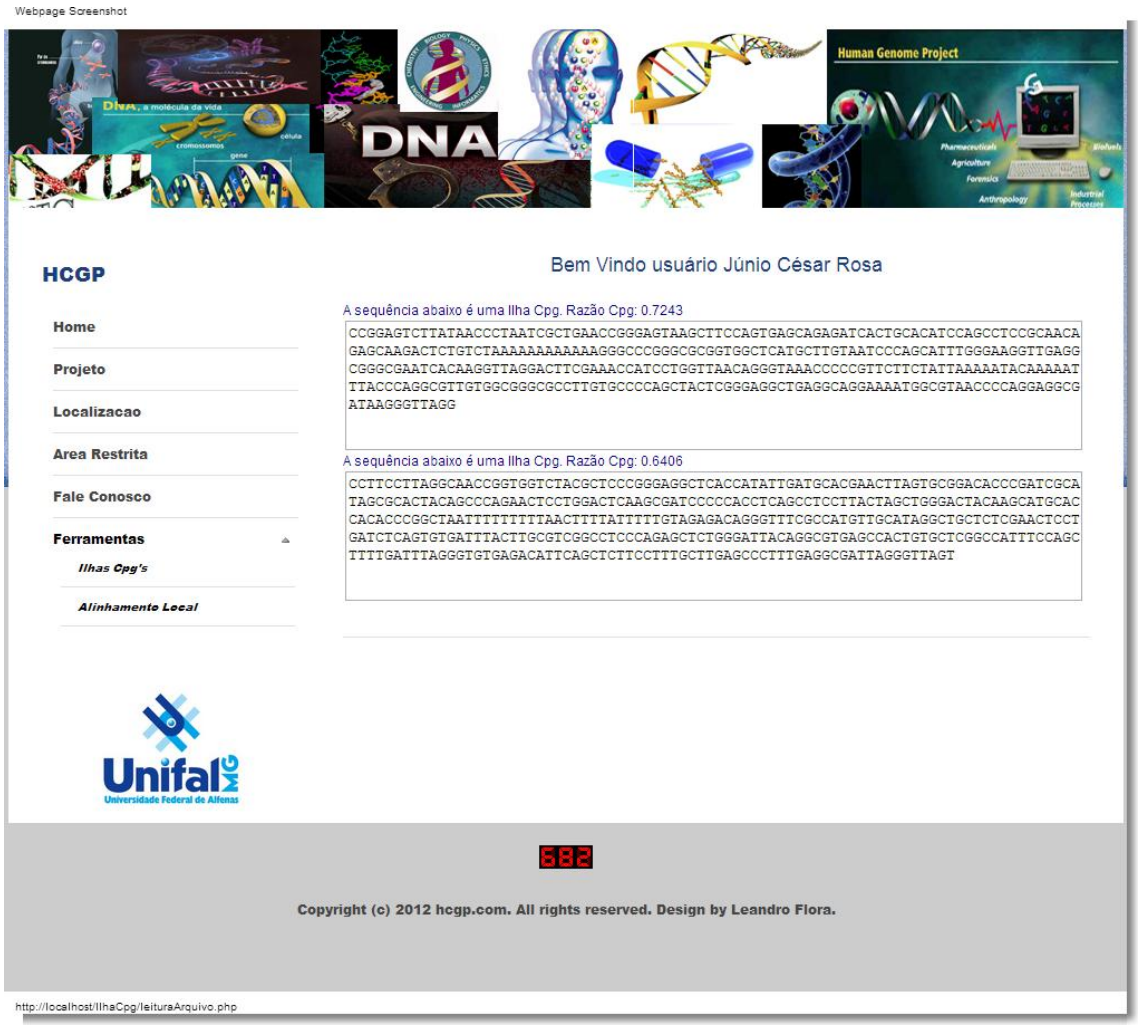


Figura 23 – Exemplo de execução da terceira funcionalidade da ferramenta de busca de ilhas CpG

9.3 Manual de funcionamento da ferramenta – Alinhamento de sequências

Essa seção consiste em mostrar o funcionamento da aba Alinhamento Local mostrada na Figura 24 a seguir.

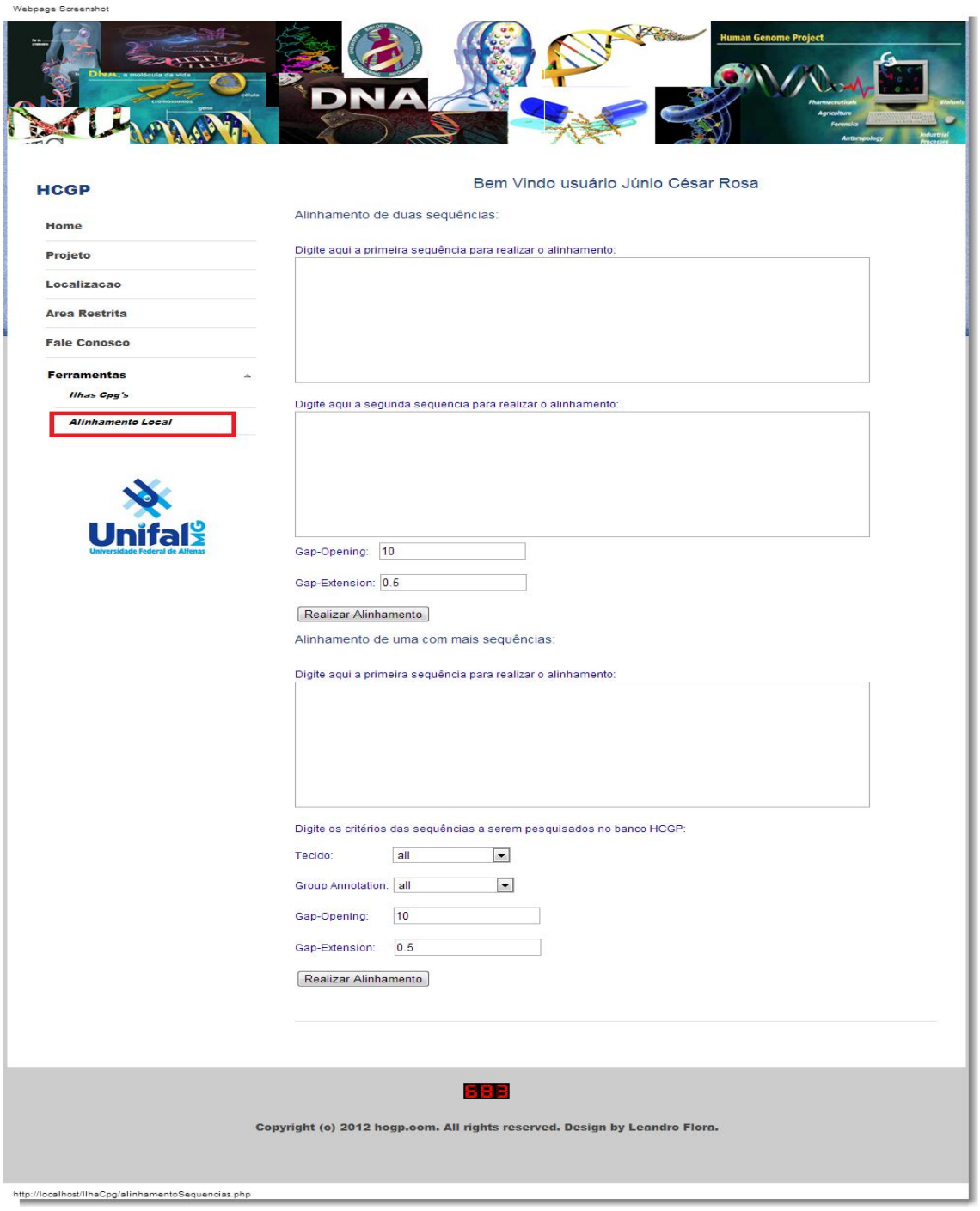


Figura 24 – Página web para alinhamento local de sequências.

Essa tela consiste de duas funcionalidades: O alinhamento local entre duas sequências ambas digitadas pelo usuário e o alinhamento local entre uma sequência digitada pelo usuário e as sequências da base de dados HCGP.

Na primeira funcionalidade o alinhamento entre duas sequências, o usuário digita as duas sequências no campos circulados na figura abaixo, preenche os campos Gap-Opening e Gap-Extension e clica em Realizar Alinhamento.

Figura 25 – Primeira funcionalidade da ferramenta de alinhamento local

Após preencher os campos e clicar em executar o site criára dois arquivos e executará o programa water e exibirá a seguinte tela de resultado para o usuário.

Figura 26 – Execução da primeira funcionalidade da ferramenta de alinhamento local

A segunda funcionalidade é muito parecida com a primeira a única diferença é que o usuário só digita uma sequência seleciona o tecido e o group annotation, preenche os campos Gap-Opening e Gap-Extension e clica em Realizar Alinhamento conforme mostra a figura abaixo.

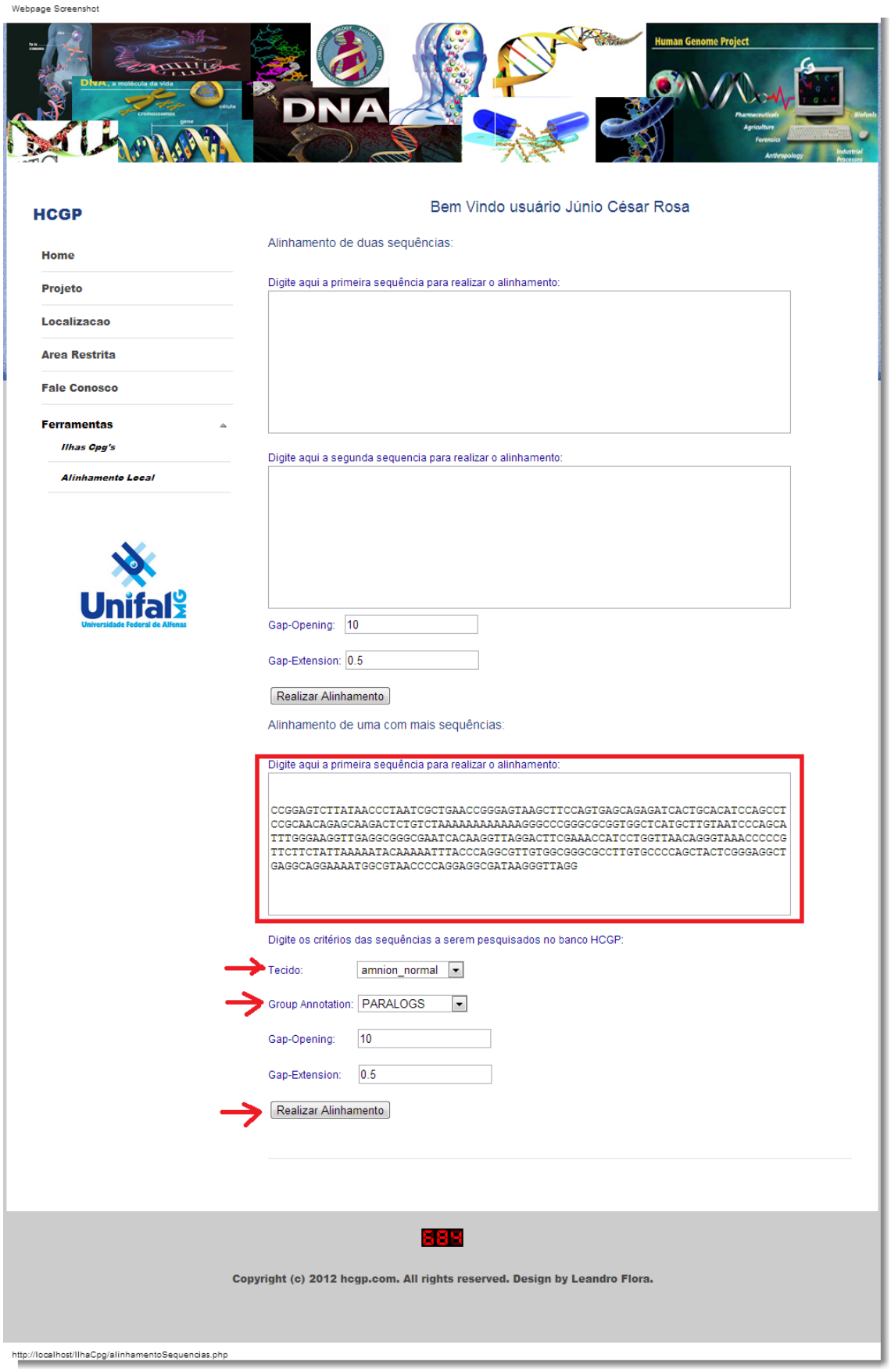


Figura 27 – Segunda funcionalidade da ferramenta de alinhamento local

Após preencher os campos e clicar em executar o site criára dois arquivos e executará o programa water e exibirá a seguinte tela de resultado para o usuário.

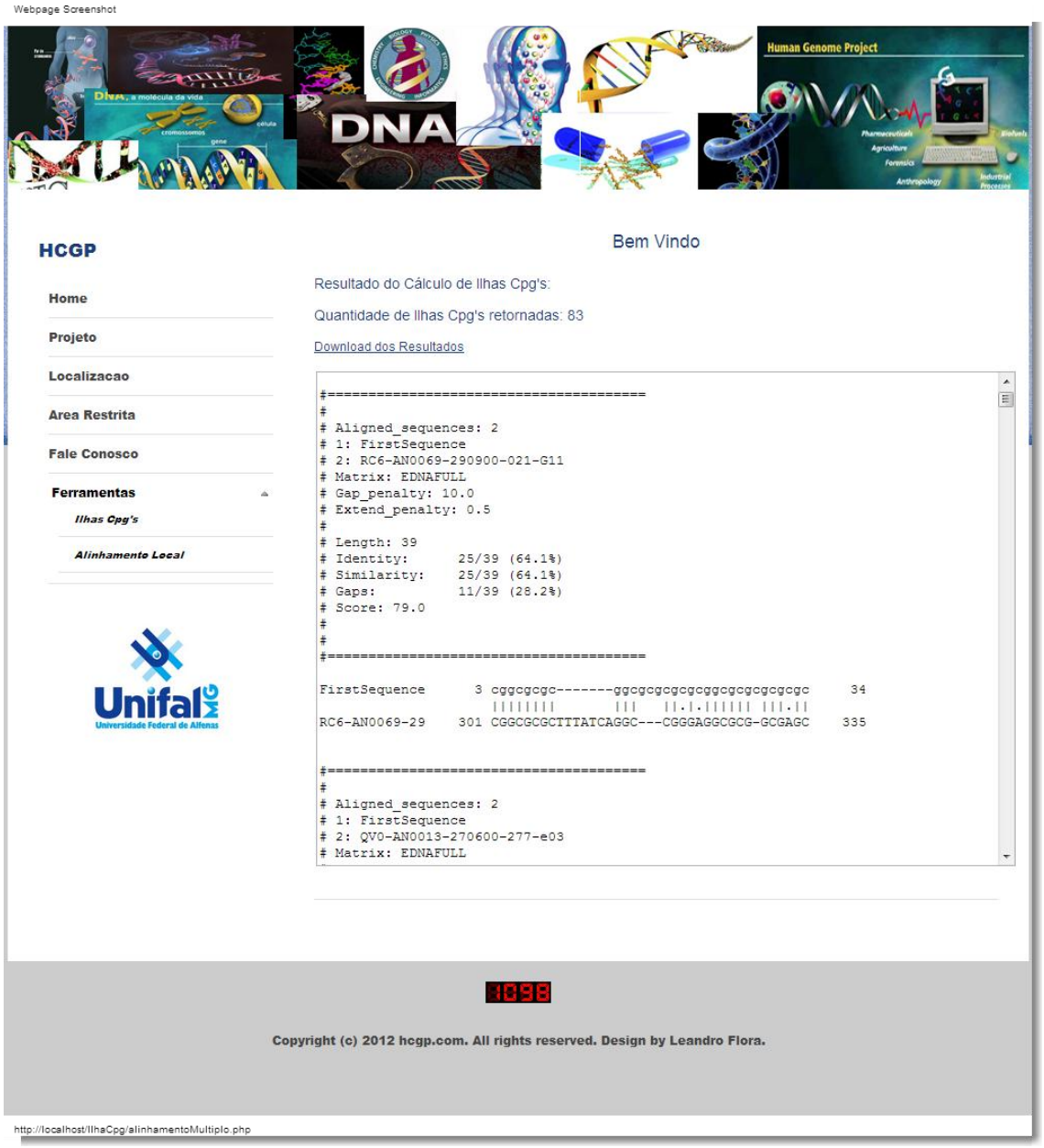


Figura 28 – Execução da segunda funcionalidade da ferramenta de alinhamento local